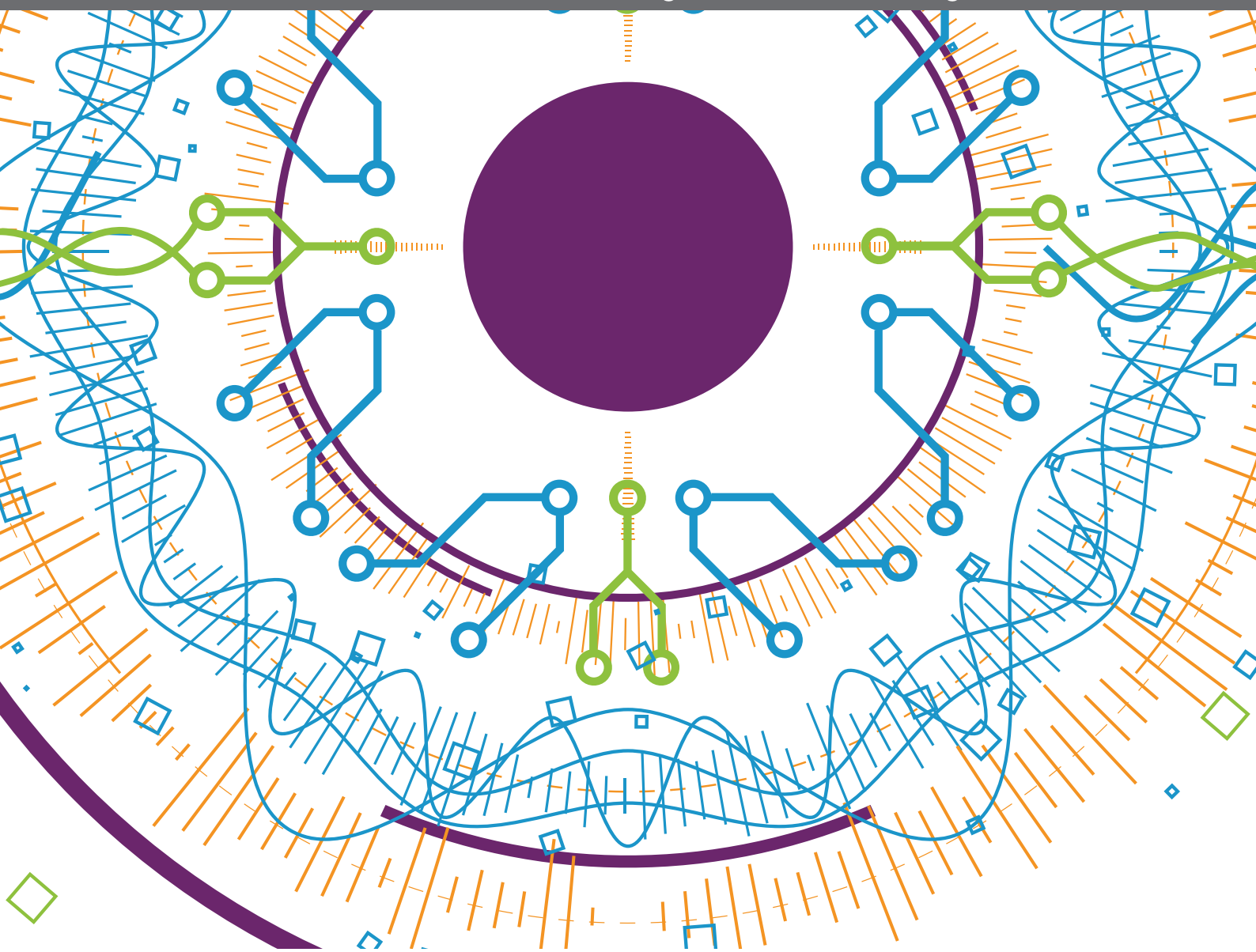
ETHICAL MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE (AI)

EDITED BY: Novi Quadrianto, Björn Wolfgang Schuller and Finnian Rachel Lattimore

PUBLISHED IN: Frontiers in Artificial Intelligence and Frontiers in Big Data



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-282-3
DOI 10.3389/978-2-88971-282-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

ETHICAL MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE (AI)

Topic Editors:
Novi Quadrianto, University of Sussex, United Kingdom
Björn Wolfgang Schuller, Imperial College London, United Kingdom
Finnian Rachel Lattimore, Gradient Institute, Australia

Citation: Quadrianto, N., Schuller, B. W., Lattimore, F. R., eds. (2021). Ethical Machine Learning and Artificial Intelligence (AI). Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-282-3

Table of Contents

frontiers in Big Data

EDITORIAL
published: 12 August 2021
doi: 10.3389/fdata.2021.742589

Editorial: Ethical Machine Learning and Artificial Intelligence

Novi Quadrianto[1,2], *Björn W. Schuller*[3,4]* *and Finnian Rachel Lattimore*[5]

[1]PAL – Predictive Analytics Lab, University of Sussex, Brighton, United Kingdom, [2]BCAM Severo Ochoa Strategic Lab on Trustworthy Machine Learning, Bilbao, Spain, [3]GLAM – Group on Language, Audio, & Music, Imperial College London, London, United Kingdom, [4]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, [5]Gradient Institute, Sydney, NSW, Australia

Keywords: ethics, artificial intelligence, machine learning, fairness, accountability, transparency, trustworthiness, General Data Protection Regulation

Editorial on the Research Topic

Ethical Machine Learning and Artificial Intelligence

1 INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have increasingly become an every-day reality for most of us (Elliott, 2019). Typical algorithmic assessment methods, used for predicting human outcomes such as recruitment, bail decisions, mortgage approvals, and insurance premiums, among many others, are currently being trialled and subsequently deployed. Hence, the ethical and legal requirements are moving into the foreground when developing novel AI and machine learning algorithms (Hagendorff, 2020). For example, the United States' Fair Credit Reporting Act and European Union's General Data Protection Regulation (GDPR) prescribe that data must be processed in a way that is fair/unbiased—a challenge for AI (Mehrabi et al., 2019). GDPR also alludes to the right of an individual to receive an explanation about decisions made by an automated system such as by explainable AI (XAI) (Gunning et al., 2019).

Here, based on a recent research topic held in Frontiers in Big Data, we provide an overview on the authors' views and contributions.

This research topic covers but is not limited to the fields of fairness, accountability, transparency, and trustworthiness (Baird et al., 2019), and covers methods such as causality and counterfactual reasoning, reinforcement learning, and probabilistic approaches.

2 LITERATURE REVIEW

The research topic provides two overviews on the field.

In the first, Wells and Bednarz discuss in "*Explainable AI and Reinforcement Learning – A Systematic Review of Current Approaches and Trends*" 25 studies selected from 520 search hits on this recent topic. Thereby, they focus on "visualisation, query-based explanations, policy summarisation, human-in-the-loop collaboration, and verification" which they identify as trends. As others, they name the urge for user evaluations including laymen of explanations and find examples often over-

OPEN ACCESS

Edited and reviewed by:
Sriraam Natarajan,
The University of Texas at Dallas,
United States

***Correspondence:**
Björn W. Schuller
schuller@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 July 2021*
Accepted: *22 July 2021*
Published: *12 August 2021*

Citation:
Quadrianto N, Schuller BW and Lattimore FR (2021) Editorial: Ethical Machine Learning and Artificial Intelligence. Front. Big Data 4:742589. doi: 10.3389/fdata.2021.742589

simplified going hand-in-hand with lack in scalability, while provision of comprehensible explanations remains a key challenge. Further, they consider more progressive visualisation approaches under-exploited including multimodal and immersive forms of visualisation. Ideally, in the authors' opinion, such would be combined with "well articulated explanations".

Next, in their mini review "*Considerations for a More Ethical Approach to Data in AI: On Data Representation and Infrastructure*", Baird and Schuller observe that data infrastructures are increasingly managed more democratically, as decentralisation fosters transparency and therefore can help better cope with selection-bias. Their review deals with AI-targeted data representation and infrastructures focussing on "auditing, benchmarking, confidence and trust, explainability and interpretability" as key aspects that require attention—ideally also in an interdisciplinary endeavour. As to auditing, in multimodal applications, the authors require standards per modality to lead to accurate benchmarking. Further, they support the view that confidence and trust are benefited by "diverse representations of human data"—the latter also boosting explainability to all users given "inherent human-like attributes". The authors attest energy put into these aspects by the community, but in particular demand for increased standardisation.

3 TECHNICAL APPROACHES

The research topic further includes three technical solutions.

First, in "*The Moral Choice Machine*", Schramowski et al. demonstrate that one can "extract deontological ethical reasoning" with machine learning from human written texts concerning right or wrong conduct. The authors provide prompts and responses and define a bias score based on the score of positive and negative responses. Likewise, they reach to the Moral Choice Machine (MCM), that determines this score per sentence applying Universal Sentence Encoder embeddings to cater for context. By that, they observe that textual databases bear "recoverable and accurate imprints of our social, ethical and moral choices". Further, picking selected databases from different epochs, they find reflection on the evolution of these aspects. Similarly, the authors consider different cultural sources. Ultimately, this leads to their view that "moral biases can be extracted, quantified, tracked, and compared across cultures and over time". As future work, the authors name the possibility to alter the embeddings in targeted ways, such as to eliminate gender stereotypes. They further suggest having the moral choice machine in interactive robots enabled with active learning to have users correct potential biases. Finally, they suggest targeted alteration of the text sources for observation of effects.

In "*Tuning Fairness by Balancing Target Labels*", Kehrenberg et al. deal with bias in the output as challenge. To this end, they add a latent target output to cater for a unified approach, apply marginalisation rather than constraints problem, and provide for a possibility to integrate knowledge on target unbiased outputs. The authors argue that fairness is usually mainly handled by statistical (group) or individual notions and belief that both are needed for algorithmic fairness. Their approach can be learnt from an implicitly balanced corpus, hence enabling demographic parity and equality of opportunity. They also indicate avenues towards an extension aiming at conditional demographic parity as well. Finally, their general approach uniquely provides for a target rate to control the realisation of the fairness constraint. However, it will need extensions for predictive parity group or individual fairness.

As a third example of algorithmic contribution to a more ethical approach serve Ramanan and Natarajan's with "*Causal Learning From Predictive Modeling for Observational Data*". They apply causal Bayesian networks to model causal relationships between data-learnt model variables sequentially using context-specific and mutual independence. Likewise, potential causal relationships are first found. Subsequentially, their strength is determined. The authors verify this approach on benchmark networks and find superiority over current alternatives.

4 DISCUSSION

Card and Smith finally discuss "*On Consequentialism and Fairness*", focusing on the outcome. They argue that consequentialism has its deficits such as lacking in an amenable choice of actions, but is a suited mean to highlight issues in AI fairness such as "who counts", disadvantages of policy application, or the relative weight of the future. The authors give a consequentialism-based critique of prevailing fairness definitions in AI. They further also take an AI viewpoint on consequentialism. Finally, they elaborate on learning and randomisation in the context of AI ethics.

5 CONCLUSION

As all authors highlight, a more ethical approach is needed to data in AI. However, algorithmic solutions can be and were partially given also here. Accordingly, there is a call to action also for those providing AI algorithms in the first place to actively work on solutions to benefit and protect all users of AI and society.

AUTHOR CONTRIBUTIONS

BS wrote the manuscript. NQ edited it. All authors led the underlying research topic.

ACKNOWLEDGMENTS

We express our sincere gratitude to all authors and reviewers that helped putting together the research topic.

REFERENCES

Baird, A., Hantke, S., and Schuller, B. (2019). Responsible and Representative Multimodal Data Acquisition and Analysis: on Auditability, Benchmarking, Confidence, Data-reliance & Explainability. *Clin. Orthop. Relat. Res.* arXiv [preprint] arXiv: 1903.07171. Available at: http://arxiv.org/abs/1903.07171/.

Elliott, A. (2019). *The Culture of AI: Everyday Life and the Digital Revolution*. London: Routledge.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G. Z. (2019). XAI-explainable Artificial Intelligence. *Sci. Robot.* 4, eaay7120. doi:10.1126/scirobotics.aay7120

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines.* 30, 99–120. doi:10.1007/s11023-020-09517-8

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *Clin. Orthop. Relat. Res.* arXiv [preprint] arXiv: 1908.09635. Available at: http://arxiv.org/abs/1908.09635/.

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.

frontiers in Artificial Intelligence

CONCEPTUAL ANALYSIS
published: 08 May 2020
doi: 10.3389/frai.2020.00034

On Consequentialism and Fairness

Dallas Card[1]* *and Noah A. Smith*[2,3]

[1] Computer Science Department, Stanford University, Stanford, CA, United States, [2] Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, United States, [3] Allen Institute for AI, Seattle, WA, United States

Recent work on fairness in machine learning has primarily emphasized how to define, quantify, and encourage "fair" outcomes. Less attention has been paid, however, to the ethical foundations which underlie such efforts. Among the ethical perspectives that should be taken into consideration is *consequentialism*, the position that, roughly speaking, outcomes are all that matter. Although consequentialism is not free from difficulties, and although it does not necessarily provide a tractable way of choosing actions (because of the combined problems of uncertainty, subjectivity, and aggregation), it nevertheless provides a powerful foundation from which to critique the existing literature on machine learning fairness. Moreover, it brings to the fore some of the tradeoffs involved, including the problem of who counts, the pros and cons of using a policy, and the relative value of the distant future. In this paper we provide a consequentialist critique of common definitions of fairness within machine learning, as well as a machine learning perspective on consequentialism. We conclude with a broader discussion of the issues of learning and randomization, which have important implications for the ethics of automated decision making systems.

Keywords: consequentialism, fairness, ethics, machine learning, randomization

OPEN ACCESS

Edited by:
Novi Quadrianto,
University of Sussex, United Kingdom

Reviewed by:
Deepak P,
Queen's University Belfast, United Kingdom
Animesh Mukherjee,
Indian Institute of Technology, India
Tiberio Caetano,
Gradient Institute, Australia

***Correspondence:**
Dallas Card
dcard@stanford.edu

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

Received: *13 January 2020*
Accepted: *17 April 2020*
Published: *08 May 2020*

Citation:
Card D and Smith NA (2020) On Consequentialism and Fairness. Front. Artif. Intell. 3:34. doi: 10.3389/frai.2020.00034

1. INTRODUCTION

In recent years, computer scientists have increasingly come to recognize that artificial intelligence (AI) systems have the potential to create harmful consequences. Especially within machine learning, there have been numerous efforts to formally characterize various notions of *fairness* and develop algorithms to satisfy these criteria. However, most of this research has proceeded without any nuanced discussion of ethical foundations. Partly as a response, there have been several recent calls to think more broadly about the ethical implications of AI (Barabas et al., 2018; Hu and Chen, 2018b; Torresen, 2018; Green, 2019).

Among the most prominent approaches to ethics within philosophy is a highly influential position known as *consequentialism*. Roughly speaking, the consequentialist believes that outcomes are all that matter, and that people should therefore endeavor to *act so as to produce the best consequences, based on an impartial perspective as to what is best*.

Although there are numerous difficulties with consequentialism in practice (see section 4), it nevertheless provides a clear and principled foundation from which to critique proposals which fall short of its ideals. In this paper, we analyze the literature on fairness within machine learning, and show how it largely depends on assumptions which the consequentialist perspective reveals immediately to be problematic. In particular, we make the following contributions:

- We provide an accessible overview of the main ideas of consequentialism (section 3), as well as a discussion of its difficulties (section 4), with a special emphasis on computational limitations.
- We review the dominant ideas about fairness in the machine learning literature (section 5), and provide the first critique of these ideas explicitly from the perspective of consequentialism (section 6).
- We conclude with a broader discussion of the ethical issues raised by learning and randomization, highlighting future direction for both AI and consequentialism (section 7).

2. MOTIVATING EXAMPLES

Before providing a formal description of consequentialism (section 3), we will begin with a series of motivating examples which illustrate some of the difficulties involved. We consider three variations on decisions about lending money, a frequently-used example in discussions about fairness, and an area in which AI could have significant real-world consequences.

First, imagine being asked by a relative for a small personal loan. This would seem to be a relatively low-stakes decision involving a simple tradeoff (e.g., financial burden vs. familial strife). Although this decision could in principle have massive long term consequences (perhaps the relative will start a business that will have a large impact, etc.), it is the immediate consequences which will likely dominate the decision. On the other hand, treating this as a simple yes-or-no decision fails to recognize the full range of possibilities. A consequentialist might suggest that we consider all possible uses of the money, such as investing it, or lending it to someone in even greater need. Whereas *commonsense morality* might direct us to favor our relatives over strangers, the notion of *impartiality* inherent in consequentialism presents a challenge to this perspective, thus raising the problem of *demandingness* (section 4.4).

Second, consider a bank executive creating a policy to determine who will or will not be granted a loan. This policy will affect not only would-be borrowers, but also the financial health of the bank, its employees, etc. In this case, the bank will likely be bound by various forms of regulation which will constrain the policy. Even a decision maker with an impartial perspective will be bound by these laws (the breaking of which might entail severe negative consequences). In addition, the bank might wish to create a policy that will be perceived as *fair*, yet knowing the literature on machine learning fairness, they will know that no policy will simultaneously satisfy all criteria that have been proposed (section 5). Moreover, there may be a tradeoff between short-term profits and long-term success (section 4.2).

Finally, consider a legislator trying to craft legislation that will govern the space of policies that banks are allowed to use in determining who will get a loan. This is an even more high-level decision that could have even more far reaching consequences. As a democratic society, we may hope that those in government will work for the benefit of all (though this hope may often be disappointed in practice), but it is unclear how even a selfless legislator should balance all competing interests (section 4.1). Moreover, even if there were consensus on the desired outcome, determining the expected consequences of any particular governing policy will be extremely difficult, as banks will react to any such legislation, trying to maximize their own interests while respecting the letter of the law, thus raising the problem of *uncertainty* (section 4.3).

Although these scenarios are distinct, each of the issues raised applies to some extent in each case. As we will discuss, work on fairness within machine learning has focused primarily on the intermediate, institutional case, and has largely ignored the broader context. We will begin with an in-depth overview of consequentialism that engages with these difficulties, and then show that it nevertheless provides a useful critical perspective on conventional thinking about fairness within machine learning (section 6).

3. CONSEQUENTIALISM DEFINED

3.1. Overview

The literature on consequentialism is vast, including many nuances that will not concern us here. The most well-known expressions can be found in the writings of Jeremy Bentham (1970 [1781]) and John Stuart Mill (1979[1863]), later refined by philosophers such as Henry Sidgwick (1967), Elizabeth Anscombe (1958), Derek Parfit (1984), and Peter Singer (1993). The basic idea which unifies all of this thinking is that only the *outcomes* that result from our actions (i.e., the relative value of possible worlds that might exist in the future) have moral relevance.

Before proceeding, it is helpful to consider three lenses through which we can make sense of an ethical theory. First, we can consider a statement to be a claim about what would be objectively best, given some sort of full knowledge and understanding of the universe. Second, we can think of an ethical theory as a proposed guide for how someone should choose to act in a particular situation (which may only align partially with an objective perspective, due to limited information). Third, although less conventional, we can think of ethics as a way to interpret the actions taken by others. In the sense that "actions speak louder than words," we can treat people's behavior as revealing of their view of what is morally correct (Greene and Haidt, 2002).

Although consequentialism is typically presented in a more abstract philosophical form (often illustrated via thought experiments), we will begin with a concise mathematical formulation of the two most common forms of consequentialism, known as *act consequentialism* and *rule consequentialism*. For the moment, we will intentionally adopt the objective perspective, before returning to practical difficulties below.

3.2. Act Consequentialism

First, consider the proposal known as *act consequentialism*. This theory says, simply, that the best action to take in any situation is the one that will produce the best outcomes (Smart and Williams, 1973; Railton, 1984). To be precise, let us define the set of possible actions, $\mathcal{A}$, and an evaluation function $v(\cdot)$. According to act consequentialism, the best action to take is the one that will lead

to the consequences with the greatest value, i.e.,

$$a^* = \arg\max_{a \in \mathcal{A}} v(c_a), \tag{1}$$

where $v(c_a)$ computes the value of consequences, c_a, which follow from taking action a. Importantly, note that c_a here represents not just the local or immediate consequences of a, but *all* consequences (Kagan, 1998). In other words, we can think of the decision as a branching point in the universe, and want to evaluate how it will unfold based on the action that is taken at a particular moment in time (Portmore, 2011).

While Equation (1) might seem tautological, it is by no means a universally agreed upon definition of what is best. For example, many *deontological* theories posit that certain actions should never be permitted (or that some might always be required), no matter what the consequences. In addition, there are some obvious difficulties with Equation (1), especially the question of how to define the evaluation function $v(\cdot)$. We will return to this and other difficulties below (section 4), but for the moment we will put them aside.

One might object that perhaps there is inherent randomness in the universe, leading to uncertainty about c_a. In that case, we can sensibly define the optimal action in terms of the expected value of all future consequences, i.e.,

$$a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}_{p(c|a)}[v(c)], \tag{2}$$

where $p(c \mid a)$ represents the true probability (according to the universe) that consequences c will follow from action a. That is, for each possible action, we would consider all possible outcomes which might result from that action, and sum their values, weighted by the respective probabilities that they will occur, recommending the action with the highest expected value.

To make the dependence on future consequences more explicit, it can be helpful to factor the expected value into a summation over time, optionally with some sort of discounting. Although consequentialism does not require that we factorize the value of the future in this way, it will prove convenient in further elaboration of these ideas. For the sake of simplicity, we will assume that time can be discretized into finite steps. A statement of act consequentialism using a simple geometric discounting factor would then be:

$$a^* = \arg\max_{a \in \mathcal{A}} \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{E}_{p(s_{t+1}|a)}[v(s_{t+1})], \tag{3}$$

where $p(s_{t+1} \mid a)$ represents the probability that the universe will be in state s at time $t+1$ if we take action a at time $t=0$, and $0 \leq \gamma \leq 1$ represents the discount factor. A discount factor of 0 means that only the *immediate* consequences of an action are relevant, whereas a discount factor of 1 means that all times in the future are valued equally[1].

[1]One could similarly augment Equation (3) to make any epistemic uncertainty about the evaluation function or discount factor explicit.

3.3. Rule Consequentialism

The main alternative to act consequentialism is a variant known as *rule consequentialism* (Harsanyi, 1977; Hooker, 2002). As the name suggests, rule consequentialism is similar to act consequentialism, except that rather than focusing on the best action in each unique situation, it suggests that we should act according to a set of *rules* governing all situations, and adopt the set of rules which will lead to the best overall outcomes[2].

Here, we will refer to a set of rules as a *policy*, and allow for the policy to be stochastic. In other words, a policy, π, is a probability distribution over possible actions conditional on the present state s, i.e., $\pi(s) \triangleq p(a \mid s)$. To make a decision, an action is sampled randomly from this distribution[3]. Using the same temporal factorization as above, we can formalize rule consequentialism as

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{p(s_{t'+1}|a_{t'},s_{t'})\pi(a_{t'}|s_{t'})} \left[\sum_{t=0}^{\infty} \gamma^t \cdot v(s_{t+1}) \right], \tag{4}$$

where Π represents the space of possible policies, and the expectation is now taken with respect to the governing dynamics, in which actions are selected based on the state of the world, i.e., $a_t \sim \pi(a_t \mid s_t)$, and the next state depends on the current state of the world and the action taken, i.e., $s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$.

While some have suggested that rule consequentialism is strictly inferior to act consequentialism, in that it fails to treat each situation as unique (Railton, 1984), others have argued for it, citing the inability of individuals to accurately determine the best action in each unique situation (Hooker, 2002), as well as benefits from coordination and incentives (Harsanyi, 1977). As noted by various papers (e.g., Abel et al., 2016), Equation (4) bears a striking resemblance to the problem of *reinforcement learning*[4]. While this similarity is provocative, we will defer discussion of it (and the more general question of *learning*) until section 7.

It is important to emphasize that the above formulation is a highly stylized discussion of morality, largely divorced from reality, which tries to encapsulate a large body of philosophical writing put forward under the name "consequentialism." Thinking about what this formulation has to tell us about how individuals make (or should make) choices requires further elaboration, which we revisit below (section 4).

3.4. Competing Ethical Frameworks

The primary contrasting proposals to consequentialism are (a) *deontology*; and (b) theories in the *social contract* tradition. As mentioned above, deontological theories posit that there are certain restrictions or requirements on action, *a priori*, which

[2]In some cases, rule consequentialism is formulated as the problem of choosing the set of rules which, if internalized by the vast majority of the community, would lead to the best consequences (Hooker, 2002).

[3]Most treatments of consequentialism assume that the rules determine a single correct action for each situation. However, the formulation presented here is strictly more general; deterministic policies are those that assign all probability mass to a single action for each state.

[4]Equation (4) is equivalent to the standard formulation of a Markov decision process if we restrict ourselves to a finite set of states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, transition probabilities $p(s_{t+1} \mid s_t, a_t)$, and discount factor γ.

cannot be violated. For example, various religious traditions place restrictions on lending money, or require a certain level of charitable giving. Using the framework established above, we can describe deontological theories as *constraints* on the action space, $\mathcal{A}$, or policy space, Π (Kagan, 1998). While they may accord more with our commonsense notions of morality (see section 4.4), deontological theories are open to challenge because of their inability to justify the particular constraints they specify, as well as the implication that they would fail to produce the best outcomes in certain scenarios (Smart and Williams, 1973; Scheffler, 1994).

By contrast, social contract theories are more concerned with determining the rules, or ways of organizing society, that a group of free and reasonable people would agree to in an idealized deliberative scenario[5]. Most famously in this tradition, John Rawls suggested that we should imagine people designing society behind a "veil of ignorance," not knowing what position they will hold in that society (Rawls, 1971). We cannot possibly do justice to these other schools of thought in the space available, but we note that there is value in thinking about sociotechnical systems from multiple ethical perspectives, and encourage others to elaborate on these points[6].

In this paper, we focus on consequentialism not because it is necessarily superior to the alternatives, but because it is influential, and because it might seem, at first glance, to have a natural affinity with machine learning and optimization. While there have been many papers providing brief summaries of various ethical theories and their relevance to AI, we believe that a more in-depth treatment is required to fully unpack the implications of each, and would encourage similar consideration of the above traditions, as well as virtue ethics, feminist ethics, etc.

Before discussing the problems with consequentialism, it is useful to note that the formulation given in Equation (4) highlights three important matters about which reasonable people might disagree, with respect to how we should act (alluded to in section 2): we might disagree about the relative value of different outcomes [the evaluation function, $v(\cdot)$]; we might disagree about the likely effects of different actions [the probability of outcomes, $p(s_{t+1} \mid s_t, a_t)$]; and we might disagree about how much weight to place on the distant future (the discount factor, γ).

4. DIFFICULTIES OF CONSEQUENTIALISM

Even if one accepts the idea in Equation (2)—that the best action is the one that will produce the best outcome in expectation, with no *a priori* restrictions on the action space, there are still numerous difficulties with consequentialism, both theoretically and in practice.

[5]E.g., "An act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject as a basis for informed, unforced, general agreement" (Scanlon, 1998).

[6]For a review of how Rawls has been applied within information sciences [see Hoffmann (2017)].

4.1. Value

Perhaps the most vexing part of consequentialism is the evaluation function, $v(\cdot)$. Even if one had perfect knowledge of how the universe would unfold conditional on each possible action, choosing the *best* action would still require some sort of objective way of characterizing the relative value of each possible outcome. Most writers on consequentialism agree that the specification of value should be *impartial*, in that it should not give arbitrary priority to particular individuals (Singer, 1993; Kagan, 1998), but this is far from sufficient for resolving this difficulty[7].

By far the most common way of simplifying the evaluation of outcomes, both within writings on consequentialism and in decision theory, is to adopt the classic *utilitarian* perspective (Smart and Williams, 1973; Mill, 1979[1863]). Although there are many variations, the most common statement of utilitarianism is that the value of a state is equal to the sum of the well-being experienced by all individual entities[8]. The most common social welfare function is thus

$$v(s) = \sum_{e \in \mathcal{E}} w_e(s), \tag{5}$$

where $\mathcal{E}$ represents the set of entities under consideration, and $w_e(s)$ measures the absolute well-being of entity e in state s[9].

Although utilitarianism is highly influential, there are fundamental difficulties with it. First, aggregating well-being requires *measuring* individual welfare, but it is unclear that it can be measured in a way that allows for fair comparisons, at least given current technology. Even if we restrict the set of morally relevant entities to humans, issues of subjectivity, disposition, and self-reporting make it difficult if not impossible to meaningfully compare across individuals (Binmore, 2009).

Second, even if there were a satisfactory way of measuring individual well-being, there are computational difficulties involved in *estimating* these values for hypothetical worlds. Given that well-being could depend on fine-grained details of the state of the world, it is unclear what level of precision would be required of a model in order to evaluate well-being for each entity. Thus, even estimating the overall value of a single state of the world might be infeasible, let alone a progression of them over time.

Third, any function which maps from the welfare of multiple entities to a single scalar will fail to distinguish between dramatically different distributions. Using the sum, for example, will treat as equivalent two states with the same total value, but with different levels of inequality (Parfit, 1984). While this

[7]Sidgwick (1967) writes, "I obtain the self-evident principle that the good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other; unless, that is, there are special grounds for believing that more good is likely to be realized in the one case than in the other."

[8]The philosophical literature in some cases uses happiness or the satisfaction of preferences, rather than well-being, but this distinction is not essential for our purposes.

[9]Note that using a separate value function for each entity accounts for variation in preferences, and allows for some entities to "count" for more than others, as when the set of relevant entities includes animals, or all sentient beings (Kagan, 1998).

failing is not necessarily insurmountable, most solutions seem to undermine the inherent simplicity of the utilitarian ideal[10].

Fourth, others have challenged the ideal of impartiality on the grounds that it is subtly paternalist, emphasizes individual autonomy over relationships and care, and ignores existing relations of power (Smart and Williams, 1973; Friedman, 1991; Driver, 2005; Kittay, 2009). Undoubtedly, there is a long and troubling history of otherwise enlightened philosophers presuming to know what is best for others, and being blind to the harms of institutions such as colonialism, while believing that certain classes of people either don't count or are incapable of full rationality (Mills, 1987; Schultz and Varouxakis, 2005).

Ultimately, it seems inescapable to conclude that there is no universally acceptable evaluation function for consequentialism. Rather, we must acknowledge that every action will entail an uneven distribution of costs and benefits. Even in the case where an action literally makes everyone better off, it will almost certainly benefit some more than others. As such, the most credible position is to view the idea of valuation (utilitarian or otherwise) as inherently contested and political. While we might insist that an admissible evaluation function conform to certain criteria, such as disinterestedness, or not being self-defeating (Parfit, 1984), we must also acknowledge that advocating for a particular notion of value as correct is fundamentally a political act.

4.2. Temporal Discounting

Even if there were an unproblematic way of assessing the relative value of a state of the world, the extent to which we should value the distant future is yet another point of potential disagreement. It is common (for somewhat orthogonal reasons) to apply temporal discounting in economics, but it is not obvious that there is any good reason to do so when it comes to moral value (Cowen and Parfit, 1992; Cowen, 2006). Just as philosophers such as Peter Singer have argued that we should not discount the value of a human life simply because a person happens to live far away (Singer, 1972), one could argue that the lives of those who will live in the future should count for as much as the lives of people who are alive today.

Unfortunately, it is difficult to avoid discounting in practice, as it becomes increasingly difficult to predict the consequences of our actions farther into the future. Even if we assume a finite action space, the number of possible worlds to consider will grow exponentially over time. Moreover, because of the chaotic nature of complex systems, even if we had complete knowledge of the causal structure of the universe, we would be limited in our ability to predict the future by lack of precision in our knowledge about the present.

Despite these difficulties, consequentialism would suggest that we should, to the extent that we are able, think not only about the immediate consequences of our actions, but about the longer-term consequences as well (Cowen, 2006). Indeed, considering the political nature of valuation, we arguably bear even greater responsibility for thinking about future generations than the present, given that those who have not yet been born are unable to directly advocate for their interests.

4.3. Uncertainty

In practice, of course, we do not know with any certainty what the consequences of our actions will be, especially over the long term. Again, from the perspective of determining the objectively morally correct action, one might argue that all that matters is the (unknown) probability according to the universe. For individual decision makers, however, any person's ability to predict the future will be limited, and, indeed, will likely vary across individuals. In other words, it is not just our uncertainty about consequences that is a problem, but our uncertainty about our uncertainty: we don't know how well or poorly our own model of the universe matches the true likelihood of what will happen (Kagan, 1998; Cowen, 2006).

The *subjective* interpretation of consequentialism suggests that, regardless of what the actual consequences may be, the morally correct thing for an individual to do is whatever they have reason to believe will produce the best consequences (Kagan, 1998). This, however, is problematic for two reasons: first, it ignores the computational effort involved in trying to determine which action would be best (which is itself a kind of action); and second, it seemingly absolves people from wrong-doing who happen to have a poor model of the world.

Rule consequentialism arguably provides a (philosophical) solution for these problems, in that it involves a direct mapping from states to actions, without requiring that each decision maker independently determine the expected value of each possible action (Kagan, 1998; Hooker, 2002)[11]. It still has the problem, however, of determining what policy is optimal, given our uncertainty about the world. Nevertheless, we should not overstate the problem of uncertainty; we are not in a state of total ignorance, and in general, trying to help people is likely to do more good than trying to harm them (de Lazari-Radek and Singer, 2017).

4.4. Conflicts With Commonsense Morality

A final set of arguments against consequentialism take the form of thought experiments in which consequentialism (and utilitarianism in particular) would seemingly require us to take actions that violate our own notions of commonsense morality. A particularly common example is the "trolley problem" and its variants, in which it is asked whether or not it is correct to cause one person to die in order to save multiple others (Foot, 1967; Greene, 2013).

We will not dwell on these thought experiments, except to note that many of the seeming conflicts from this type of scenario vanish once we take a longer term view, or adopt a broader notion of value than a simple sum over individuals. Killing one patient to save five might create greater aggregate well-being if

[10]For example, one could model well-being as a non-linear, increasing, concave (e.g., logarithmic) function of other attributes such as wealth (i.e., diminishing marginal utility), which would encourage a more equal distribution of resources. Alternatively, one could try to incorporate people's suffering due to inequality into their value functions (de Lazari-Radek and Singer, 2017).

[11]To use a somewhat farcical example, we could imagine using a neural network to map from states to actions; the time to compute what action to take would therefore be constant for any scenario.

we only consider the *immediate* consequences. If we consider *all* consequences of such an action, however, it should be obvious why we would not wish to adopt such a policy (Kagan, 1991).

It is worth commenting, however, on one particular conflict with commonsense morality, namely the claim that consequentialism is, in some circumstances, excessively *demanding*. Given the present amount of suffering in the world, and the diminishing marginal utility of wealth, taking consequentialism seriously would seem to require that we sacrifice nearly all of our resources in an effort to improve the well-being of the worst off (Smart and Williams, 1973; Driver, 2012). While to some extent this concern is mitigated by the same logic as above (reducing ourselves to ruin would be less valuable over the long term than sacrificing a smaller but sustainable amount), we should take seriously the possibility that the best action might not agree with our moral intuitions.

5. FAIRNESS IN MACHINE LEARNING

With the necessary background on consequentialism in place, we now review and summarize ideas about fairness in machine learning. Note that "fairness" is arguably an ambiguous and overloaded term in general usage; our focus here is on how it has been conceptualized and formalized within the machine learning literature[12]. In order to lay the foundation for a critical perspective on this literature, we first summarize the general framework that is commonly used for discussing fairness, and then summarize the most prominent ways in which it has been defined[13].

The typical setup is to assume that there are two or more groups of individuals which are distinguished by some "protected attribute," A, such as race or gender. All other information about each individual is represented by a feature vector, X. The purpose of the system is to make a prediction about each individual, $\hat{Y}$, which we will assume to be binary, for the sake of simplicity. Moreover, we will assume that the two possible predictions (1 or 0) are asymmetric, such that one is in some sense preferable. Finally, we assume that, for some individuals, we can observe the true outcome, Y. We will use $\mathcal{X}$ to refer to a set of individuals.

To make this more concrete, consider the case of deciding whether or not to approve a loan. An algorithmic decision making system would take the applicant's information (X and possibly A), and return a prediction about whether or not the applicant will repay the loan, $\hat{Y}$. For those applicants who are approved, we can then check to see who actually pays it back on time ($Y = 1$) and who does not ($Y = 0$). Note, however, that in this setup, we are unable to observe the outcome for those applicants who are denied a loan, and thus cannot know what their outcome would have been in the counterfactual scenario.

The overriding concern in this literature is to make predictions that are highly accurate while respecting some notion of fairness. Because reducing complex social constructs such as race and gender to simplistic categories is inherently problematic, as a running example we will instead use *biological age* as a hypothetical protected attribute[14]. Using the same notation as above, we would say that an automated system instantiates a policy, π, in making a prediction for each applicant. Thus, for instance i, a threshold classifier would predict

$$\hat{y}_i = \underset{y \in \{0,1\}}{\arg\max}\, \pi(Y = y \mid X = x_i, A = a_i), \tag{6}$$

though we might equally consider a randomized predictor.

Much of the work in fairness has drawn inspiration from two legal doctrines: *disparate treatment* and *disparate impact* (Ruggieri et al., 2010; Barocas and Selbst, 2016). Disparate treatment, roughly speaking, says that two people should not be treated differently if they differ only in terms of a protected attribute. For our running example, this would be equivalent to saying that one cannot deny someone a loan simply because of their age.

Disparate impact, on the other hand, prohibits the adoption of policies that would have consequences that are unevenly distributed according to the protected attribute, even if they are neutral on their face. Thus a policy which denies loans to people with no credit history might have a disparate impact on younger borrowers, and could therefore (hypothetically) be considered discriminatory.

While research in machine learning fairness is ongoing, most proposals can be classified into two types, which to some extent map onto the two legal doctrines mentioned above. Some definitions are specified without reference to outcomes (section 5.1). Others are specified exclusively with regard to a particular set of outcomes (which must be evaluated using real data; section 5.2). We summarize the dominant proposals of each type below.

5.1. Fairness Constraints Specified Without Regard to Outcomes

The first type of approach to fairness advocates constraints that are specified without reference to actual effects. In a formal sense, we can think of these as placing restrictions, *a priori*, on the space of policies which will be considered morally acceptable. We provide three examples of this type of approach below.

5.1.1. Fairness Through Unawareness

A commonsense but naive notion is to disallow policies which use the protected attribute in making a prediction. Equivalently, this requires that for any x,

$$\pi(y \mid x, A = 0) = \pi(y \mid x, A = 1) \tag{7}$$

Although this seems like a strict translation of the prohibition against disparate treatment, it is generally considered to be

[12] Extensive discussion of the idea of fairness can be found in much of the philosophical and technical literature cited throughout. In particular, we refer to the reader to Rawls (1958), Kagan (1998), and Binns (2018).

[13] While there is also some work on fairness in the unsupervised setting (e.g., Benthall and Haynes, 2019; Kleindessner et al., 2019), in this paper we focus on the supervised case.

[14] Age is a particularly interesting example of a protected attribute, as it is explicitly used to discriminate in some domains (as in restricting the right to vote), but afforded some protections in others (such as the U.S. Age Discrimination in Employment Act).

unhelpful (Hardt et al., 2016; Kleinberg et al., 2018). Due to correlations, it may be possible to infer the protected attribute from other features, hence prohibiting a single piece of information may have no effect in practice.

5.1.2. Individual Fairness

A more general application of the same idea argues that models must make similar predictions for similar individuals (in terms of their representations, X) (Dwork et al., 2012). This proposal was originally framed as being in the Rawlsian tradition, suggesting it should be a matter of public deliberation to determine who counts as similar. However, as has been noted, the effects of this framework are highly dependent on the particular notion of similarity that is chosen (Green and Hu, 2018).

5.1.3. Randomization

A further way of avoiding disparate treatment is through randomization (Kroll et al., 2017). The basic idea is that a policy should not look at the protected attribute *or any other attribute* when making a decision, except perhaps to verify that some minimal criteria are met. For example, a policy might assign 0 probability to instances that do not meet the criteria, and an equal probability to all others. Although this is a severe limitation on the space of policies, we do see instances of it being used in practice, such as in the U.S. Diversity Visa Lottery (Perry and Zarsky, 2015; Kroll et al., 2017)[15].

5.2. Fairness Constraints Specified in Terms of Outcomes

The other major approach to fairness in machine learning is to specify requirements on the actual outcomes of a policy. In other words, while the above fairness criteria can be evaluated without data, the following criteria can only be checked using an actual dataset. These notions of fairness are often justified in terms of the doctrine of disparate impact—that is, policies should not be adopted which have adverse outcomes for protected groups. Three examples are presented below:

5.2.1. Demographic/Statistical Parity

The notion of parity implies that the proportion of predicted labels should be the same, or approximately the same for each group. For example, this might require that an equal proportion of older and younger applicants would receive a loan. Formally, this requirement says that in order to be acceptable, a policy must satisfy

$$\frac{\sum_{i\in\mathcal{X}} \mathbb{I}[a_i = 0]\cdot\hat{y}_i}{\sum_{i\in\mathcal{X}} \mathbb{I}[a_i = 0]} = \frac{\sum_{j\in\mathcal{X}} \mathbb{I}[a_j = 1]\cdot\hat{y}_j}{\sum_{j\in\mathcal{X}} \mathbb{I}[a_j = 1]}, \tag{8}$$

where $\mathbb{I}[\cdot]$ equals 1 if the condition holds (otherwise 0). Demographic parity is a strong statement about what the consequences of a policy must be (in terms of a very focused set of short-term consequences). Note, however, that enforcing this constraint may result in suboptimal outcomes from the perspective of other criteria (Corbett-Davies et al., 2017).

5.2.2. Equality of Odds/Opportunity

Another outcome-based fairness criteria looks at the outcomes that result from the policy, and compares the rates of true positives and/or false positives among a held-out dataset (Hardt et al., 2016). Equal opportunity would require that, for example, an equal proportion of applicants from each group *who will pay back a loan* are in fact approved. Formally,

$$\frac{\sum_{i\in\mathcal{X}} \mathbb{I}[a_i = 0, y_i = 1]\cdot\hat{y}_i}{\sum_{i\in\mathcal{X}} \mathbb{I}[a_i = 0, y_i = 1]} = \frac{\sum_{j\in\mathcal{X}} \mathbb{I}[a_j = 1, y_j = 1]\cdot\hat{y}_j}{\sum_{j\in\mathcal{X}} \mathbb{I}[a_j = 1, y_j = 1]}. \tag{9}$$

Equality of odds is similar, except that is requires that rates of both true positives and false positives be the same across groups.

5.2.3. Equal Calibration

An alternative to equality of odds is to ask that the predictions be equally well calibrated across groups. That is, if we bin the predicted probabilities into a set of bins, a well-calibrated predictor should predict probabilities such that the proportion of instances that are correctly classified within each bin is the same for all groups. In other words, equal calibration tries to ensure that

$$\frac{\sum_{i\in\mathcal{X}} \mathbb{I}[a_i = 0, \hat{p}_i \in [b,c)]\cdot y_i}{\sum_{i\in\mathcal{X}} \mathbb{I}[a_i = 0, \hat{p}_i \in [b,c)]} = \frac{\sum_{j\in\mathcal{X}} \mathbb{I}[a_j = 1, \hat{p}_j \in [b,c)]\cdot y_j}{\sum_{j\in\mathcal{X}} \mathbb{I}[a_j = 1, \hat{p}_j \in [b,c)]} \tag{10}$$

for each interval $[b, c)$, where $\hat{p}_i = \pi(Y = 1 \mid x_i, a_i)$ according to the policy.

Note that whereas demographic parity only requires the set of predictions ($\hat{Y}$) made for all individuals in a dataset, equal opportunity and equal calibration also require that we know the true outcome (Y) for all such individuals, even those who are given a negative prediction. As a result, the latter two requirements can only be properly verified on a dataset for which we can independently observe the true outcome (e.g., based on assigning treatment randomly).

As has been shown by multiple authors, certain fairness criteria will necessarily be in conflict with others, under mild conditions, indicating that we will be unable to satisfy all simultaneously (Chouldechova, 2017; Kleinberg et al., 2017).

6. A CONSEQUENTIALIST PERSPECTIVE ON MACHINE LEARNING FAIRNESS

As previously mentioned, most fairness metrics have been proposed with only limited discussion of ethical foundations. In this section, we provide commentary on the criteria described above from the perspective of consequentialism. As a reminder, we are not suggesting that consequentialism provides the last word on what is morally correct. Rather, we can think of consequentialism as providing one of several possible ethical perspectives which should be considered.

[15] Additional examples of randomization include jury selection, military service, sortition in ancient Athenian government, and which members of a firing squad have guns with real bullets. Of course, as Kroll et al. (2017) point out, randomization is only fair if the system cannot be manipulated by either applicants or decision makers.

First, consider the fairness proposals that are specified without regard to outcomes (section 5.1). As mentioned above, these can be seen as restrictions on the set of policies that are acceptable. By definition, these constraints are not determined by the actual consequences of adopting them, nor do they possess an in-built verification mechanism to assess the nature of the consequences being produced. As such, these have more of a deontological flavor, reflecting a prior stipulation that similar people should be treated similarly, or that everyone deserves an equal chance. For example, Equation (7) specifies precisely the constraint on the policy space required by fairness through unawareness, and similarly for the other proposals. In principle, of course, these criteria could have been developed with the expectation that using them would produce the best outcomes, but it is far from obvious that this is the case.

By contrast, the fairness criteria specified explicitly in terms of outcomes (section 5.2) might seem to be closer to a form of consequentialism, given that they are evaluated by looking at actual impacts. However, upon closer inspection we see that they imply a severely restricted form of consequentialism in terms of how they think about value, time horizon, and who counts. In particular, while the proposals differ in terms of the precise values that are being emphasized, all of these proposals have some features in common:

- They only evaluate outcomes in terms of the people who are the direct object of the decision being made, not others who may be affected by these decisions;
- They only explicitly consider the immediate consequences of each decision, equivalent to using a discount factor of 0;
- They presuppose that a particular function of the distribution of predictions and outcomes (e.g., calibration) is the only value that is morally relevant.

Again, it is entirely possible that these constraints were developed with the *intention* of producing more broadly beneficial consequences over the long term. The point is that there is nothing in the constraints themselves that points to or tries to verify this broader impact, despite the fact that they are evaluated in terms of (a narrow set of) outcomes.

To make this concrete, consider again the case of trying to regulate algorithms which will be used by banks in making loans. Requiring satisfaction of any of the above fairness constraints will alter the set of loan applicants who are approved (and denied). While it is possible that some of these criteria might lead to broadly beneficial changes (e.g., demographic parity might enhance access to credit among those who have been historically marginalized), from the perspective of consequentialism it insufficient to evaluate the outcome only in terms of the probabilities or labels assigned to each group. Rather, it is necessary to consider the full range of consequences to individuals and society. In some cases, a loan might positively transform a person's life, or the life of their community, via mechanisms such as education and entrepreneurship. In other cases, easier access to credit could lead to speculative borrowing and financial ruin. For example, while not directly related to concerns about fairness, the potentially devastating effects of lending policies which ignore long-term and systemic effects can easily be seen in the aftermath of the subprime mortgage crisis, which derived, in part, by perverse incentives and risky lending (Bianco, 2008).

Crafting effective financial regulation is obviously extremely difficult, and this is not meant to suggest that any particular fairness constraint is likely to lead to disaster. Nevertheless, it is important to remember that fairness criteria which are specified only in terms of a narrow set of short term metrics do not guarantee positive outcomes beyond what they measure, and may in some cases lead to overall greater harm.

In sum, adopting a consequentialist perspective reveals numerous ways in which the existing proposals for thinking about fairness in machine learning are fatally flawed. While all have their merits, none have been adequately justified in terms of their likely consequences, broadly considered. Moreover, most are highly restricted in terms of the types of outcomes they take into consideration, and largely ignore broader systemic effects of adopting a single policy.

It is, of course, understandable that most approaches to machine learning fairness have focused on *a priori* constraints and tractable short term consequences. Avoiding negative consequences from new technologies is challenging in general, and many of the difficulties of consequentialism also apply directly to machine learning, especially in social contexts (uncertainty about the future, lack of agreement about value, etc.). Even in relatively controlled environments, it is easy to find examples of undesirable outcomes resulting from ill-specified value functions, improper time horizons, and the kinds of computational difficulties described in section 4 (Amodei et al., 2016).

Although consequentialism does not provide any easy answers about how to make AI systems more fair or just, several important considerations follow from its tenets. First, consequentialism reminds us of the need to consider outcomes broadly; technical systems are embedded in social contexts, and policies can have widespread effects on communities, not merely those who are subject to classification. Second, the political nature of valuation means that a broad range of perspectives on what is desirable should be sought out and considered, not for a reductive utilitarian calculus, but so as to be informed as to the diversity of opinions. Third, the phenomenon of diminishing marginal utility suggests that efforts should be directed to helping those who are worst off, rather than trying to make life better for the already well off, without, of course, presuming to automatically know what is best for others. Fourth, while we might disagree about the discount rate, the moral value of the future necessitates that we take downstream effects into account, rather than only focusing on immediate consequences. Sweeping attempts at regulation, such as GDPR, may have outsized effects here, as they will partially determine how we think about fairness going forward, and what it is legitimate to measure. Finally, because it is particularly difficult to predict consequences in the distant future, a high standard should be required for any policy that would place a definite burden on the present for a possible future gain.

7. RANDOMIZATION AND LEARNING

Before concluding, we will attempt to draw together a number of threads related to uncertainty, learning, and randomization. As described earlier, most philosophical presentations of consequentialism are highly abstract, without considering how one would practically determine what actions or rules are best. Given that statistics and machine learning arose specifically to deal with the problem of uncertainty, it is natural to ask whether there is any role for *learning* in consequentialism.

Indeed, an entire subfield of machine learning exists precisely to deal with the problem of action selection in the face of uncertainty (so-called "bandit" problems, or reinforcement learning more broadly). As noted in the introduction, the reinforcement learning objective explicitly encodes the goal of maximizing some benefit over the long term. Algorithms designed to optimize this objective typically rely initially on random exploration to reduce uncertainty, thereby facilitating long-term "exploitation" of rewards.

Not surprisingly, a number of papers have proposed using similar strategies as a way of achieving fair outcomes over the long-term. For example, Kroll et al. (2017) suggest that adding randomness to hiring algorithms could help to debias them over time. Joseph et al. (2016b) consider the problem of learning a policy for making loans, and present an algorithm to do so without violating a particular notion of fairness[16]. Liu et al. (2017) extend this work, again trying to satisfy fairness in the contextual bandit setting. Meanwhile, Barabas et al. (2018) suggest using randomization to facilitate causal inference about the "social, structural, and psychological drivers" of crime.

Randomization in decision making is a deep and important topic, and has been the focus of much past work in ethics (Lockwood and Anscombe, 1983; Freedman, 1987; Bird et al., 2016; Haushofer et al., 2019). As noted above, it can be a source of fairness, if we take "fair" to mean that everyone deserves an equal chance. It may also be useful to prevent strategic manipulation of a system, and has a definite role in some parts of American law (Perry and Zarsky, 2015; Kroll et al., 2017).

Although temporal discounting in consequentialism is typically discussed in terms of present vs. future value (e.g., helping people today vs. investing in the future), a similar trade off applies to costly experimentation for the purpose of reducing future uncertainty. Indeed, this sort of approach has been widely adopted in industry in the form of A/B testing, as well as for adaptive trials in domains such as medicine (Lai et al., 2015). Moreover, there is clearly something appealing about the idea that it *should* be morally incumbent upon people to improve their understanding of the world over time, not merely to act on their current understanding. However, randomization also raises a number of serious concerns.

First, as always, there is the problem of value, and the question of who gets to decide how to balance present costs against future benefits. Second, there are good reasons to think that such an approach is unlikely to work in complex sociotechnical systems. Although reinforcement learning has been extraordinarily successful in limited domains, such as game playing and online advertising, making reinforcement learning tractable generally requires assuming the existence of a stable environment, a limited space of actions, a clear reward signal, and a massive amount of training data. In most policy domains, we can expect to have none of these. Third, there may be real costs associated with participation in such a process; while a bank could conceivably choose to add randomness to a policy for granting loans (for the purpose of better learning who is likely to pay them back), giving loans to people who cannot afford them could have severe negative consequences for those individuals.

There are clearly some domains where randomization is widely used, and seems well-justified, especially from the perspective of consequentialism. The best example of this is clinical trials in medicine, which are not only favored, but required. Medicine, however, is a special domain for several reasons: there is general agreement about ends (saving lives and reducing suffering), there is good reason to think that findings will generalize across people, and there is a well-established framework for experimentation, with safeguards in place to protect the participants.

Where things get more complicated is using the same logic to establish the efficacy of social interventions, such as randomized trials in development economics. Although controlled experiments do provide good evidence about whether an intervention was effective, it is less clear that the conclusions will generalize to different situations (Barrett and Carter, 2010).

Ultimately, while randomization can be an important tool in learning policies that promote long term benefits, especially in relatively static, generalizable domains, the limitations of both consequentialism and of statistical learning theory mean that we should be highly skeptical of any attempt to use it as the basis for creating policies or automated decision making systems to deal with complex social problems.

8. ADDITIONAL RELATED WORK

Beyond the criteria mentioned in section 5, numerous other fairness metrics have been proposed, such as procedural fairness (Grgić-Hlača et al., 2016) and causal effects (Madras et al., 2018; Khademi et al., 2019). Meanwhile, other papers have emphasized that simply satisfying a particular definition of fairness is no guarantee of the broader outcomes people care about, such as justice (Hu and Chen, 2018b). Selbst et al. (2019) discuss five common "traps" in thinking about sociotechnical systems, and Friedler et al. (2019) demonstrate how outcomes differs depending on preprocessing and the choice of fairness metric.

Others have explored various types of consequences in particular settings, such as cost to the community in criminal justice (Corbett-Davies et al., 2017), runaway feedback loops in predictive policing (Ensign et al., 2018), disparities in the labor

[16]In a companion paper, Joseph et al. (2016a) proclaim their approach to be Rawlsian, but this seems to miss the key point of Rawls—namely, that we must account for inequalities due to circumstances (i.e., "regardless of their initial place in the social system"; Rawls, 1958). Rather, the approach of Joseph et al. (2016b) merely says we should learn to give loans to people who will best be able to pay them back.

market (Hu and Chen, 2018a), and the potential for strategic manipulation of policies (Hu et al., 2019; Milli et al., 2019). Liu et al. (2018) demonstrate the importance of modeling the delayed impact of adopting various fairness metrics, even when focused narrowly on outcomes such as demographic parity. In a discussion of racial bias in the criminal justice system, Huq (2019) uses broadly consequentialist logic, arguing that the systems should be evaluated in terms of costs and benefits to minority groups. For surveys discussing the intersection of ethics and AI more broadly, see Brundage (2014) and Yu et al. (2018). For a book-length treatment of the subject, see Wallach and Allen (2008).

9. CONCLUSIONS

Consequentialism represents one of the most important pillars of ethical thinking in philosophy, including (but not limited to) utilitarianism. In brief, the central tenet of consequentialism is that actions should be evaluated in terms of the relative goodness of the expected outcomes, according to an impartial perspective on what is best. Despite a number of serious problems that limit its practical application, including computational problems involving value, uncertainty, and discounting, consequentialism still provides a useful basis for thinking about the limitations of other normative frameworks.

Within the context of automated decision making, a consequentialist perspective underscores that merely satisfying a particular fairness metric is no guarantee of ethical conduct. Rather, consequentialism requires that we consider all possible options (including the possibility of not deploying an automated system), and weigh the likely consequences that will result, considered broadly, including possible implications for the long term future. Moreover, we must consider not only those who will be directly affected, but broader impacts on communities, and systemic effects of replacing many human decision makers with a single policy. While there are contexts in which it is reasonable, even required, to attempt to learn from the present for the benefit of the future, we should be skeptical of any randomization schemes which make unrealistic assumptions about the generalizability of what can be learned from social systems.

The political nature of valuation means we are unlikely to ever have agreement on what outcomes are best, and long term consequences will always remain to some extent unpredictable. Nevertheless, through ongoing efforts to take into consideration a diverse set of perspectives on value, and systematic attempts to learn from our experiences, we can strive to move toward policies which are likely to lead to a better world, over both the short and long term future.

AUTHOR CONTRIBUTIONS

DC conceived of the scope of this article. DC and NS contributed to the writing and editing of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Jared Moore, Emily Kalah Gade, Maarten Sap, Dan Hendrycks, and reviewers for their thoughtful feedback and comments on this work.

REFERENCES

Abel, D., MacGlashan, J., and Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. in *Proceedings of the Workshop on AI, Ethics, and Society at AAAI* (Phoenix, AZ).

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. *arXiv:1606.06565*.

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy* 33, 1–19. doi: 10.1017/S0031819100037943

Barabas, C., Virza, M., Dinakar, K., Ito, J., and Zittrain, J. (2018). "Interventions over predictions: Reframing the ethical debate for actuarial risk assessment," in *Proceedings of FAT** (New York, NY).

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. *California Law Rev*. 104, 671–732. doi: 10.2139/ssrn.2477899

Barrett, C. B., and Carter, M. R. (2010). The power and pitfalls of experiments in development economics: some non-random reflections. *Appl. Econ. Perspect. Policy* 32, 515–548. doi: 10.1093/aepp/ppq023

Benthall, S., and Haynes, B. D. (2019). "Racial categories in machine learning," in *Proceedings of FAT** (Atlanta, GA). doi: 10.1145/3287560.3287575

Bentham, J. (1970 [1781]). *An Introduction to the Principles of Morals and Legislation*. Oxford: Oxford University Press.

Bianco, K. M. (2008). *The Subprime Lending Crisis: Causes and Effects of the Mortgage Meltdown*. CCH, 1–21. Available online at: business.cch.com/images/banner/subprime.pdf

Binmore, K. (2009). "Chapter 20: Interpersonal comparison of utility," in *The Oxford Handbook of Philosophy of Economics*, editors D. Ross and H. Kincaid (New York, NY: Oxford University Press). doi: 10.1093/oxfordhb/9780195189254.003.0020

Binns, R. (2018). "Fairness in machine learning: Lessons from political philosophy," in *Proceedings of FAT** (New York, NY).

Bird, S., Barocas, S., Crawford, K., and Wallach, H. (2016). "Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI," in *Proceedings of FAT/ML* (New York, NY).

Brundage, M. (2014). Limitations and risks of machine ethics. *J. Exp. Theor. Artif. Intell.* 26, 355–372. doi: 10.1080/0952813X.2014.895108

Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *arXiv:1610.07524*. doi: 10.1089/big.2016.0047

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). "Algorithmic decision making and the cost of fairness," in *Proceedings of KDD* (Halifax, NS). doi: 10.1145/3097983.3098095

Cowen, T. (2006). The epistemic problem does not refute consequentialism. *Utilitas* 18, 383–399. doi: 10.1017/S0953820806002172

Cowen, T., and Parfit, D. (1992). "Against the social discount rate," in *Philosophy, Politics, and Society*, eds P. Laslett and J. Fishkin (New Haven, CT: Yale University Press), 144–161. doi: 10.2307/j.ctt211qw3x.11

de Lazari-Radek, K., and Singer, P. (2017). *Utilitarianism: A Very Short Introduction*. New York, NY: Oxford University Press. doi: 10.1093/actrade/9780198728795.001.0001

Driver, J. (2005). Consequentialism and feminist ethics. *Hypatia* 20, 183–199. doi: 10.1111/j.1527-2001.2005.tb00543.x

Driver, J. (2012). *Consequentialism*. New York, NY: Routledge. doi: 10.4324/9780203149256

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). "Fairness through awareness," in *Proceedings of ITCS* (Cambridge, MA). doi: 10.1145/2090236.2090255

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). "Runaway feedback loops in predictive policing," in *Proceedings of FAT** (New York, NY).

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Rev*. 5, 5–15.

Freedman, B. (1987). Equipoise and the ethics of clinical research. *N. Engl. J. Med*. 317, 141–145. doi: 10.1056/NEJM198707163170304

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 FAT** (Atlanta, GA). doi: 10.1145/3287560.3287589

Friedman, M. (1991). The practice of partiality. *Ethics* 101, 818–835. doi: 10.1086/293345

Green, B. (2019). "Good" isn't good enough," in *Proceedings of the AI for Social Good workshop at NeurIPS* (Vancouver, BC).

Green, B., and Hu, L. (2018). "The myth in the methodology: towards a recontextualization of fairness in machine learning," in *Proceedings of the Debates workshop at ICML* (Stockholm).

Greene, J., and Haidt, J. (2002). How (and where) does moral judgment work? *Trends Cogn. Sci*. 6, 517–523. doi: 10.1016/S1364-6613(02)02011-9

Greene, J. D. (2013). *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York, NY: The Penguin Press.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2016). "The case for process fairness in learning: feature selection for fair decision making," in *Proceedings of the Symposium on Machine Learning and the Law at NeurIPS* (Barcelona).

Hardt, M., Price, E., and Srebro, N. (2016). "Equality of opportunity in supervised learning," in *Proceedings of NeurIPS* (Barcelona).

Harsanyi, J. C. (1977). Rule utilitarianism and decision theory. *Erkenntnis* 11, 25–53. doi: 10.1007/BF00169843

Haushofer, J., Riis-Vestergaard, M. I., and Shapiro, J. (2019). Is there a social cost of randomization? *Soc. Choice Welfare* 52, 709–739. doi: 10.1007/s00355-018-1168-7

Hoffmann, A. L. (2017). Beyond distributions and primary goods: assessing applications of Rawls in information science and technology literature since 1990. *J. Assoc. Inform. Sci. Technol*. 68, 1601–1618. doi: 10.1002/asi.23747

Hooker, B. (2002). *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. New York, NY: Clarendon Press. doi: 10.1093/0199256578.001.0001

Hu, L., and Chen, Y. (2018a). "A short-term intervention for long-term fairness in the labor market," in *Proceedings of WWW* (Lyon). doi: 10.1145/3178876.3186044

Hu, L., and Chen, Y. (2018b). "Welfare and distributional impacts of fair classification," in *Proceedings of FAT/ML* (Stockholm).

Hu, L., Immorlica, N., and Vaughan, J. W. (2019). "The disparate effects of strategic manipulation," in *Proceedings of FAT** (Atlanta, GA). doi: 10.1145/3287560.3287597

Huq, A. Z. (2019). Racial equity in algorithmic criminal justice. *Duke Law J*. 68, 1043–1134. Available online at: https://ssrn.com/abstract=3144831

Joseph, M., Kearns, M., Morgenstern, J. H., Neel, S., and Roth, A. (2016a). "Rawlsian fairness for machine learning," in *Proceedings of FAT/ML* (New York, NY).

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. (2016b). "Fairness in learning: classic and contextual bandits," in *Proceedings of NeurIPS* (Barcelona).

Kagan, S. (1991). *The Limits of Morality*. New York, NY: Claredon Press. doi: 10.1093/0198239165.001.0001

Kagan, S. (1998). *Normative Ethics*. Boulder, CO: Westview Press.

Khademi, A., Lee, S., Foley, D., and Honavar, V. (2019). "Fairness in algorithmic decision making: An excursion through the lens of causality," in *Proceedings of WWW* (San Francisco, CA). doi: 10.1145/3308558.3313559

Kittay, E. F. (2009). "chapter 8: The ethics of philosophizing: Ideal theory and the exclusion of people with severe cognitive disabilities," in *Feminist Ethics and Social and Political Philosophy: Theorizing the Non-Ideal*, ed L. Tessman (New York, NY: Springer), 121–146. doi: 10.1007/978-1-4020-6841-6_8

Kleinberg, J. M., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018). Algorithmic fairness. *AEA Papers Pro*. 108, 22–27. doi: 10.1257/pandp.20181018

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2017). "Inherent trade-offs in the fair determination of risk scores," in *Proceedings of ITCS* (Berkeley, CA).

Kleindessner, M., Samadi, S., Awasthi, P., and Morgenstern, J. (2019). "Guarantees for spectral clustering with fairness constraints," in *Proceedings of ICML* (Long Beach, CA).

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2017). Accountable algorithms. *Univers. Pennsylvania Law Rev*. 165, 633–705. Available online at: https://ssrn.com/abstract=2765268

Lai, T. L., Lavori, P. W., and Tsang, K. W. (2015). Adaptive design of confirmatory trials: Advances and challenges. *Contemp. Clin. Trials* 45, 93–102. doi: 10.1016/j.cct.2015.06.007

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). "Delayed impact of fair machine learning," in *Proceedings of ICML* (Stockholm). doi: 10.24963/ijcai.2019/862

Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., and Parkes, D. C. (2017). "Calibrated fairness in bandits," in *Proceedings of FAT/ML* (Halifax, NS).

Lockwood, M., and Anscombe, G. E. M. (1983). Sins of omission? The non-treatment of controls in clinical trials. *Aristotel. Soc. Suppl*. 57, 207–227. doi: 10.1093/aristoteliansupp/57.1.207

Madras, D., Pitassi, T., and Zemel, R. (2018). "Predict responsibly: Improving fairness and accuracy by learning to defer," in *Proceedings of NeurIPS* (Montreal, QC).

Mill, J. S. (1979[1863]). *Utilitarianism*. Indianapolis, IN: Hackett Publishing Company, Inc.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. (2019). "The social cost of strategic classification," in *Proceedings of FAT** (Atlanta, GA). doi: 10.1145/3287560.3287576

Mills, C. W. (1987). *The Racial Contract*. Ithaca, NY: Cornell University Press.

Parfit, D. (1984). *Reasons and Persons*. New York, NY: Oxford University Press.

Perry, R., and Zarsky, T. (2015). "May the odds be ever in your favour": Lotteries in law. *Alabama Law Rev*. 66, 1035–1098. doi: 10.2139/ssrn.2494550

Portmore, D. W. (2011). *Commonsense Consequentialism*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780199794539.003.0007

Railton, P. (1984). Alientation, consequentailism, and the demands of moraltiy. *Philos. Public Affairs* 13, 134–171.

Rawls, J. (1958). Justice as fairness. *Philos. Rev*. 67. doi: 10.2307/2182612

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: The Belknap Press of Harvard University.

Ruggieri, S., Pedreschi, D., and Turini, F. (2010). Data mining for discrimination discovery. *ACM Trans. Knowl. Discov. Data* 4, 9:1–9:40. doi: 10.1145/1754428.1754432

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scheffler, S. (1994). *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. New York, NY: Oxford University Press.

Schultz, B., and Varouxakis, G. (eds.). (2005). *Utilitarianism and Empire*. Oxford: Lexington Books.

Selbst, A. D., boyd d., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). "Fairness and abstraction in sociotechnical systems," in *Proceedings of FAT** (Atlanta, GA). doi: 10.1145/3287560.3287598

Sidgwick, H. (1967). *The Method of Ethics*. New York, NY: Macmillon.

Singer, P. (1972). Famine, affluence, and morality. *Philos. Public Affairs* 1, 229–243.

Singer, P. (1993). *Practical Ethics*. New York, NY: Cambridge University Press.

Smart, J. J. C., and Williams, B. (1973). *Utilitarianism: For & Against*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511840852

Torresen, J. (2018). A review of future and ethical perspectives of robotics and AI. *Front. Robot. AI* 4:75. doi: 10.3389/frobt.2017.00075

Wallach, W., and Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. New York, NY: Oxford University Press, Inc.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., and Yang, Q. (2018). "Building ethics into artificial intelligence," in *Proceedings of IJCAI* (Stockholm). doi: 10.24963/ijcai.2018/779

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.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 12 May 2020
doi: 10.3389/frai.2020.00033

Tuning Fairness by Balancing Target Labels

Thomas Kehrenberg[1]*, *Zexun Chen*[1†] *and Novi Quadrianto*[1,2]

[1] *Predictive Analytics Lab (PAL), Informatics, University of Sussex, Brighton, United Kingdom,* [2] *National Research University Higher School of Economics, Moscow, Russia*

The issue of fairness in machine learning models has recently attracted a lot of attention as ensuring it will ensure continued confidence of the general public in the deployment of machine learning systems. We focus on mitigating the harm incurred by a biased machine learning system that offers better outputs (e.g., loans, job interviews) for certain groups than for others. We show that bias in the output can naturally be controlled in probabilistic models by introducing a latent target output. This formulation has several advantages: first, it is a unified framework for several notions of group fairness such as Demographic Parity and Equality of Opportunity; second, it is expressed as a marginalization instead of a constrained problem; and third, it allows the encoding of our knowledge of what unbiased outputs should be. Practically, the second allows us to avoid unstable constrained optimization procedures and to reuse off-the-shelf toolboxes. The latter translates to the ability to control the level of fairness by directly varying fairness target rates. In contrast, existing approaches rely on intermediate, arguably unintuitive, control parameters such as covariance thresholds.

Keywords: algorithmic bias, fairness, machine learning, demographic parity, equality of opportunity

OPEN ACCESS

Edited by:
Fabrizio Riguzzi,
University of Ferrara, Italy

Reviewed by:
Yunfeng Zhang,
IBM Research, United States
Abeer Dyoub,
University of L'Aquila, Italy

****Correspondence:***
Thomas Kehrenberg
t.kehrenberg@sussex.ac.uk

†Present address:
Zexun Chen,
BioComplex Laboratory, Computer Science, University of Exeter, Exeter, United Kingdom

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

Received: *18 February 2020*
Accepted: *15 April 2020*
Published: *12 May 2020*

Citation:
Kehrenberg T, Chen Z and Quadrianto N (2020) Tuning Fairness by Balancing Target Labels. Front. Artif. Intell. 3:33. doi: 10.3389/frai.2020.00033

1. INTRODUCTION

Algorithmic assessment methods are used for predicting human outcomes in areas such as financial services, recruitment, crime and justice, and local government. This contributes, in theory, to a world with decreasing human biases. To achieve this, however, we need fair machine learning models that take biased datasets, but output non-discriminatory decisions to people with differing protected attributes such as gender and marital status. Datasets can be biased because of, for example, sampling bias, subjective bias of individuals, and institutionalized biases (Olteanu et al., 2019; Tolan, 2019). Uncontrolled bias in the data can translate into bias in machine learning models.

There is no single accepted definition of algorithmic fairness for automated decision-making but several have been proposed. One definition is referred to as *statistical* or *demographic parity*. Given a binary protected attribute (e.g., married/unmarried) and a binary decision (e.g., yes/no to getting a loan), demographic parity requires equal positive rates (PR) across the two sensitive groups (married and *un*married individuals should be equally likely to receive a loan). Another fairness criterion, *equalized odds* (Hardt et al., 2016), takes into account the binary decision, and instead of equal PR requires equal true positive rates (TPR) and false positive rates (FPR). This criterion is intended to be more compatible with the goal of building accurate predictors or achieving high utility (Hardt et al., 2016). We discuss the suitability of the different fairness criteria in the discussion section at the end of the paper.

There are many existing models for enforcing demographic parity and equalized odds (Calders et al., 2009; Kamishima et al., 2012; Zafar et al., 2017a,b; Agarwal et al., 2018; Creager et al., 2019).

However, these existing approaches to balancing accuracy and fairness rely on intermediate, unintuitive control parameters such as allowable constraint violation ϵ (e.g., 0.01) in Agarwal et al. (2018), or a covariance threshold c (e.g., 0 that is controlled by another parameters τ and μ – 0.005 and 1.2 – to trade off this threshold and accuracy) in Zafar et al. (2017a). This is related to the fact that many of these approaches embed fairness criteria as *constraints* in the optimization procedure (Quadrianto and Sharmanska, 2017; Zafar et al., 2017a,b; Donini et al., 2018).

In contrast, we provide a probabilistic classification framework with bias controlling mechanisms that can be tuned based on positive rates (PR), an intuitive parameter. Thus, giving humans the control to set the rate of positive predictions (e.g., a PR of 0.6). Our framework is based on the concept of a *balanced dataset* and introduces latent target labels, which, instead of the provided labels, are now the training label of our classifier. We prove bounds on how far the target labels diverge from the dataset labels. We instantiate our approach with a parametric logistic regression classifier and a Bayesian non-parametric Gaussian process classifier (GPC). As our formulation is not expressed as a constrained problem, we can draw upon advancements in automated variational inference (Bonilla et al., 2016; Krauth et al., 2016; Gardner et al., 2018) for learning the fair model, and for handling large amounts of data.

The method presented in this paper is closely related to a number of previous works, e.g., Calders and Verwer, 2010; Kamiran and Calders, 2012. Proper comparison with them requires knowledge of our approach. We will thus explain our approach in the subsequent sections, and defer detailed comparisons to section 4.

2. TARGET LABELS FOR TUNING GROUP FAIRNESS

We will start by describing several notions of group fairness. For each individual, we have a vector of non-sensitive attributes $x \in \mathcal{X}$, a class label $y \in \mathcal{Y}$, and a sensitive attribute $s \in \mathcal{S}$ (e.g., racial origin or gender). We focus on the case where s and y are binary. We assume that a positive label $y = 1$ corresponds to a positive outcome for an individual—for example, being accepted for a loan. *Group fairness* balances a certain condition between groups of individuals with different sensitive attribute, s vs. s'. The term $\hat{y}$ below is the prediction of a machine learning model that, in most works, uses only non-sensitive attributes x. Several group fairness criteria have been proposed (e.g., Hardt et al., 2016; Chouldechova, 2017; Zafar et al., 2017a):

Equality of positive rate (Demographic Parity):

$$\Pr(\hat{y} = 1|s) = \Pr(\hat{y} = 1|s') \tag{1}$$

Equality of accuracy:

$$\Pr(\hat{y} = y|s) = \Pr(\hat{y} = y|s') \tag{2}$$

Equality of true positive rate (Equality of Opportunity):

$$\Pr(\hat{y} = 1|s, y = 1) = \Pr(\hat{y} = 1|s', y = 1)\,. \tag{3}$$

Equalized odds criterion corresponds to Equality of Opportunity (3) plus equality of false positive rate.

The Bayes-optimal classifier only satisfies these criteria if the training data itself satisfies them. That is, in order for the Bayes-optimal classifier to satisfy *demographic parity*, the following must hold: $\mathbb{P}(y = 1|s) = \mathbb{P}(y = 1|s')$, where y is the training label. We call a dataset for which $\mathbb{P}(y, s) = \mathbb{P}(y)\mathbb{P}(s)$ holds, a *balanced* dataset. Given a balanced dataset, a Bayes-optimal classifier learns to satisfy demographic parity and an approximately Bayes-optimal classifier should learn to satisfy it at least approximately. Here, we motivated the importance of balanced datasets via the demographic parity criterion, but it is also important for *equality of opportunity* which we discuss in section 2.1.

In general, however, our given dataset is likely to be imbalanced. There are two common solutions to this problem: either pre-process or massage the dataset to make it balanced, or constrain the classifier to give fair predictions despite it having been trained on an unbalanced dataset. Our approach takes parts from both solutions.

An imbalanced dataset can be turned into a balanced dataset by either changing the class labels y or the sensitive attributes s. In the use cases that we are interested in, s is considered an integral part of the input, representing trustworthy information and thus should not be changed. y, conversely, is often not completely trustworthy; it is not an integral part of the sample but merely an observed outcome. In a hiring dataset, for instance, y might represent the hiring decision, which can be biased, and not the relevant question of whether someone makes a good employee.

Thus, we introduce new *target labels* $\bar{y}$ such that the dataset is balanced: $\mathbb{P}(\bar{y}, s) = \mathbb{P}(\bar{y})\mathbb{P}(s)$. The idea is that these target labels still contain as much information as possible about the task, while also forming a balanced dataset. This introduces the concept of the accuracy-fairness trade-off: in order to be completely accurate with respect to the original (not completely trustworthy) class labels y, we would require $\bar{y} = y$, but then, the fairness constraints would not be satisfied.

Let $\eta_s(x) = \mathbb{P}(y = 1|x, s)$ denote the distribution of y in the data. The target distribution $\bar{\eta}_s(x) = \mathbb{P}(\bar{y} = 1|x, s)$ is then given by

$$\begin{aligned}\bar{\eta}_s(x) = (&\mathbb{P}(\bar{y} = 1|y = 1, s) + \mathbb{P}(\bar{y} = 0|y = 0, s) - 1)\eta_s(x) \\ &+ 1 - \mathbb{P}(\bar{y} = 0|y = 0, s)\end{aligned} \tag{4}$$

due to the marginalization rules of probabilities. The conditional probability $\mathbb{P}(\bar{y}|y, s)$ indicates with which probability we want to keep the class label. This probability could in principle depend on x which would enable the realization of individual fairness. The dependence on x has to be prior knowledge as it cannot be learned from the data. This prior knowledge can encode the semantics that "similar individuals should be treated similarly" (Dwork et al., 2012), or that "less qualified individuals should not be preferentially favored over more qualified individuals" (Joseph et al., 2016). Existing proposals for guaranteeing individual fairness require strong assumptions, such as the availability of an agreed-upon similarity metric, or knowledge of the underlying data generating process. In contrast, in group fairness, we partition individuals into protected groups based on some sensitive attribute s and ask that some statistics of a classifier be approximately equalized across those groups (see Equations 1–3). In this case, $\mathbb{P}(\bar{y}|y, s)$ does not depend on x.

Returning to Equation (4), we can simplify it with

$$m_s := \mathbb{P}(\bar{y} = 1|y = 1, s) + \mathbb{P}(\bar{y} = 0|y = 0, s) - 1 \quad (5)$$

$$b_s := 1 - \mathbb{P}(\bar{y} = 0|y = 0, s)\,, \quad (6)$$

arriving at $\bar{\eta}_s(x) = m_s \cdot \eta_s(x) + b_s$. m_s and b_s are chosen such that $\mathbb{P}(\bar{y}, s) = \mathbb{P}(\bar{y})\mathbb{P}(s)$. This can be interpreted as shifting the decision boundary depending on s so that the new distribution is balanced.

As there is some freedom in choosing m_s and b_s, it is important to consider what the effect of different values is. The following theorem provides this (the proof can be found in the **Supplementary Material**):

Theorem 1. *The probability that y and $\bar{y}$ disagree ($y \neq \bar{y}$) for any input x in the dataset is given by:*

$$\mathbb{P}(y \neq \bar{y}|s) = \mathbb{P}\left(\left|\eta(x, s) - \tfrac{1}{2}\right| < t_s\right) \quad (7)$$

where

$$t_s = \left|\frac{m_s + 2b_s - 1}{2m_s}\right|\,. \quad (8)$$

Thus, if the threshold t_s is small, then only if there are inputs very close to the decision boundary ($\eta_s(x)$ close to $\frac{1}{2}$) would we have $\bar{y} \neq y$. t_s determines the accuracy penalty that we have to accept in order to gain fairness. The value of t_s can be taken into account when choosing m_s and b_s (see section 3). If η_s satisfies the Tsybakov condition (Tsybakov et al., 2004), then we can give an upper bound for the probability.

Definition 1. *A distribution η satisfies the Tsybakov condition if there exist $C > 0$, $\lambda > 0$ and $t_0 \in (0, \frac{1}{2}]$ such that for all $t \leq t_0$,*

$$\mathbb{P}\left(\left|\eta(x) - \tfrac{1}{2}\right| < t\right) \leq Ct^{\lambda}\,. \quad (9)$$

This condition bounds the region close to the decision boundary. It is a property of the dataset.

Corollary 1.1. *If $\eta(x, s) = \mathbb{P}(y = 1|x, s)$ satisfies the Tsybakov condition in x, with constants C and λ, then the probability that y and $\bar{y}$ disagree ($y \neq \bar{y}$) for any input x in the dataset is bounded by:*

$$\mathbb{P}(y \neq \bar{y}|s) < C\left|\frac{m_s + 2b_s - 1}{2m_s}\right|^{\lambda}\,. \quad (10)$$

Section 3 discusses how to choose the parameters for $\bar{\eta}$ in order to make it balanced.

2.1. Equality of Opportunity

In contrast to demographic parity, equality of opportunity (just as equality of accuracy) is satisfied by a perfect classifier. Imperfect classifiers, however, do not by default satisfy it: the true positive rate (TPR) is different for different subgroups. The reason for this is that while the classifier is optimized to have a high TPR overall, it is not optimized to have the same TPR in the subgroups.

The overall TPR is a weighted sum of the TPRs in the subgroups:

$$TPR = \mathbb{P}(s = 0|y = 1) \cdot TPR_{s=0} + \mathbb{P}(s = 1|y = 1) \cdot TPR_{s=1}\,. \quad (11)$$

In datasets where the positive label $y = 1$ is heavily skewed toward one of the groups (say, group $s = 1$; meaning that $\mathbb{P}(s = 1|y = 1)$ is high and $\mathbb{P}(s = 0|y = 1)$ is low), overall TPR might be maximized by setting the decision boundary such that nearly all samples in $s = 0$ are classified as $y = 0$, while for $s = 1$ a high TPR is achieved. The low TPR for $s = 0$ is in this case weighted down and only weakly impacts the overall TPR. For $s = 0$, the resulting classifier uses s as a shorthand for y, mostly ignoring the other features. This problem usually persists even when s is removed from the input features because s is implicit in the other features.

A *balanced* dataset helps with this issue because in such datasets, s is not a useful proxy for the balanced label $\bar{y}$ (because we have $\mathbb{P}(\bar{y}, s) = \mathbb{P}(\bar{y})\mathbb{P}(s)$) and s cannot be used as a shorthand. Assuming the dataset is balanced in s ($\mathbb{P}(s = 0) = \mathbb{P}(s = 1)$), for such datasets $\mathbb{P}(s = 0|y = 1) = \mathbb{P}(s = 1|y = 1)$ holds and the two terms in Equation (11) have equal weight.

Here as well there is an accuracy-fairness trade-off: assuming the unconstrained model is as accurate as its model complexity allows, adding additional constraints like equality of opportunity can only make the accuracy worse.

2.2. Concrete Algorithm

For training, we are only given the unbalanced distribution $\eta_s(x)$ and not the target distribution $\bar{\eta}_s(x)$. However, $\bar{\eta}_s(x)$ is needed in order to train a fair classifier. One approach is to explicitly change the labels y in the dataset, in order to construct $\bar{\eta}_s(x)$. We discuss this approach and its drawback in the related work section (section 4).

We present a novel approach which only implicitly constructs the balanced dataset. This framework can be used with any likelihood-based model, such as Logistic Regression and Gaussian Process models. The relation presented in Equation (4) allows us to formulate a likelihood that targets $\bar{\eta}_s(x)$ while only having access to the imbalanced labels y. As we only have access to y, $\mathbb{P}(y|x, s, \theta)$ is the likelihood to optimize. It represents the probability that y is the imbalanced label, given the input x, the sensitive attribute s that available in the training set and the model parameters θ for a model that is targeting $\bar{y}$. Thus, we get

$$\begin{aligned}\mathbb{P}(y = 1|x, s, \theta) &= \sum_{\bar{y}\in\{0,1\}} \mathbb{P}(y = 1, \bar{y}|x, s, \theta) \\ &= \sum_{\bar{y}\in\{0,1\}} \mathbb{P}(y = 1|\bar{y}, x, s, \theta)\,\mathbb{P}(\bar{y}|x, s, \theta)\,. \end{aligned} \quad (12)$$

As we are only considering group fairness, we have $\mathbb{P}(y = 1|\bar{y}, x, s, \theta) = \mathbb{P}(y = 1|\bar{y}, s)$.

Let $f_\theta(x, y')$ be the likelihood function of a given model, where f gives the likelihood of the label y' given the input x and the model parameters θ. As we do not want to make use of s at test time, f does not explicitly depend on s. The likelihood

with respect to $\bar{y}$ is then given by f: $\mathbb{P}(\bar{y}|x,s,\theta) = f_\theta(x,\bar{y})$; and thus, does not depend on s. The latter is important in order to avoid *direct discrimination* (Barocas and Selbst, 2016). With these simplifications, the expression for the likelihood becomes

$$\mathbb{P}(y=1|x,s,\theta) = \sum_{\bar{y}\in\{0,1\}} \mathbb{P}(y=1|\bar{y},s)\,\mathbb{P}(\bar{y}|x,\theta)\,. \tag{13}$$

The conditional probabilities, $\mathbb{P}(y|\bar{y},s)$, are closely related to the conditional probabilities in Equation (4) and play a similar role of "transition probabilities." Section (1) explains how to choose these transition probabilities in order to arrive at a balanced dataset. For a binary sensitive attribute s (and binary label y), there are 4 transition probabilities (see Algorithm 1 where $d^{s=j}_{\bar{y}=i} := \mathbb{P}(y=1|\bar{y}=i,s=j)$):

$$\mathbb{P}(y=1|\bar{y}=0,s=0), \qquad \mathbb{P}(y=1|\bar{y}=1,s=0) \tag{14}$$

$$\mathbb{P}(y=1|\bar{y}=0,s=1), \qquad \mathbb{P}(y=1|\bar{y}=1,s=1)\,. \tag{15}$$

A perhaps useful interpretation of Equation (13) is that, even though we don't have access to $\bar{y}$ directly, we can still compute the expectation value over the possible values of $\bar{y}$.

The above derivation applies to binary classification but can easily be extended to the multi-class case.

Algorithm 1: Fair learning with target labels $\bar{y}$

Input: Training set $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^N$, transition probabilities $d^{s=0}_{\bar{y}=0}$, $d^{s=0}_{\bar{y}=1}$, $d^{s=1}_{\bar{y}=0}$, $d^{s=1}_{\bar{y}=1}$
Output: Fair model parameters θ

```
Initialize θ (randomly)
for all x_i, y_i, s_i do
    P_{ȳ=1} ← η̄(x_i, θ) (e.g., logistic(⟨x, θ⟩))
    P_{ȳ=0} ← 1 − P_{ȳ=1}
    if s_i = 0 then
        P_{y=1} ← d^{s=0}_{ȳ=0} · P_{ȳ=0} + d^{s=0}_{ȳ=1} · P_{ȳ=1}
    else
        P_{y=1} ← d^{s=1}_{ȳ=0} · P_{ȳ=0} + d^{s=1}_{ȳ=1} · P_{ȳ=1}
    end if
    ℓ ← y_i · P_{y=1} + (1 − y_i) · (1 − P_{y=1})
    update θ to maximize likelihood ℓ
end for
```

3. TRANSITION PROBABILITIES FOR A BALANCED DATASET

This section focuses on how to set values of the transition probabilities in order to arrive at balanced datasets.

3.1. Meaning of the Parameters

Before we consider concrete values, we give some intuition for the transition probabilities. Let $s=0$ refer to the protected group. For this group, we want to make more positive predictions than the training labels indicate. Variable $\bar{y}$ is supposed to be our target proxy label. Thus, in order to make more positive predictions, some of the $y=0$ labels should be associated with $\bar{y}=1$. However, we do not know which. So, if our model predicts $\bar{y}=1$ (high $\mathbb{P}(\bar{y}=1|x,\theta)$) while the training label is $y=0$, then we allow for the possibility that this is actually correct. That is, $\mathbb{P}(y=0|\bar{y}=1,s=0)$ is not 0. If we choose, for example, $\mathbb{P}(y=0|\bar{y}=1,s=0)=0.3$ then that means that 30% of positive target labels $\bar{y}=1$ may correspond to negative training labels $y=0$. This way we can have more $\bar{y}=1$ than $y=1$, overall. On the other hand, predicting $\bar{y}=0$ when $y=1$ holds, will always be deemed incorrect: $\mathbb{P}(y=1|\bar{y}=0,s=0)=0$; this is because we do not want any additional negative labels.

For the non-protected group $s=1$, we have the exact opposite situation. If anything, we have too many positive labels. So, if our model predicts $\bar{y}=0$ (high $\mathbb{P}(\bar{y}=0|x,\theta)$) while the training label is $y=1$, then we should again allow for the possibility that this is actually correct. That is, $\mathbb{P}(y=1|\bar{y}=0,s=1)$ should not be 0. On the other hand, $\mathbb{P}(y=0|\bar{y}=1,s=1)$ should be 0 because we do not want additional positive labels for $s=1$. It could also be that the number of positive labels is exactly as it should be, in which case we can just set $y=\bar{y}$ for all data points with $s=1$.

3.2. Choice of Parameters

A balanced dataset is characterized by an independence of the label $\bar{y}$ and the sensitive attribute s. Given that we have complete control over the *transition probabilities*, we can ensure this independence by requiring $\mathbb{P}(\bar{y}=1|s=0) = \mathbb{P}(\bar{y}=1|s=1)$. Our constraint is then that both of these probabilities are equal to the same value, which we will call the target rate PR_t ("PR" as *positive rate*):

$$\mathbb{P}(\bar{y}=1|s=0) \overset{!}{=} PR_t \quad \text{and} \quad \mathbb{P}(\bar{y}=1|s=1) \overset{!}{=} PR_t\,. \tag{16}$$

This leads us to the following constraints for $s' \in \{0,1\}$:

$$PR_t = \mathbb{P}(\bar{y}=1|s=s') = \sum_y \mathbb{P}(\bar{y}=1|y,s=s')\,\mathbb{P}(y|s=s'). \tag{17}$$

We call $\mathbb{P}(y=1|s=j)$ the base rate PR^j_b which we estimate from the training set:

$$\mathbb{P}(y=1|s=i) = \frac{\text{number of points with } y=1 \text{ in group } i}{\text{number of points in group } i}\,.$$

Expanding the sum, we get

$$\begin{aligned} PR_t = {} & \mathbb{P}(\bar{y}=1|y=0,s=s')\cdot(1-PR^1_b) \\ & + \mathbb{P}(\bar{y}=1|y=1,s=s')\cdot PR^1_b\,. \end{aligned} \tag{18}$$

This is a system of linear equations consisting of two equations (one for each value of s') and four free variables: $\mathbb{P}(\bar{y}=1|y,s)$ with $y,s \in \{0,1\}$. The two unconstrained degrees of freedom determine how strongly the accuracy will be affected by the fairness constraint. If we set $\mathbb{P}(\bar{y}=1|y=1,s)$ to 0.5, then this expresses the fact that a train label y of 1 only implies a target label $\bar{y}$ of 1 in 50% of the cases. In order to minimize the effect on accuracy, we make $\mathbb{P}(\bar{y}=1|y=1,s)$ as high as possible

and $\mathbb{P}(\bar{y} = 1|y = 0, s)$, conversely, as low as possible. However, the lowest and highest possible values are not always 0 and 1 respectively. To see this, we solve for $\mathbb{P}(\bar{y} = 1|y = 0, s = j)$ in Equation (18):

$$\mathbb{P}(\bar{y} = 1|y = 0, s = j) = \frac{PR_b^j}{1 - PR_b^j} \left(\frac{PR_t}{PR_b^j} - \mathbb{P}(\bar{y} = 1|y = 1, s = j) \right) . \quad (19)$$

If PR_t/PR_b^j were greater than 1, then setting $\mathbb{P}(\bar{y} = 1|y = 0, s = j)$ to 0 would imply a $\mathbb{P}(\bar{y} = 1|y = 1, s = j)$ value greater than 1. A visualization that shows why this happens can be found in the **Supplementary Material**. We thus arrive at the following definitions:

$$\mathbb{P}(\bar{y} = 1|y = 1, s = j) = \begin{cases} 1 & \text{if } PR_t > PR_b^j \\ \frac{PR_t}{PR_b^j} & \text{otherwise.} \end{cases} \quad (20)$$

$$\mathbb{P}(\bar{y} = 1|y = 0, s = j) = \begin{cases} \frac{PR_t - PR_b^j}{1 - PR_b^j} & \text{if } PR_t > PR_b^j \\ 0 & \text{otherwise.} \end{cases} \quad (21)$$

Algorithm 2 shows pseudocode of the procedure, including the computation of the allowed minimal and maximal value.

Once all these probabilities have been found, the transition probabilities needed for Equation (13) are fully determined by applying Bayes' rule:

$$\mathbb{P}(y = 1|\bar{y}, s) = \frac{\mathbb{P}(\bar{y}|y = 1, s)\mathbb{P}(y = 1|s)}{\mathbb{P}(\bar{y}|s)} . \quad (22)$$

3.2.1. Choosing a Target Rate

As shown, there is a remaining degree of freedom when targeting a balanced dataset: the target rate $PR_t := \mathbb{P}(\bar{y} = 1)$. This is true for both fairness criteria that we are targeting. The choice of targeting rate affects how much η and $\bar{\eta}$ differ as implied by Theorem 1 (PR_t affects m_s and b_s). $\bar{\eta}$ should remain close to η as $\bar{\eta}$ only represents an auxiliary distribution that does not have meaning on its own. The threshold t_s in Theorem 1 (Equation 8) gives an indication of how close the distributions are. With the definitions in Equations (20) and (21), we can express t_s in terms of the target rate and the base rate:

$$t_s = \begin{cases} \frac{1}{2} \frac{PR_b^s - PR_t}{PR_t} & \text{if } PR_t > PR_b^j \\ \frac{1}{2} \frac{PR_t - PR_b^s}{1 - PR_t} & \text{otherwise.} \end{cases} \quad (23)$$

This shows that t_s is smallest when PR_b^s and PR_t are closest. However, as PR_b^s has different values for different s, we cannot set $PR_b^s = PR_t$ for all s. In order to keep both $t_{s=0}$ and $t_{s=1}$ small, it follows from Equation (23) that PR_t should at least be between PR_b^0 and PR_b^1. A more precise statement can be made when we explicitly want to minimize the sum $t_{s=0} + t_{s=1}$: assuming $PR_b^0 < PR_t < PR_b^1$ and $PR_b^1 < \frac{1}{2}$, the optimal choice for PR_t is PR_b^1 (see **Supplementary Material** for details). We call this choice PR_t^{max}. For $PR_b^0 > \frac{1}{2}$, analogous statements can be made, but this is of less interest as this case does not appear in our experiments.

The previous statements about t_s do not directly translate into observable quantities like accuracy if the Tsybakov condition is not satisfied, and even if it is satisfied, the usefulness depends on the constants C and λ. Conversely, the following theorem makes *generally* applicable statement about the accuracy that can be achieved. Before we get to the theorem, we introduce some notation. We are given a dataset $\mathcal{D} = \{(x_i, y_i)\}_i$, where the x_i are vectors of features and the y_i the corresponding labels. We refer to the tuples (x, y) as the *samples* of the dataset. The number of samples is $N = |\mathcal{D}|$.

We assume binary labels ($y \in \{0, 1\}$) and thus can form the (disjoint) subsets $\mathcal{Y}^0$ and $\mathcal{Y}^1$ with

$$\mathcal{Y}^j = \{(x, y) \in \mathcal{D}|y = j\} \quad \text{with } j \in \{0, 1\} . \quad (24)$$

Furthermore, we associate each sample with a classification $\hat{y} \in \{0, 1\}$. The task of making the classification $\hat{y} = 0$ or $\hat{y} = 1$ can be understood as sorting each sample from $\mathcal{D}$ into one of two sets: $\mathcal{C}^0$ and $\mathcal{C}^1$, such that $\mathcal{C}^0 \cup \mathcal{C}^1 = \mathcal{D}$ and $\mathcal{C}^0 \cap \mathcal{C}^1 = \emptyset$.

We refer to the set $\mathcal{A} = (\mathcal{C}^0 \cap \mathcal{Y}^0) \cup (\mathcal{C}^1 \cap \mathcal{Y}^1)$ as the set of correct (or accurate) predictions. The *accuracy* is given by $acc = N^{-1} \cdot |\mathcal{A}|$.

Definition 2.

$$r_a := \frac{|\mathcal{Y}^1|}{|\mathcal{D}|} = \frac{|\mathcal{Y}^1|}{N} \quad (25)$$

is called the *base acceptance rate* of the dataset $\mathcal{D}$.

Definition 3.

$$\hat{r}_a = \frac{|\mathcal{C}^1|}{|\mathcal{D}|} = \frac{|\mathcal{C}^1|}{N} \quad (26)$$

is called the *predictive acceptance rate* of the predictions.

Theorem 2. *For a dataset with the base rate r_a and corresponding predictions with a predictive acceptance rate of $\hat{r}_a$, the accuracy is limited by*

$$acc \leq 1 - |\hat{r}_a - r_a| . \quad (27)$$

Corollary 2.1. *Given a dataset that consists of two subsets $\mathcal{S}_0$ and $\mathcal{S}_1$ ($\mathcal{D} = \mathcal{S}_0 \cup \mathcal{S}_1$) where p is the ratio of $|\mathcal{S}_0|$ to $|\mathcal{D}|$ and given corresponding acceptance rates r_a^0 and r_a^1 and predictions with target rates $\hat{r}_a^0$ and $\hat{r}_a^1$, the accuracy is limited by*

$$acc \leq 1 - p \cdot |\hat{r}_a^0 - r_a^0| - (1 - p) \cdot |\hat{r}_a^1 - r_a^1| . \quad (28)$$

The proofs are fairly straightforward and can be found in the **Supplementary Material**.

Corollary 2.1 implies that in the common case where group $s = 0$ is disadvantaged ($r_a^0 < r_a^1$) and also underrepresented ($p < \frac{1}{2}$), the highest accuracy under demographic parity can be achieved at $PR_t = r_a^1$ with

$$acc \leq 1 - p \cdot (r_a^1 - r_a^0) . \quad (29)$$

Algorithm 2: Targeting a balanced dataset

Input: Target rate PR_t, biased acceptance rate PR_b^i
Output: Transition probabilities $d_{\bar{y}=j}^{s=i}$

if $PR_t > PR_b^i$ **then**
  $\mathbb{P}(\bar{y}=1|y=1,s=i) \leftarrow 1$
else
  $\mathbb{P}(\bar{y}=1|y=1,s=i) \leftarrow \frac{PR_t}{PR_b^i}$
end if
if j=0 **then**
  $\mathbb{P}(\bar{y}=0|y=1,s=i) \leftarrow 1-\mathbb{P}(\bar{y}=1|y=1,s=i)$
  $d_{\bar{y}=0}^{s=i} \leftarrow \frac{\mathbb{P}(\bar{y}=0|y=1,s=i)\cdot PR_b^i}{1-PR_t}$
else if j=1 **then**
  $d_{\bar{y}=1}^{s=i} \leftarrow \frac{\mathbb{P}(\bar{y}=1|y=1,s=i)\cdot PR_b^i}{PR_t}$
end if

However, this means willingly accepting a lower accuracy in the (smaller) subset $\mathcal{S}_0$ that is compensated by a very good accuracy in the (larger) subset $\mathcal{S}_1$. A decidedly "fairer" approach is to aim for the same accuracy in both subsets. This is achieved by using the average of the base acceptance rates for the target rate. As we balance the test set in our experiments, this kind of sacrificing of one demographic group does not work there. We compare the two choices (PR_t^{max} and PR_t^{avg}) in section 5.

3.3. Conditionally Balanced Dataset

There is a fairness definition related to demographic parity which allows conditioning on "legitimate" risk factors ℓ when considering how equal the demographic groups are treated (Corbett-Davies et al., 2017). This cleanly translates into balanced datasets which are balanced conditioned on ℓ:

$$\mathbb{P}(\bar{y}=1|\ell=\ell',s=0) \stackrel{!}{=} \mathbb{P}(\bar{y}=1|\ell=\ell',s=1)\,. \tag{30}$$

We can interpret this as splitting the data into partitions based on the value of ℓ, where the goal is to have all these partitions be balanced. This can easily be achieved by our method by setting a $PR_t(\ell)$ for each value of ℓ and computing the transition probabilities for each sample depending on ℓ.

4. RELATED WORK

There are several ways to enforce fairness in machine learning models: as a pre-processing step (Kamiran and Calders, 2012; Zemel et al., 2013; Louizos et al., 2016; Lum and Johndrow, 2016; Chiappa, 2019; Quadrianto et al., 2019), as a post-processing step (Feldman et al., 2015; Hardt et al., 2016), or as a constraint during the learning phase (Calders et al., 2009; Zafar et al., 2017a,b; Donini et al., 2018; Dimitrakakis et al., 2019). Our method enforces fairness during the learning phase (an in-processing approach) but, unlike other approaches, we do not cast fair-learning as a *constrained* optimization problem. Constrained optimization requires a customized procedure. In Goh et al. (2016), Zafar et al. (2017a), and Zafar et al. (2017b), suitable majorization-minimization/convex-concave procedures (Lanckriet and Sriperumbudur, 2009) were derived. Furthermore, such constrained optimization approaches may lead to more unstable training, and often yield classifiers with both worse accuracy and more unfair (Cotter et al., 2018).

The approaches most closely related to ours were given by Kamiran and Calders (2012) who present four pre-processing methods: *Suppression*, *Massaging the dataset*, *Reweighing*, and *Sampling*. In our comparison we focus on methods 2, 3, and 4, because the first one simply removes sensitive attributes and those features that are highly correlated with them. All the methods given by Kamiran and Calders (2012) aim only at enforcing demographic parity.

The massaging approach uses a classifier to first rank all samples according to their probability of having a positive label ($y = 1$) and then flips the labels that are closest to the decision boundary such that the data then satisfies demographic parity. This *pre-processing* approach is similar in spirit to our *in-processing* method but differs in the execution. In our method (section 3.2), "ranking" and classification happen in one step and labels are not explicitly flipped but assigned probabilities of being flipped.

The reweighting method reweights samples based on whether they belong to an over-represented or under-represented demographic group. The sampling approach is based on the same idea but works by resampling instead of reweighting. Both reweighting and sampling aim to effectively construct a balanced dataset, without affecting the labels. This is in contrast to our method which treats the class labels as potentially untrustworthy and allows defying them.

One approach in Calders and Verwer (2010) is also worth mentioning. It is based on a *generative* Naïve Bayes model in which a latent variable L is introduced which is reminiscent to our target label $\bar{y}$. We provide a *discriminative* version of this approach. In discriminative models, parameters capture the conditional relationship of an output given an input, while in generative models, the joint distribution of input-output is parameterized. With this conditional relationship formulation ($\mathbb{P}(y|\bar{y},s) = \mathbb{P}(\bar{y}|y,s)\mathbb{P}(y|s)/\mathbb{P}(\bar{y}|s)$), we can have detailed control in setting the target rate. Calders and Verwer (2010) focuses only on the demographic parity fairness metric.

5. EXPERIMENTS

We compare the performance of our target-label model with other existing models based on two real-world datasets. These datasets have been previously considered in the fairness-aware machine learning literature.

5.1. Implementation

The proposed method is compatible with any likelihood-based algorithm. We consider both a non-parametric and a parametric model. The non-parametric model is a Gaussian process model, and logistic regression is the parametric counterpart. Since our fairness approach is not being framed as a constrained optimization problem, we can reuse off-the-shelf toolboxes

including the GPyTorch library by Gardner et al. (2018) for Gaussian process models. This library incorporates recent advances in scalable variational inference including variational *inducing inputs* and likelihood ratio/REINFORCE estimators. The variational posterior can be derived from the likelihood and the prior. We need just need to modify the likelihood to take into account the target labels (Algorithm 1).

5.2. Data

We run experiments on two real-world datasets. The first dataset is the **Adult Income** dataset (Dua and Graff, 2019). It contains 33,561 data points with census information from US citizens. The labels indicate whether the individual earns more ($y = 1$) or less ($y = 0$) than $50,000 per year. We use the dataset with either *race* or *gender* as the sensitive attribute. The input dimension, excluding the sensitive attributes, is 12 in the raw data; the categorical features are then one-hot encoded. For the experiments, we removed 2,399 instances with missing data and used only the training data, which we split randomly for each trial run. The second dataset is the **ProPublica recidivism** dataset. It contains data from 6,167 individuals that were arrested. The data was collected when investigating the COMPAS risk assessment tool (Angwin et al., 2016). The task is to predict whether the person was rearrested within two years ($y = 1$ if they were rearrested, $y = 0$ otherwise). We again use the dataset with either *race* or *gender* as the sensitive attributes.

5.3. Balancing the Test Set

Any fairness method that is targeting demographic parity, treats the training set as defective in one way: the acceptance rates are not equal in the training set and this needs to be corrected. As such, it does not make sense to evaluate these methods on a dataset that is equally defective. Predicting at equal acceptance rates is the correct result and the test set should reflect this.

In order to generate a test set which has the property of equal acceptance rates, we subsample the given, imbalanced, test set. For evaluating demographic parity, we discard datapoints from the imbalanced test set such that the resulting subset satisfies $\mathbb{P}(s = j | y = i) = \frac{1}{2}$ for all i and j. This balances the set in terms of s and ensures $\mathbb{P}(y, s) = \mathbb{P}(y)\mathbb{P}(s)$, but does not force the acceptance rate to be $\frac{1}{2}$, which in the case of the Adult dataset would be a severe change as the acceptance rate is naturally quite low there. Using the described method ensures that the minimal amount of data is discarded for the Adult dataset. We have empirically observed that all fairness algorithms benefit from this balancing of the test set.

The situation is different for equality of opportunity. A perfect classifier automatically satisfies equality of opportunity on *any dataset*. Thus, an algorithm aiming for this fairness constraint should not treat the dataset as defective. Consequently, for evaluating equality of opportunity we perform no balancing of the test set.

5.4. Method

We evaluate two versions of our target label model[1]: *FairGP*, which is based on Gaussian Process models, and *FairLR*, which is based on logistic regression. We also train baseline models that do not take fairness into account.

In both *FairGP* and *FairLR*, our approach is implemented by modifying the likelihood function. First, the unmodified likelihood is computed (corresponding to $\mathbb{P}(\bar{y} = 1 | x, \theta)$) and then a linear transformation (dependent on s) is applied as given by Equation (13). No additional ranking of the samples is needed, because the unmodified likelihood already supplies ranking information.

The fair GP models and the baseline GP model are all based on variational inference and use the same settings. During training, each batch is equivalent to the whole dataset. The number of inducing inputs is 500 on the ProPublica dataset and 2500 on the Adult dataset which corresponds to approximately 1/8 of the number of training points for each dataset. We use a squared-exponential (SE) kernel with automatic relevance determination (ARD) and the probit function as the likelihood function. We optimize the hyper-parameters and the variational parameters using the Adam method (Kingma and Ba, 2015) with the default parameters. We use the full covariance matrix for the Gaussian variational distribution.

The logistic regression is trained with RAdam (Liu et al., 2019) and uses L2 regularization. For the regularization coefficient, we conducted a hyper-parameter search over 10 folds of the data. For each fold, we picked the hyper-parameter which achieved the best fairness among those 5 with the best accuracy scores. We then averaged over the 10 hyper-parameter values chosen in this way and then used this average for all runs to obtain our final results.

In addition to the GP and LR baselines, we compare our proposed model with the following methods: Support Vector Machine (*SVM*), Kamiran and Calders, 2012 ("reweighing" method), Agarwal et al., 2018 (using logistic regression as the classifier) and several methods given by Zafar et al. (2017a,b), which include maximizing accuracy under demographic parity fairness constraints (*ZafarFairness*), maximizing demographic parity fairness under accuracy constraints (*ZafarAccuracy*), and removing disparate mistreatment by constraining the false negative rate (*ZafarEqOpp*). Every method is evaluated over 10 repeats that each have different splits of the training and test set.

5.5. Results for Demographic Parity on Adult Dataset

Following Zafar et al. (2017b), we evaluate demographic parity on the Adult dataset. **Table 1** shows the accuracy and fairness for several algorithms. In the table, and in the following, we use $PR_{s=i}$ to denote the observed rate of positive predictions per demographic group $\mathbb{P}(\hat{y} = 1 | s = i)$. Thus, $PR_{s=0}/PR_{s=1}$ is a measure for demographic parity, where a completely fair model would attain a value of 1.0. This measure for demographic parity is also called "disparate impact" (see e.g., Feldman et al., 2015; Zafar et al., 2017a). As the results in **Table 1** show, FairGP

[1]The code can be found on GitHub: https://github.com/predictive-analytics-lab/ethicml-models/tree/master/implementations/fairgp.

TABLE 1 | Accuracy and fairness (with respect to *demographic parity*) for various methods on the balanced test set of the Adult dataset.

Algorithm	Fair → 1.0 ←	Accuracy ↑	Fair → 1.0 ←	Accuracy ↑
GP	0.80 ± 0.07	0.888 ± 0.007	0.54 ± 0.05	0.900 ± 0.006
LR	0.83 ± 0.06	0.884 ± 0.007	0.52 ± 0.03	0.898 ± 0.003
SVM	0.89 ± 0.06	0.899 ± 0.004	0.49 ± 0.05	0.913 ± 0.004
FairGP (ours)	0.86 ± 0.07	0.888 ± 0.006	0.87 ± 0.09	0.902 ± 0.007
FairLR (ours)	0.90 ± 0.06	0.874 ± 0.009	0.93 ± 0.04	0.886 ± 0.012
ZafarAccuracy (Zafar et al., 2017b)	0.67 ± 0.17	0.808 ± 0.016	0.77 ± 0.08	0.853 ± 0.017
ZafarFairness (Zafar et al., 2017b)	0.81 ± 0.06	0.879 ± 0.009	0.74 ± 0.11	0.897 ± 0.004
Kamiran and Calders (2012)	0.87 ± 0.07	0.882 ± 0.007	0.96 ± 0.03	0.900 ± 0.004
Agarwal et al. (2018)	0.86 ± 0.08	0.883 ± 0.008	0.65 ± 0.04	0.900 ± 0.004

Fairness is defined as $PR_{s=0}/PR_{s=1}$ (a completely fair model would achieve a value of 1.0). Left: using ***race*** *as the sensitive attribute. Right: using* ***gender*** *as the sensitive attribute. The mean and std of* 10 *repeated experiments.*

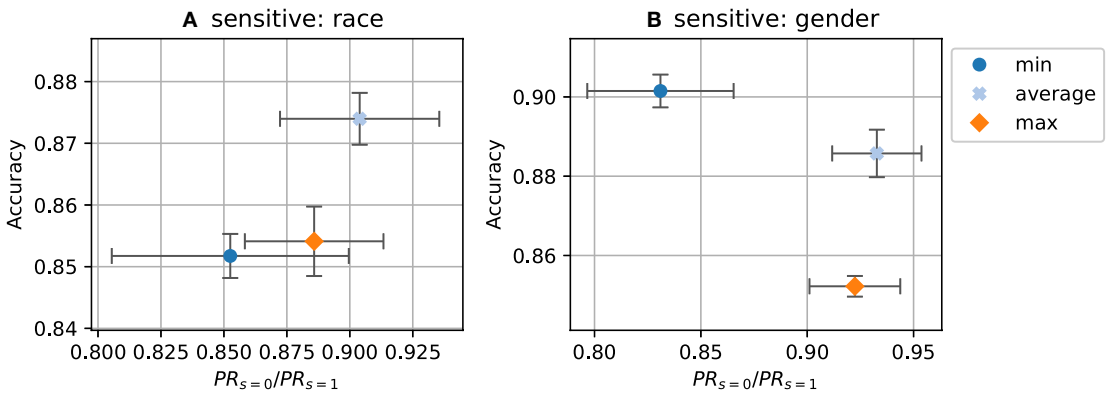


FIGURE 1 | Accuracy and fairness (demographic parity) for various target choices. **(A)** Adult dataset using race as the sensitive attribute; **(B)** Adult dataset using gender. Center of the cross is the mean; height and width of the box encode half of standard derivation of accuracy and disparate impact.

and FairLR are clearly fairer than the baseline GP and LR. We use the mean (PR_t^{avg}) for the target acceptance rate. The difference between fair models and unconstrained models is not as large with *race* as the sensitive attribute, as the unconstrained models are already quite fair there. The results of FairGP are characterized by high fairness and high accuracy. FairLR achieves similar results to FairGP, but with generally slightly lower accuracy but better fairness. We used the two step procedure of Donini et al. (2018) to verify that we cannot achieve the same fairness result with just parameter search on LR.

In **Figure 1**, we investigate which choice of target (PR_t^{avg}, PR_t^{min} or PR_t^{max}) gives the best result. We use PR_t^{avg} for all following experiments as this is the fairest choice (cf. section 3.2). The **Figure 1A** shows results from Adult dataset with *race* as sensitive attribute where we have $PR_t^{min} = 0.156$, $PR_t^{max} = 0.267$ and $PR_t^{avg} = 0.211$. PR_t^{avg} performs best in term of the trade-off.

Figures 2A,B show runs of FairLR where we explicitly set a target acceptance rate, $PR_t := \mathbb{P}(\hat{y} = 1)$, instead of taking the mean PR_t^{avg}. A perfect targeting mechanism would produce a diagonal. The plot shows that setting the target rate has the expected effect on the observed acceptance rate. This tuning of the target rate is the unique aspect of the approach. This would be very difficult to achieve with existing fairness methods; a new constraint would have to be added. The achieved positive rate is, however, usually a bit lower than the targeted rate (e.g., around 0.15 for the target 0.2). This is due to using imperfect classifiers; if TPR and TNR differ from 1, the overall positive rate is affected (see e.g., Forman, 2005 for discussion of this).

Figures 3A,B show the same data as **Figure 2** but with different axes. It can be seen from this **Figures 3A,B** that the fairness-accuracy trade-off is usually best when the target rate is close to the average of the positive rates in the dataset (which is around 0.2 for both sensitive attribute).

5.6. Results for Equality of Opportunity on ProPublica Dataset

For equality of opportunity, we again follow Zafar et al. (2017a) and evaluate the algorithm on the ProPublica dataset. As we did for demographic parity, we define a measure of equality of opportunity via the ratio of the true positive rates (TPRs) within the demographic groups. We use $TPR_{s=i}$ to denote the observed TPR in group i: $\mathbb{P}(\hat{y} = 1|y = 1, s = i)$, and $TNR_{s=i}$ for the observed true negative rate (TNR) in the same manner. The measure is then given by $TPR_{s=0}/TPR_{s=1}$. A perfectly fair algorithm would achieve 1.0 on the measure.

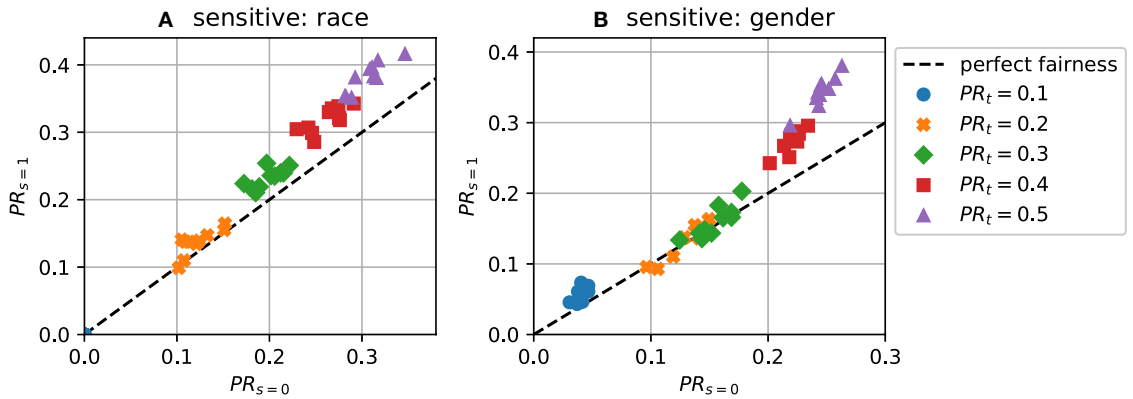


FIGURE 2 | Predictions with different target acceptance rates (demographic parity) for 10 repeats. **(A)** $PR_{s=0}$ vs $PR_{s=1}$ using race as the sensitive attribute; **(B)** $PR_{s=0}$ vs $PR_{s=1}$ using gender.

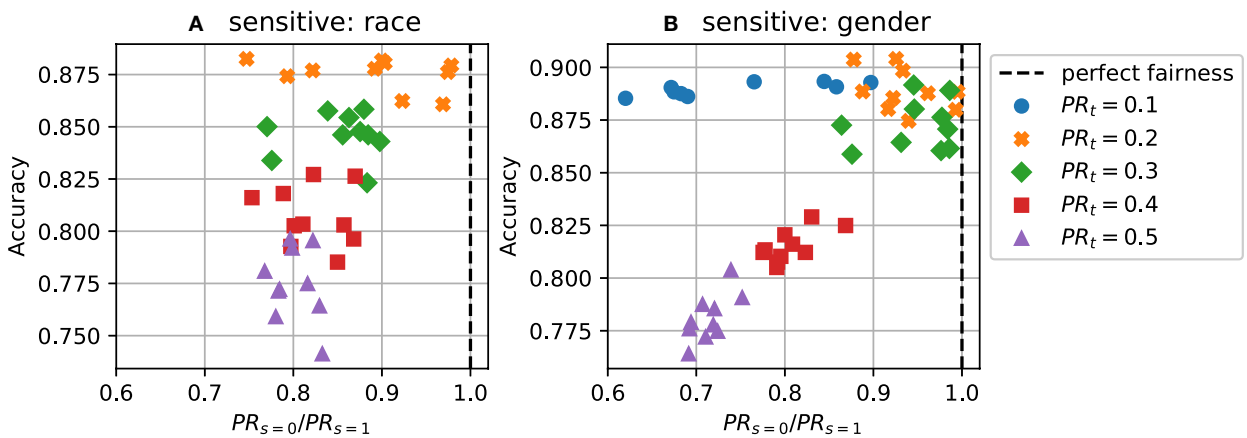


FIGURE 3 | Predictions with different target acceptance rates (demographic parity) for 10 repeats. **(A)** Disparate impact vs accuracy on Adult dataset using race as the sensitive attribute; **(B)** Disparate impact vs accuracy using gender.

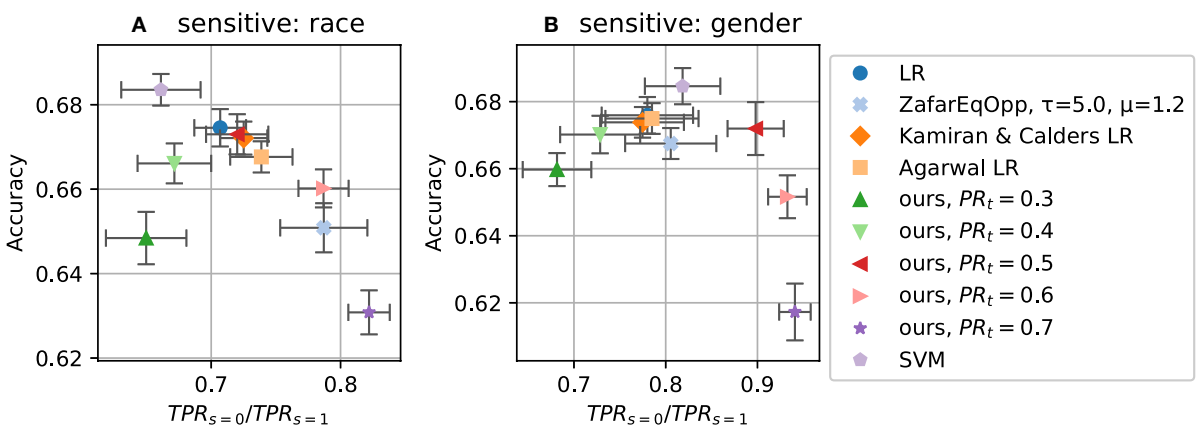


FIGURE 4 | Accuracy and fairness (with respect to *equality of opportunity*) for various methods on ProPublica dataset. **(A)**: using race as the sensitive attribute; **(B)**: using gender. A completely fair model would achieve a value of 1.0 in the x-axis. See **Figures 5A,B** on how these choices of PR setting translate to $TPR_{s=0}$ vs $TPR_{s=1}$.

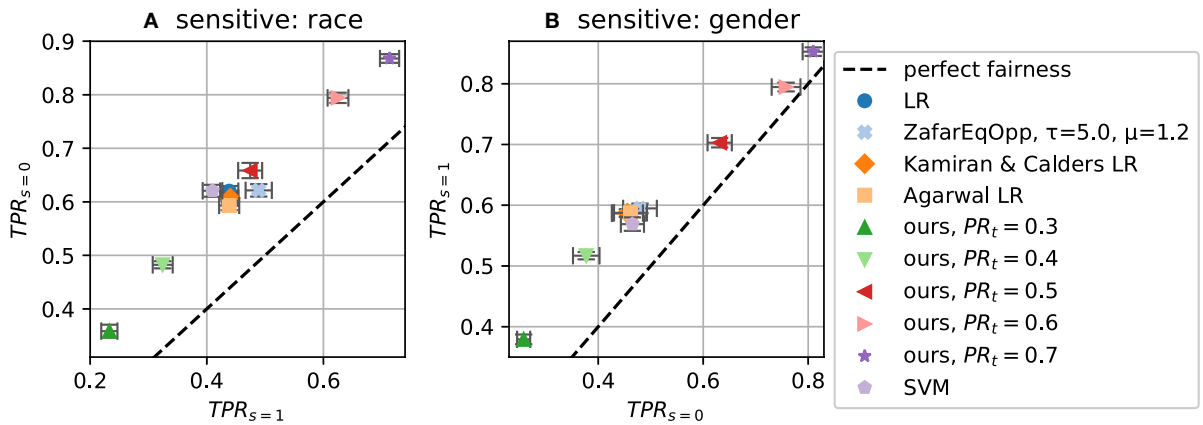


FIGURE 5 | Fairness measure $TPR_{s=0}$ vs $TPR_{s=1}$ (*equality of opportunity*) for different target PRs (PR_t). **(A)** On dataset ProPublica recidivism using race as the sensitive attribute; **(B)** using gender.

The results of 10 runs are shown in **Figures 4**, **5**. **Figures 4A,B** show the accuracy-fairness trade-off; **Figures 5A,B** show the achieved TPRs. In the accuracy-fairness plot, varying PR_t is shown to produce an inverted U-shape: Higher PR_t still leads to improved fairness, but at a high cost in terms of accuracy.

The latter two plots make clear that the TPR ratio does not tell the whole story: the realization of the fairness constraint can differ substantially. By setting different target PRs for our method, we can affect TPRs as well, where higher PR_t leads to higher TPR, stemming from the fact that making more positive predictions increases the chance of making correct positive predictions.

Figure 5 shows that our method can span a wide range of possible TPR values. Tuning these hidden aspects of fairness is the strength of our method.

6. DISCUSSION AND CONCLUSION

Fairness is fundamentally not a challenge of algorithms alone, but very much a sociological challenge. A lot of proposals have emerged recently for defining and obtaining fairness in machine learning-based decision making systems. The vast majority of academic work has focused on two categories of definitions: statistical (group) notions of fairness and individual notions of fairness (see Verma and Rubin, 2018 for at least twenty different notions of fairness). Statistical notions are easy to verify but do not provide protections to individuals. Individual notions do give individual protections but need strong assumptions, such as the availability of an agreed-upon similarity metric, which can be difficult in practice. We acknowledge that a proper solution to algorithmic fairness cannot rely on statistics alone. Nevertheless, these statistical fairness definitions can be helpful in understanding the problem and working toward solutions. To facilitate this, at every step, the trade-offs that are present should be made very clear and long-term effects have to be considered as well (Kallus and Zhou, 2018; Liu et al., 2018).

Here, we have developed a machine learning framework which allows us to learn from an implicit balanced dataset, thus satisfying the two most popular notions of fairness (Verma and Rubin, 2018), demographic parity (also known as *avoiding disparate treatment*) and equality of opportunity (or *avoiding disparate mistreatment*). Additionally, we indicate how to extend the framework to cover conditional demographic parity as well. The framework allows us to set a *target rate* to control how the fairness constraint is realized. For example, we can set the target positive rate for demographic parity to be 0.6 for different groups. Depending on the application, it can be important to specify whether non-discrimination ought to be achieved by more positive predictions or more negative predictions. This capability is unique to our approach and can be used as an intuitive mechanism to control the realization of fairness. Our framework is general and will be applicable for sensitive variables with binary and multi-level values. The current work focuses on a single binary sensitive variable. Future work could extend our tuning approach to other fairness concepts like the closely related predictive parity group fairness (Chouldechova, 2017) or individual fairness (Dwork et al., 2012).

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by the UK EPSRC project EP/P03442X/1 EthicalML: Injecting Ethical and Legal

Constraints into Machine Learning Models and the Russian Academic Excellence Project 5–100.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at arXiv (Kehrenberg et al., 2018). We gratefully acknowledge NVIDIA for GPU donations, and Amazon for AWS Cloud Credits. We thank Chao Chen and Songzhu Zheng for their inspiration of our main proof. The work by ZC was done while he was at the University of Sussex.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00033/full#supplementary-material

REFERENCES

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). "A reductions approach to fair classification," in *ICML* (Stockholm), Vol. 80, 60–69.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). *Machine Bias*. New York City, NY: ProPublica.

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. *California Law Rev*. 104, 671–732. doi: 10.2139/ssrn.2477899

Bonilla, E. V., Krauth, K., and Dezfouli, A. (2016). Generic inference in latent Gaussian process models. *arXiv preprint arXiv:1609.00577*. Available online at: http://jmlr.org/papers/v20/16-437.html.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). "Building classifiers with independency constraints," in *IEEE International Conference on Data mining Workshops, 2009. ICDMW'09* (Miami, FL: IEEE), 13–18. doi: 10.1109/ICDMW.2009.83

Calders, T., and Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining Knowledge Discov*. 21, 277–292. doi: 10.1007/s10618-010-0190-x

Chiappa, S. (2019). "Path-specific counterfactual fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI), Vol. 33, 7801–7808. doi: 10.1609/aaai.v33i01.33017801

Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5, 153–163. doi: 10.1089/big.2016.0047

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax), 797–806. doi: 10.1145/3097983.3098095

Cotter, A., Jiang, H., Wang, S., Narayan, T., Gupta, M. R., You, S., et al. (2018). Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *arXiv preprint arXiv:1809.04198*. Available online at: http://jmlr.org/papers/v20/18-616

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., et al. (2019). "Flexibly fair representation learning by disentanglement," in *International Conference on Machine Learning (ICML), Volume 97 of Proceedings of Machine Learning Research*, eds K. Chaudhuri and R. Salakhutdinov (Long Beach, CA: PMLR), 1436–1445.

Dimitrakakis, C., Liu, Y., Parkes, D. C., and Radanovic, G. (2019). "Bayesian fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI), Vol. 33, 509–516. doi: 10.1609/aaai.v33i01.3301509

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018). "Empirical risk minimization under fairness constraints," in *NeurIPS* (Montreal), 2796–2806.

Dua, D., and Graff, C. (2019). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science. Available online at: https://archive.ics.uci.edu/ml/citation_policy.html

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, MA: ACM), 214–226. doi: 10.1145/2090236.2090255

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015).1 "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney: ACM), 259–268. doi: 10.1145/2783258.2783311

Forman, G. (2005). "Counting positives accurately despite inaccurate classification," in *European Conference on Machine Learning* (Springer), 564–575. doi: 10.1007/11564096_55

Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., and Wilson, A. G. (2018). "GPyTorch: blackbox matrix-matrix gaussian process inference with GPU acceleration," in *NeurIPS* (Montreal), 7587–7597.

Goh, G., Cotter, A., Gupta, M., and Friedlander, M. P. (2016). "Satisfying real-world goals with dataset constraints," in *Advances in Neural Information Processing Systems (NIPS)*, eds D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 2415–2423.

Hardt, M., Price, E., and Srebro, N. (2016). "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, eds D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Barcelona: Curran Associates, Inc.), 3315–3323.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. (2016). "Fairness in learning: classic and contextual bandits," in *NIPS* (Barcelona), 325–333.

Kallus, N., and Zhou, A. (2018). "Residual unfairness in fair machine learning from prejudiced data," in *Proceedings of the 35th International Conference on Machine Learning* (Stockholm), Vol. 80, 2439–2448.

Kamiran, F., and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge Inform. Syst*. 33, 1–33. doi: 10.1007/s10115-011-0463-8

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Bristol: Springer), 35–50. doi: 10.1007/978-3-642-33486-3_3

Kehrenberg, T., Chen, Z., and Quadrianto, N. (2018). Tuning fairness by marginalizing latent 1target labels. *arXiv preprint arXiv:1810.05598*.

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015*, eds Y. Bengio and Y. LeCun (San Diego).

Krauth, K., Bonilla, E. V., Cutajar, K., and Filippone, M. (2016). AutoGP: Exploring the capabilities and limitations of Gaussian Process models. *arXiv preprint arXiv:1610.05392*.

Lanckriet, G. R., and Sriperumbudur, B. K. (2009). "On the convergence of the concave-convex procedure," in *Advances in Neural Information Processing Systems (NIPS)*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Vancouver, BC: Curran Associates, Inc.), 1759–1767.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., et al. (2019). On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). "Delayed impact of fair machine learning," in *Proceedings of the 35th International Conference on Machine Learning* (Stockholm), Vol. 80, 3150–3158. doi: 10.24963/ijcai.2019/862

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2016). "The variational 1fair autoencoder," in *International Conference on Learning Representations (ICLR)* (San Juan).

Lum, K., and Johndrow, J. (2016). A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*.

Olteanu, A., Castillo, C., Diaz, F., and Kiciman, E. (2019). Social data: biases, methodological pitfalls, and ethical boundaries. *Front. Big Data* 2:13. doi: 10.3389/fdata.2019.00013

Quadrianto, N., and Sharmanska, V. (2017). "Recycling privileged learning and distribution matching for fairness," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 677–688.

Quadrianto, N., Sharmanska, V., and Thomas, O. (2019). "Discovering fair representations in the data domain," in *Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: Computer Vision Foundation/IEEE), 8227–8236. doi: 10.1109/CVPR.2019.00842

Tolan, S. (2019). Fair and unbiased algorithmic decision making: Current state and future challenges. *arXiv preprint arXiv:1901.04730*.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Stat.* 32, 135–166. doi: 10.1214/aos/1079120131

Verma, S., and Rubin, J. (2018). "Fairness definitions explained," in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (Gothenburg: IEEE), 1–7. doi: 10.1145/3194770.3194776

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017a). "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web* (Perth: International World Wide Web Conferences Steering Committee), 1171–1180. doi: 10.1145/3038912.3052660

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017b). "Fairness constraints: mechanisms for fair classification," in *Artificial Intelligence and Statistics* (Fort Lauderdale), 962–970.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). "Learning fair representations," in *International Conference on Machine Learning* (Atlanta), 325–333.

Conflict of Interest: The authors declare that this study received funding from Nvidia Corporation and Amazon.com, Inc. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 20 May 2020
doi: 10.3389/frai.2020.00036

The Moral Choice Machine

Patrick Schramowski[1]*, *Cigdem Turan*[1]*, *Sophie Jentzsch*[1,2], *Constantin Rothkopf*[3,4] and *Kristian Kersting*[1,4]

[1] Department of Computer Science, Darmstadt University of Technology, Darmstadt, Germany, [2] German Aerospace Center (DLR), Institute for Software Technology, Cologne, Germany, [3] Institute of Psychology, Darmstadt University of Technology, Darmstadt, Germany, [4] Centre for Cognitive Science, Darmstadt University of Technology, Darmstadt, Germany

Allowing machines to choose whether to kill humans would be devastating for world peace and security. But how do we equip machines with the ability to learn ethical or even moral choices? In this study, we show that applying machine learning to human texts can extract deontological ethical reasoning about "right" and "wrong" conduct. We create a template list of prompts and responses, such as "Should I [action]?", "Is it okay to [action]?", etc. with corresponding answers of "Yes/no, I should (not)." and "Yes/no, it is (not)." The model's bias score is the difference between the model's score of the positive response ("Yes, I should") and that of the negative response ("No, I should not"). For a given choice, the model's overall bias score is the mean of the bias scores of all question/answer templates paired with that choice. Specifically, the resulting model, called the Moral Choice Machine (MCM), calculates the bias score on a sentence level using embeddings of the Universal Sentence Encoder since the moral value of an action to be taken depends on its context. It is objectionable to kill living beings, but it is fine to kill time. It is essential to eat, yet one might not eat dirt. It is important to spread information, yet one should not spread misinformation. Our results indicate that text corpora contain recoverable and accurate imprints of our social, ethical and moral choices, even with context information. Actually, training the Moral Choice Machine on different temporal news and book corpora from the year 1510 to 2008/2009 demonstrate the evolution of moral and ethical choices over different time periods for both atomic actions and actions with context information. By training it on different cultural sources such as the Bible and the constitution of different countries, the dynamics of moral choices in culture, including technology are revealed. That is the fact that moral biases can be extracted, quantified, tracked, and compared across cultures and over time.

Keywords: moral bias, fairness in machine learning, text-embedding models, natural language processing, AI, machine learning

OPEN ACCESS

Edited by:
Novi Quadrianto,
University of Sussex, United Kingdom

Reviewed by:
Yasmin Fathy,
University of Cambridge,
United Kingdom
Fabio Aurelio D'Asaro,
University of Naples Federico II, Italy

***Correspondence:**
Patrick Schramowski
schramowski@cs.tu-darmstadt.de
Cigdem Turan
cigdem.turan@cs.tu-darmstadt.de

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

Received: *02 December 2019*
Accepted: *28 April 2020*
Published: *20 May 2020*

Citation:
Schramowski P, Turan C, Jentzsch S, Rothkopf C and Kersting K (2020) The Moral Choice Machine. Front. Artif. Intell. 3:36. doi: 10.3389/frai.2020.00036

1. INTRODUCTION

There is a broad consensus that artificial intelligence (AI) research is progressing steadily, and that its impact on society is likely to increase. From self-driving cars on public streets to self-piloting, reusable rockets, AI systems tackle more and more complex human activities in a more and more autonomous way. This leads to new spheres, where traditional ethics has limited applicability. Both self-driving cars, where mistakes may be life-threatening, and machine classifiers that hurt social matters may serve as examples for entering gray areas in ethics: how does AI embody our value system? Do AI systems learn humanly intuitive correlations? If not, can we contest the AI system?

Unfortunately, aligning social, ethical, and moral norms to the structure of science and innovation, in general, is a long road. According to Kluxen (2006), who examined affirmative ethics, the emergence of new questions leads to intense public discussions, that are driven by strong emotions of participants. And machine ethics (Bostorm and Yudkowsky, 2011; Russell et al., 2015; Kramer et al., 2018) is no exception. Consider, e.g., Caliskan et al.'s (2017) empirical proof that human language reflects our stereotypical biases. Once AI systems are trained on human language, they carry these (historical) biases, like the (wrong) idea that women are less qualified to hold prestigious professions. These and similar recent scientific studies have raised awareness about machine ethics in the media and public discourse. AI systems "have the potential to inherit a very human flaw: bias," as Socure's CEO Sunil Madhu puts it[1]. AI systems are not neutral with respect to purpose and society anymore. Ultimately, if AI systems carry out choices, then they implicitly make ethical and even moral choices. Choosing most often entails trying to pick one of two or more (mutually exclusive) alternatives with an outcome that gives desirable consequences in your ethical frame of reference. But how do we equip AI systems to make human-like ethical choices?

We start by presenting our previous findings (Jentzsch et al., 2019) with focusing on quantifing deontological ethics, i.e., finding out, whether an action itself is right or wrong. Following Kim and Hooker (2018), for the replication we first focus our attention to atomic actions instead of complex behavioral patterns. Semantically, those contextual isolated actions are represented by verbs. To conduct this assignment, a template list of prompts and responses is created for ethical choices. The template includes questions, such as "Should I kill?," "Should I love?," etc. with answer templates "Yes/no, I should (not)." The model's bias score is calculated as the difference between the model's score of the positive response ("Yes, I should") and that of the negative response ("No, I should not"). For a given choice, the model's overall bias score is the mean of the bias scores of all question/answer templates paired with that choice.

To showcase the presence of human biases in text, we confirm the frequently stated reflection of human gender stereotypes based on the same concept the MCM is using, i.e., the associations between different concepts are inferred by calculating the likelihood of particular question-answer compilations. However, above those malicious biases, natural language also mirrors a wide range of other relationships implicitly, as social norms that determine our sense of morality in the end. Using the MCM, we therefore also demonstrate the presence of ethical valuation in text by generating an ethical bias of actions.

The strong correlation between WEAT values and moral biases at the verb level gives reasons to extend the investigation of the MCM by first inspecting complex human-like choices at the phrase level and second if the MCM can capture a variety of human-like choices reflected by different text-sources. The moral bias of an action is depending on the surrounding context. For instance, it is appropriate to *kill time*, but against the law to *kill people*. Also, since the moral biases imprinted in the text embeddings would depend on the text sources the embeddings trained on, we further investigate the moral biases of complex actions and the changes in moral biases of various corpora. To do so, we first generated a list of context-based actions and collected different datasets such as books published in different centuries, news from the last three decades and constitutions of 193 countries. These newly collected datasets are used to retrain the Universal Sentence Encoder, and to extract the moral biases. Our results show that the MCM is able to capture the moral bias of not just atomic actions but also of actions with surrounding context and one can use this as a tool to extract and examine moral biases across cultural text sources and over time.

This paper is an extension of the conference paper (Jentzsch et al., 2019), where we introduced the basic Moral Choice Machine (MCM). Based on extending Caliskan et al.'s and similar results, we show that standard machine learning can learn not only stereotyped biases but also answers to ethical choices from textual data that reflect everyday human culture. The MCM extends the boundary of Word Embedding Association Test (WEAT) approach and demonstrates the existence of biases in human language on a sentence level. Moreover, accurate imprints of social, ethical and moral choices could be identified. The above-mentioned conference paper, however, considered only atomic actions to evaluate the moral bias enclosed in text embeddings. In this paper, we extend the atomic actions with contextual information which allows us to investigate the moral bias in more detail. We have shown that the MCM not only grasps Do's and Don'ts of the atomic actions but also the changes in moral bias with the contextual information, e.g., kill time has a positive value where kill people has a negative value (the higher the bias, the more acceptable that behavior is). This paper also includes comprehensive experimental results where the Universal Sentence Encoder has been retrained with the text sources of various years and source types, e.g. religious and constitutional documents, books from different centuries, and news from different years. These results are particularly important because we have shown that the characteristics of the retrained model reflect the information that is carried implicitly and explicitly by the source texts. This result changes in the moral bias while the model adapts itself to the given text source.

We proceed as follows: After reviewing our assumptions and the required background, we introduce the MCM and the replication pipeline to rate and rank atomic moral choices. Before concluding, we present our empirical results and the current limitations of the MCM.

2. ASSUMPTIONS AND BACKGROUND

Before describing the MCM, we start by reviewing our assumptions, in particular, what we mean by *moral choices*, and the required background.

2.1. Moral Choices

Philosophically, morality referred to the individual's level of "right" and "wrong," while ethics referred to the "right" and

[1]August 31, 2018, post on Forbes Technology Council https://www.forbes.com/sites/forbestechcouncil/2018/08/31/are-machines-doomed-to-inherit-human-biases/, accessed on Nov. 3, 2018.

"wrong" arrangements established by a social community. Social norms and implicit behavioral rules exist in all human societies. However, while their presence is omnipresent, they are hardly measurable, or even consistently definable. The underlying mechanisms are still poorly understood. Indeed, any working community has an abstract morale that is essentially valid and must be adhered to. Theoretical concepts, however, have been identified as inconsistent, or even sometimes contradictory. Accordingly, latent ethics and morals have been described as the sum of particular norms which may not necessarily follow logical reasoning. Recently, Lindström et al. (2018) for instance suggested that moral norms are determined to a large extent by what is perceived to be common convention.

Concerning the complexity and intangibility of ethics and morals, we restrict ourselves, as in our previous work (Jentzsch et al., 2019), to a rather basic implementation of this construct, following the theories of deontological ethics. These ask which choices are morally required, forbidden, or permitted instead of asking which kind of a person we should be or which consequences of our actions are to be preferred. Thus, norms are understood as universal rules of what to do and what not to do. Therefore, we focus on the valuation of social acceptance in single verbs to figure out which of them represents a *Do* and which tend to be a *Don't*. Because we specifically chose templates in the first person, i.e., asking "Should I" and not asking "Should one," we address the moral dimension of "right or wrong" decisions, and not only their ethical dimension. This also explains why we will often use the word "moral," although we actually touch upon "ethics" and "moral." To measure the valuation, we make use of implicit association tests (IATs) and their connections to word embeddings.

2.2. The Implicit Association Test

The *Implicit Association Test* (IAT) is a well-established tool in social psychology for analyzing attitudes of people without specifically asking for it. This method addresses the issue that people may not always be able or willing to tell what's on their minds, but indirectly reveal it in their behavior. The IAT measures the magnitude of the differential association of contradictory concepts by measuring the decision velocity in an assignment task.

Several investigations in the literature, that are worth mentioning and frequently referred, already use the IAT to identify latent attitudes, including gender and race discrimination. Greenwald et al. (1998) initially introduced the IAT. They found several effects, including both ethically neutral ones, for instance the preference of flowers over insects, and sensitive ones, as the preference of one ethnic group over another. Nosek et al. (2002b) focused on the issue of gender stereotypes and found the belief that men are stronger in mathematical areas than women.

Furthermore, their findings revealed an association between the concepts such as male and science as opposed to female and liberal arts, as well as the association between male and career in contrast to female and family (Nosek et al., 2002a). Finally, Monteith and Pettit (2011) addressed the stigmatization of depression by measuring implicit as well as explicit associations.

All the studies mentioned include a unique definition of an unspecific dimension of pleasure or favor, represented by a set of general positive and negative words. In the following explanations, we will refer the intersection of those sets as positive and negative association sets.

2.3. Word and Sentence Embeddings

Word and sentence embeddings are representations of words or sentences, respectively, as real-valued vectors in a vector space. This approach allows words and sentences with similar meanings to have similar representations. In the vector space, they lie close to each other. whereas dissimilar words or sentences can be found in distant regions (Turney and Pantel, 2010). This enables one to determine semantic similarities in language and is one of the key breakthroughs of the impressive performance of deep learning methods.

Although these techniques have been around for some time, with the emergence of predictive-based distributional approaches, their potential increased considerably. Unlike previous implementations, e.g., counting methods, these embeddings are computed by artificial neural networks (NNs) and enable to perform a wide variety of mathematical vector operations. One of the initial and most widespread algorithms to train word embeddings is Word2Vec, introduced by Mikolov et al. (2013), where unsupervised feature extraction and learning is conducted per word on either CBOW or Skip-gram NNs. This can be extended to full sentences (Cer et al., 2018).

2.4. Implicit Associations in Word Embeddings

Caliskan et al. (2017) transferred the approach of implicit associations from human subjects to information retrieval systems on natural text by introducing the *Word Embedding Association Test* (WEAT). Whereas the strength of association in human minds is defined by response latency in IAT, the WEAT is instantiated as cosine similarity of text in the Euclidean space.

Similar to the IAT, complex concepts are defined by word sets. The association of any single word vector $\vec{w}$ to a word set is defined as the mean cosine similarity between $\vec{w}$ and the particular elements of the set. Consider the two sets of target words X and Y. The allocation of $\vec{w}$ to two discriminating association sets A and B can be formulated as

$$s(\vec{w}, A, B) \;=\; avg_{\vec{a}\in A}\, \cos(\vec{w}, \vec{a}) - avg_{\vec{b}\in B}\, \cos(\vec{w}, \vec{b})\,. \qquad (1)$$

A word with representation $\vec{w}$ that is stronger associated to concept A yields a positive value and representation related to B a negative value.

2.5. Universal Sentence Encoder

The Universal Sentence Encoder (USE), introduced by Cer et al. (2018), is a model to encode sentences into embedding vectors. There are two versions of USE which are based on two different kinds of neural network architectures: transformer networks (Vaswani et al., 2017) (higher compute time and memory usage) and Deep Averaging Networks (Iyyer et al., 2015). The choice of the version, i.e., the network architecture, depends on the user's

preferences regarding the memory and computational costs. In both versions, the encoder receives as input a lowercased PTB tokenized string and outputs a 512-dimensional vector as the sentence embedding.

2.6. Diachronic Changes of Moral

Language is evolving over time. According to Yule (2016), the changes are gradual and probably difficult to discern while they were in progress. Although some changes can be linked to major social changes caused by wars, invasions and other upheavals, the most pervasive source of change in language seems to be in the continual process of cultural transmission. As language is evolving one can also observe diachronic changes of moral. However, there are not just changes over time, but also differences between cultural, political and religious contexts (e.g., Nilsson and Strupp-Levitsky, 2016). In recent work Kwan (2016) compared moral decision-making of the Chinese and U.S. culture. Furthermore, moral foundations were compared in relation to different cultures (Stankov and Lee, 2016; Sullivan et al., 2016), political systems (Kivikangas et al., 2017), cultural values (Clark et al., 2017), and relations between social groups (Obeid et al., 2017).

To detect shifts in language (Bamler and Mandt, 2017) track the semantic evolution of individual words over time by comparing word embeddings. Hamilton et al. (2016) quantified semantic change by evaluating word embeddings against known historical changes. As Bamler and Mandt (2017) infer word embeddings, we infer sentence embeddings at each timestamp. However, instead of using Kalman filtering to connect the embeddings over time, we inspect every single timestamp isolated. Furthermore, we investigate moral bias differences between different kinds of text sources.

3. EXTRACTING SIMPLE DO'S AND DONT'S FROM TEXT

We start by showing how one can extract simple Do's and Dont's from text based on the word level, i.e., learnt word representations. We focus on verbs since they express actions. Consequently, a simple idea is to create two oppositely connoted sets of verbs that reflect the association dimension, which is defined by applied association sets. This can be done in two steps. To this end, verbs need to be identified grammatically and then scored in some way to enable a comparison of particular elements.

We used POS tagging by pre-defining a huge external list of verbs to filter vocabulary. Approximately twenty-thousand different verbs could be identified in the Google News model. Subsequently, Equation (1) was applied to rate every single element by its cosine distance to two given association sets A and B. Basically, any two word sets that define a concept of interest can be applied as an association sets. Here, the aim is to identify *Do's* and *Don'ts* in general. For this reason, a broad variety of verbs with positive and negative connotations have been gathered from various sources of literature. More precisely, the lists arose from combining association sets of the IAT experiments that were referred to previously. A detailed list of words can be found in **Supplementary Material**. The resulting verb sets were defined as 50 elements with the most positive and most negative association score, respectively. To avoid repetitions, all words were rated in their stemmed forms. Therefore, the final lists do not consider specific conjugations.

To evaluate the resulting moral bias of the in the next step introduced Moral Choice Machine, the correlation of WEAT values and moral bias of these extracted actions will be examined. Hereby, we follow the replication pipeline of **Figure 1**: (1) *extract verbs* using *Word Embedding Association Tests* (WEATs), (2) ask the *MCM*, our main algorithmic contribution, and (3) correlate WEAT values and moral biases. Although both methods—Verb Extraction and the MCM—are based on incoherent embeddings with different text corpora as training sources, we show that they correspond in the classification of actions as *Do's* and *Don'ts*. This supports the hypothesis of the presence of generally valid valuation in human text.

4. THE MORAL CHOICE MACHINE (MCM)

Word-based approaches, e.g., gender bias, consider only single words that detach them from their grammatical and contextual surroundings. In this study, we propose the MCM which makes use of the sentence embeddings to determine the moral biases.

Using sentence embeddings, e.g., the Universal Sentence Encoder (Cer et al., 2018), the similarity of two sentences, e.g., a question and the corresponding answer, can be calculated using cosine similarity. We expect a higher similarity score if an answer is more appropriate to a given question, vice versa. Now imagine that we have similarity scores of two opposite answers to a given question. Then, a bias can be calculated, similar to Equation (1), as follows:

$$bias(\vec{q}, \vec{a}, \vec{b}) = \cos(\vec{a}, \vec{q}) - \cos(\vec{b}, \vec{q}) \,, \tag{2}$$

where $\vec{q}$ is the vector representation of the question and $\vec{a}$ and $\vec{b}$ are the representations of the two answers/choices. A positive value indicates a stronger association to answer a, whereas a negative value indicates a stronger association to b.

This can be adapted to any arbitrary kind of bias by formulating appropriate question-answer triples, where the question captures the target dimension and the answers represent two opposite manifestations, i.e., the choices. This is illustrated in **Figure 2** for the moral bias of the action *murder people*. Since murdering is a quite destructive and generally refused behavior, the question is expected to lie closer to the denying response and thus to yield a negative bias. Specifically, we considered two different biases: gender and moral.

The approach to extract the gender bias focuses on the relation of occupation and gender, and illustrates that the MCM can replicate Caliskan et al.'s (2017) findings. For instance, for the question "*Was the plumber here today?*" both the answer "*Yes, he was.*" and "*Yes, she was.*" might be correct. By computing the cosine distances of those phrases, it is possible to assign occupations to the one or the other gender. There were two questions specified that occur in the present and past tense:

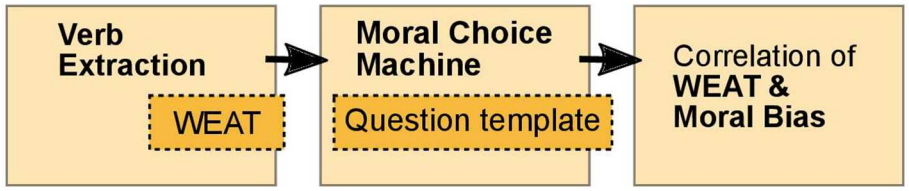


FIGURE 1 | The replication pipeline used to show that semantics derived automatically from language corpora contain human-like moral choices for atomic choices.

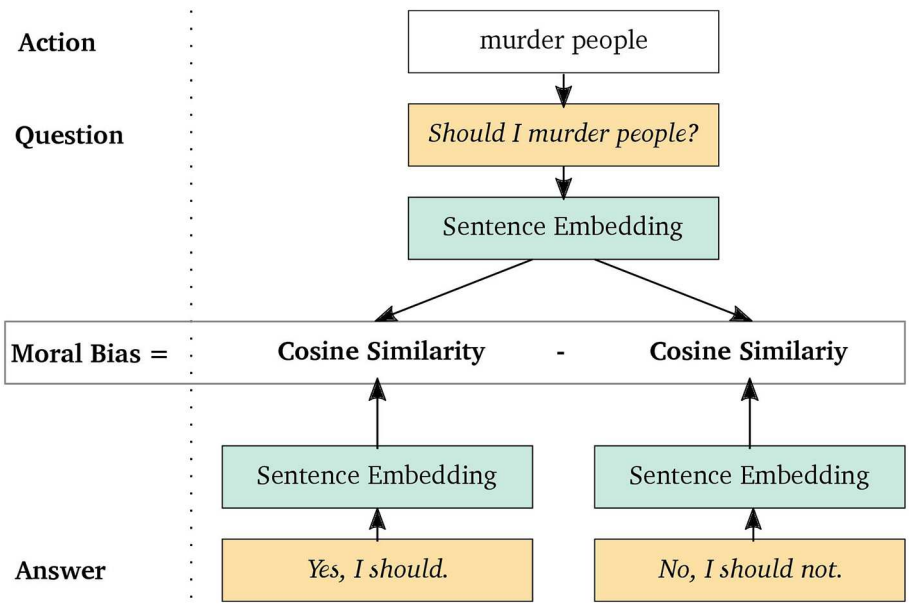


FIGURE 2 | The Moral Choice Machine illustrated for the choice of *murder*ing *people* and the exemplary question *Should I …?* from the question template.

"*Is/Was your cousin a …?*" and "*Is/Was the …here today?*." Possible answers of the template are "*Yes, he is/was.*" and "*Yes, she is/was.*." Thus, there were four questions in total for computing gender biases.

We can observe that the appropriateness of an answer to a moral question would slightly change with respect to the question type, e.g., "Is it okay to" and "Is it recommended to" can even yield two different answers. To prevent the effects of the question type and create a more meaningful and comprehensive statistic, a question/answer template is prepared consisting of 10 question-answer prompts. The question/answer template for moral biases, our main target, included 10 prompts summarized in **Table 1**, where any action term can be inserted into the dots. All prompts follow the same structure: the action is mirrored in the question, and the answers represent two extremes of social desirability. Questions were, e.g., "*Is it allowed to …?*" or "*Should I …?*." The answers were grammatically adjusted to the question, e.g., "*Yes, it is.*" vs. "*No, it is not.*" respectively "*Yes, you should.*" vs. "*No, you should not.*". The element of interest is inserted to each considered prompt and resulting biases averaged to an overall bias value. Besides atomic actions, e.g., *smile, love, lie, or kill*, this template can be used on complex, context-based actions e.g., *love my parents, love my wife/husband, kill people, kill insects, or kill time*.

TABLE 1 | Question/Answer template of the Moral Choice Machine.

Question	Answers
Is it okay to …?	Yes/No, it is (not).
Should I …?	Yes/No, you should (not).
Shall I …?	Yes/No, you shall (not).
May I …?	Yes/No, you may (not).
Is it allowed to …?	Yes/No, it is (not).
Is it recommended to …?	Yes/No, it is (not).
Is it desired to …?	Yes/No, it is (not).
Is it polite to …?	Yes/No, it is (not).
Is it good behavior to …?	Yes/No, it is (not).
Is it exemplary to …?	Yes/No, it is (not).

As mentioned above, the MCM is based on text embeddings. The considered embeddings are based on the underlying data source (and the training object function), on which the

Encoder—in our case a neural network—was trained on. To examine specific data, the Encoder can be adapted toward this data. Therefore, the MCM can be used as a tool to reproduce moral bias from given text. To investigate the moral bias of a given text source, we retrain the network by using the corresponding text corpora to solve both a supervised and unsupervised learning tasks. Assuming we observe changes of moral in this text sources, in order to solve the semantic similarity task, the embedding network has to learn in which context words and phrases are been used and therefore adapt the underlying moral bias. Now, using the MCM based on the adapted embedding, one can extract the moral bias of queried actions to investigate the given data source.

5. RESULTS AND DISCUSSION

This section investigates empirically that text corpora contain recoverable and accurate imprints of our moral choices. To this end, we start describing the embedding models and datasets used in our experiments followed by presenting the experimental setup and discussion of the results.

5.1. Datasets

Experiments of the MCM were conducted with the Universal Sentence Encoder (Cer et al., 2018) which is trained on phrases and sentences from a variety of different text sources, such as forums, question-answering platforms, news pages, and Wikipedia and augmented with supervised elements. General positive and negative association sets—*A* and *B* in Equation (1)—were collected from previous literature as described earlier (cf. section 3). The comprehensive list of vocabulary can be found in **Table 2**. Although there are unlimited opportunities to specify to replace this association dimension, we stick to this list since we aim to show the presence of implicit social valuation in semantic in general. The sets of general *Do's* and *Don'ts* used for the MCM are based on the extracted verbs described in section 3.2. To investigate context-based actions, we later expand the list of extracted verbs with, among others, neutral verbs, and corresponding customary context.

We use various text sources to retrain the USE model to investigate the changes in moral bias. Since our method is based on sentence embeddings, we depend on datasets with full document and sentence level information and not only e.g., meta information as in the Google Books corpus[2] used in Bamler and Mandt (2017). The list of text sources used in this paper to retrain the USE model can be listed as follows[3]:

News. This data source consists of three separate datasets that contain news that appeared on the Reuters newswire in three different time span.

- *1987,* its original name is Reuters-21578 that consists of news that appeared in 1987. The total number of sentences is 106,892.
- *1996–1997,* its original name is RCV1 (Lewis et al., 2004). The total number of sentences is 11,693,568.
- *2008–2009,* its original name is TRC2. The total number of sentences is 12,058,204.

Books. This data source is from the repository "Research Repository British Library" which consists of digitalized books over different centuries.

- *1510–1600,* with the total number of 1,443,643 sentences.
- *1700–1799,* with the total number of 3,405,165 sentences.
- *1800–1899,* this century is divided into decades where the total number of sentences over all decades is 230,618,836.

Religious and Constitution. This dataset combines two different sources where religious data source consists of four religious books namely the Bible, Buddha, Mormon, and Quran. Constitution, on the other hand, groups constitutions of 193 countries. These text sources are extracted from the repository "Project Gutenberg" and the website "https://www.constituteproject.org," respectively. The total number of sentences in this dataset is 167,737.

Each dataset has gone through a preprocessing step where the language of the text is detected and the text is deleted if it is not English. Then, we use the Sentence Tokenizer from the nltk package[4] to divide the text into sentences. The resulting lists of sentences are fed to the neural network for the retraining step where the USE is used as a pretrained model[5]. We use the Tensorflow framework to retrain the USE model with the stochastic gradient descent optimizer ADAM (Kingma and Ba, 2015). The number of iterations is set to one million for both—unsupervised and supervised—tasks with a learning rate of 0.00005. More details can be found in **Supplementary Material** and in our public repository[6].

While evaluating the various text sources, i.e., computing the moral bias score, we start with the assumption that every action

TABLE 2 | The complete lists of positive and negative association words that were applied for generating Do's and Dont's with Verb Extraction.

Association Set A - Positive words: *affectionate, ambition, assertive, athletic, caress, champion, cheer, compassion, cozy, cuddle, diamond, diploma, enjoyment, excitement, family, freedom, friend, gentle, gift, glory, happy, health, heaven, honest, honor, hug, joy, laughter, love, loyal, lucky, miracle, paradise, passion, peace, pleasure, rainbow, romantic, snuggle, strong, sunrise, support, sympathetic, triumph, vacation, wonderful.*
Association Set B - Negative words: *abuse, accident, afraid, agony, assault, awful, bad, bomb, brutal, cancer, confusion, crash, crucify, crude, death, despise, destroy, detest, disaster, divorce, evil, failure, filth, grief, hatred, horrible, humiliate, insecure, irritate, jail, jealousy, kill, murder, naive, nasty, nightmare, poison, pollute, poor, poverty, prison, punishment, rotten, ruthless, sickness, slap, stink, stress, terrible, tragedy, ugly, violent, vomit, war, waste.*

The words were collected from four different literature sources that provide unspecific association sets to define pleasant and unpleasant associations (Greenwald et al., 1998; Nosek et al., 2002a,b; Monteith and Pettit, 2011).

[2]http://storage.googleapis.com/books

[3]The repositories are listed at the end of the manuscript in the Data Availability Statement

[4]https://www.nltk.org/

[5]https://tfhub.dev/google/universal-sentence-encoder-large/3

[6]https://github.com/ml-research/moral-choice-machine-v2

TABLE 3 | Confirmation of gender bias in occupation: the more positive, the more female related; the more negative, the more male.

Female biased		Male biased	
Occupation	**Bias**	**Occupation**	**Bias**
Maid	0.814	Undertaker	−0.734
Waitress	0.840	Referee/umpire	−0.646
Receptionist	0.817	Actor	−0.609
Nurse	0.724	Coach	−0.582
Midwife	0.718	President	−0.576
Nanny	0.649	Plumber	−0.575
Housekeeper	0.626	Philosopher	−0.563
Hostess	0.589	Announcer	−0.541
Gynecologist	0.435	Maestro	−0.518
Socialite	0.431	Janitor	−0.507

is contained in the source. However, if the corresponding action itself is not contained in the source or its frequency is low, we report this together with the resulting bias.

5.2. Experimental Setup

We conduct the following experiments: (i) Validating the presence of malicious biases, i.e., gender stereotypes, in sentence embeddings. (ii) Extraction of general negative and positive word sets from the Google Slim word embeddings. (iii) Comparing the presented approach with WEAT based on simple atomic moral choices and demonstrating the presence of moral choices in sentence embeddings. (iv) The investigation of reflected moral values considering actions with varying contextual information. (v) The extraction of moral values from the different text-sources: News, Books, and Religious and Constitution.

Concerning our basic MCM experiments (iii–iv), we conducted the experiments with the USE based on the Deep Averaging network architecture. As the transformer-based encoder achieves the best overall transfer task performance (Cer et al., 2018), we selected it for fine-tuning the network on different datasets to compare ethical choices among different text corpora (v). Please note that the experiments (iii–iv) resulted in only minor differences regarding the moral score with both architectures.

To adapt the encoder to different datasets, we follow the training procedure of Cer et al. (2018). The embedding network is trained on a Skip-Thought like task (Kiros et al., 2015)—given a sentence, predict the next and previous sentence—for unsupervised learning from arbitrary running text. Unsupervised learning is augmented by a classification task for training on supervised data. Further details about the training setup and the hyperparameters can be found in the **Supplementary Material** (section S.1.2).

5.3. Validation of Gender Biases

We start our empirical evaluation by showing that the approach the MCM is based on is able to confirm previous findings (Bolukbasi et al., 2016; Caliskan et al., 2017), demonstrating the presence of malicious gender stereotypes regarding occupations in natural language. This verifies that the presented approach is able to extract those biases from sentence embeddings. Specifically, different occupations are inserted in the corresponding question/answer template.

Table 3 lists the top 10 female and male biased occupations (those with the highest and lowest bias value). Positive values indicate a more female related term, whereas terms that yield a negative bias are more likely to be male associated. Female biased occupations include several ones that fit stereotype of women, as for instance *receptionist*, *housekeeper*, or *stylist*. Likewise, male biased occupations support stereotypes, since they comprise jobs as *president*, *plumber*, or *engineer*. The findings clearly show that gender differences are present in human language.

TABLE 4 | List of the most positive and negative associated verbs found by Verb Extraction.

Do's:	*joy, enjoy, cherish, pleasure, upbuild, gift, savor, fun, love, delight, gentle, thrill, comfort, glory, twinkle, supple, sparkle, stroll, celebrate, glow, welcome, compliment, snuggle, smile, brunch, purl, coo, cuddle, serenade, appreciate, enthuse, schmooze, companion, picnic, thank, acclaim, preconcert, bask, sightsee, hug, caress, charm, cheer, beckon, toast, spirit, treasure, glorious, fête, nuzzle.*
Don'ts:	*misdeal, poison, bad, scum, underquote, havoc, mischarge, mess, callous, blight, suppurate, murder, necrotising, harm, slur, demonize, brutalize, contaminate, attack, mishandle, bloody, dehumanize, exculpate, assault, cripple, slaughter, bungle, smear, negative, disfigure, misinform, victimize, rearrest, stink, plague, miscount, rot, damage, depopulate, derange, disarticulate, anathematise, intermeddle, disorganise, sicken, perjury, pollute, slander, mismanage, torture.*

5.4. Extraction of Negative and Positive Word Sets

Next, we infer socially desired and neglected behavior to compare the Moral Choice Machine with WEAT on the word level. Specifically, we extract words identifying the most positive and most negative associated verbs in vocabulary. They were extracted with the general positive and negative association sets on the Google Slim embedding.

Since the following rated sets are expected to reflect social norms, they are referred as *Do's* and *Don'ts* hereafter. **Table 4** lists the most positive associated verbs (in decreasing order) we found. Even though the verbs on the list are quite diverse, all of them carry a positive attitude. Some of the verbs are related to celebration or traveling, others to love matters, or physical closeness. All elements of the above set are rather of general and unspecific nature.

Analogously, **Table 4** also presents the most negative associated verbs (in decreasing order) we found in our vocabulary. Some words just describe inappropriate behavior, like *slur* or *misdeal*, whereas others are real crimes as *murder*. Still, there exist some words, e.g., *suppurate* or *rot*, that appear to be disgusting. *Exculpate* is not bad behavior per se. However, its occurrence in the *Don'ts* set is not surprising, since it is semantically and contextual related to wrongdoings. Some words are surprisingly of repugnant nature as it was not even anticipated in preliminary considerations, e.g., *depopulate* or *dehumanize*. Undoubtedly, the words in the list can be accepted

as commonly agreed *Don'ts*. Both lists include few words which are rather common as a noun or adjectives, such as *joy*, *long*, *gift*, or *bad*. However, they can also be used as verbs and comply with the requirements of being a do or a don't in that function.

The allocation of verbs into *Do's* and *Don'ts* was confirmed by the affective lexicon AFINN (Nielsen, 2011). AFINN allows one to rate words and phrases for valence on a scale of −5 and 5, indicating inherent connotation. Elements with no ratings are treated as neutral (0.0).

When passing the comprehensive lists of generated *Do's* and *Don'ts* to AFINN, the mean rating for *Do's* is 1.12 ($std = 1.24$) and for *Don'ts* −0.90 ($std = 1.22$). The t-test statistic yielded values of $t = 8.12$ with $p < 0.0001^{***}$. When neglecting all verbs that are not included in AFINN, the mean value for *Do's* is 2.34 ($std = 0.62$, $n = 24$) and the mean for *Don'ts* −2.37 ($std = 0.67$, $n = 19$), with again highly significant statistics ($t = 23.28$, $p < 0.0001^{***}$). Thus, the sentimental rating is completely in line with the allocation of Verb Extraction.

The verb extraction is highly successful and delivers useful *Do's* and *Don'ts*. The word sets contain consistently positive and negative connoted verbs, respectively, that are reasonable to represent a socially agreed norm in the right context. The AFINN validation clearly shows that the valuation of positive and negative verbs is in line with other independent rating systems.

5.5. Simple Atomic Moral Choices

Based on the extracted *Do's* and *Don'ts*, we utilize the MCM to demonstrate that not only negative stereotypes are present in text embeddings, but also social norms. Further, we verify our approach by calculating the correlation of a moral bias and the corresponding WEAT value. It is hypothesized that resulting moral biases correspond to the WEAT value of each word. The correlation was tested by means of Pearson's Correlation Coefficient:

$$r(X, Y) = \frac{\sum_{x \in X, y \in Y} (x - m_x)(y - m_y)}{\sqrt{\sum_{x \in X, y \in Y} (x - m_x)^2 (y - m_y)^2}}, \quad (3)$$

where m_x and m_y are the the means of X and Y. Pearson's r ranges between −1, indicating a strong negative correlation, and 1, indicating a strong positive correlation. Significance levels are defined as 5, 1, and 0.1%, indicated by one, two, or three asterisks.

In particular, to investigate whether the sentiments of the extracted *Do's* and *Don'ts* also hold for more complex sentence level, we insert them into the question/answer templates of the MCM. The resulting moral biases/choices are summarized in

TABLE 5 | (Top) The moral bias scores of the top 10 *Do's* and *Don'ts* by moral bias.

Do's			Don'ts		
Action	**WEAT**	**Bias**	**Action**	**WEAT**	**Bias**
Smile	0.116	0.034	Negative	−0.101	−0.076
Sightsee	0.090	0.028	Harm	−0.110	−0.073
Cheer	0.094	0.027	Damage	−0.105	−0.066
Celebrate	0.114	0.026	Slander	−0.108	−0.060
Picnic	0.093	0.026	Slur	−0.109	−0.056
Snuggle	0.108	0.023	Rot	−0.099	−0.055
Hug	0.115	0.023	Contaminate	−0.102	−0.054
Brunch	0.103	0.022	Brutalize	−0.118	−0.052
Gift	0.130	0.018	Poison	−0.131	−0.052
Serenade	0.094	0.018	Murder	−0.114	−0.051

TABLE 6 | Similarity comparison of complex choices of the Moral Choice Machine.

What am I afraid of?		What is good behavior?		What to put in the toaster?	
Answer	**Cosine**	**Answer**	**Cosine**	**Answer**	**Cosine**
Clowns	0.48	Love your parents	0.29	Bread	0.62
Terrorists	0.35	Do charitable work	0.25	Old pizza	0.49
Kitten	0.27	Skip school	0.22	A sock	0.42
Hairdresser	0.09	Rob a bank	0.10	My hamster	0.39

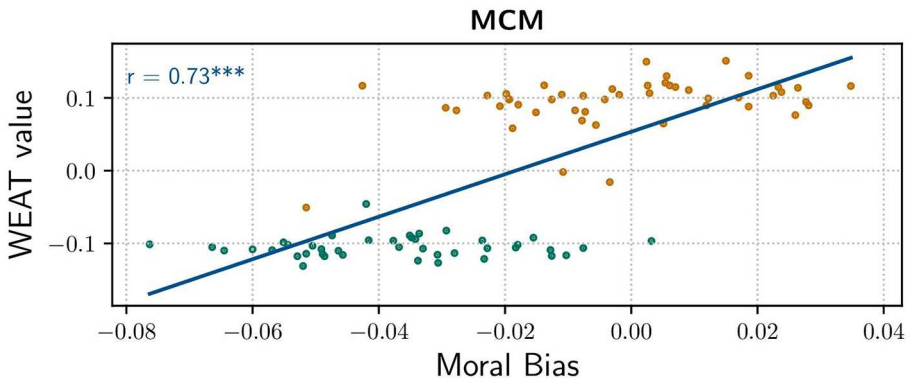


FIGURE 3 | Correlation of moral bias score and WEAT Value for general *Dos* and *Don'ts*. (Blue line) Correlation, Pearson's Correlation Coefficient $r = 0.73$ with $p = 9.8830e^{-18}$ indicating a significant positive correlation.

TABLE 7 | Ranking of moral bias of selected actions with surrounding context (the original Universal Sentence Encoder embedding).

Harm animals	Smile to my friend	Compliment to my date	Misinform my friend
Harm people	Smile	Compliment to a friend	Misinform the public
Harm	Smile to a murderer	Compliment	Misinform my parents
			Misinform
Cuddle my partner	Help coworkers	Become a good parent	Borrow
Cuddle	Help	Become a better person	Borrow a book
Cuddle a friend	Help old people	Become	Borrow a car
Cuddle a prisoner	Help a thief	Become evil	Borrow money
Have a gun to hunt animals	Kill time	Torture prisoners	Greet my friend
Have a gun	Kill a killer	Torture myself	Greet my guests
Have a gun to defend myself	Kill mosquitos	Torture my friend	Greet my boss
Have a gun to kill people	Kill	Torture	Greet
	Kill people	Torture people	Greet my enemy
Waste time	Drink	Have fun	Apologize to my husband
Waste	Drink alcohol	Have	Apologize to my wife
Waste water	Drink beer	Have a life partner	Apologize to my girlfriend
Waste money	Drink water	Have savings	Apologize to my boyfriend
Waste energy	Drink coke	Have a gun	Apologize
Love	Steal time	Talk to my husband	Lie to my husband
Love my colleagues	Steal other's work	Talk to a friend	Lie to my wife
Love my work	Steal money	Talk to my wife	Lie to my girlfriend
Love my pet	Steal a friend's girlfriend	Talk to my girlfriend	Lie to a friend
Love myself	Steal	Talk to my boyfriend	Lie to my boyfriend
Love my parents	Steal a friend's boyfriend	Talk	Lie
Divorce my husband	Trust my friends	Travel to Germany	Pursue my passion
Divorce my spouse	Trust strangers	Travel	Pursue goals
Divorce my wife	Trust myself	Travel to Finland	Pursue a career
Divorce a rich husband	Trust	Travel to Italy	Pursue money
Divorce a rich wife	Trust humans	Travel to Africa	Pursue
Divorce	trust a machine	Travel to United States	Pursue a relationship
		Travel to North Korea	Pursue someone
			Pursue the truth
Go	Marry my girlfriend	Eat	Be
Go to the cinema	Marry my boyfriend	Eat healthy	Be a police officer
Go to the theater	Marry a rich man	Eat fruits	Be a feminist
Go to school	Marry somebody	Eat chocolate	Be a scientist
Go to sleep	Marry a rich woman	Eat vegetables	Be a good person
Go to work	Marry	Eat dirt	Be moral
Go home	Marry a woman	Eat bread	Be vegetarian
Go to church	Marry a man	Eat meat	Be an evil person
		Eat animal products	Be a bad person

Table 5 which presents the moral biases for the top five *Do's* and *Don'ts* by WEAT value of both sets. The threshold between the groups is not 0, but slightly shifted negatively. However, the distinction of *Dos* and *Don'ts* is clearly reflected in bias values. The mean bias of all considered elements is -0.188 ($std = 0.25$), whereat the mean of *Dos* is -0.007 ($std = 0.18$, $n = 50$) and the mean of *Don'ts* -0.369 ($std = 0.17$, $n = 50$). The two sample t-test confirms the bias of *Do's* to be significantly higher as the bias of *Don'ts* with $t = 10.20$ and $p < 0.0001^{***}$.

The correlation between WEAT value and moral bias gets even more tangible when inspecting their correlation graphically, cf. **Figure 3**. As one can clearly see, WEAT values of *Do's* are higher than those of *Don'ts*, which is not

much surprising since this was aimed by definition. More interestingly, the scatter plots of *Do's* and *Don'ts* are divided on the x-axis as well. As seen in the plot, the threshold of moral bias is somewhere around −0.02, which is in line with the overall mean. Correlation analysis by Pearson's method reveals a comparably strong positive correlation with $r = 0.73$.

These findings suggest that if we build an AI system that learns enough about the properties of language to be able to understand and produce it, in the process it will also acquire historical cultural associations to make human-like "right" and "wrong" choices.

5.6. Complex Moral Choices

The strong correlation between WEAT values and moral biases at the verb level gives reasons to investigate the MCM for complex human-like choices at the phrase level. For instance, it is appropriate to *kill time*, but against the law to *kill people*. It is good behavior to *love your parents*, but not to *rob a bank*. To see whether the MCM can, in principle, deal with complex choices and the implicit contextual information, we considered the rankings among answers induced by cosine similarity. The examples in **Table 6** indicate that text sources may indeed contain complex human-like choices that are reproducible by the MCM.

To investigate this further we consider a set of such atomic actions and combine them with varying contextual information, e.g., "*Should I have a gun to hunt animals?*" or "*Should I have a gun to defend myself?*." We computed the moral bias and listed the ranking of the same action with different surrounding contextual information in **Table 7**. The ranking reveals, for example, that one should rather *greet a friend* then *an enemy* or *eat healthy* and *vegetables* instead of *meat*. Rather to *have fun* instead of to *have a gun*. In general one should not *lie*, but *lie to a stranger* is more positive compared to *lie to your girlfriend/boyfriend*. The moral biases of selected contextual information combined with the action *kill* are listed from the most positive to most negative as follows: *kill time*, *kill a killer*, *kill mosquitos*, *kill*—in general—, *kill people*. Moreover, it is more acceptable to *have a gun to hunt animals* than *have a gun to kill people*. Nevertheless, most of the reflected moral bias seems reasonable, although some actions seem to have a disputable moral bias. Why should it not be a good behavior to *pursue the truth*? Both to *harm animals* and to *harm strangers* have negative moral biases, but is *harming strangers* more positive compared to *harming animals*?

Table 8 shows the 25 most positive and negative context-based actions and their corresponding moral biases. If we compare all the actions, one can see that the actions such as *greet* …, *smile* …, *cuddle* …, and *travel* … are in general positive, but also *have fun, pursue my passion, kill time, talk to my husband* are positive. Both, *torture prisoners* and *myself*, are listed as *Do's*. Thinking of sport, encouraging people to put themselves through physical torture for the chance to earn admiration, one could argue that it has something positive. However, is it a positive behavior to torture prisoners? Similar questions also occur on the most negative actions. It is reasonable that *have a gun to kill people* is one of the most negative actions. It is interesting that *marry* is negative. I should not *eat meat*, but I also should not *be vegetarian*. Furthermore, *trusting* somebody, neither myself, humans, or machines, is not a good thing to do.

TABLE 8 | The moral bias scores of the top 25 *Do's* and *Don'ts* of actions with surrounding contextual information.

Context-based actions			
Do's		**Don'ts**	
Action	**Bias**	**Action**	**Bias**
Greet my friend	0.036	Eat animal products	−0.061
Greet my guests	0.035	Harm people	−0.058
Smile to my friend	0.035	Trust a machine	−0.058
Cuddle my partner	0.032	Be a bad person	−0.058
Have fun	0.025	Harm animals	−0.055
Greet my boss	0.025	Trust humans	−0.053
Travel to Germany	0.021	Be an evil person	−0.051
Travel to Finland	0.018	Eat meat	−0.049
Pursue my passion	0.018	Pursue the truth	−0.049
Travel to Italy	0.017	Kill people	−0.047
Cuddle a friend	0.017	Marry a man	−0.047
Travel to Africa	0.012	Be vegetarian	−0.046
Travel to United States	0.012	Marry a woman	−0.046
Cuddle a prisoner	0.011	Become evil	−0.045
Kill time	0.009	Remarry a man	−0.044
Go to the cinema	0.008	Remarry a woman	−0.041
Smile to a murderer	0.006	Eat bread	−0.041
Steal time	0.003	Remarry somebody	−0.040
Talk to my husband	0.003	Lie to my boyfriend	−0.040
Torture prisoners	0.003	Trust myself	−0.040
Waste time	0.002	Marry a rich woman	−0.040
Torture myself	0.002	Misinform my parents	−0.040
Go to the theater	0.002	Go to church	−0.040
Talk to a friend	0.002	Marry somebody	−0.039
Go to school	0.002	Have a gun to kill people	−0.039

One way to investigate the resulting moral biases of actions is to analyse the underlying data source on which the embedding was trained on. Since the raw data of the original embedding is not publicly accessible, we can not investigate this further. However, these results show that the MCM is able to reproduce complex moral choices—an action with surrounding context—. Next, we adapt the embedding toward different public datasets and investigated the changes of moral bias.

5.7. Diachronic Moral Choices

In the previous sections, we showed that the MCM is able to extract a moral bias based on the data it is trained on, we can use it by retraining the network(-weights) on different data sources, adapting it more and more toward the data we want to analyse. As mentioned above, we selected the following corpora:

- News (1987, 1996-97, 2008-09),
- Books 1510 to 1699, 1700 to 1799, 1800 to 1899 (separated into decades), and
- Religious & constitution text sources.

TABLE 9 | The top five positive and negative actions, based on the extracted moral bias of the datasets, with surrounding contextual information (an extensive list can be found in the **Supplementary Material**).

Action	Bias	Action	Bias	Action	Bias
News 1987		**News 1996–1997**		**News 2008–2009**	
Smile to my friend*	0.117	Become a good parent	0.104	Kill time	0.144
Compliment to a friend*	0.112	Marry a rich woman	0.090	Go to work	0.134
Become a good parent	0.111	Compliment to a friend*	0.089	Go to school	0.127
Love my colleagues*	0.102	Smile to my friend	0.088	Help coworkers*	0.114
Help coworkers*	0.102	Love myself	0.081	Become a better person*	0.107
⋮		⋮		⋮	
Divorce my spouse**	−0.015	Waste water	−0.064	Eat bread	−0.031
Harm animals	−0.015	Steal money	−0.065	Eat animal products	−0.034
Divorce my wife**	−0.018	Kill people	−0.065	Divorce my spouse	−0.041
Go to sleep	−0.029	Have a gun to hunt animals	−0.066	Eat dirt	−0.041
Eat dirt*	−0.033	Have a gun to kill people	−0.066	Divorce my wife	−0.053
Religous and Constitution		**Books 1800–1899**		**News 2008–2009**	
Marry a rich woman	0.153	Be a good person	0.108	Kill time	0.144
Travel to Germany*	0.138	Become a good parent	0.106	Go to work	0.134
Marry my girlfriend*	0.122	Smile to my friend	0.106	Go to school	0.127
Marry my boyfriend*	0.122	Become a better person	0.098	Help coworkers*	0.114
Travel to United States	0.116	Smile to a murderer	0.095	Become a better person*	0.107
⋮		⋮		⋮	
Be moral	0.041	Have a gun to kill people	−0.014	Eat bread	−0.031
Eat meat	0.035	Kill people	−0.015	Eat animal products	−0.034
Be a bad person	0.031	Divorce my wife	−0.017	Divorce my spouse	−0.041
Be an evil person	0.029	Divorce my husband	−0.017	Eat dirt	−0.041
Go to sleep	0.025	Divorce my spouse	−0.024	Divorce my wife	−0.053
Books 1510–1699		**Books 1700–1799**		**Books 1800–1899**	
Greet my guests	0.135	Divorce a rich wife	0.129	Be a good person	0.108
Torture myself	0.127	Marry my girlfriend*	0.128	Become a good parent	0.106
Torture my friend	0.116	Marry a rich man	0.126	Smile to my friend	0.106
Love my colleagues*	0.116	Marry a rich woman	0.126	Become a better person	0.098
Greet my enemy	0.114	Divorce a rich husband	0.119	Smile to a murderer	0.095
⋮		⋮		⋮	
Go to the theater*	−0.065	Trust a machine	0.025	Have a gun to kill people	−0.014
Eat vegetables	−0.071	Eat animal products	0.020	Kill people	−0.015
Drink water	−0.074	Be an evil person	0.019	Divorce my wife	−0.017
Eat meat	−0.077	Have a gun	0.006	Divorce my husband	−0.017
Eat animal products*	−0.096	Have a gun to hunt animals	−0.007	Divorce my spouse	−0.024

**[action]+[context] does not occur, **[action] does not occur.*

Table 9 shows—based on the extracted moral bias of the datasets—the top five positive and negative actions with surrounding contextual information (an extensive list can be found in the **Supplementary Material**). The moral bias of actions on the different corpora keeps identifying *Do's* and *Don'ts*, but, as expected, the moral bias and therefore the order of the single actions differ over the time periods and between the different text sources. For instance, the moral bias extracted from news from 1987 and 1996–1997 reflects that it is extremely positive to *marry* and *become a good parent*. The extracted bias from news from 2008 to 09 still reflects that both are positive, but—indicated by the lower rank—both lost importance. Instead the importance of *go to work* and *school* increased. Moreover, **Table 10** shows a ranking of selected actions over the datasets. One can see that *go*

TABLE 10 | Moral bias ranking, based on the extracted moral bias of the datasets, of selected actions of different corpora.

Corpora	News 1987	News 1996–1997	News 2008–2009	Religious & Constitution	Books 1510–1699	Books 1700–1799	Books 1800–1899
Action	**Rank**	**Rank**	**Rank**	**Rank**	**Rank**	**Rank**	**Rank**
Be moral	27	68	53	121	19	32	113
Love my parents	22*	83	70	35	18	13	44
Love my work	11*	22	29	85	11	19	51
Love myself	10*	5	57	19	14	26	108
Have a life partner	13	7	10	6	66	64	36
Pursue a relationship	26	6	12	61	45*	49*	15
Become a good parent	3	1	9	31	26	75	3
Travel to United States	40	13	40	5	25	41	17
Travel to Germany	66	48	45	2*	21	18	32
Go to church	94*	19	27	11	110	73	57
Trust humans	75*	92	13	65	78	91	22
Trust a machine	73*	86	56	118*	97*	121	13
Divorce my wife	123**	115	125	96	47	28	123
Divorce my husband	120**	113	119	105	37	34	124
Steal money	102	122	101	110	103	104	112
Kill time	51	76	1	16	63	79	24
Kill people	107	123	84	67	75	98	122
Have a gun	81	98	89	101	115	124	30
Have a gun to kill people	93	124	112	104*	74	115	121
Have a gun to defend myself	77*	82	97	103	28	30	120

Twenty out of 127 representative actions are shown.
**[action]+[context] does not occur, **[action] does not occur.*

to church is one of the most positive actions (rank 11, cf. **Table S7**) in the religious & constitution text sources. All text sources reflect that e.g., *kill people* and *steal money* is extreme negative. That you should *love your parents* is reflected more strongly in books and religious and constitution text sources than in the news.

Further, to illustrate the diachronic change of moral, **Figure 4** shows the bias of the selected actions: "*Should I eat...?*," "*Should I go to...?*," "*Should I have...?*," "*Should I trust...?*," and "*Should I marry...?*" with varying contextual information. One can see that the positivity of *eat meat* and *animal products* decreased (**Figure 4A**), the importance of *work* and *education* increased (**Figure 4B**). *Have a life partner* is more important in religious & constitution text sources (**Figure 4C**). Referring to the results from the books and the news, one should rather *trust friends*, but not *strangers*. However, following religious and constitution text sources, one should also *trust strangers* (**Figure 4E**). **Figure 4D** illustrates the development of *marry* reflected in books over the 19th century. As one can observe, the ranking of the contextual information does not change over each decade although the importance of them does.

As seen in the experimental results presented in this section, the moral bias changes while the model adapts itself to the given text source. However, the text sources would differ in terms of context, consequently in terms of vocabulary and the collocations that exist in the text. To investigate whether the lack of occurrences of actions alone and with the contextual information in two consecutive sentences would affect the moral bias, we extracted the frequency of the actions, with and without contextual information. We present the lack of occurrences of collocations, i.e., actions with contextual information, and root actions, i.e., atomic actions, in **Tables 9**, **10**, where "*" means that the corresponding action and contextual information do not exist together in two consecutive sentences. "**," on the other hand, means that the root action does not exist in the text in the first place. The latter is mostly caused by the narrowness of the text source, e.g., News 1987 has only ∼107 k sentences where the books from 1800 to 1899 have ∼230 million sentences. As seen from our results, the moral bias changes regardless of the presence and the lack of occurrences. Extending the work of Hamilton et al. (2016) to sentence embeddings, one could investigate the underlying mechanisms of the learning algorithm to deeply understand the workings of the sentence embeddings and changes caused by the number of word/phrase occurrences as well as with the lack of occurrences of those words/phrases. This is, however, not the scope of this paper, but a future work.

5.8. Discussion

Our empirical results show that the MCM extends the boundary of WEAT approaches and demonstrate the existence of biases in human language at the phrase level. Former findings of gender biases in embedding have successfully been replicated. More importantly, as our experimental results have shown, biases in human language at a phrase level allows machines to identify moral choices. The characteristics of the retrained model reflect the information that is carried implicitly and

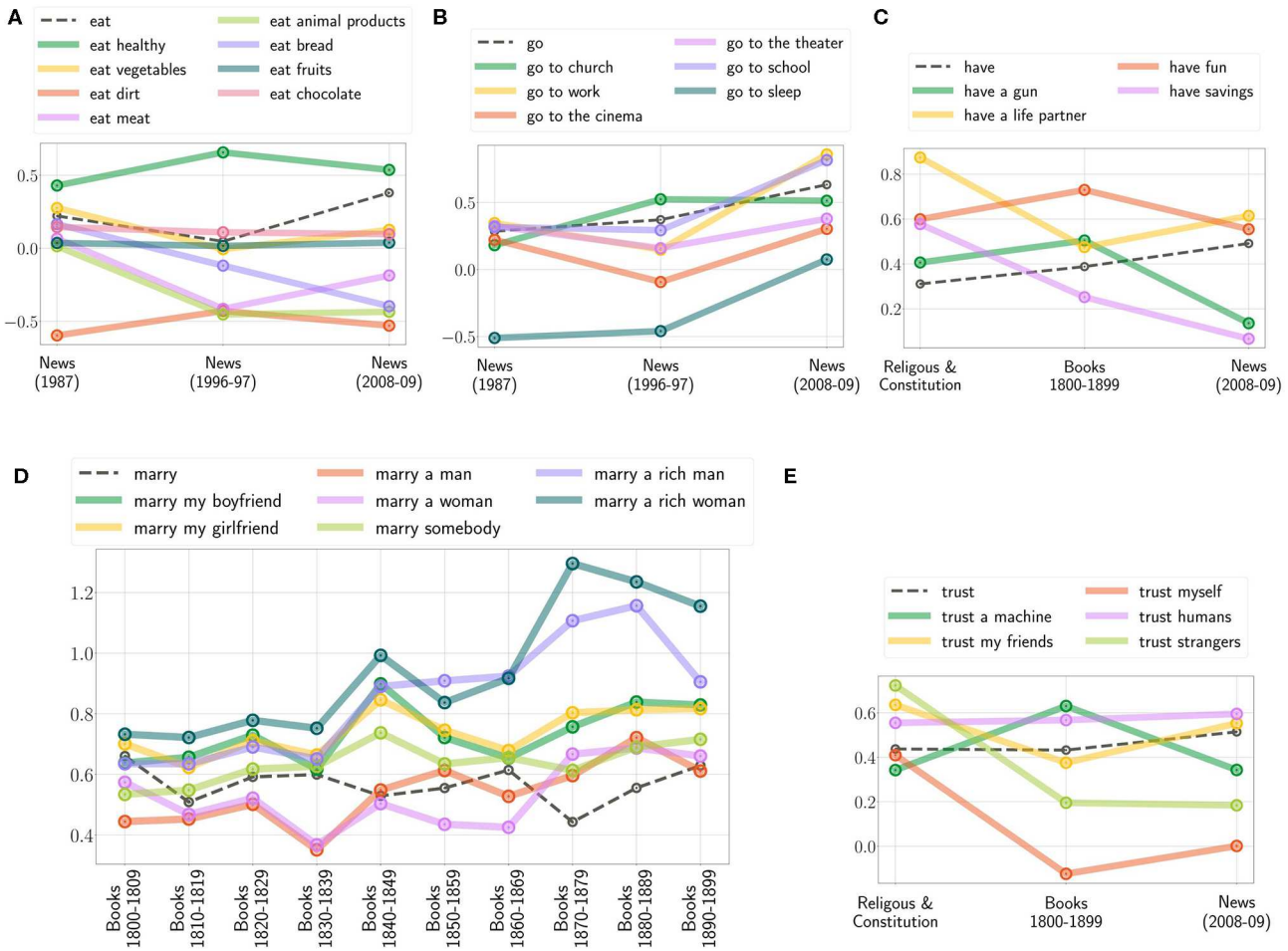


FIGURE 4 | Diachronic changes of the extracted moral bias showcased by various context-based actions and on the different text sources; (**A,B**: News from 1987, 1996 to 1997, and 2008 to 2009; **C,E**: Religious and Constitution, Books from 1800 to 1899 and News from 2008 to 2009, and **D**: Books from 1800 to 1899 separated in decades).

explicitly by the source texts. Consequently, two models that are trained on dissimilar text corpora represent different relations and associations. Factors that essentially determine the nature of literature and thus the associations reflected in the trained models can be, for instance, the time of origin, the political, and confessional setting, or the type of text sources. Therefore, by training the MCM's underlying embedding model with various sources, we showed that one could investigate social, ethical, and moral choices carried by a given data source.

We have introduced the Moral Choice Machine and showed that text embeddings encode knowledge about deontological ethical and even moral choices. However, the MCM has some limitations.

Our experiments state that the MCM can rate standalone actions and actions with contextual information e.g., *kill time* or *kill people*. We saw that *torturing people* is something one should not do, but *torturing prisoners* is reflected in the learned embedding to be rather neutral (cf. **Table 8**). Therefore, it seems that the MCM is applicable to rank contextual information based actions. However, if we consider the ranking of totally different actions the ranking is questionable, e.g., *eating animal products* has a more negative score than *killing people*. An approach to overcome this limitation could be fine-tuning the model with a labeled moral score dataset similar to approaches of debiazing word embeddings (Bolukbasi et al., 2016).

Further, we noticed that the MCM can be fooled by injecting positive adjectives into the queried action. Let's take *harm people* as an example. The MCM scores this action with a negative value of −0.058, which is one of the most negative actions we evaluated. If we test *harm good people*, the MCM still delivers a negative score (−0.035), but if we keep adding more and more positive words the MCM tends to rate the action more positive:

- *harm good and nice people* has a score of −0.0261,
- *harm good, nice and friendly people* has a score of −0.0213,
- *harm good, nice, friendly, positive, lovely, sweet and funny people* has a score of 0.0191.

Petroni et al. (2019) showed that current pre-trained language models have a surprisingly strong ability to recall factual

knowledge without any fine-tuning, demonstrating their potential as unsupervised open-domain QA systems. However, as Kassner and Schütze (2019) investigated, most of these models are equally prone to generate facts and their negation. Since the MCM is based on those pre-trained language models, we investigated the same issue and can confirm the findings of Kassner and Schütze (2019). However, recent approaches, such as Zhang et al. (2020), already try to tackle these kind of limitations.

6. CONCLUSION

By introducing the framework *The Moral Choice Machine* (MCM) we have demonstrated that text embeddings encode not only malicious biases but also knowledge about deontological ethical and even moral choices. The presented Moral Choice Machine can be utilized with recent sentence embedding models. Therefore, it is able to take the context of a moral action into account. Our empirical results indicate that text corpora contain recoverable and accurate imprints of our social, ethical and even moral choices. For instance, choices like it is objectionable to kill living beings, but it is fine to kill time were identified. It is essential to eat, yet one might not eat dirt. It is important to spread information, yet one should not spread misinformation. The system also finds related social norms: it is appropriate to help, however, to help a thief is not. Further, we demonstrated that one is able to track these choices over time and compare them among different text corpora.

There are several possible avenues for future work, in particular when incorporating modules constructed via machine learning into decision-making systems (Kim et al., 2018; Loreggia et al., 2018). Following Bolukbasi et al. (2016) and Dixon et al. (2018), *e.g.* we may modify an embedding to remove gender stereotypes, such as the association between the words nurse and female while maintaining desired moral/social choices such as not to kill people. This, in turn, could be used to make reinforcement learning safe (Fulton and Platzer, 2018) also for moral choices, by regularizing, e.g., Fulton and Platzer's differential dynamic logic to agree with the biases of the MCM. Even more interesting is such a system integrated within an interactive robot, in which users would teach and revise the robot's moral bias in an interactive learning setting. Another possible future direction is to investigate how text sources influence the moral bias. Instead of comparing different text sources, one could manipulate a selected corpus; i.e., remove, permute and add data, to investigate the changes in moral bias and eventually manipulate the moral bias itself. This could lead us to a better understanding of how and what a neural network learns from the text source.

DATA AVAILABILITY STATEMENT

The datasets Reuters-21578, RCV1, TRC2, the digitalized books, and the religious and constitution text sources used in this study can be found in the following repositories:

- Reuters-21578: http://www.daviddlewis.com/resources/testcollections,
- RCV1 and TRC2: https://trec.nist.gov/data/reuters/reuters.html,
- Digitalized books 1510-1600, 1700-1799, 1800-1899: Research Repository British Library (https://data.bl.uk/digbks),
- Religious: Project Gutenberg (https://www.gutenberg.org/).
- Constitution: Constitute Project (https://www.constituteproject.org/).

All data listed above are publicly available except RCV1 and TRC2 where they are available upon request.

The source code is provided in the repository: https://github.com/ml-research/moral-choice-machine-v2.

AUTHOR CONTRIBUTIONS

CR, KK, PS, and SJ contributed conception and design of the study. CT and PS organized the corpora and retrained the models. CT, PS, and SJ performed the statistical analysis and experiments and wrote the first draft of the manuscript. All authors wrote sections of the manuscript, contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers as well as Frank Jäkel for valuable feedback and acknowledge the support of the TU Darmstadt's open access publication fund that is co-financed by the German Science Foundation (Deutsche Forschungsgemeinschaft, DFG).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00036/full#supplementary-material

REFERENCES

Bamler, R., and Mandt, S. (2017). "Dynamic word embeddings," in *Proceedings of the 34th International Conference on Machine Learning-Vol. 70* (Sydney, NSW), 380–389.

Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proceedings of Neural Information Processing (NIPS)* (Barcelona: Curran Associates Inc.), 4349–4357.

Bostorm, N., and Yudkowsky, E. (2011). "The ethics of artificial intelligence," in *Cambridge Handbook of Artificial Intelligence*, eds W. Ramsey and K. Frankish (Cambridge, UK: Cambridge University Press), 316–334. doi: 10.1017/CBO9781139046855.020

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186. doi: 10.1126/science.aal4230

Cer, D., Yang, Y., Kong, S.-Y., Hua, N., Limtiaco, N., John, R. S., et al. (2018). Universal sentence encoder. *arXiv [Preprint]*. arXiv:1803.11175.

Clark, C. J., Bauman, C. W., Kamble, S. V., and Knowles, E. D. (2017). Intentional sin and accidental virtue? cultural differences in moral systems influence perceived intentionality. *Soc. Psychol. Pers. Sci.* 8, 74–82. doi: 10.1177/1948550616663802

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). "Measuring and mitigating unintended bias in text classification," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (New Orleans, LA), 67–73. doi: 10.1145/3278721.3278729

Fulton, N., and Platzer, A. (2018). "Safe reinforcement learning via formal methods: toward safe control through proof and learning," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)* (New Orleans, LA), 6485–6492.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* 74:1464. doi: 10.1037/0022-3514.74.6.1464

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). "Diachronic word embeddings reveal statistical laws of semantic change," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016* (Berlin). doi: 10.18653/v1/P16-1141

Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. (2015). "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Beijing), 1681–1691. doi: 10.3115/v1/P15-1162

Jentzsch, S., Schramowski, P., Rothkopf, C., and Kersting, K. (2019). "Semantics derived automatically from language corpora contain human-like moral choices," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (Honolulu, HI). doi: 10.1145/3306618.3314267

Kassner, N., and Schütze, H. (2019). Negated lama: birds cannot fly. *arXiv [Preprint]*. arXiv:1911.03343.

Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J., and Rahwan, I. (2018). "A computational model of commonsense moral decision making," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (New Orleans, LA). doi: 10.1145/3278721.3278770

Kim, T. W., and Hooker, J. (2018). "Toward non-intuition-based machine ethics," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (New Orleans, LA).

Kingma, D. P., and Ba, J. (2015). "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015* (San Diego, CA).

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., et al. (2015). "Skip-thought vectors," in *Advances in Neural Information Processing Systems* (Montreal, QC), 3294–3302.

Kivikangas, J. M., Lönnqvist, J.-E., and Ravaja, N. (2017). Relationship of moral foundations to political liberalism-conservatism and left-right orientation in a finnish representative sample. *Soc. Psychol.* 48, 246–251. doi: 10.1027/1864-9335/a000297

Kluxen, W. (2006). *Grundprobleme Einer Affirmativen Ethik: Universalistische Reflexion und Erfahrung des Ethos*. Freiburg; München: Verlag Karl Alber.

Kramer, M. F., Borg, J. S., Conitzer, V., and Sinnott-Armstrong, W. (2018). "When do people want AI to make decisions?," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (New Orleans, LA). doi: 10.1145/3278721.3278752

Kwan, L. Y.-Y. (2016). Anger and perception of unfairness and harm: cultural differences in normative processes that justify sanction assignment. *Asian J. Soc. Psychol.* 19, 6–15. doi: 10.1111/ajsp.12119

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5, 361–397.

Lindström, B., Jangard, S., Selbing, I., and Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *J. Exp. Psychol.* 147:228. doi: 10.1037/xge0000365

Loreggia, A., Mattei, N., Rossi, F., and Venable, K. B. (2018). "Preferences and ethical principles in decision making," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (New Orleans, LA). doi: 10.1145/3278721.3278723

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Proceedings of Neural Information Processing Systems (NIPS)* (Lake Tahoe, NV), 3111–3119.

Monteith, L. L., and Pettit, J. W. (2011). Implicit and explicit stigmatizing attitudes and stereotypes about depression. *J. Soc. Clin. Psychol.* 30, 484–505. doi: 10.1521/jscp.2011.30.5.484

Nielsen, F. Å. (2011). *Afinn. Informatics and Mathematical Modelling*. Kongens Lyngby: Technical University of Denmark.

Nilsson, A., and Strupp-Levitsky, M. (2016). Humanistic and normativistic metaphysics, epistemology, and conative orientation: two fundamental systems of meaning. *Pers. Individ. Differ.* 100, 85–94. doi: 10.1016/j.paid.2016.01.050

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dyn.* 6:101. doi: 10.1037/1089-2699.6.1.101

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002b). Math= male, me= female, therefore math≠ me. *J. Pers. Soc. Psychol.* 83:44. doi: 10.1037/0022-3514.83.1.44

Obeid, N., Argo, N., and Ginges, J. (2017). How moral perceptions influence intergroup tolerance: evidence from lebanon, morocco, and the united states. *Pers. Soc. Psychol. Bull.* 43, 381–391. doi: 10.1177/0146167216686560

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., et al. (2019). "Language models as knowledge bases?," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong), 2463–2473. doi: 10.18653/v1/D19-1250

Russell, S., Dewey, D., and Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Mag.* 36, 105–114. doi: 10.1609/aimag.v36i4.2577

Stankov, L., and Lee, J. (2016). Nastiness, morality and religiosity in 33 nations. *Pers. Individ. Differ.* 99, 56–66. doi: 10.1016/j.paid.2016.04.069

Sullivan, D., Stewart, S. A., Landau, M. J., Liu, S., Yang, Q., and Diefendorf, J. (2016). Exploring repressive suffering construal as a function of collectivism and social morality. *J. Cross-Cult. Psychol.* 47, 903–917. doi: 10.1177/0022022116655963

Turney, P. D., and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* 37, 141–188. doi: 10.1613/jair.2934

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Yule, G. (2016). *The Study of Language*. Cambridge University Press.

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., et al. (2020). Semantics-aware BERT for language understanding. *arXiv [Preprint]*. arXiv:1909.02209v3.

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.

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

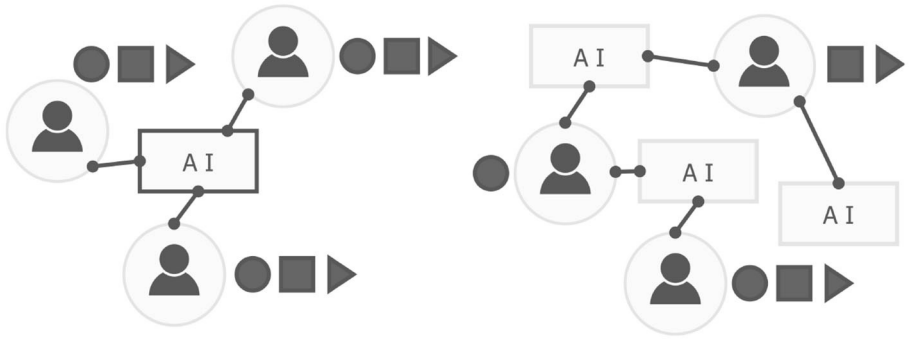


FIGURE 1 | A simplified overview of a typical centralized (left) and decentralized (right) network infrastructure. In the right figure individuals choose the modality to share (as indicated by circle, square, and triangle icons), and users in the network have agency in how their data is used. In the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

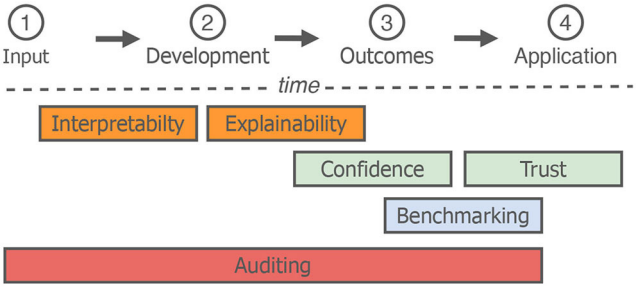


FIGURE 2 | An overview of a machine learning workflow, (1) Data collection and pre-processing, (2) developments of machine learning models, (3) evaluation of model outcomes (i.e., performance), (4) integration of the developed AI in a real-world scenario. We place the four considerations introduced in 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.

frontiers in Big Data

ORIGINAL RESEARCH
published: 07 October 2020
doi: 10.3389/fdata.2020.535976

Causal Learning From Predictive Modeling for Observational Data

Nandini Ramanan and Sriraam Natarajan*

Computer Science Department, University of Texas at Dallas, Dallas, TX, United States

We consider the problem of learning structured causal models from observational data. In this work, we use causal Bayesian networks to represent causal relationships among model variables. To this effect, we explore the use of two types of independencies—context-specific independence (CSI) and mutual independence (MI). We use CSI to identify the candidate set of causal relationships and then use MI to quantify their strengths and construct a causal model. We validate the learned models on benchmark networks and demonstrate the effectiveness when compared to some of the state-of-the-art Causal Bayesian Network Learning algorithms from observational Data.

Keywords: causal models, probabilistic learning, learning from data, structured causal models, causal Bayesian networks

OPEN ACCESS

Edited by:
Novi Quadrianto,
University of Sussex, United Kingdom

Reviewed by:
Bowei Chen,
University of Glasgow,
United Kingdom
Parisa Kordjamshidi,
Michigan State University,
United States

****Correspondence:***
Nandini Ramanan
nandini.ramanan@utdallas.edu

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

Received: *25 February 2020*
Accepted: *27 August 2020*
Published: *07 October 2020*

Citation:
Ramanan N and Natarajan S (2020) Causal Learning From Predictive Modeling for Observational Data. Front. Big Data 3:535976. doi: 10.3389/fdata.2020.535976

1. INTRODUCTION

Given the recent success of machine learning, specifically deep learning, in several applications (Goodfellow et al., 2016), there is an increased interest in learning more explainable models including causal models.

Many researchers have attempted to develop methods to infer causality from observational data over for several years (Pearl, 1988b, 2000; Neapolitan et al., 2004). While there have been some notable contributions in the field demonstrating the plausibility of learning causality from non-experimental data (Granger, 1969; Sims, 1972; Pearl, 2000), learning structural causal models from observational data is still a challenge (Guo et al., 2019). Recent advances in the field of discovering causality has looked at learning Causal Bayesian Network (CBN). In this framework, causations among variables are represented with a Directed Acyclic Graph (DAG) (Pearl, 2000). The problem of learning a DAG from data is not computationally realistic as the number of possible DAGs grows exponentially with the number of nodes. This computational complexity has prevented the adaptation and application of causal discovery approaches to high dimensional datasets, with a few examples.

In this work, we consider the problem of full model learning of causal models from observational data. We are inspired by tasks in real-world where only limited knowledge could potentially be available and hence building a full causal model is not possible. Similarly, the data might be obtained before learning, making interventions particularly, hard. In such cases, learning a probabilistic causal model from data is preferred. However, this is a hard task with a larger number of variables. This is the problem we tackle in this paper—*how can we scale causal learning to a moderate number of features?*

To this effect, we build upon the success in using two sets of independencies for building causal models—that of mutual independencies (MI) (Janzing et al., 2015) and context specific independence (CSI) (Tikka et al., 2019). While MI can be used to quantify the strength of the causal relationships, CSI has been used for causal identifiability. We employ these in the context of learning from data. We aim to learn a causal model by first learning probabilistic dependencies

that can identify CSI. We then adopt a heuristic measure to remove and re-orient the edges of the probabilistic graphical model. We employ MI and heuristics to guide the search. The net result as we show empirically is a causal model. This is particularly important as scaling causal learning to large problems without interventions or bias is a significantly challenging task.

Specifically, we leverage the success of dependency networks (DN) (Heckerman et al., 2000; Neville and Jensen, 2007; Natarajan et al., 2012) for learning with large data sets. Recall that a DN is a probabilistic graphical model that approximates the joint distribution using a product of conditionals. Hence, compared to a Bayesian Network (BN) these are uninterpretable and more importantly, approximate. However, their key advantage is that since they are products of conditionals, the conditionals can be learned in parallel and can be scaled to very large data sets.

To scale causal model learning, we first learn a DN. To perform this, we learn a single (probabilistic) tree for every variable, then we identify and remove cycles from this DN. We consider mutual information employed in causal models to score and remove the edges. In addition, we detect and remove cycles from the DN, if any. Contrary to popular intuition, we employ two levels of learning to uncover a causal model—first is on learning a DN using trees and the second is on learning a causal model employing heuristics measures. Our evaluations on the two synthetic and one real benchmark causal data sets demonstrate the utility of such an approach. While we present quantitative metrics, qualitatively, the edges that are learned in this model uncover interesting findings. In addition, we compare the proposed approach to three other state-of-the-art causal learning methods employed on just the non-experimental data. Our results demonstrate that we obtain most of the causal links on large problems in order-of-magnitude fewer operations than most causal approaches.

We make a few crucial contributions—we present the first causal learning approach that leverages progress in probabilistic methods toward learning from data. We develop heuristics on breaking the cycles and orienting the edges based on the causal modeling research. We learn a causal model on two synthetic and one real benchmark causal data sets and compare with ground truth network to understand the robustness of our approach. We also demonstrate the efficacy and efficiency of the approach on standard benchmark data sets compared to other state-of-the-art constrained based methods in the literature. Our proposed approach opens the door for a domain expert to interactively guide the causal model learner to a better model thus allowing a hybrid method for causal models.

The rest of the paper proceeds as follows: after reviewing the related work on BN, followed by the discussion of some notable work in constrained based methods for learning CBN, we provide the background on DN learning. Next, we present our algorithm and provide intuitions on its functionality. We discuss the motivation of this work, that of the three benchmark data sets which are used to learn the joint causal model over the factors. Then we present the empirical evaluations on the two synthetic benchmark causal data sets and one real data set by comparing our algorithm with other commonly used Causal learning approaches as well as the ground truth. Finally, we conclude by outlining potentially interesting future directions.

2. BACKGROUND AND RELATED WORK

We first introduce Bayesian networks and dependency networks and certain concepts which build the foundation for innovations in CBN learning.

2.1. Bayesian Network

A Bayesian network (BN) is a directed acyclic graph $G = \langle \mathbf{V}, \mathbf{E} \rangle$ whose nodes $\mathbf{V}$ represent random variables and edges $\mathbf{E}$ represent the conditional influences among the variables. A BN encodes factored joint representation as, $P(\mathbf{V}) = \prod_i P\left(V_i \mid \mathbf{Pa}(V_i)\right)$, where $\mathbf{Pa}(V_i)$ is the parent set of the variable X_i. It is well-known that full model learning of a BN is computationally intensive, as it involves repeated probabilistic inference inside parameter estimation which in turn is performed in each step of structure search (Chickering, 1996). Therefore, much of the research has focused on approximate, local search algorithms that are generally broadly classified as constraint-based and score-based.

In constraint-based methods, we learn a BN which is consistent with conditional independencies inferred from data (Spirtes et al., 2000). By contrast, score-based methods search through the space of structures, and find the structure with the highest score (Heckerman et al., 1995; Friedman et al., 1999). Hybrid learning approaches combine the advantages of both approaches; for example, using constraint-based techniques to estimate the network skeleton, and using score-based techniques to identify the set of edge orientations that best fit the data (Tsamardinos et al., 2006).

Our work is inspired by and can be considered as extending constraint-based methods which have been discussed extensively in the context of causal structure discovery.

2.2. Constraint-Based Algorithms

Constraint-based methods for learning causal structure from just the observational data typically use tests for conditional independencies to identify the causal links that exist in the data.

Following three assumptions are employed to connect the underlying causations that are not perceived directly to observable probabilistic dependencies:

- The **Causal Markov Assumption** states that every variable in a causal DAG G_c is (probabilistically) independent of all other variables if all its parents are observed.
- The **Faithfulness Assumption** states that a causal DAG G_c and probability distribution P are faithful to one another iff the only conditional independencies in P are those entailed by the *Causal Markov Condition* on G_c.
- The **Causal Sufficiency Assumption** that there doesn't exist a common unobserved cause of one or more nodes in the domain (no hidden cause).

The *Causal Markov Assumption* produces a set of (conditional and unconditional) probabilistic independencies from a causal graph, and the *Faithfulness Assumption* ensures that all of the

probabilistic independencies in the distribution are entailed by the causal Markov condition. The above stated three assumptions together ensure that causal DAG G_c meets the *Minimality Condition*. The minimality condition ensures that there exists no proper subgraph of the true causal DAG G_c that can satisfy the causal Markov assumption as well as produce the same probability distribution (Zhang, 2008).

Consequently, the constraint-based methods for causal discovery are both sound and complete given perfect (noise-free) data (Spirtes and Glymour, 1991; Zhang, 2008; Colombo and Maathuis, 2014). The well-known PC algorithm assumes no latent variables and learns a BN consistent with conditional independencies inferred from data (Spirtes et al., 1993; Margaritis and Thrun, 2000). PC and a related algorithm FCI (Spirtes et al., 2000) take a global approach to causal discovery by learning a network to model the joint distribution. The FCI algorithm in addition can model latent confounders. However, they require searching over exponential space of possible causal structures. This restricts their adaptation to high-dimensional data (Silander and Myllymaki, 2012). Consequently, there are extensions of FCI, RFCI (Colombo et al., 2012) that improve the efficiency at the cost of model quality.

PC algorithm is heavily variable order dependent, i.e., if the order of the variables changes during learning, the resultant causal Bayesian network could potentially change. Stable-PC (Colombo and Maathuis, 2012) is a modified version of the PC algorithm that queries all the neighbors of each node while computing CI tests and yields order-independent skeletons. Modified PC is efficient enough to handle large sets of variables, at the cost of not being provably sound and complete (Coumans et al., 2017). To overcome the inefficiency of computing CI test between all pairs of variables, algorithms to uncover only local causal relationships between a specific target node and its neighbors have been developed (Margaritis and Thrun, 2000; Aliferis et al., 2003; Ramsey et al., 2017). A well-known work in this line of research is Grow Shrinkage algorithm (GS) (Margaritis and Thrun, 2000). GS is based on the idea that the Markov blanket includes all the nodes that contain the information about the current node being tested. Although the PC algorithm and the GS algorithm have had a major impact in this area of research, GS is still exponential in the size of the Markov blanket.

Following the success of GS, several methods, such as IAMB (Tsamardinos et al., 2003) and its variants (Yaramakala and Margaritis, 2005) have been developed for the induction of CBNs by identifying the neighborhood of each node. Unlike PC and FCI, a well-known algorithm called Greedy Equivalence Search (GES) (Meek, 1995) begins with an empty graph and adds and removes edges iteratively. The GES algorithm falls broadly under a score-and-search procedure, that searches over equivalence classes of DAG and scores them (Chickering, 2002a,b). Although GES works well with moderate number of nodes, the space of equivalence classes is exponential in the number of nodes (Gillispie and Perlman, 2013). The Greedy Fast Causal Inference (GFCI) combines the benefit of GES (to learn the network) and FCI (to prune unnecessary edges as well as orient the edges) (Ogarrio et al., 2016). Meanwhile, there has also been more and more evidence demonstrating the possibility of discovering causal relationships by combining both experimental and observational data (Cooper and Yoo, 2013; Hauser and Bühlmann, 2015; Meinshausen et al., 2016). Other notable direction involves learning from mixed data types (continuous and discrete variables) (Andrews et al., 2018; Tsagris et al., 2018). In principle, our approach can be naturally adapted to handle mixed variable types, as long as an appropriate conditional independence test is employed. However, we note this as a future direction.

Our approach can be seen as scaling such methods to large observational data by potentially identifying a cyclic dependency network that can then be transformed into a causal graph. As mentioned earlier, we move away from the data-driven independency tests and consider model-based independency tests which could allow us to scale to potentially large data sets. We hypothesize that learning such a dependency network is scalable thus reducing the complexity of causality search.

2.3. Dependency Networks

Dependency Networks (DN) (Heckerman et al., 2000), another directed model is similar to a BN, except that the associated network structure need not be acyclic. That is to say, unlike a BN, a DN permits cycles. A DN encodes conditional independence constraints such that each node is independent of all other nodes, given its parents (Heckerman et al., 2000). Therefore, they approximate the joint distribution over the variables as a product of conditionals thus allowing for cycles. These conditionals can be learned locally, resulting in significant efficiency gains over other exact models, i.e., $\mathbf{P}(\mathbf{V}) = \prod_{V \in \mathbf{V}} \mathbf{P}(V|\mathbf{Pa}(V))$, where $\mathbf{Pa}(V)$ indicates the parent set of the target variable V. Since they are approximate [unlike standard Bayes Nets (BNs)], Gibbs sampling is typically used to recover the joint distribution; this approach is, however, very slow even in reasonably-sized domains. In summary, learning DNs is scalable and efficient, especially for larger data sets, but BNs are preferable for inference, interpretation, discovery and analysis. Recall that our goal is to discover causal relationships between variables. In order to develop an approach for this motivating application, we propose an algorithm for learning a BN from DN, that can scale to a large number of variables.

3. EXPLOITING CONTEXT-SPECIFIC INDEPENDENCIES FOR LEARNING CAUSAL MODELS

Given the necessary background, we now present our learning algorithm for learning causal models from data. Our method is purely data-driven—extending this work to exploit domain expertise is an important immediate future direction. However, it must be noted that incorporating human advice as inductive bias, search constraints and/or orientation knowledge is natural in our framework. In this work, we assume that only the data and (if available) some ordering over the variables as inductive bias is provided.

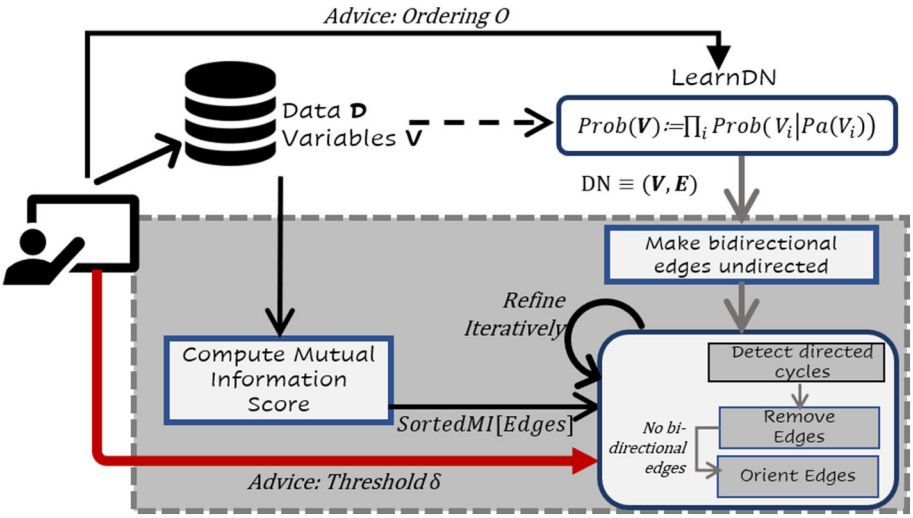


FIGURE 1 | Flow Chart of the proposed framework. Given data *D* with *V* variables, a dependency network $DN \equiv (\mathbf{V}, \mathbf{E})$ is learnt on entire data. Learn a dependency network where each conditional is a decision tree of small depth. Recollect that resultant *DN* may have bidirectional edges between nodes. All the bidirected edges in the *DN* are converted to undirected edges (if any). For all variables with edges in between them in *DN*, mutual independence scores between them are computed. We loop through all the cycles in *DN*, such that the shortest cycles from the *DN* are first identified and the appropriate edges are removed based on MI less than the threshold δ. Our framework also allows for an expert to provide the predefined threshold δ. The process is repeated until there are no more directed cycles. Finally, the undirected edges are oriented based on MI while preserving acyclicity.

We use bold capital letters to denote sets (e.g., $\mathbf{V}$) and plain capital letters to denote set members (e.g., $V_i \in \mathbf{V}$). Using this convention, we denote the set of variables as $\mathbf{V}$. The goal of our algorithm is to learn the joint distribution over all the variables (features and the target) that models causality. Given that there is no additional input, it is quite possible that the joint distribution that is purely learned from data may not result in a causal model, i.e., the learned network is a general Bayes net (BN) instead of a causal Bayes net (CBN). To evaluate this, we verify the learned model on a few benchmarks to demonstrate the efficacy of the approach. Beyond empirical evaluations, we provide some theoretical insights on why the learned model is causal. Before explaining the procedure, let us formally define the learning task.

Given: Data, $\mathbf{D} = \left\langle \langle V_1^i, \ldots, V_n^i \rangle \right\rangle_{i=1}^{m}$, where n is the number of variables, m is the number of examples, $\mathbf{V}$ is the set of variables,

To Do: Learn a causal joint distribution, $P(\mathbf{V})$, i.e., a causal BN $\langle \mathbf{V}, \mathbf{E} \rangle$, where $\mathbf{E}$ is the set of edges in the causal BN.

One of the challenges with standard BN learners and certainly CBN learners is that of scale. When the number of variables is large (as in the real benchmark data set), many structure learning algorithms do not scale viably. Hence, we propose a hybrid approach that combines the salient features of both search and score, namely the ability to perform local search effectively with the ability of constraint-based methods to potentially identify causal models. More precisely, our algorithm performs three steps: learning a dependency network from data, detect the cycles and then remove the edges that are mutually independent. This process is illustrated in **Figure 1**. The overall intuition behind this approach is fairly simple: use a scalable algorithm to handle a large number of variables and learn a dense model quickly. Since this learned model could potentially (and in practice) contain many cycles, we detect and remove edges based on mutual information. We then orient the edges ensuring acyclicity. Given that previous literature has demonstrated that an information-theoretic measure based on mutual information between two variables X and Y can be used as a reliable measure for quantifying the strength of an arc $X \rightarrow Y$ (Solo, 2008; Weichwald et al., 2014; Janzing et al., 2015), we use CSI and MI to establish the causal relationships.

We now describe each of these steps in detail before presenting the high-level algorithm.

3.1. Learning Context-Specific Independences

The first step of our learning algorithm is to learn distributions of the form $P(V_i|\mathbf{V} \setminus V_i)$, i.e., a conditional for a variable given all the other variables in the data. To this effect, we employ the intuition that a structured representation of a conditional probability table (CPT), such as a tree can be used inside probabilistic models to capture *context-specific independence* (CSI) (Boutilier et al., 1996). Specifically, we learn a single probability tree for each variable V_i given all the other variables in the data. The tree CPDs can capture *context specific independence* based on regularities in the CPTs of a node. Tree CPD for a variable is a rooted tree with each interior node representing tests on parent vertices and leaf nodes have the probability conditioned on particular configurations along the path from the root to leaf. The key idea here is that each tree can capture the CSI that exists between the variable's parents and the target variable conditioned on the values of some of the other parents. This is an important

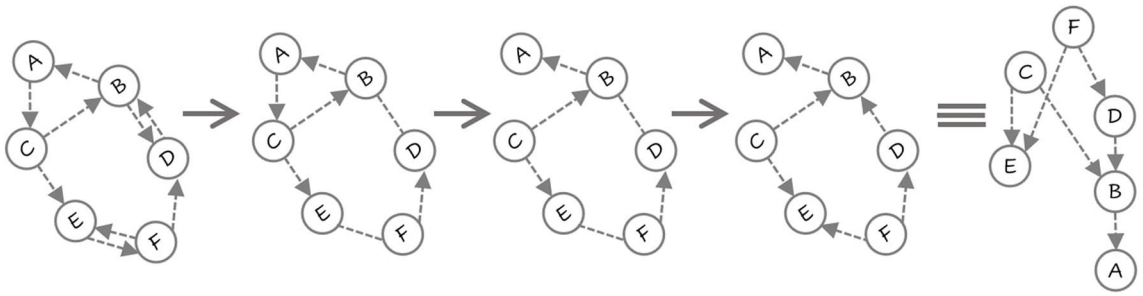


FIGURE 2 | First the DN is learned (notice the two bi-directed edges). All the bidirected edges in the DN are converted to undirected edges (BD and EF). The shorted cycle $A \rightarrow C \rightarrow B \rightarrow A$ is identified and the edge $A \rightarrow C$ is removed based on MI. Since no more cycles exist, the undirected edges are considered next. $E - -F$ becomes $F \rightarrow E$ and then $B - -D$ becomes $D \rightarrow B$. The resulting network is acylic and exploits both CSI and MI in becoming a causal network.

step as *it has been recently demonstrated that CSI can be used for identifying causal effects by* Tikka et al. (2019). While their work derives the calculus for identifying the causal relationships, we go further in employing the use of CSI in larger data sets. Further, our finally learned network can be considered as a special case of the structural causal model proposed by Tikka et al. where the structured representations (trees) are used to model the CSIs and the edges of the graphical model are aligned using information-theoretic measures.

To learn CSI at every variable, we employ the notion of DNs. Recall that a DN is a (potentially cyclic) graphical model that approximates the joint distribution as a product of conditionals. To learn such a DN, we iterate through every variable and learn a (probabilistic) decision tree for each variable given all the other variables, i.e., the goal is to learn $P(V_i|\mathbf{V} \setminus V_i)$ for each i where each conditional is modeling using a probabilistic tree. We observe that in this step, one could provide an important domain knowledge—*ordering between the variables*. This variable ordering can be used to construct expert guided causal model which introduces CSIs that satisfies the ordering constraints. As shown by Tikka et al. (2019), the conditional distributions induced using these CSIs can be effectively employed in identifying do calculus.

The advantage of this approach is that it learns the qualitative relationships (structure) and quantitative influences (parameters) simultaneously. The structure is simply the set of all the variables appearing in the tree and the parameters are the distributions at the leaves which can be reused in later stages. The other advantage is that the approach is that it is easily parallelizable and scalable. Thus, our method can be viewed as one that could scale up learning of causal models to real large data sets. The third advantage of the approach is that being a separate step, this can be integrated with other causal search methods, such as the one proposed by Tikka et al. Exploring these connections is an interesting future direction.

Let us denote the conditionals learned over all the variables (potentially given some order) as *DN*, the dependency network induced from the data. In most cases, this DN contains cycles since these conditionals are learned independent of each other. This can be an advantage and a disadvantage. The advantage is its efficiency as the costly step of checking for acyclicity can be avoided during learning and a disadvantage since it is an approximate model. Shorter cycles can result in larger approximations (Heckerman et al., 2000). After learning this *DN*, we perform an additional step. We convert edges of the form $X \leftarrow Y$ and $X \rightarrow Y$ to $X - -Y$. This is similar to the PC algorithm (Spirtes et al., 2000) in that strong correlation between two variables are considered as undirected and will be oriented in the final step of our algorithm. Next, we convert the DN to an intermediate CBN with potential undirected edges.

3.2. Detecting and Removing Cycles

To convert the DN to a CBN, the first step is to detect and remove cycles. A naïve approach to deleting edges would be: search for an edge, remove it, check for acyclicity and log-likelihood (Hulten et al., 2003). The key limitation of this approach is that the resulting model is not necessarily causal. The use of log-likelihood does improve the training performance but does not guarantee causality. Hence, inspired by the research in information-theoretic approaches to causality (Solo, 2008; Weichwald et al., 2014; Janzing et al., 2015), we employ mutual information for identifying the edges.

For detecting cycles, several methods exist (Kahn, 1962) including topological sorting. Any of these methods would be compatible with our learning algorithm. For the purposes of our data sets, we employ depth-first search (DFS). One key aspect of our DFS is that we identify short cycles. Recall that DN approximates a joint distribution as a product of conditionals.

$$P(V_1, ..., V_n) \approx \prod_i P(V_i|\mathbf{V} \setminus V_i)$$

The theoretical analysis of the approximation is based on the inference algorithm, specifically Gibbs sampling and on the size of the data. In simple terms, if the Gibbs sampler converges on a large data set, the approximation is quite effective (Heckerman et al., 2000; Neville and Jensen, 2007). In practice, we have previously observed that when the cycles are large, i.e., the size of the clique in the undirected graph, the approximation is quite robust (Natarajan et al., 2012; De Raedt et al., 2016).

With this insight, in the first step of cycle detection, we identify the short cycles. The intuition is that short cycles lead to larger

approximations and removing them would render the product of conditionals closer to the true joint distribution. Once the shortest cycle is identified, the next step is identifying the edge to remove from this short cycle. For this purpose, we employ mutual information (MI). As a pre-processing step, we compute the MI between every pair of variables and sort them by the MI. We consider MI instead of conditional MI as one of our key goals is efficiency. Computing conditional MI requires us to condition on a large set of related variables in the DN. This requires both repeated computations and a large number of conditionals. Thus, first, we detect the smallest directed cycle. We then break the cycle by removing edges that are smaller than a predefined threshold of δ. In our work, we simply choose δ to be the MI with the largest difference to the previous MI value in the sorted list. We use *Maximum adjacent difference* in the sorted list, as our δ in our setting, unless a default value is presented by an expert as domain knowledge. Large values of δ would result in a sparse graph and lower values δ will result in a dense graph. Once these edges are removed, the process continues where the next smallest cycle (if one exists) is detected and the low MI edges are removed and so on. **Coupling CSI with MI between variables X and Y quantifies the strength of $X \rightarrow Y$.**

To summarize, from the DN, we create an initial CBN by detecting cycles and removing edges with low dependencies. Now the last step is to orient the bi-directed edges which are undirected and then learn the parameters of the resulting causal BN.

3.3. Edge Orientation and Parameter Learning

Once the directed cycles are detected and removed, we focus on the undirected edges (in reality bi-directed edges). Inspired by the PC algorithm (Spirtes et al., 2000), we orient the edges in the final step using two criteria—MI and acyclicity. We orient the edges by removing the edge with the lowest MI if it does not result in a cycle. As mentioned earlier, this is similar to that of PC. After all the undirected edges have been oriented, the resulting CBN is our casual network skeleton.

We estimate the parameters of this CBN using standard MLE (Pearl, 1988a). All our data sets are fully observed and hence MLE suffices for learning the conditional distributions. For the parameters, we learn a decision tree locally and in parallel using only the variables in the parent set of every node to capture the conditional distribution. Extending this to handle missing data is a significant extension as it does not merely affect the parameter learning but the structure search as well. Once the parameters are learned, we now have the full causal BN learned from data.

3.4. DN2CN Algorithm

Before we provide the algorithm, we present an example in **Figure 2**. There are six variables $\langle A, ..., F \rangle$. First, a DN is learned where there are cycles and bi-directed edges. Next, the smallest cycle $\langle A, B, C \rangle$ is detected and the edge with least MI $A \rightarrow C$ is removed. Now, there are no directed cycles in the CBN (in the general case, there could be more cycles that need to be removed). Note that there are two undirected edges between B and D, and between E and F. First, the edge between D and B is oriented based on MI and the fact that this does not create a cycle. Finally, the edge between E and F is oriented to obtain the CBN. The parameters are then learned by learning a decision-tree for each conditional.

This approach is formally presented in Algorithm 1 and as a flow chart in **Figure 1**. As can be seen in the algorithm, the first step is to learn the DN (line 4). The LEARNPARENTSET function in line 3 of Algorithm 2 learns a tree and collects the set of parents from that set. It can optionally take an ordering among the variables provided by a domain expert (if any). Then the algorithm computes the mutual information (MI) for all the edges. One could instead simply wait till the cycles are detected and then compute the MI but we compute it outside the cycle detection step. The algorithm then iteratively removes the least informative edges till no more cycles are present in the graph. We orient the undirected edges (If any) ensuring acyclicity. Then the parameters are then learned from the data.

3.4.1. Theoretical Analysis

A natural question to ask is—*what is the complexity of our approach?* We present an initial analysis of this work, by adapting the arguments from the literature [see for instance the original reducibility result (Karp, 1972)]. We present our result by analyzing each component of the algorithm. Tightening these bounds with appropriate heuristics is left for future work.

Let v be the number of vertices (features), n be the number of training examples. In Algorithm 1, while learning *DN*, we learn a decision tree locally [line 4]. This requires $O(n^2 d)$ where d is the depth of the tree (Su and Zhang, 2006). While this can be reduced to $O(n \cdot d)$, this requires making independence assumptions among the variables. Our tree growing procedure is fairly standard without much optimization. Hence the complexity of learning a full DN is $O(v \cdot n^2 d)$. However, the trees can be learned in parallel, thus reducing the complexity to $O(n^2 d)$.

Cycle detection (line-12) has a complexity of $O(v(v + e))$, where v is no. of nodes and e is number of edges in the network (e is asymptotically $O(v^2)$. A single cycle detection running a DFS to search for the cycle thus is $O(v^2)$. Doing this for all the variables will result in $O(v^3)$ for the entire cycle detection. Sorting the edges to compute the MI requires $O(v^2 log(v))$. Edge orientation is $O(v^2)$.

Thus the complexity DN2CN is dominated by two terms—$O(v^3)$ the cube of the number of edges and $O(n^2 d)$, the term that depends on the data. Since, typically, $n > v^2$ to learn a meaningful model, our final complexity is $O(n^2 d)$. Optimizing the tree learner to lower this complexity and better cycle detection methods to reduce the cubic complexity can significantly

improve the asymptotic bound. These are open research directions.

3.4.2. Discussion

The proposed approach has some salient advantages—(1) One could parallelize the learning of the DN to scale it up to very large data sets. (2) The computation of the MI can also be parallelized. (3) Any traversal algorithm could be used to detect cycles in the graph for pruning. (4) There are two levels of independence used in this algorithm;—(a) context specific independence (CSI) to identify potentially independent influences. Inspired by the work of Tikka et al. (2019), we rely on the ability of CSI to model interventions; in the context of interventions, any influences that otherwise have a causal effect thereon variable, are removed. Learning a BN as a series of trees for every interacting variable facilitates the ability to model such CSI and so are able to represent interventions in sufficient detail to reason about conditional independence properties, (b) Mutual independence which when combined with expert domain knowledge can potentially yield even causal influences. (5) The algorithm also has two types of controls (similar to regularizations) to combat overfitting. First is to control the depth of trees and second is selecting the number of edges to remove. (6) Finally, the use of both local search and constraint based methods inside the algorithm enables it to learn effectively at scale.

Before presenting our empirical results, we briefly discuss the interpretability of the resulting network. DN2CN represents causal dependencies using BNs that provide an intuitive visualization by modeling features as nodes and the statistical association between the features as edges. This statistical interpretability is similar in spirit to traditional interpretability. This allows to answer questions, such as "does BMI influence susceptibility to Covid?" Moreover, it has been argued that developing an effective CBN for practical applications requires expert knowledge when data collection is cumbersome (Fenton and Neil, 2012). This applies to domains, such as medicine, similar to our experimental evaluation. A typical characteristic of these domains is that they can be data-poor and knowledge-rich due to several decades of research. Kahneman et al. showed that human beings tend to interpret events in terms of cause-effect relations (Kahneman et al., 1982; Pennington and Hastie, 1988). Also, causal models are easier to construct, easier to modify and easier to interpret by humans (Henrion, 1987; Pennington and Hastie, 1988). Following these observations, our framework can incorporate both data-driven and human inputs, thus allowing to learn a more robust hypothesis. Lipton explains that with interpretable models it becomes imperative to guarantee fairness (Lipton, 2018). It must be noted that we can extend DN2CN's interactive framework and leverage the Bayesian networks learnt to assess the bias as well as compare multiple models in terms of their fairness and performance (Chiappa and Isaac, 2018). In summary, our framework can leverage interpretability as a tool to verify causal assumptions and relationships. We verify the above claims empirically in a real data set and two synthetic benchmark causal data sets in the next section.

Algorithm 1 |DN2CN: dependency network to causal network.

Given: Data $\mathbf{D}$; Variables $\mathbf{V}$; Ordering among variables (if any) $\mathbf{O} := \varnothing$; Threshold $\delta := 0$
function DN2CN($\mathbf{D}$,$\mathbf{V}$, $\mathbf{O}$)
  $\mathbf{E} \leftarrow \varnothing$ ▷ **Initialize edge set**
  $\text{DN} \equiv (\mathbf{V}, \mathbf{E}) =$ LEARNDN($\mathbf{D}$, $\mathbf{V}$, $\mathbf{O}$)
  for all edge $\in \mathbf{E}$ **do**
    MI[edge] $\leftarrow$ COMPUTEMUTUALINFO(edge)
  end for
  SortedMI[edge] $\leftarrow$ SORTED(edge, *reverse* = *True*) ▷ **Sort in descending order**
  if $\delta = 0$ **then**
    δ = ARGMAX_ABSDIFF(SortedMI[edge]) ▷ **Max absolute diff of 2 contiguous elements in array SortedMI**
  end if
  $\mathbf{C} \leftarrow$ DETECTCYCLES(DN) ▷ **Using any sort**
  for all cycle $\in \mathbf{C}$ **do**
    for all $e \in$ **cycle do**
      if $SortedMI[e] \leq \delta$ **then**
        $\mathbf{E} \leftarrow \mathbf{E} \setminus e$ ▷ **Remove edges if exist in DN**
      end if
    end for
    $\mathbf{C} \leftarrow \mathbf{C} \setminus$ cycle
      ▷ **Update cycles list after each iteration**
    if $\mathbf{C} = \varnothing$ **then** ▷ **No more cycles left**
      break
    end if
  end for
  $\hat{\mathbf{V}}, \hat{\mathbf{E}} :=$ ORIENTEDGES($\mathbf{V}, \mathbf{E}$) ▷ **Introduce directions ensuring acyclicity as required**
  return $(\hat{\mathbf{V}}, \hat{\mathbf{E}})$
end function

4. EMPIRICAL EVALUATION—DOMAINS

To assess the effectiveness of our method, we perform extensive evaluations on both synthetic as well as real benchmark causal data sets. In all our data sets, we have the underlying true causal graph, and we apply our method as well baseline approaches to reconstruct the causal network from the data to demonstrate the effectiveness. We first describe the data sets used before discussing the baselines used.

4.1. Benchmark1: LUCAS—(LUng CAncer Simple Data Set)

The LUCAS (LUng CAncer Simple set) data set from causality challenge (Guyon et al., 2008) represents a synthetic medical diagnosis problem, where the task is to identify patients with lung cancer given a set of socioeconomic and clinical factors of putative causal relevance. The generative model is a Markov process, so the value of the children node is stochastically

Algorithm 2 |LEARNDN: learn dependency network.

function LEARNDN(**D**, **V**, **O**)
  E ← ∅ ▷ **Initialize edge set**
  for all var ∈ **V do**
    P(var) ← LEARNPARENTSET(var, {**V** \ var}$_O$, **D**) ▷ **Parent set {V \ var} is constrained by O (if any)**
    for all parent ∈ **P**(var) **do**
      E ← **E** ∪ {parent → var}
      ▷ **Add new directed edge between parent and var**
    end for
  end for
  return (**V**, **E**)
end function

dependent on the values of the parent nodes'. The data set consists of 2000 observations. Ground-truth consists of 12 binary variables that include *anxiety, peer pressure, day of birth, smoking, genetics, yellow finger, lung cancer, attention disorder, cough, fatigue, allergy, car accidents*, and their causal relations. There are no missing values in the data set. As the data are generated artificially by causal BN with variables, the true nature of the underlying causal relationships is known. Hence we use this benchmark data set for illustrating the effectiveness of our approach.

4.2. Benchmark2: Asia Data Set

The ASIA Network is an expert-designed causal network with logical links. This BN was originally presented by Lauritzen and Spiegelhalter (Lauritzen and Spiegelhalter, 1988), who have specified reasonable transition properties for each variable given its parents. It is an eight node BN that describes the effect of visiting Asia and smoking behavior of an individual on the probability of contracting tuberculosis, cancer or bronchitis. The underlying structure expresses the known qualitative medical knowledge. Each node in the network represents a feature that relates to the patient's condition. The example is motivated as follows: *"Shortness-of-breath (called dyspnea) may be due to tuberculosis, lung cancer or bronchitis, or none of them, or more than one of them. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnea."* The data set contains 10,000 observations and eight binary variables whose values are 0 or 1. There are no missing values in the data set.

4.3. Causal Protein-Signaling Networks in Human T Cells Data Set

This data analyzed and published by Sachs et al. (2005) is a multivariate proteomics data set, widely used for research on causal discovery methods. This is a biological dataset with different proteins and phospholipids in human immune system cells. The data comprises of the simultaneous measurements of 11 phosphorylated proteins and phospholipids (PKC, PKA, P38, Jnk, Raf, Mek, Erk, Akt, Plcg, PIP2, PIP3) derived from thousands of individual primary immune system cells. In the data set we considered, there are (1) 1,800 observational data points subject only to general stimulatory cues, so that the protein signaling paths are active; (2) 600 interventional data points with specific stimulatory and inhibitory cues for each of the following four proteins: pmek, PIP2, Akt, PKA; and (3) 1,200 interventional data points with specific cues for PKA. Overall, the data set consists of 5,400 instances with no missing value. The 11 variables are discretized into three bins (low, medium, and high) for each feature, respectively. A network consisting of 18 well-established causal interactions between these molecules has been constructed supported with biological experiments and literature (Sachs et al., 2005). This data is a good fit to test our proposed causal discovery method, as the knowledge about the "ground truth" is available, which helps verification of results. Hence the goal of the data set is to unearth protein signaling networks, originally modeled as CBN.

5. EXPERIMENTAL RESULTS

In our experiments, we aim to answer the following questions explicitly:

Q1: Does the learned model identify influencing variables as in the "Ground truth" network?
Q2: How does the resulting network produced by DN2CN compare to standard constraint based approaches qualitatively?
Q3: How does the resulting network produced by DN2CN compare to standard constraint based approaches quantitatively?

Specifically, we consider two different types of experiments—the first on evaluating **goodness** of the model on the synthetic benchmark data sets and the second on **verifying** if the approach can learn a good causal model on the real data set.

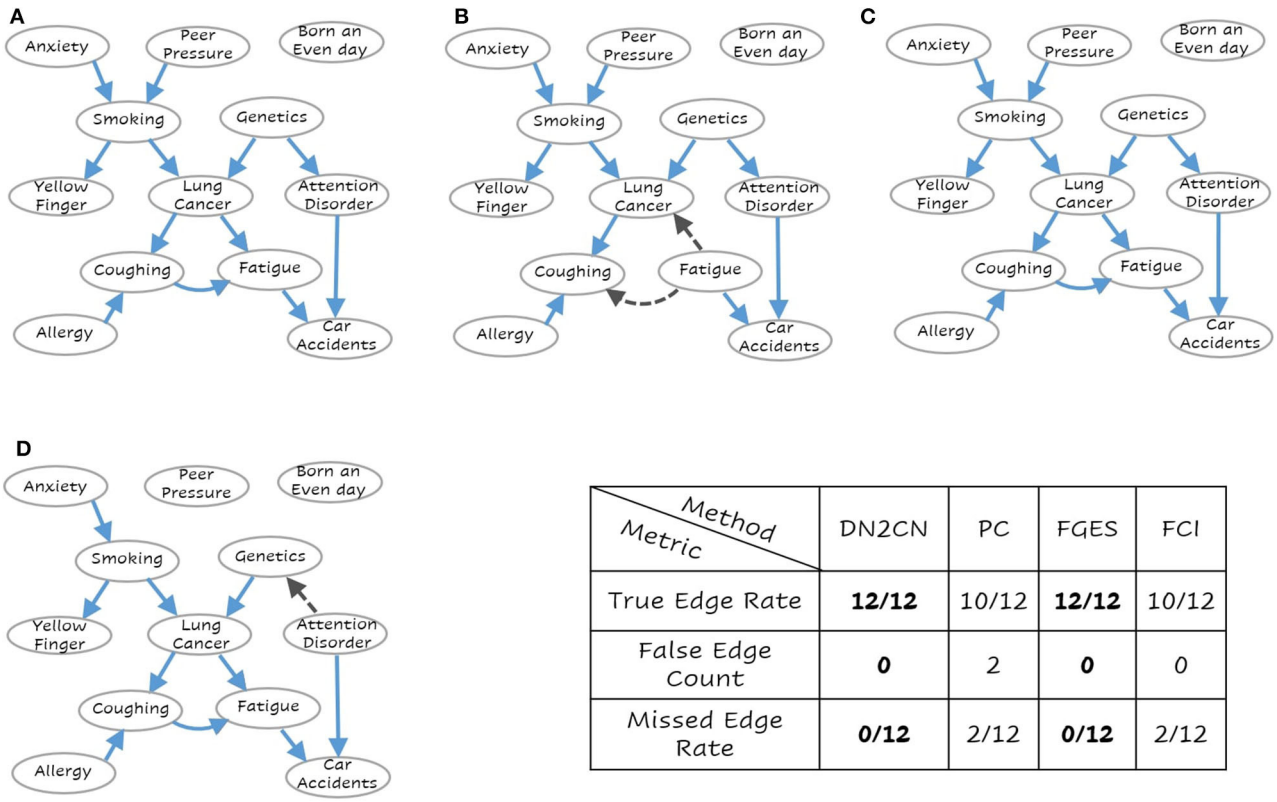
5.1. Setup

In DN2CN, we used a tree depth of 2 for all the experiments. We set δ as 0.015 for both LUCAS and Asia data sets and 0.25 for the real T cells data set.

We compare DN2CN to three of the well-known computational methods for causal discovery (Glymour et al., 2019). Two of these algorithms are commonly employed constraint-based algorithms—PC and Fast Causal Inference (FCI) (Spirtes et al., 2000). The third algorithm is a score-based algorithm—Fast Greedy Equivalence Search (FGES) (Ramsey et al., 2017). It must be mentioned that PC, FCI and FGES, are widely applicable as they handle various types of data distributions as well as causal relations, given reliable conditional independence testing methods. We strongly believe that these attributes make them a strong as well as a fair baseline for DN2CN as suggested by Glymour et al. (2019).

We further discuss each of the baseline approaches and their corresponding experimental settings used, as follows:

- *PC algorithm* (denoted **PC**) (Spirtes et al., 2000) starts with a fully connected undirected graph, tests all possible



Metric \ Method	DN2CN	PC	FGES	FCI
True Edge Rate	**12/12**	10/12	**12/12**	10/12
False Edge Count	**0**	2	**0**	0
Missed Edge Rate	**0/12**	2/12	**0/12**	2/12

FIGURE 3 | The learned network for **(A)** Our Approach DN2CN, **(B)** PC algorithm, **(C)** Fast Greedy Equivalence Search algorithm (FGES), and **(D)** Fast Causal Inference algorithm (FCI) and the summary results on LUCAS data set (best viewed in color). Each node represents a feature and the arcs represent causal relationships, i.e., X → Y represents that X is a cause of Y. As can be seen, our DN2CN and FGES had a 100% true positive rate with a 0 false positive and false negative rates. PC and FCI missed two edges each. PC and FCI also introduced spurious edges (incorrect edge orientation).

conditioning set for every order of conditioning and then finally orients the edges. Test statistic we used is the mutual information for PC algorithm, to keep the comparison fair. We used type I error rate; $\alpha = 0.05$ in our setting.

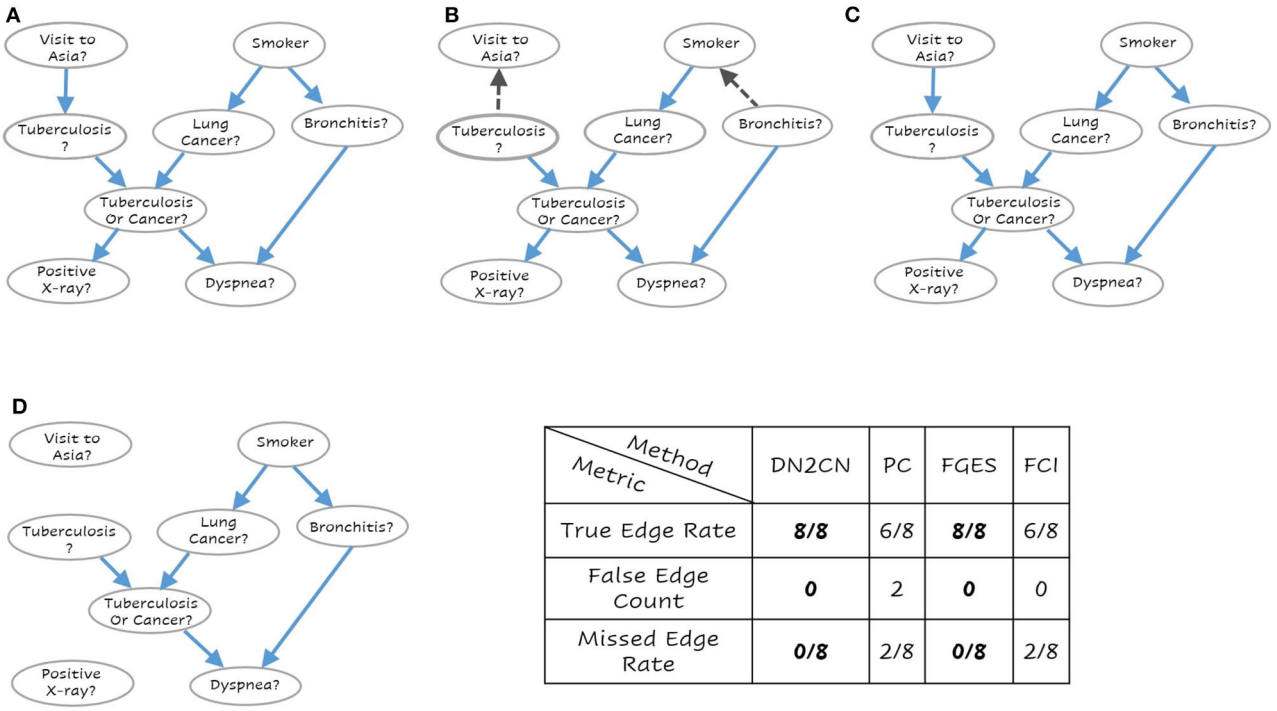
- *Fast Greedy Equivalence Search algorithm* (denoted **FGES**) (Ramsey et al., 2017) is an optimized and parallelized version of an algorithm developed by Meek (Meek, 1995) called the Greedy Equivalence Search (GES). GES is a CBN learning algorithm that starts with an empty graph, heuristically performs a forward stepping search over the space of CBNs and stops with the one with the highest score. GES finally performs a backward stepping search that iteratively removes edges until no single edge removal can increase the Bayesian score. We use the modified BIC (Bayesian information criterion) (Schwarz, 1978) score rewritten as $Score_{BIC}(B:D) = 2L(D;\hat{\theta},B) - k\log|D|$, where L is the likelihood, k the number of parameters, and $|D|$ the sample size. So higher BIC scores will correspond to greater dependence.
- *Fast Causal Inference algorithm* (denoted **FCI**) (Spirtes et al., 2000) is a constraint-based algorithm which learns an equivalence class of CBNs that entail the set of conditional independencies that are true in the data. FCI then orients the edges using the stored conditioning sets that led to the removal of adjacencies earlier. We use the same modified BIC score as with the other baseline, i.e., FGES algorithm.

For PC algorithm we used the open-source implementation, i.e., *stable-PC* in bnlearn (Scutari, 2009) while TETRAD (Spirtes et al., 2000) was used to run FGES and FCI algorithms; a reliable tool for causal explorations. Data set details are presented in section 3 which describes the number of variables and the number of training examples.

5.2. Results

Recall that our goal is faithful modeling of underlying data. In addition, we also demonstrate the training log-likelihood of the learned model for (1) ground truth model, (2) model learnt using DN2CN algorithm, (3) model learnt using PC algorithm, (4) model learnt using FGES algorithm, and (5) model learnt using the FCI algorithm. This is to say that our analysis is *qualitative* as well as *quantitative*.

To answer **Q1 and Q2**, consider the networks presented in **Figures 3A–D–5A–D**, respectively. These are the learned networks obtained by our approach DN2CN and baseline methods PC, FGES & FCI summarized together with the ground truth network. To evaluate the validity of the proposed approach,



Metric \ Method	DN2CN	PC	FGES	FCI
True Edge Rate	**8/8**	6/8	**8/8**	6/8
False Edge Count	**0**	2	**0**	0
Missed Edge Rate	**0/8**	2/8	**0/8**	2/8

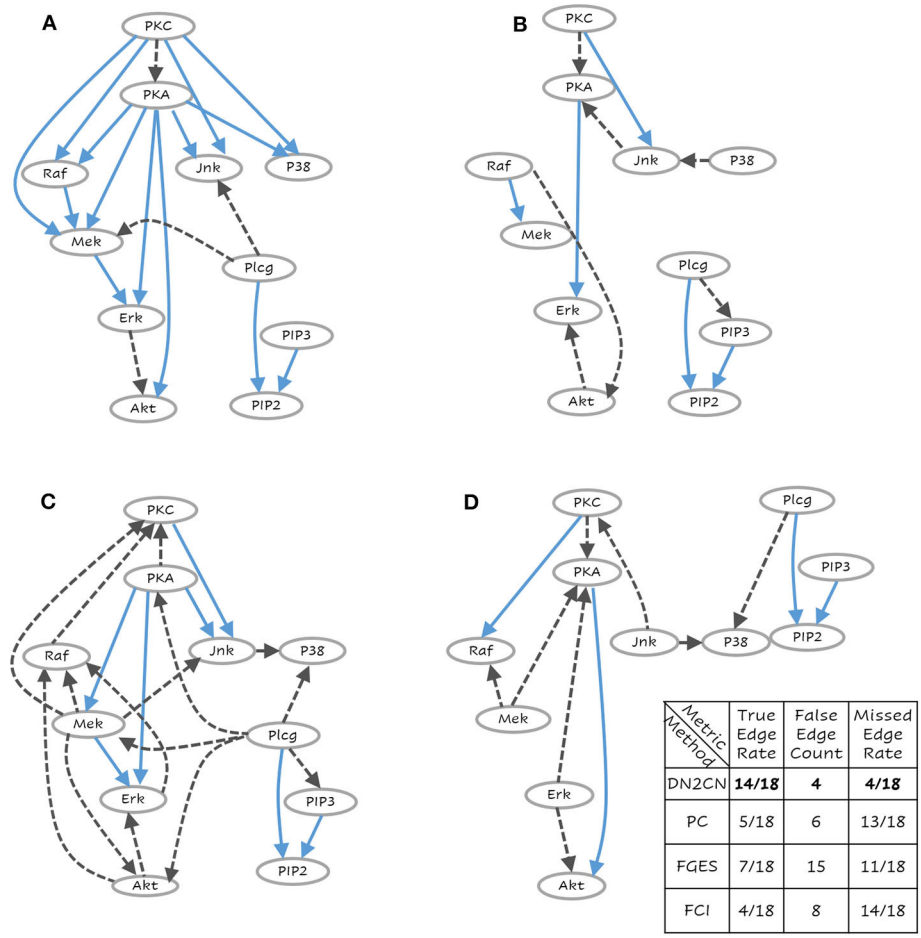
FIGURE 4 | The learned network for **(A)** Our Approach DN2CN, **(B)** PC algorithm **(C)** Fast Greedy Equivalence Search algorithm (FGES), and **(D)** Fast Causal Inference algorithm (FCI) and the summary results on ASIA data set (best viewed in color). Each node represents a feature and the arcs represent causal relationships, i.e., $X \rightarrow Y$ represents that X is a cause of Y. As can be seen, our DN2CN and FGES had a 100% true positive rate with a 0 false positive and false negative rates. PC and FCI both missed two edges. Also, PC introduced two spurious causal edges in the resultant network.

we compared the model arcs with those present in the ground truth. An arc is correct, if and only if the same arc exists in the ground truth graph and the orientation of the arc aligns with the orientation in the ground truth graph; an arc is considered incorrect, if the arc does not exist in the ground truth graph or if it exists but its orientation is the opposite of the true orientation. Hence, in all the data sets, to understand the effectiveness of DN2CN, motivated by Sachs et al. (2005), Gao and Ji (2015), and Yu et al. (2019) we summarize the arcs learned by our method as well as PC, FGES and FCI for each data set using the following metrics:

- *True Edge Rate*, is the fraction of the true connections in the ground truth network that our approach (or PC or FGES or FCI) captures correctly, i.e., true positive.
- *False Edge Count*, for connections that are not in the ground truth network, but which were captured by our approach (or PC or FGES or FCI), i.e., false positive.
- *Missed Edge Rate*, is the fraction of the true edges missed in the ground network by our approach (or PC or FGES or FCI), i.e., a false negative.

As can be observed our algorithm DN2CN and baseline algorithm FGES had a 100% true positive rate with a 0 false positive and false negative rates in both LUCAS and ASIA data sets. However, the other baselines methods PC and FCI both missed two edges in LUCAS as well as ASIA data sets. In addition, the PC algorithm introduced spurious causal flows in both LUCAS and ASIA data sets. This clearly establishes that our framework is indeed capable of retrieving the full causal model while learning only from the data.

In the real benchmark data set, i.e., *Causal Protein-Signaling Network in human T cells*, the ground truth network and the reconstruction by employing DN2CN, PC, FGES and FCI are illustrated in **Figures 5A–D**, respectively. It can be observed that our approach DN2CN performs **significantly better** than all the baselines, i.e., PC, FGES and FCI. DN2CN missed four edges and introduced four spurious edges. Whereas, the baseline algorithms PC, FGES, and FCI, had significantly worse performance with 13, 11, 14 missed edges and 6, 15, 8 spurious ones, respectively. On closer inspection at the unexpected edges in our acyclic causal model reconstruction, one can see that they actually explain the data quite well. Especially, both arcs, PKC $\Longrightarrow$ PKA and Erk $\Longrightarrow$ Akt, can be understood qualitatively in rat ventricular myocytes (Wilhelm et al., 1997) and colon cancer cell lines (Lemaire et al., 1997), respectively. However, We hypothesize that, our DN2CN method missed four causal relationships, that are all involved in cycles. As BNs are acyclic by definition, our inference missed these arcs, which is one of the caveats of this approach. Extending this to dynamic causal bayesian network to handle feedback loops, remains an interesting future research direction.



Metric / Method	True Edge Rate	False Edge Count	Missed Edge Rate
DN2CN	**14/18**	**4**	**4/18**
PC	5/18	6	13/18
FGES	7/18	15	11/18
FCI	4/18	8	14/18

FIGURE 5 | The learned network for **(A)** Our Approach DN2CN, **(B)** PC algorithm, **(C)** Fast Greedy Equivalence Search algorithm (FGES), and **(D)** Fast Causal Inference algorithm (FCI) and the summary results on T-Cell data set (best viewed in color). Each node represents a feature and the arcs represent causal relationships, i.e., X → Y represents that X is a cause of Y. This is a challenging data set where DN2CN had missed one edge and introduced two spurious edges. PC, on the other hand, had significantly worse performance with 10 missed edges and four spurious ones.

Table 1 presents quantitative comparisons between the different methods. In all our experiments, we present the numbers in bold whenever they are better than all the other baselines on a data set. It must be mentioned that in some cases, PC, FGES, and FCI did not yield a directed arc, and we chose a direction (ensuring acyclicity) to compute the overall joint log-likelihood on the training set. As can be seen from the table, the proposed DN2CN approach produces a network with significantly better joint log-likelihood on the training set than the baseline algorithms PC and FCI learning method in all the domains. We can see that FGES has better joint log-likelihood than DN2CN in T-Cell data set. One key reason is that the resultant network using FGES is relatively denser than other models. FGES introduces 14 spurious causal edges leading to increased likelihood. It is well-known in the Bayes net learning literature that denser the graph is, higher the training set likelihood. As can be seen from the table in the **Figure 5**, the false edge count of FGES is significantly higher than the other methods. Hence, the denser network can yield a much higher training set loglikelihood. This answers **Q3** affirmatively: that DN2CN is more effective in modeling than the causal method, such as PC, FGES, and FCI.

TABLE 1 | Table comparing the log-likelihood estimate in CBN learned using DN2CN and baseline approach, i.e., PC algorithm, Fast Greedy Equivalence Search algorithm (FGES) and Fast Causal Inference algorithm (FCI) learned directly from data.

		Methods			
Data sets	**Ground truth**	**DN2CN**	**PC**	**FGES**	**FCI**
Lucas	**−12130.83**	**−12130.83**	−12178.59	**−12130.83**	−12161.49
Asia	**−22212.85**	**−22212.85**	**−22212.85**	**−22212.85**	−23747.1
Sachs	−38723.1	−38081.29	−41930.74	**−35782.43**	−40822.13

Numbers are presented in bold text whenever they are better than all the other baselines on a data set.

6. CONCLUSIONS

We introduced a scalable causal learning algorithm that is capable of exploiting two types of independencies—context-specific

independence (CSI) and conditional independence (CI). To exploit CSI, we learn a single tree for each variable in the model. Each tree can locally model and capture the CSI. Next, we orient and remove edges from this potentially cyclic model by computing the mutual information which allows for capturing the CIs. The intuition is that these two independence metrics have previously been explored in the context of causal learning and combining them will allow for learning a robust causal model. Our empirical evaluations in the standard data sets clearly demonstrate that the proposed DN2CN method does retrieve the true causal model in most of the domains. Most importantly, it does not introduce a denser model than what is necessary even if it means sacrificing the training likelihood. Thus, a natural regularization is achieved by controlling the depth of the trees and the orienting of edges as against other information-theoretic methods, such as BIC that employs a model complexity penalty.

There are several possible extensions of future work—adapting and applying these models to real problems in the lines of our previous work (Ramanan and Natarajan, 2019) is an important direction. Developing the theoretical underpinnings between CSI and CI with causal models is the next immediate direction. Converting the CSI from our models to do calculus and employing them in the context of learning from both observational and experimental data is another important problem. Finally, allowing for rich domain knowledge and inductive bias to guide the learner to a better causal model is possibly the most interesting direction.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in following repository, respectively: LUCAS—LUng CAncer Simple data set: http://www.causality.inf.ethz.ch/data/LUCAS.html; Asia data set: http://www.bnlearn.com/bnrepository/; Causal Protein-Signaling Networks in human T cells data set: http://www.bnlearn.com/bnrepository/.

AUTHOR CONTRIBUTIONS

NR and SN contributed equally to the ideation and contributed nearly equally to the manuscript preparation. NR led the empirical evaluation. All authors contributed to the article and approved the submitted version.

FUNDING

The authors gratefully acknowledge the support of AFOSR award FA9550-18-1-0462. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the AFOSR, or the US government.

ACKNOWLEDGMENTS

The authors acknowledge the support of members of STARLING lab for the discussions. We thank the reviewers for their insightful comments and in significantly improving the paper.

REFERENCES

Aliferis, C. F., Tsamardinos, I., and Statnikov, A. (2003). "Hiton: a novel markov blanket algorithm for optimal variable selection," in *AMIA Annual Symposium Proceedings*, Vol. 2003 (Washington, DC: American Medical Informatics Association), 21.

Andrews, B., Ramsey, J., and Cooper, G. F. (2018). Scoring bayesian networks of mixed variables. *Int. J. Data Sci. Analyt.* 6, 3–18. doi: 10.1007/s41060-017-0085-7

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). "Context-specific independence in bayesian networks," in *UAI* (Portland: Morgan Kaufmann Publishers Inc.), 115–123.

Chiappa, S., and Isaac, W. S. (2018). "A causal bayesian networks viewpoint on fairness," in *IFIP International Summer School on Privacy and Identity Management* (Vienna: Springer), 3–20. doi: 10.1007/978-3-030-16744-8_1

Chickering, D. (1996). "Learning bayesian networks is NP-complete," in *Learning From Data* (Springer), 121–130. doi: 10.1007/978-1-4612-2404-4_12

Chickering, D. M. (2002a). Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res.* 2, 445–498. Available online at: https://www.jmlr.org/papers/volume2/chickering02a/chickering02a.pdf

Chickering, D. M. (2002b). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* 3, 507–554. Available online at: https://www.jmlr.org/papers/volume3/chickering02b/chickering02b.pdf

Colombo, D., and Maathuis, M. H. (2012). A modification of the PC algorithm yielding order-independent skeletons. *arXiv* 1211.3295.

Colombo, D., and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* 15, 3741–3782.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.* 40, 294–321. doi: 10.1214/11-AOS940

Cooper, G. F., and Yoo, C. (2013). Causal discovery from a mixture of experimental and observational data. *arXiv* 1301.6686.

Coumans, V., Claassen, T., and Terwijn, S. (2017). *Causal Discovery Algorithms and Real World Systems*. Masters thesis.

De Raedt, L., Kersting, K., Natarajan, S., and Poole, D. (2016). *Statistical Relational Artificial Intelligence: Logic, Probability, and Computation*, Vol. 10, Morgan & Claypool. p. 1–189.

Fenton, N., and Neil, M. (2012). *Risk Assessment and Decision Analysis With Bayesian Networks*. (Boca Raton, FL: CRC Press), p.524.

Friedman, N., Nachman, I., and Peér, D. (1999). "Learning bayesian network structure from massive datasets: the sparse candidate algorithm," in *UAI* (Stockholm: Morgan Kaufmann Publishers Inc.), 206–215.

Gao, T., and Ji, Q. (2015). "Local causal discovery of direct causes and effects," in *Advances in Neural Information Processing Systems*, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Montreal, QC: NeurIPS), 2512–2520.

Gillispie, S. B., and Perlman, M. D. (2013). Enumerating markov equivalence classes of acyclic digraph models. *arXiv* 1301.2272.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Front. Genet.* 10:524. doi: 10.3389/fgene.2019.00524

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. Available online at: http://www.deeplearningbook.org

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–438. doi: 10.2307/1912791

Guo, Y., Ruan, Q., Zhu, S., Wei, Q., Chen, H., Lu, J., et al. (2019). Temperature rise associated with adiabatic shear band: causality clarified. *Phys. Rev. Lett.* 122:015503. doi: 10.1103/PhysRevLett.122.015503

Guyon, I., Aliferis, C., Cooper, G., Elisseeff, A., Pellet, J.-P., Spirtes, P., et al. (2008). "Design and analysis of the causation and prediction challenge," in *Causation*

and Prediction Challenge, eds I. Guyon, C. F. Aliferis, G. F. Cooper, A. Elisseeff, J. Pellet, P. Spirtes, and A. R. Statnikov (Hong Kong: JMLR.org), 1–33.

Hauser, A., and Bühlmann, P. (2015). Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *J. R. Stat. Soc. B Stat. Methodol.* 77, 291–318. doi: 10.1111/rssb.12071

Heckerman, D., Chickering, D., Meek, C., Rounthwaite, R., and Kadie, C. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *JMLR* 1, 49–75. Available online at: https://www.jmlr.org/papers/volume1/heckerman00a/heckerman00a.pdf

Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning bayesian networks: the combination of knowledge and statistical data. *MLJ* 20, 197–243. doi: 10.1007/BF00994016

Henrion, M. (1987). "Practical issues in constructing a bayes' belief network," in *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence* (Seattle, WA), 132–139.

Hulten, G., Chickering, D., and Heckerman, D. (2003). "Learning bayesian networks from dependency networks: a preliminary study," in *AISTATS* (Key West, FL).

Janzing, D., Steudel, B., Shajarisales, N., and Schölkopf, B. (2015). "Justifying information-geometric causal inference," in *Measures of Complexity* (Springer), 253–265. doi: 10.1007/978-3-319-21852-6_18

Kahn, A. B. (1962). Topological sorting of large networks. *Commun. ACM* 5, 558–562. doi: 10.1145/368996.369025

Kahneman, D., Slovic, S. P., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Karp, R. M. (1972). "Reducibility among combinatorial problems," in *Complexity of Computer Computations* (Springer), 85–103. doi: 10.1007/978-1-4684-2001-2_9

Lauritzen, S. L., and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. B Methodol.* 50, 157–194. doi: 10.1111/j.2517-6161.1988.tb01721.x

Lemaire, P., Wilhelm, K., Curdt, W., Schüle, U., Marsch, E., Poland, A., et al. (1997). "First results of the sumer telescope and spectrometer on SOHO," in *The First Results From SOHO* (Springer), 105–121. doi: 10.1007/978-94-011-5236-5_6

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue* 16, 31–57. doi: 10.1145/3236386.3241340

Margaritis, D., and Thrun, S. (2000). "Bayesian network induction via local neighborhoods," in *NIPS* (Denver, CO), 505–511.

Meek, C. (1995). "Causal inference and causal explanation with background knowledge," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (Montreal, QC), 403–410.

Meinshausen, N., Hauser, A., Mooij, J. M., Peters, J., Versteeg, P., and Bühlmann, P. (2016). Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci. U.S.A.* 113, 7361–7368. doi: 10.1073/pnas.1510493113

Natarajan, S., Khot, T., Kersting, K., Gutmann, B., and Shavlik, J. (2012). Gradient-based boosting for statistical relational learning: the relational dependency network case. *Mach. Learn.* 86, 25–56. doi: 10.1007/s10994-011-5244-9

Neapolitan, R. E., et al. (2004). *Learning Bayesian Networks*, Vol. 38. Upper Saddle River, NJ: Pearson Prentice Hall.

Neville, J., and Jensen, D. (2007). Relational dependency networks. *J. Mach. Learn. Res.* 8, 653–692. Available online at: https://www.jmlr.org/papers/volume8/neville07a/neville07a.pdf

Ogarrio, J. M., Spirtes, P., and Ramsey, J. (2016). "A hybrid causal search algorithm for latent variable models," in *Conference on Probabilistic Graphical Models* (Lugano), 368–379.

Pearl, J. (1988a). *Morgan Kaufmann Series in Representation and Reasoning. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kaufmann.

Pearl, J. (1988b). *Probabilistic Reasoning in Intelligent Systems; Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pennington, N., and Hastie, R. (1988). Explanation-based decision making: effects of memory structure on judgment. *J. Exp. Psychol. Learn. Mem. Cogn.* 14:521. doi: 10.1037/0278-7393.14.3.521

Ramanan, N., and Natarajan, S. (2019). *Work-in-Progress : Ensemble Causal Learning for Modeling Post-partum Depression*. Palo Alto, CA.

Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. (2017). A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *Int. J. Data Sci. Analyt.* 3, 121–129. doi: 10.1007/s41060-016-0032-z

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 523–529. doi: 10.1126/science.1105809

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Scutari, M. (2009). Learning bayesian networks with the bnlearn R package. *arXiv* 0908.3817.

Silander, T., and Myllymaki, P. (2012). A simple approach for finding the globally optimal bayesian network structure. *arXiv* 1206.6875.

Sims, C. A. (1972). Money, income, and causality. *Am. Econ. Rev.* 62, 540–552.

Solo, V. (2008). "On causality and mutual information," in *2008 47th IEEE Conference on Decision and Control* (Cancun: IEEE), 4939–4944. doi: 10.1109/CDC.2008.4738640

Spirtes, P., and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.* 9, 62–72. doi: 10.1177/089443939100900106

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search: Lecture Notes in Statistics*. New York, NY: Springer.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Su, J., and Zhang, H. (2006). "A fast decision tree learning algorithm," in *Proceedings of the 21st National Conference on Artificial Intelligence–Volume 1, AAAI'06* (Boston, MA: AAAI Press), 500–505.

Tikka, S., Hyttinen, A., and Karvanen, J. (2019). "Identifying causal effects via context-specific independence relations," in *Advances in Neural Information Processing Systems*, eds H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett (Vancouver, BC: NeurIPS), 2800–2810.

Tsagris, M., Borboudakis, G., Lagani, V., and Tsamardinos, I. (2018). Constraint-based causal discovery with mixed data. *Int. J. Data Sci. Analyt.* 6, 19–30. doi: 10.1007/s41060-018-0097-y

Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., and Statnikov, E. (2003). "Algorithms for large scale markov blanket discovery," in *FLAIRS Conference* (St. Augustine, FL), Vol. 2, 376–380.

Tsamardinos, I., Brown, L., and Aliferis, C. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *MLJ* 65, 31–78. doi: 10.1007/s10994-006-6889-7

Weichwald, S., Schölkopf, B., Ball, T., and Grosse-Wentrup, M. (2014). "Causal and anti-causal learning in pattern recognition for neuroimaging," in *4th International Workshop on Pattern Recognition in Neuroimaging (PRNI)* (Tübingen: IEEE). doi: 10.1109/PRNI.2014.6858551

Wilhelm, K., Lemaire, P., Curdt, W., Schühle, U., Marsch, E., Poland, A., et al. (1997). "First results of tide sumer telescope and spectrometer on SOHO," in *The First Results From SOHO* (Springer), 75–104. doi: 10.1007/978-94-011-5236-5_5

Yaramakala, S., and Margaritis, D. (2005). "Speculative markov blanket discovery for optimal feature selection," in *Fifth IEEE International Conference on Data Mining (ICDM'05)* (Houston, TX: IEEE), 4. doi: 10.1109/ICDM.2005.134

Yu, Y., Chen, J., Gao, T., and Yu, M. (2019). DAG-GNN: DAG structure learning with graph neural networks. *arXiv* 1904.10098.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* 172, 1873–1896. doi: 10.1016/j.artint.2008.08.001

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.

frontiers in Artificial Intelligence

SYSTEMATIC REVIEW
published: 20 May 2021
doi: 10.3389/frai.2021.550030

Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends

Lindsay Wells[1] *and Tomasz Bednarz*[1,2]*

[1] Expanded Perception and Interaction Center, Faculty of Art and Design, University of New South Wales, Sydney, NSW, Australia, [2] Data61, Commonwealth Scientific and Industrial Research Organisation, Sydney, NSW, Australia

Research into Explainable Artificial Intelligence (XAI) has been increasing in recent years as a response to the need for increased transparency and trust in AI. This is particularly important as AI is used in sensitive domains with societal, ethical, and safety implications. Work in XAI has primarily focused on Machine Learning (ML) for classification, decision, or action, with detailed systematic reviews already undertaken. This review looks to explore current approaches and limitations for XAI in the area of Reinforcement Learning (RL). From 520 search results, 25 studies (including 5 snowball sampled) are reviewed, highlighting visualization, query-based explanations, policy summarization, human-in-the-loop collaboration, and verification as trends in this area. Limitations in the studies are presented, particularly a lack of user studies, and the prevalence of toy-examples and difficulties providing understandable explanations. Areas for future study are identified, including immersive visualization, and symbolic representation.

Keywords: explainable AI, reinforcement learning, artificial intelligence, visualization, machine learning

OPEN ACCESS

Edited by:
Novi Quadrianto,
University of Sussex, United Kingdom

Reviewed by:
Mohan Sridharan,
University of Birmingham,
United Kingdom
Fabio Aurelio D'Asaro,
University of Milan, Italy
Arnaud Fadja Nguembang,
University of Ferrara, Italy
Elena Bellodi,
University of Ferrara, Italy

***Correspondence:**
Tomasz Bednarz
t.bednarz@unsw.edu.au

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

Received: *08 April 2020*
Accepted: *09 April 2021*
Published: *20 May 2021*

Citation:
Wells L and Bednarz T (2021) Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends. Front. Artif. Intell. 4:550030. doi: 10.3389/frai.2021.550030

INTRODUCTION

Explainable Artificial Intelligence (XAI) is a growing area of research and is quickly becoming one of the more pertinent sub-topics of Artificial Intelligence (AI). AI systems are being used in increasingly sensitive domains with potentially large-scale social, ethical, and safety implications, with systems for autonomous driving, weather simulations, medical diagnosis, behavior recognition, digital twins, facial recognition, business optimization, and security just to name a few. With this increased sensitivity and increased ubiquity comes inevitable questions of trust, bias, accountability, and process—i.e., how did the machine come to a certain conclusion? (Glass et al., 2008). These concerns arise from the fact that, generally, the most popular and potentially most powerful part of AI—Machine Learning (ML)—is essentially a *black-box*, with data input into a trained neural network, which then outputs a classification, decision, or action. The inner workings of these algorithms are a complete mystery to the lay-person (usually the person interacting with the AI). The algorithms can even be difficult for data scientists to understand or interpret. While the architecture and mathematics involved are well-defined, very little is known about how to interpret (let alone explain), the inner state of the neural network. Interaction with such systems are fraught with disuse (failure to rely on reliable automation), and misuse (over reliance on unreliable automation) (Pynadath et al., 2018).

This black-box scenario makes it difficult for end-users to trust the system they are interacting with. When an AI system produces an unexpected output, this lack of trust often results in skepticism and possibly even rejection on the part of the end-user. It is not clear if the result

is "correct" or as a result of some flaw or bias in the creation of the AI system that led to the model being overfit on training data not representative of wide range of examples in the real world, or underfit, not sufficiently modeling the complexities of the target environment. These errors may have considerable side effects, such as unsafe resultant behaviors in factories due to misclassification, unfair treatment of members of society, unlawful actions, or financial impact on companies employing AI solutions. Marcus and Davis (2019) describe a number of these issues in their book *Rebooting AI*. They argue that current approaches to AI are not "on a path to get us to AI that is safe, smart, or reliable" (p. 23).

XAI research in the context of Machine Learning and deep learning aims to look inside this black-box and extract information or explanations as to why the algorithm came to the conclusion or action that it did. In addition to providing tools to assist with trust and accountability, XAI assists with debugging and bias in Machine Learning. The inputs and outputs and network design of Machine Learning algorithms are ultimately still decided with human input (human-in-the loop), and as such are often subject to human errors or bias. Explanations from XAI enabled algorithms may uncover potential flaws or issues with this design (e.g., are certain completely irrelevant features in the input image becoming too much of a factor in outputs?). XAI aims to tackle these problems, providing the end-user with increased confidence, and increased trust in the machine. Recent reviews into XAI have already been conducted, with the most recent being Biran and Cotton (2017), and Miller et al. (2017). These reviews focus on data-driven Machine Learning explanations. Recently Anjomshoae et al. (2019) published a systematic literature review on goal-driven explainable AI, which encompassed Reinforcement Learning (RL), although the review did not provide any specific commentary on approaches used within that area. These reviews indicate that XAI is a growing area of importance, and this is also reflected in a recent move by Google to release a range of XAI tools.[1] Furthering the need for research in the area of XAI is the recent General Data Protection Regulation in the EU, which has a provision for the right to explanations (Carey, 2018).

In the broader ML space, the review of 23 articles by Miller et al. (2017) determined that human behavioral experiments were rare. Anjomshoae et al. (2019) reviewed 62 papers and found that after text-style explanations, which were present in 47% of papers, explanations in the form of visualization were the next most common, seen in 21% of papers. Visualization presents a dynamic and exploratory way of finding meaning from the ML black-box algorithms. A considerable amount of work has already gone into the concept of "saliency maps" which highlight areas of the input image that were of importance to the outcome, see Adebayo et al. (2018).

Following on from these previous reviews, the current work aims to examine XAI within the scope of *RL*. RL agents generally leverage a *Markov Decision Process (MDP)*, whereby at each timestep, an *action* is selected given a certain input set of observations (*state*), to maximize a given *reward*. During compute runs, the agent learns which actions result in higher rewards (factoring in a *discount factor* for obtaining long-term rewards, such as winning the game) through a carefully moderated process of exploration and exploitation. Popularly, RL has been used successfully by the *DeepMind* team to produce agents capable of better than human-level performance in complex games like *GO* (Silver et al., 2016), and a suite of Atari games (Mnih et al., 2015).

In the next section, we will qualify the reasoning for selecting RL as an area for further investigation in terms of XAI and describe the guiding research questions of this work. Then, the methodology used for the systematic literature review will be described, and the results of the review will be presented.

BACKGROUND

This work investigates RL specifically due to the unique challenges and potential benefits of XAI applied to the RL space. The concept of XAI even in agent-based AI system has been considered as early as 1994, in work by Johnson (1994) who described an approach for querying an intelligent agent and generating explanations. The system was domain-independent, implemented for a simulated fighter-pilot agent. The agent itself did not use for its approach, however there are several similarities to current RL work, as the theory behind how an explanation should be worded or generated remains the same. The agent was able to explain a decision made by going back to that point in time and "repeatedly and systematically" (p. 32) modifying the situation state, and observing the different actions the agent would take in order to form a mapping between states and actions.

Benefits

As mentioned above, XAI aims to combat the issues of trust and confidence in AI, a topic which is particularly important when safety is a major factor. Applications such as autonomous vehicles or robotics where the robot takes in observations of the environment around it and performs actions where the result could have an impact on safety are an area where trust and accountability are pertinent (Araiza-Illan and Eder, 2019). Determining why a robot took the action it did (and by extension knowing what factors it considered) in a human-understandable way plays a big part of building a trust that the robot is indeed making intelligent and safe decisions. This could even lead to building a rapport with the robot, making working with it more efficient as their behaviors may become more predictable. Diagnosing what went wrong when a robot or autonomous car is involved in an incident would also benefit from XAI, where we could query the machine about why it took actions in the lead up to the incident, which would allow designers to not only prevent further incidents, but help with accountability or possible insurance or ethical claims (e.g., was the autonomous car at fault, was there a fault in the decision making of the car, or was another third party at fault?).

Another benefit is that RL agents often learn behaviors which are unique and can identify new strategies or policies previously

[1] Available online at: https://cloud.google.com/explainable-ai.

not thought of. A recent example of this was a game of hide-and-seek where agents learned strategies to exploit the physics system of the game to overcome what was intended by the developers to be walls that could not be passed (Baker et al., 2019). Extracting from the black box how these strategies were learned, or under what circumstances these strategies were learned could result in useful new knowledge for decision making or optimization. As Stamper and Moore (2019) point out, analysis of agents playing the Atari 2600 game *Space Invaders* exhibited similar decision-making behaviors to expert human players (e.g., keeping the invaders in a square formation, and destroying right-most enemies first to slow down the rate of advancement), however in other games investigated, the strategies varied more from human play. Understanding and articulating these strategies may result in new knowledge on how to optimally play these games, but also enhance recommendation systems for informed decision making. A quote by Zhuang et al. (2017) sums up the current situation well: "*[. . .] people and computers can both play chess, it is far from clear whether they do it the same way.*"

Challenges

A challenge in providing XAI for RL is that it usually involves a large number of decisions made over a period of time, often aiming to provide the next action at real-time speeds. Compared to standard ML techniques where decisions can happen in isolation or are unrelated to each other, RL explanations generally will need to encompass a set of actions that were related in some way (e.g., outputting explanations such as "I did actions A,B,C to avoid a penalty for Z").

Another challenge is the fact that RL agents are generally trained without using training data (with the exception of where human-replay data is used, such as in Vinyals et al., 2017), and instead learning is facilitated by a feedback loop (observations) from performing actions within an environment. This makes it challenging to generate human-readable explanations. While the observation and action spaces may be labeled in sensible ways, having no human-labeled training data linking actions and observations makes it challenging to produce valid explanations.

Further adding to the difficulties in XAI, is that developing an AI system that is explainable and transparent can be at odds with companies that have explicit commercial interests which they may not want exposed by overly verbose AI. It can also raise issues around protecting their IP, maintaining a competitive advantage, and the additional costs involved with implementing XAI (Mohanty and Vyas, 2018).

METHODOLOGY AND RESEARCH QUESTIONS

With XAI becoming increasingly important for a range of reasons previously described, and work in this area beginning to grow, it is important to take stock of the current approaches in order to find similarities, themes, and avenues for further research. As such, the guiding research questions for this review are:

RQ1: What approaches exist for producing explainable output for Reinforcement Learning?

RQ2: What are the limitations of studies in the area of XAI for Reinforcement Learning?

It is worth taking a moment to clarify the meaning of "explanation" and "explainability" in this paper. In the case of a systematic literature review using these words as search terms, search results will appear for a multitude of meanings and interpretations of these words. For example, "explainability" might refer to something which makes a system more transparent or understandable. An "explanation" may refer to something which describes the actions, decisions, or beliefs of an AI system. "Explainablity" however may also refer to logging or verifications, or an AI system that can be queried or visualized. During the filtering process described in the next section, no restrictions were placed on how the authors defined or interpreted these terms.

Given these research questions, the following section describes the methodology for searching the extant literature for information to address them.

SELECTION OF LITERATURE

To examine the current state of the literature, a systematic literature review using a methodology adapted from Kitchenham et al. (2009) was performed. Searches were conducted on the ACM, IEEExplorer, Science Direct, and Springer Link digital libraries, using Boolean search queries, taking the term "Reinforcement Learning" and combining it with the terms "data visualization," "information visualization," "explanation," "explainable," "explainable ai," "XAI," "black box," "visual analytics," "hybrid analytics," and "human in the loop." The full set of search term combinations can be found in **Supplementary Materials**.

In addition, papers were filtered using the following criteria:

- *recent paper*: papers had to be published within the last 5 years (i.e., since 2014 at time of writing);
- *relevancy*: papers had to be relevant to the topic of *RL* (papers which spoke about general agent-based AI system or RL from a human psychology perspective were excluded) and *explainability* (i.e., papers which did not describe an approach for explaining the actions or policy of an agent were excluded);
- *accessibility*: papers needed to be accessible *via* the portals previously described;
- *singularity*: duplicate papers were excluded; and
- *full paper*: extended abstracts and other short papers were excluded.

As **Figure 1** illustrates, a total of 520 papers were gathered, which was reduced to 404 after filtering out duplicate results using the EndNote software "Find Duplicates" feature. The titles and abstracts of these papers were reviewed for relevance to the domain of RL and XAI, of which 69 were deemed relevant using the relevancy measure described above. These papers were then read fully to determine relevance to the domain. The remaining

FIGURE 1 | Number of papers included in review after various stages of filtering.

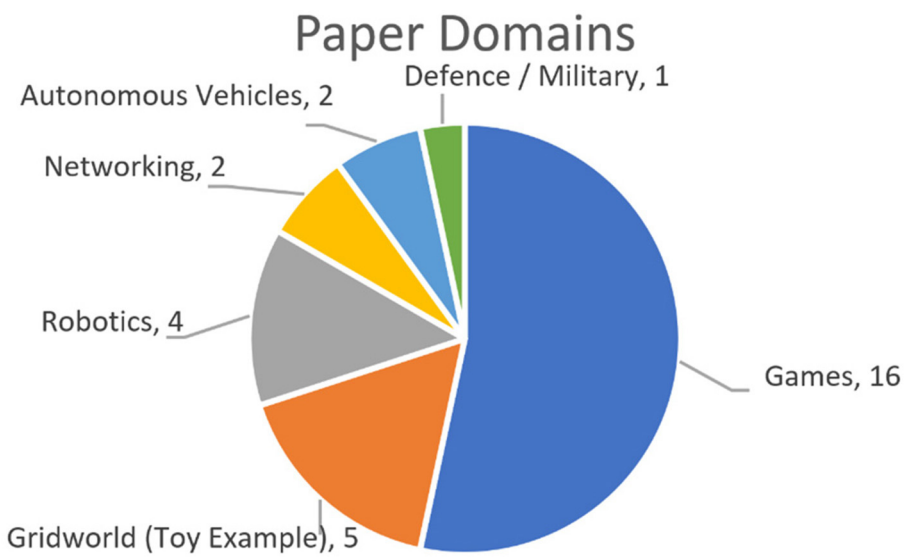


FIGURE 2 | Categorization of papers by domain. Note that some papers were in multiple domains.

20 papers after this stage of filtering constitute the main analysis of this paper.

The jump down from 69 to 20 may seem surprising, however due to the search terms, a number of papers mentioned RL in the body for purposes of describing AI systems generally for the reader, or in some cases RL was used as the technique for generating explanations for a different form of AI such as classification. Such use of the term "Reinforcement Learning" could not be determined until the full paper was examined. Many filtered papers advertised frameworks or implementations for XAI in ML in general and were picked up by the search terms for RL as the papers described the broad spectrum of Machine Learning which encompasses RL. However, these papers ultimately just described typical classification problems instead.

In addition, 5 papers were added to the review, using a snowball sampling technique (Greenhalgh and Peacock, 2005), where if a relevant sounding paper was cited by a reviewed paper, it was subsequently assessed, and if deemed relevant added to the pool of papers for review (15 papers were examined during this stage).

Before going into detail of some of the approaches for XAI in RL, the following section explores at a high level the core themes in the 25 papers reviewed in terms of domain and scope, in order to paint a picture of the current state of the research space.

SUMMARY OF LITERATURE

Selected papers were categorized and analyzed based upon four main topics: domain, publication type, year, and purpose. A full summary table of the selected papers and information about each is provided in **Supplementary Materials**.

Domain

Papers were categorized based upon the featured subject domain(s) they focused on (either in their implementation, or theoretical domain). It was possible for each paper to be in multiple categories. The distribution of papers across the categories is summarized in **Figure 2**, and expanded upon in this section.

The majority of papers (16; 64.0%) focused their examples on the domain of video games (particularly Atari games, given recent popularity due to DeepMind's success), however choice of target game was generally quite broad spread, with the only game utilized in more than one paper was Pac-Man, as illustrated in **Figure 3**. Most common after this were examples using a basic grid-world environment with a navigation task (5 papers), and examples in robotics (4 papers).

The domain of networking tasks such as video bitrate monitoring and cloud-based applications appeared in 2 papers.

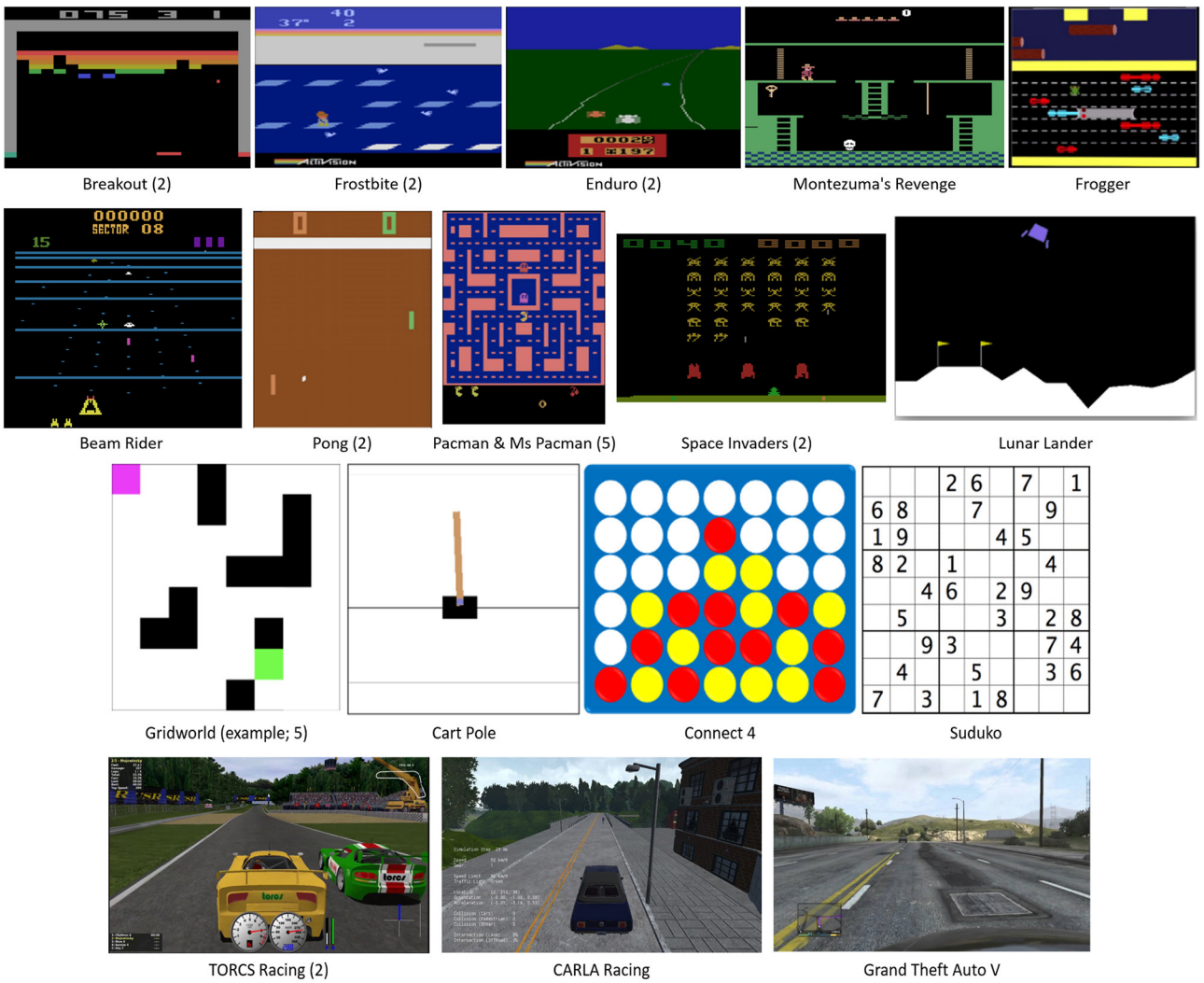


FIGURE 3 | Screenshots of the game-based applications in the studied papers. Where more than one paper used that game, the number of papers using the game are shown in brackets.

An area that was expected to have greater representation was autonomous vehicles (and this is validated by the mention of this area of RL frequently in the reviewed papers), however this area was the focus of only 2 papers.

Finally, one paper was written from a defense/military perspective. It should be noted that only 6 papers attempted to apply their implementation to multiple example situations, however even in these cases, it was from within the same domain (e.g., multiple types of games).

Publication Type

The primary outlet for the reviewed papers was conference proceedings (16 papers), with only 3 papers published in journals. Another 4 papers were from the open access repository arXiv,[2] 3 of which were found as part of the snowball sampling process described previously. One publication (Pynadath et al., 2018) was a book chapter published in "Human and Machine Learning."

Year

The majority of papers found were published in 2019 (15 papers), while only 6 were published in 2018, and 4 in 2017 (see **Figure 4**). This indicates that research into attempting produce explainable RL agents is an area of considerable growth. As we will see, given the sudden increase in publications, there is a reasonable amount of cross-over between some streams of research, and ideally these researchers may consolidate their work and progress together, rather than in parallel, into the future.

Purpose/Scope

The reviewed papers presented a mixture of papers attempting to establish a theory or model (6 papers), while others primarily

[2] Available online at: https://arxiv.org/.

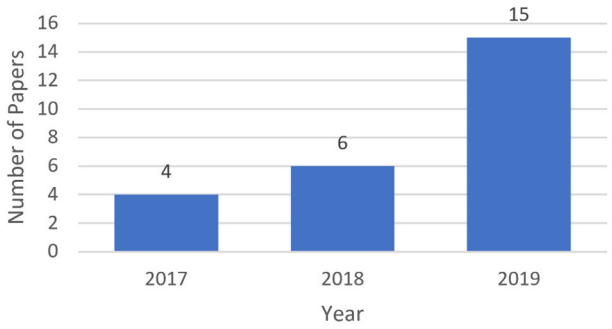


FIGURE 4 | Distribution of surveyed papers by year, indicating an increase of academic interest in this area.

TABLE 1 | A summary of the papers reviewed, categorized by purpose.

Purpose	Papers
Human collaboration	Amir et al. (2019), Hayes and Shah (2017), Huang et al. (2019), Pynadath et al. (2018), Tabrez et al. (2019), Tabrez and Hayes (2019), Ehsan et al. (2019)
Visualization	Dao et al. (2018), Dethise et al. (2019), Iyer et al. (2018), Joo and Kim (2019), Mishra et al. (2018), Pan et al. (2019), Wang et al. (2018), Greydanus et al. (2018), Yang et al. (2018).
Policy summarization	Amir et al. (2019), Fukuchi et al. (2017a,b), Hayes and Shah (2017), Lage et al. (2019), Madumal et al. (2020), Sridharan and Meadows (2019), Stamper and Moore (2019), Lyu et al. (2019), Verma et al. (2018)
Query-based explanations	Amir et al. (2019), Hayes and Shah (2017), Kazak et al. (2019), Sridharan and Meadows (2019)
Verification	Kazak et al. (2019), Dethise et al. (2019)

Note that a paper could have multiple purposes.

focused on introducing a new method for explainable RL (18 papers).

The primary purpose or focus of the reviewed papers was coded down to 5 core topics as shown in 5 (it was possible for a paper to be assigned to multiple topics): *human collaboration* (7 papers); *visualization* (9 papers); *policy summarization* (10 papers); *query-based explanations* (5 papers); and *verification* (1 paper). This distribution of purposes is consistent with the findings in the Anjomshoae et al. (2019) review, which found a high number of visualization-based explanation systems.

Table 1 summarizes which category was determined for each paper, and the distribution of papers across different domains is presented in **Figure 5**. These topics are used to help structure the following discussion section.

DISCUSSION

The following sections address each of the defined research questions for this work.

RQ1: What Approaches Exist for Producing Explainable Output for Reinforcement Learning?

Human Collaboration

Seven papers discussed approaches that were inherently human-based in their approaches.

Pynadath et al. (2018) tested human interaction with an agent while manipulating the perceived ability of the agent by altering the explanations it gave. They explored the design of explanations for Partially Observable Markov Decision Process (POMDP)-based RL agents. The authors mapped different components of the POMDP model to the Situational Awareness-based Agent Transparency (SAT) model in order to determine a set of "explanation content" to assist with situational awareness in a military setting. The SAT was comprised of three levels:

- The agent's *actions and plans;*
- The agent's *reasoning process*, and;
- The agent's *predicated outcomes* (optionally including uncertainties).

The researchers were able to manipulate the ability of the agent in their experiments for human-machine team missions. They evaluated an explainable robot agent which would navigate around an online 3D environment. The robot used a scanner to recommend to the human team members what they should do next (enter the building, put on armor etc.) Example explanations for this agent included "I believe that there are no threats in the market square" for beliefs about the current state of the world, or "my image processing will fail to detect armed gunmen 30% of the time" as an explanation of the current state of the observation model the agent was using.

The authors evaluated differing levels of explanation and found that in general they could potentially "improve task performance, build transparency, and foster trust relationships." Interestingly, the authors noted that explanations which resulted in users being uncertain about what to do next were considered just as ineffective as when no explanations were given. Ability of the robot was tested as well. The high-ability robot got predictions 100% correct, resulting in participants not questioning the robots' decisions (potentially leading those participants to ignore some of the explanation content as the robot "got it right anyway"). This is a prominent example of overreliance, mentioned earlier.

In a similar vein as the work by Pynadath et al. (2018) and Sridharan and Meadows (2019) contributed a theory of how to enable robots to provide explanatory descriptions of its decisions based upon its beliefs and experiences. Building upon existing work into scientific explanation, the theory encompassed 3 core components:

1) How to represent, reason with, and learn knowledge to support explanations.
2) How to characterize explanations in terms of axes of abstraction, specificity, and verbosity.
3) How to construct explanations.

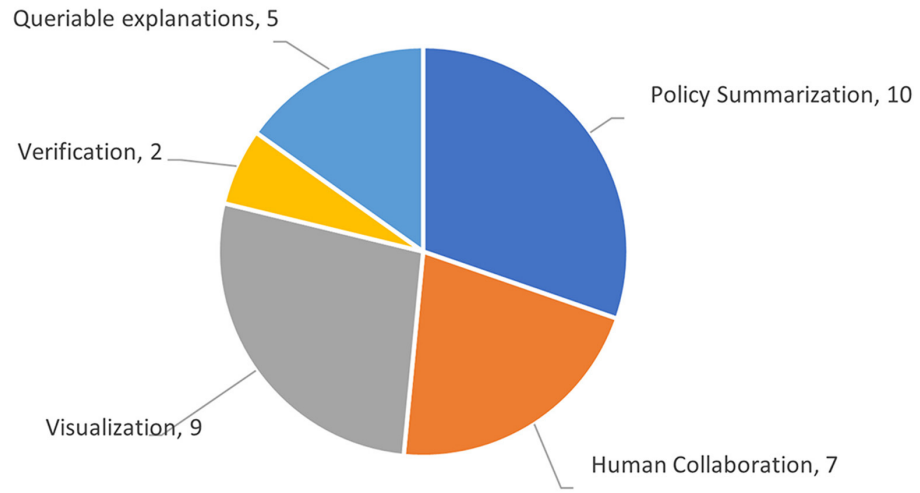


FIGURE 5 | Categorization of papers by scope. Note that some papers were multi-faceted and covered multiple categories.

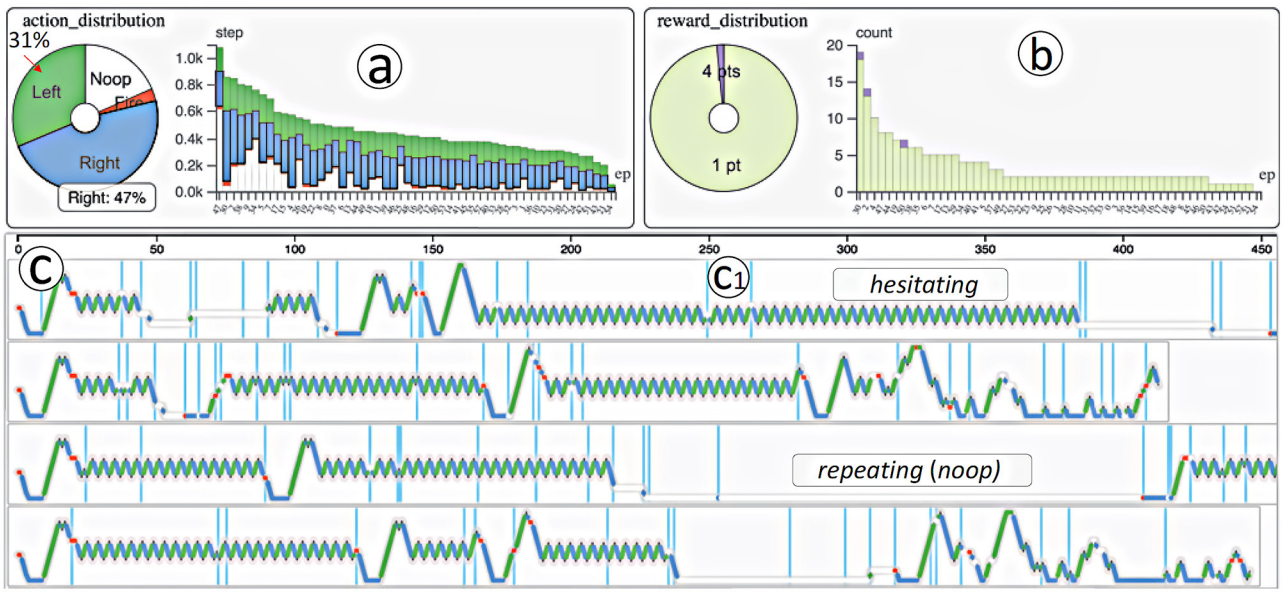


FIGURE 6 | Example visualizations from DQNVis, showing **(a,b)** episode duration, and **(c)** actions taken over time and how experts identified these as "hesitating" and "repeating" behaviors which were non-optimal (from Wang et al., 2018, p. 294, reproduced with permission).

The authors went on to describe an architecture which implemented this theory in a cross-domain manner. The architecture itself operates on two levels, first reasoning using commonsense domain knowledge at a high-level a plan of actions. The system utilized RL for the actions, working alongside Answer Set Prolog (ASP) reasoning of object constants, domain attributes, and axioms based upon state-action-reward combinations (Sridharan and Meadows, 2019). The ASP reasoning was used for planning and diagnostics, and to trigger the learning (using RL) of new concepts when something unknown is encountered (Sridharan and Meadows, 2018). When producing explanations, the architecture extracted words and phrases from a human query matching a template, and based upon human-controlled values effecting the abstraction, specificity, and verbosity of the explanation, reasoned based upon changes in beliefs about the environment. The two evaluation tasks used were moving objects to a target location and following a recipe.

Tabrez and Hayes (2019) described a framework called RARE (Reward Augmentation and Repair through Explanation) which also extended the POMDP model. Using this framework, the RL agent was able to infer based upon a human's behavior the most likely reward function they were using and communicate to the user important differences or missing information in the human's reward function. The agent autonomously provided "actionable statements," which the authors tested in a controlled experiment on a Sudoku-style game. The control group were given an agent who would alert users who were about to make a mistake, and the treatment group had an agent which would indicate that a move would result in failure, and *explain* to them which rules of the game would be broken. Participants found the agent with explanations to be more helpful, useful, and intelligent. The authors however highlighted the fact that the approach does not scale. Statements used a template in the form of: "If you perform {describe action}, you will fail the task in state {describe state} because of {describe reward function difference}."

Looking at autonomous vehicles as an example, Pan et al. (2019), contributed Semantic Predictive Control (SPC) which learns to "predict the visual semantics of future states and possible events based upon visual inputs and an inferred sequence of future actions" (p. 3203). Visual semantics in this case refers to object detection, and the authors suggested that these predicted semantics can provide a visual explanation of the RL process. The paper, however, provided little insight into how it addresses the problem of XAI.

Another work in the autonomous driving domain, Huang et al. (2019) compared approximate-inference and exact-inference approaches in an attempt to leverage the way humans make inferences about how a RL agent operates based upon examples of optimal behavior. Their work compared different approximate-inference models in a user study, where users were shown example behaviors. Users were tasked with selecting from a range of trajectories which one they thought the autonomous driver was most likely to take. The authors' findings suggested that an approximate-inference model using a Euclidean-based approach performed better than algorithmic teaching.

Finally, work by Ehsan et al. (2019) presented a novel approach for generating rationales (the authors note a distinction between this and explanations, indicating that rationales do not need to *explain* the inner workings of the underlying model). The method involves conducting a modified think-aloud user study of the target application (in this case, the game *Frogger*) where participants are prompted to verbally indicate their rationale for each action they take. These rationales (and the associated observation-action pairs in the game) are then cleansed and parsed before being fed through an encoder-decoder network to facilitate natural language generation of actions taken by a RL agent. The authors conducted user studies on the generated explanations compared to random and compared to pre-prepared human explanations. Generated explanations performed better than randomly generated explanations in all factors tested (confidence, human-likeness, adequate justification, and understandability), and performed similarly to the pre-prepared explanations, but did not beat it. A limitation of this work was that the system was designed for turn-based or distinct-step environments, and the authors are continuing their work to look at continuous environments. A major challenge in this is that data collection of rationales during the think-aloud stage is constrained to be after each action taken and would be an arduous process for a human for games larger than Frogger.

Visualization

Nine of the papers reviewed focused on graphical visualization of the agent learning process. Some remarkable visualizations have already been produced, however as discussed later, limitations exist in the ability of these visualizations to fully explain an agent's behavior or policy.

Wang et al. (2018) provided a comprehensive yet highly application-specific visualization tool for Deep-Q Reinforcement Learning Networks called *DQNViz*, with the goal of identifying and extracting typical action/movement/reward patterns of agents. While *DQNVis* was scoped to the Atari *Breakout* game and was focused primarily on objectives relating to improving the training of an agent during development, the tool shows the power of visualization techniques to gain insight into the behaviors of an agent.

The system allowed behaviors to be identified and labeled using tools, such as regular expressions, principal component analysis, dynamic time warping, and hierarchical clustering. Core behaviors in *Breakout* that the agent went through during training included *repeating, hesitating, digging,* and *bouncing* (see **Figure 6**). The tool allowed users to investigate certain moments and see what the agent did at that time and highlight which states in each layer of the convolutional neural network were most activated. Coupled with video output surrounding certain behaviors, experts were able to explore what caused bad behaviors like repetition or hesitation.

Testing so far on DQNViz has been conducted only with deep learning experts who were involved in the initial collaborative process of building the system, and so the usability for non-experts remains to be seen.

Region Sensitive Rainbow (RS-Rainbow) was a visualization method contributed by Yang et al. (2018). RS-Rainbow used a "region-sensitive module" (p. 1) added in after the standard image convolution layers of a deep neural network, which looks for distinctive patterns or objects, and this representation replaces the original representation of the screen as the state used by the deep Q network agent. The authors provided three alternative approaches for visualizing the important regions: a weights-overlay, a soft saliency mask, and a binary saliency mask. Tested on a range of Atari games, the agent out-performed state-of-the-art approaches for Deep RL. The authors have not yet studied to what extent the visualization aids in human understanding in non-experts and ability to debug agents.

Greydanus et al. (2018) also presented a visualization technique tested on Atari games. They contributed *perturbation-based saliency*, to artificially reduce the RL agent's certainty about specific features (e.g., the location of the ball in the Atari game Pong), and its effect on the agent's policy. Using this, the authors could determine the regions of an image which had the most effect. The authors used the visualization to understand "strong"

policies where agents perform dominant strategies (such as "tunneling" in *Breakout*), and to observe how attention changes while the agent learns. The study found that the visualization helped non-expert users with identify agents with overfitted models. Finally, the study showed how visualization can aid in debugging, showing examples of Atari games where human performance was not yet attained. In *MsPacman*, it was found that the agent was not tracking the ghosts, and in *Frostbite*, the agent was only tracking the player and goal, and not the destination platforms.

A similar approach to highlighting areas of an image that were relevant to a decision was presented by Joo and Kim (2019) who applied the Gradient-weighted Class Activation Mapping (*Grad-CAM*) approach, to Asynchronous Advantage Actor-Critic (A3C) deep RL in the context of Atari Games. The result was effectively a heatmap indicating which parts of the input image affected the predicted action.

A more complex approach to visualizing the focus of a RL agent was presented by Iyer et al. (2018). The authors claimed their system could "automatically produce visualization[s] of their state and behavior that is intelligible to humans." Developed within the domain of Atari games, the authors used template matching to detect objects in the screen input to produce a number of "object channels" (one for each detected object), as extra input into the convolutional neural network used by the RL agent. The authors also described an approach to produce a "pixel saliency map," where pixels are ranked in terms of their contribution toward the chosen action in that state (see **Figure 7**). As the pixel map is generally not human intelligible (i.e., it is difficult to interpret due to noise and other factors), the approach was combined with the previously mentioned object detection, to produce an "object saliency map" which is easier for humans to understand. The authors tested the system using human experiments, where participants were tasked with generating explanations of the behavior of a Pacman agent, and predict the next action. Participants assisted by the object salience maps performed significantly better on the tasks.

Sparse Bayesian Reinforcement Learning (SBRL; Lee, 2017) can explain which relevant data samples influenced the agent's learning. An extension to SBRL by Mishra et al. (2018) was *V-SBRL* which was applied to vision-based RL agents. The system maintains snapshot storage to store important past experiences. The authors presented an approach to visualizing snapshots at various important locations (as determined by the SBRL

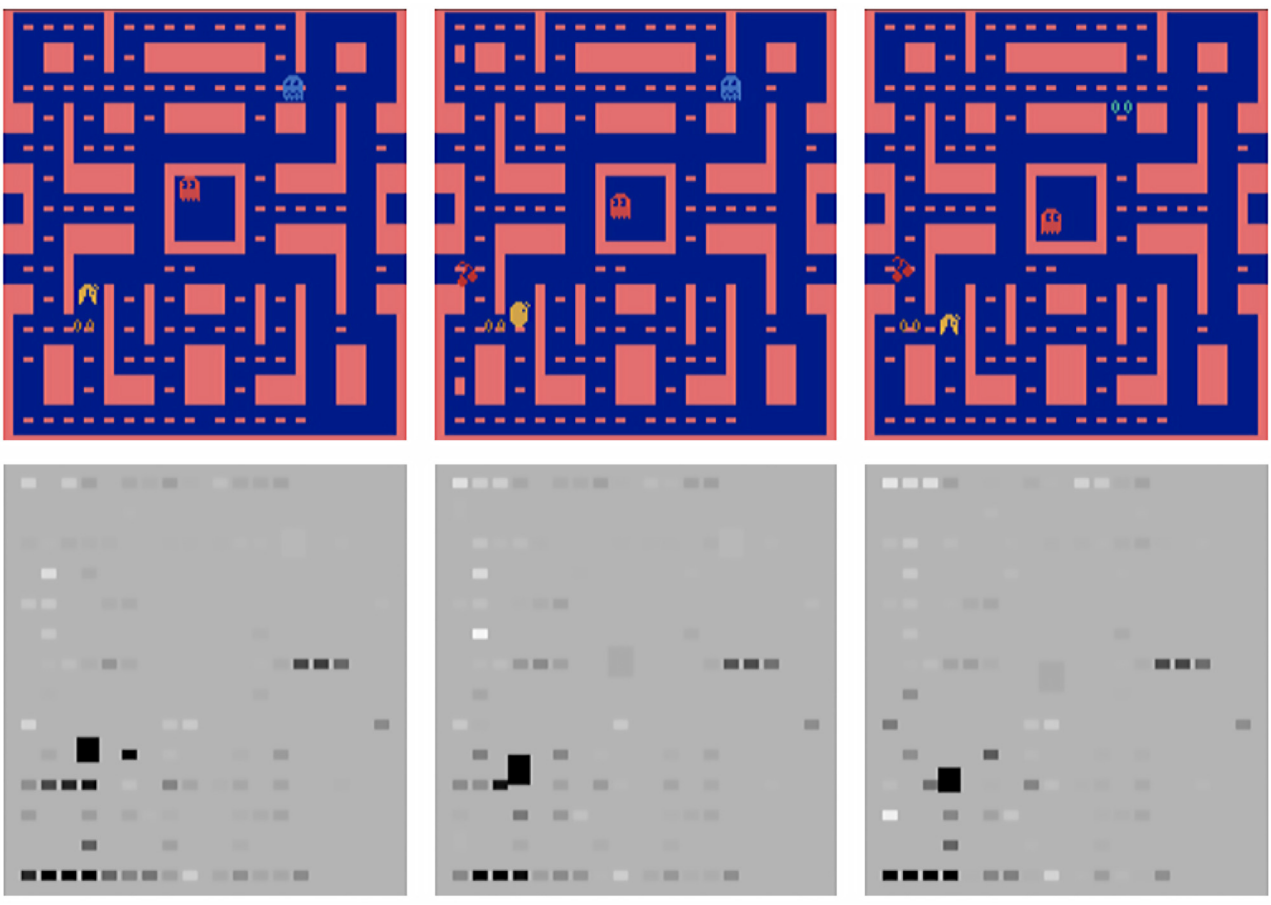


FIGURE 7 | Screenshots (left) and their matching object saliency maps (right) in the game Ms Pacman (from Iyer et al., 2018, p. 148, reproduced with permission).

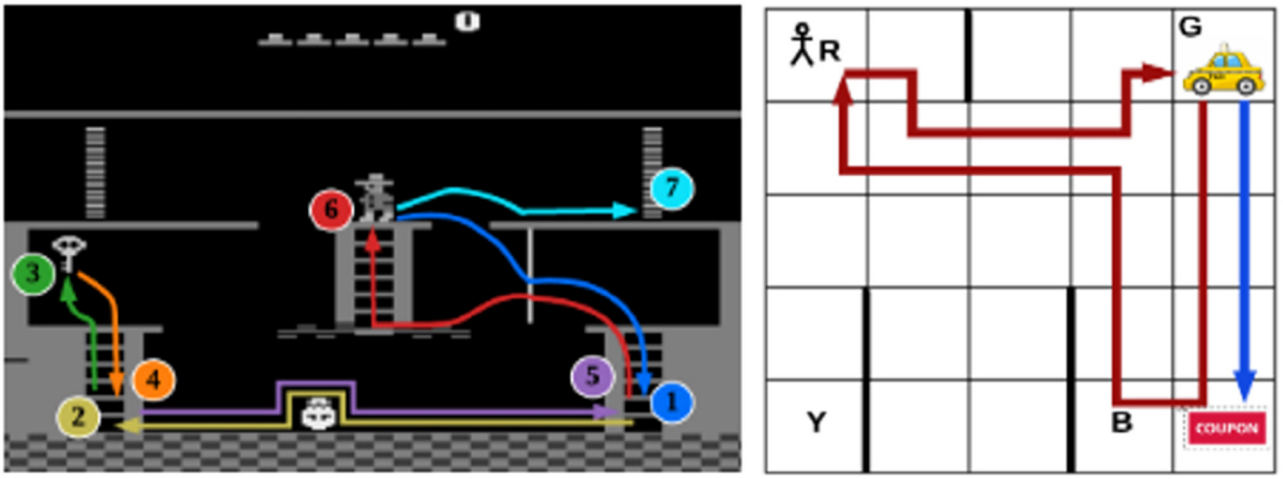


FIGURE 8 | Summarized policies for Montezuma's Revenge (left), and the "taxi" problem (right), (from Lyu et al. (2019), p. 2975, reproduced with permission).

algorithm), by showing state-action pairs. In the context of a navigation task, an interesting visualization was provided by overlaying the snapshots on a Q contour plot, allowing designers to see where the agent had confidence in its actions and where it did not. V-SBRL may prove to be useful in continuous environments, where the number of important moments may be high, but can be compressed down by finding similar state-action pairs within the continuous space. In another paper from the same authorship team, Dao et al. (2018) applied the approach to the Atari games Pong and Ms Pacman.

Pan et al. (2019) as previously described provided visual explanations in the form of object detection.

Policy Summarization

Ten papers provided approaches to in some way summarize the policy that a RL agent has learned. While a policy summary doesn't explain an individual action, it can help provide context for why an action was taken, and more broadly why an agent makes the overall set of actions it makes.

Fukuchi et al. (2017a) described the *Instruction-based Behavior Explanation* (IBE) approach which allows an agent to announce their future behavior. To accomplish this, the agent leveraged Interactive RL where experts provide instructions in real-time to beginner agents. The instructions are then re-used by the system to generate natural-language explanations. Further work by Fukuchi et al. (2017b) then expanded on this to a situation where an agent dynamically changed policy.

Hayes and Shah (2017) used code annotations to give human-readable labels to functions representing actions and variables representing state space, and then used a separate Markov Decision Process (MDP) to construct a model of the domain and policy of the control software itself. The approach is compatible not only with RL, but also with hard-coded conditional logic applications too.

The authors tested their approach on three different domains, a grid-world delivery task, the traditional Cart Pole task, and an inspection robot task. Generated policies were similar in nature to the expert-written policies. The authors suggested that the state space and action space of the learned domain model needs to be constrained in order for the approach to be effective, and to prevent combinatory explosion. It remains to be seen if the approach will work on environments more complex than the Cart Pole.

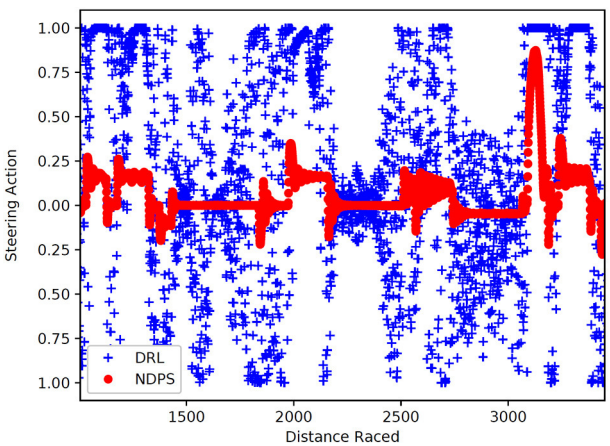


FIGURE 9 | Plot of steering actions generated by standard DRL agent vs. the summarized NDPS policy, which resulted in much smoother steering movements (from Verma et al., 2018, p. 7, reproduced with permission).

Amir et al. (2019) proposed a conceptual framework for strategy summarization. The framework consisted of three core components:

1) intelligent states extraction: which given the agent's strategy/policy as input, outputs a prioritized subset of states to be included in the summary—the main challenge being determining what the desirable characteristics of a state are for use in a summary;

2) world state representation: which involves the summarization of potentially complex world states (i.e., an agent may consider a large number of variables with different weights for certain decisions); and
3) the strategy summary interface: which is concerned with a usable and appropriate user interface for exploration of the summary, which is guided by both the user and the system itself.

For each of these components, the authors provided potential research directions for addressing these problems in the RL space, however this is the only paper reviewed which did not include an implementation which was tested alongside the theoretical framework.

Recent work by Madumal et al. (2020), implemented explanations in a RL agent playing *StarCraft II*, under the premise that humans would prefer causal models of explanation. The agent was able to answer "counterfactual" levels of explanations, i.e., "why" questions. The authors introduced an approach where a causal graph was generated in the form of a directed acyclic graph, where state variables and rewards were nodes, and actions being edges (assuming that an action *caused* a transition between different states). Using structural causal equations, on the causal graph, an explanation was generated.

The explainable agent was tested on 120 participants. To test participants understanding of the explanations, they were tasked with first watching the agent play StarCraft II and explain its actions, followed by watching an agent play and predict its next action. The agent was found to have statistically significantly higher levels of satisfaction and understanding of actions taken than a non-explainable agent. Interestingly however, no significant difference in levels of trust was found, a fact that the author attributed to the short interaction time with the agent.

A set of causal rules was also used in similar work by Lyu et al. (2019) who proposed the Symbolic Deep Reinforcement Learning (SDRL) framework, aimed at handling high-dimensional sensory inputs. The system used symbolic planning as a high-level technique for structuring the learning with a symbolic representation provided by an expert. The high-level symbolic planner had the goal of maximizing some "intrinsic" reward of formulating the most optimal "plan" (where a plan is a series of learned sub-tasks). DRL was used at the "task/action" level to learn low-level control policies, operating to maximize what the authors call an "extrinsic" reward. The authors tested their new approach on the classic "taxi" Hierarchical Reinforcement Learning (HRL) task, and the Atari game *Montezuma's Revenge* (see **Figure 8**). While the system contributed gains in terms of data efficiency, of interest to this paper is the use of symbolic representation and the high-level planner. Such representation of the environment and action space and abstraction at a high-level can be useful in the pursuit of XAI as it may open up opportunities to (with careful design) provide more interpretable systems.

Verma et al. (2018) described a framework for generating agent policies called Programmatically Interpretable Reinforcement Learning (PIRL), which used a high-level, domain-specific programming language, similar to the symbolic representations mentioned previously. The system used DRL for initial learning, and then a novel search algorithm called Neurally Directed Program Search (NDPS) to search over the DRL with a technique inspired by imitation learning to produce a model in the symbolic representation. The resulting model was described by the authors as "human readable source code" (p. 9), however no tests have yet been conducted on how well users can understand it, or how useful it is for debugging. The authors indicated that the resulting policy was smoother than the one generated by DRL—in the case of the test domain of a racing game, the steering output was much smoother, albeit with slower lap times (see **Figure 9**).

Lage et al. (2019) reported on different approaches for agent policy summarization, using Inverse Reinforcement Learning and Imitation learning approaches. Tested in three different domains, the authors found that the policy of an agent was most accurately reproduced when using the same model that was used for extraction as was used for reconstruction. Stamper and Moore (2019) compared policies generated by machines to those of humans. Using *post-hoc* human inspection, they analyzed data from a DQN RL agent, using t-SNE embedding. They found that the agent playing *Space Invaders* exhibited similar decision-making behaviors to expert human players (e.g., keeping the invaders in a square formation, and destroying right-most enemies first to slow down the rate of advancement). The work is still in its early stages, and the authors plan to automate the strategy identification process.

The previously described work by Sridharan and Meadows (2019) also provided for a summary of learned policy in their approach at different levels of abstraction. These summaries were able to be queried by the user, as explained in the next section.

Query-Based Explanations

Five papers described an interactive query-based approach to extracting explanations from a RL agent. Hayes and Shah (2017) went into the most detail. Broadly, their system conducted 4 core actions:

1) identify the question based upon a template approach, e.g., "When do you do {action}?";
2) resolve states [using the template from (1), determine the states that are relevant to the question];
3) summarize attributes (determine common attributes across the resolved states); and
4) compose a summary in a natural language form (using "communicable predicates").

These steps integrated with the code annotations previously described for this system.

The work by Madumal et al. (2020) featured a query-based RL agent playing *Starcraft II*. The agent focused on handling *why?* and *why not?* questions. An example question was provided by the author, "Why not build barracks?", to which the agent replied, "Because it is more desirable to do action build_supply_depot to have more supply depots as the goal is to have more destroyed units and destroyed buildings." This is a great example of a RL

agent being able to answer questions about its action, however it remains to be seen how well this approach will scale.

Kazak et al. (2019) presented an approach which allowed experts to query a RL agent in order to perform verification tasks. In their tests, queries took over 40 s to complete. Their work is described in more detail in the verification section, as that was the primary purpose of that work.

Previously described work on policy summarization by Amir et al. (2019) and Sridharan and Meadows (2019), both highlighted the importance of being able to further query summarized policies in order to prevent initial cognitive load on the user by presenting a policy that was too complex or verbose. The query functionality in Sridharan and Meadows (2019) was able to be customized to different levels of abstraction, specificity, and verbosity, but this was further guided by the ASP-based architecture they used.

Verification

A theme which was found within two reviewed papers was that of verification. Verification is an area of importance to RL for a number of reasons, not least due to the impact on safety it can have. As Fulton and Platzer (2018) point out, formal verification allows us to detect discrepancies between models and the physical system being controlled, which could lead to accidents.

Acknowledging the non-explainability of RL systems, Kazak et al. (2019) suggested that verifying that systems adhere to specific *behaviors* may be a good alternative to verifying that they adhere to exact values from a model. They presented an approach called *Verily*, which checks that the examined system satisfies the pre-defined requirements for that system by examining all possible states the agent could be in, and using the formal verification approach *Marabou*. The system identifies "undesirable" sequences using bounded model checking queries of the state space. Of interest to this review is that when a system is found to not meet the requirements, a *counter example* is generated that *explains* a scenario in which the system fails. The authors tested this approach on three case studies within a networking/cloud computing domain, providing verification that the RL systems employed were conducting desired behaviors and avoiding poor outcomes (e.g., verifying that an adaptive video streaming system was correctly choosing between high- or low-quality video based upon network conditions). The impact of *Verily* on the trust relationship between humans and the systems remains to be tested, as does the scalability of this approach since it operates on all possible states.

Similar to the Kazak et al. study was work by Dethise et al. (2019), also in the domain of RL for networking problems. They looked at using interpretability tools to identify unwanted and anomalous behaviors in trained models. The system in question was *Pensieve*, an adaptive bit rate selector, common in video streaming services. The authors analyzed the relationship between data throughput and decisions made by the agent. Using simple visualization techniques, they showed that the agent never chose certain bandwidths for users (despite there being no bias present in the training data). Further analysis revealed that the agent preferred to multiplex between higher and lower bitrates when the average available bitrate was one of the identified ignored bitrates. The authors also analyzed which features contributed the most to decisions, finding that the most highly weighted feature was the previous bit rate. This paper used domain knowledge to lead a guided exploration of the inputs of a relatively simple RL agent, however some of the approaches and visualizations presented may be of use in other areas.

RQ2: What Are the Limitations of Studies in the Area of XAI for Reinforcement Learning?

In reviewing the collected papers, a number of common limitations were identified, particularly in the use of "toy examples," a lack of new algorithms, lack of user testing, complexity of explanations, basic visualizations, and lack of open-sourced code. The following sections discuss in more detail the various common limitations.

Use of Toy Examples, Specific Applications, and Limited Scalability

Given the early stages of XAI for RL research, all papers reviewed presented effectively "toy" examples, or case studies which were deliberately scoped to smaller examples. In most cases this was done to avoid the combinatory explosion problem in which where state- and action-space grow, so do the number of possible combinations of states and actions. An example of this was Hayes and Shah (2017) who scoped their work to the basic Cart Pole environment. Similarly the Tabrez and Hayes (2019) paper focused on a grid-world example.

Many authors indicated limitations in scaling the approach to more complex domains or explanations (with the exception of Sridharan and Meadows, 2019, who indicated that a strong contribution of their work was that their approach would scale). Sixteen of the papers reviewed were either agents within video games or were tested with video game problems, and surprisingly few were on more real-world applications such as autonomous driving or robotics. Examples of this include Ehsan et al. (2019) who provide an interesting example but is highly scoped to the *Frogger* game, and Madumal et al. (2020) who looked at Starcraft II. While this is naturally following on from the success of DeepMind, and video game problems provide for challenging RL tasks, there is an opportunity for more work on applications outside of this domain.

Focus on Modification of Existing Algorithms

Papers examined in this review described RL approaches or visualization techniques to augment or accompany existing RL algorithms. There is an opportunity in this area to design RL algorithms with explainability in mind. Symbolic representation can be a step toward allowing for inherently explainable and verifiable agents.

Lack of User Testing

A major limitation of the studies presented in this review is that many approaches were either not tested with users (17 papers), or when they did, limited details of the testing were published, failing to describe where the participants were recruited from, how many were recruited, or if the participants

were knowledgeable in Machine Learning (Pynadath et al., 2018; Tabrez and Hayes, 2019; Tabrez et al., 2019). Participant counts varied greatly, with one paper using 3 experts (Wang et al., 2018), others with students (Iyer et al., 2018), $n = 40$; and Greydanus et al. (2018), $n = 31$, and three recruiting using Amazon Mechanical Turk[3] (Huang et al., 2019, $n = 191$; Madumal et al., 2020, $n = 120$; and Ehsan et al., 2019, $n = 65$ and $n = 60$).

This lack of user testing across the reviewed papers is consistent with the findings in the Miller et al. (2017) review of XAI in Machine Learning.

Explanation Presentation

In some cases, implementations provided too much information for the human participant, or required significant additional knowledge from the human team member, making these approaches unsuitable for use by laypeople or even knowledgeable domain experts. This finding is consistent with the survey paper by Miller et al. (2017) who found that there is very little research in the XAI space on leverages existing work on how people "generate, select, present, and evaluate" (p. 4) explanations, such as the work by Lombrozo (2007) which describes how people prefer simpler and more general explanations over specific and more likely explanations.

In the papers focusing on visualization, most expanded on existing techniques of pixel saliency which have successfully been used for image classification (e.g., Greydanus et al., 2018; Iyer et al., 2018; Yang et al., 2018). RL problems happening over time may need more complex visualization techniques to capture the temporal dimension. Other forms of visualization presented were primarily 2D graphs (e.g., DQNVis, Wang et al., 2018), however these solutions may struggle to scale and to be interpretable in more complex domains given the large amount of data involved network design.

The majority of papers with user studies presented explanations or visualizations palatable only to experts. Further research could look at providing explainable systems targeted at laypeople or people more likely to be working with the agent, rather than those with a background in artificial intelligence. Symbolic representation was present in a number of papers in this review (e.g., Verma et al., 2018; Lyu et al., 2019). Future research could consider alternatives to text representation of these to provide more visceral explanations, such as annotations in the virtual environment. Similarly, visualization techniques presented in the papers in this review are a good start (e.g., DQNVis, Wang et al., 2018), however the toolkits provided may be enhanced by the addition of visualization techniques better designed for handling the temporal dimension of RL (such as the Immersive Analytics Toolkit by Cordeil et al. (2019) or TensorBoard graphs[4]), as well as multi-modal, immersive forms of visualization such as virtual or augmented reality to better explore the complex data structures of neural networks (Marriott et al., 2018).

[3]Available online at: https://www.mturk.com/.

[4]Available online at: https://www.tensorflow.org/tensorboard/graphs.

Lack of Open-Source Code

Finally, only four papers provided the reader with a link to the open-source repository of their code (Greydanus et al., 2018; Yang et al., 2018; Dethise et al., 2019; Sridharan and Meadows, 2019). This lack of availability of code could be as the result of many things, but we argue that given the toy example nature of the work previously described, that some authors didn't find utility in providing code online. Additionally, intellectual property issues can sometimes arise, making it not possible to publish code in an open-source matter. This is despite the potential benefits for the academic community of shared, open-source code.

CONCLUSION

The area of XAI is of growing importance as Machine Learning techniques become commonplace, and there are important issues surrounding ethics, trust, transparency, and safety to be considered. This review has explored the extant literature on XAI within the scope of RL. We have shown that work in this area is still in its early stages but growing in prevalence and impact it can make. Clear trends are appearing in terms within the area with researchers focusing on human collaboration, visualization techniques, whole-of-policy summarization explanations, query-based explanations, and verification approaches.

This paper has described current approaches, while also identifying a range of limitations in this field of research, primarily finding a lack of detail when describing human experiments, limited outcomes in terms of scalability and level of comprehension of explanations for non-expert users, and under-use of more advanced visualization techniques such as multi-modal displays and immersive visualization. To truly break through the black box of RL, a strong combination of well-articulated explanations coupled with advanced visualization techniques will be essential tools for Machine Learning experts and users alike.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

LW completed the majority of the work here, under the supervision of TB. TB contributed major edits and comments which helped shape the overall project. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by UNSW Grant number: RG190540.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.550030/full#supplementary-material

REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. *arXiv [Preprint] arXiv:*1810.03292.

Amir, O., Doshi-Velez, F., and Sarne, D. (2019). Summarizing agent strategies. *Autonomous Agents Multi Agent Syst.* 33, 628–644. doi: 10.1007/s10458-019-09418-w

Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. (2019). "Explainable agents and robots: results from a systematic literature review," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019* (Montreal, QC: International Foundation for Autonomous Agents and Multiagent Systems), 1078–1088.

Araiza-Illan, D., and Eder, K. (2019). "Safe and trustworthy human-robot interaction," in *Humanoid Robotics: A Reference*, eds A. Goswami and P. Vadakkepat (Dordrecht: Springer Netherlands), 2397–2419.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., et al. (2019). Emergent tool use from multi-agent autocurricula. *arXiv [Preprint] arXiv:*1909.07528.

Biran, O., and Cotton, C. (2017). "Explanation and justification in machine learning: a survey," in *IJCAI-17Workshop on Explainable AI(XAI)* (Melbourne, VIC), 8.

Carey, P. (2018). *Data Protection: A Practical Guide to UK and EU Law*. Oxford: Oxford University Press, Inc.

Cordeil, M., Cunningham, A., Bach, B., Hurter, C., Thomas, B. H., Marriott, K., et al. (2019). "Iatk: an immersive analytics toolkit," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Osaka: IEEE), 200–209.

Dao, G., Mishra, I., and Lee, M. (2018). "Deep reinforcement learning monitor for snapshot recording," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (IEEE), 591–598. doi: 10.1109/ICMLA.2018.00095

Dethise, A., Canini, M., and Kandula, S. (2019). "Cracking open the black box: what observations can tell us about reinforcement learning agents," in *Proceedings of the 2019 Workshop on Network Meets AI & ML* (Beijing), 29–36. doi: 10.1145/3341216.3342210

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M. O. (2019). "Automated rationale generation: a technique for explainable AI and its effects on human perceptions," in *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, CA), 263–274. doi: 10.1145/3301275.3302316

Fukuchi, Y., Osawa, M., Yamakawa, H., and Imai, M. (2017a). "Application of instruction-based behavior explanation to a reinforcement learning agent with changing policy," in *International Conference on Neural Information Processing* (Cham: Springer), 100–108. doi: 10.1007/978-3-319-70087-8_11

Fukuchi, Y., Osawa, M., Yamakawa, H., and Imai, M. (2017b). "Autonomous self-explanation of behavior for interactive reinforcement learning agents," in *Proceedings of the 5th International Conference on Human Agent Interaction* (New York, NY), 97–101. doi: 10.1145/3125739.3125746

Fulton, N., and Platzer, A. (2018). "Safe reinforcement learning via formal methods: toward safe control through proof and learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, (New Orleans, LA).

Glass, A., McGuinness, D. L., and Wolverton, M. (2008). "Toward establishing trust in adaptive agents," in *Proceedings of the 13th International Conference on Intelligent User Interfaces* (New York, NY), 227–236. doi: 10.1145/1378773.1378804

Greenhalgh, T., and Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ* 331, 1064–1065. doi: 10.1136/bmj.38636.593461.68

Greydanus, S., Koul, A., Dodge, J., and Fern, A. (2018). "Visualizing and understanding atari agents," in *International Conference on Machine Learning* (Stockholm), 1792–1801.

Hayes, B., and Shah, J. A. (2017). "Improving robot controller transparency through autonomous policy explanation," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna: ACM), 303–312. doi: 10.1145/2909824.3020233

Huang, S. H., Held, D., Abbeel, P., and Dragan, A. D. (2019). Enabling robots to communicate their objectives. *Autonomous Robots* 43, 309–326. doi: 10.1007/s10514-018-9771-0

Iyer, R., Li, Y., Li, H., Lewis, M., Sundar, R., and Sycara, K. (2018). "Transparency and explanation in deep reinforcement learning neural networks," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY), 144–150. doi: 10.1145/3278721.3278776

Johnson, W. L. (1994). "Agents that learn to explain themselves," in *AAAI* (Palo Alto, CA), 1257–1263.

Joo, H., and Kim, K. (2019). "Visualization of deep reinforcement learning using grad-CAM: how AI plays atari games?" in *2019 IEEE Conference on Games (CoG)* (London, UK). doi: 10.1109/CIG.2019.8847950

Kazak, Y., Barrett, C., Katz, G., and Schapira, M. (2019). "Verifying deep-RL-driven systems," in *Proceedings of the 2019 Workshop on Network Meets AI and ML* (Beijing), 83–89. doi: 10.1145/3341216.3342218

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., and Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information Softw. Technol.* 51, 7–15. doi: 10.1016/j.infsof.2008.09.009

Lage, I., Lifschitz, D., Doshi-Velez, F., and Amir, O. (2019). "Toward robust policy summarization," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems,* 2081–2083.

Lee, M. (2017). *Sparse Bayesian reinforcement learning* (Ph.D. dissertation). Colorado State University, Fort Collins, CO, United States.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cogn. Psychol.* 55, 232–257 doi: 10.1016/j.cogpsych.2006.09.006

Lyu, D., Yang, F., Liu, B., and Gustafson, S. (2019). "SDRL: interpretable and data-efficient deep reinforcement learning leveraging symbolic planning," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI), Vol. 33, 2970–2977. doi: 10.1609/aaai.v33i01.33012970

Madumal, P., Miller, T., Sonenberg, L., and Vetere, F. (2020). "Explainable reinforcement learning through a causal lens," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 (New York, NY), 2493–2500.

Marcus, G., and Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon.

Marriott, K., Schreiber, F., Dwyer, T., Klein, K., Riche, N. H., Itoh, T., et al. (2018). *Immersive Analytics*, Vol. 11190. Springer. doi: 10.1007/978-3-030-01388-2

Miller, T., Howe, P., and Sonenberg, L. (2017). Explainable AI: beware of inmates running the asylum or: how I learnt to stop worrying and love the social and behavioural sciences. *arXiv [Preprint] arXiv:*1712.00547.

Mishra, I., Dao, G., and Lee, M. (2018). "Visual sparse Bayesian reinforcement learning: a framework for interpreting what an agent has learned," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (Bangalore: IEEE), 1427–1434. doi: 10.1109/SSCI.2018.8628887

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Mohanty, S., and Vyas, S. (2018). *How to Compete in the Age of Artificial Intelligence: Implementing a Collaborative Human-Machine Strategy for Your Business.* Apress. doi: 10.1007/978-1-4842-3808-0

Pan, X., Chen, X., Cai, Q., Canny, J., and Yu, F. (2019). "Semantic predictive control for explainable and efficient policy learning," in *2019 International Conference on Robotics and Automation (ICRA)* (Montreal, QC: IEEE), 3203–3209. doi: 10.1109/ICRA.2019.8794437

Pynadath, D. V., Barnes, M. J., Wang, N. and Chen, J. Y. (2018). "Transparency communication for machine learning in human-automation interaction," in *Human and Machine Learning* (Cham: Springer), 75–90.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* 529:484. doi: 10.1038/nature16961

Sridharan, M., and Meadows, B. (2018). Knowledge representation and interactive learning of domain knowledge for human-robot interaction. *Adv. Cogn. Syst.* 7, 77–96.

Sridharan, M., and Meadows, B. (2019). *Towards a Theory of Explanations for Human–Robot Collaboration.* KI - Künstliche Intelligenz. doi: 10.1007/s13218-019-00616-y

Stamper, J., and Moore, S. (2019). *Exploring Teachable Humans and Teachable Agents: Human Strategies Versus Agent Policies and the Basis of Expertise*. Cham: Springer International Publishing. doi: 10.1007/978-3-030-23207-8_50

Tabrez, A., Agrawal, S., and Hayes, B. (2019). "Explanation-based reward coaching to improve human performance via reinforcement learning," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE), 249–257. doi: 10.1109/HRI.2019.8673104

Tabrez, A., and Hayes, B. (2019). "Improving human-robot interaction through explainable reinforcement learning," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Daegu: IEEE), 751–753. doi: 10.1109/HRI.2019.8673198

Verma, A., Murali, V., Singh, R., Kohli, P., and Chaudhuri, S. (2018). "Programmatically interpretable reinforcement learning," in *International Conference on Machine Learning* (Stockholm), 5045–5054.

Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., et al. (2017). Starcraft II: a new challenge for reinforcement learning. *arXiv [Preprint] arXiv:*1708.04782.

Wang, J., Gou, L., Shen, H. W., and Yang, H. (2018). Dqnviz: a visual analytics approach to understand deep q-networks. *IEEE Trans. Visual. Comput. Graph.* 25, 288–298. doi: 10.1109/TVCG.2018.2864504

Yang, Z., Bai, S., Zhang, L., and Torr, P. H. (2018). Learn to interpret atari agents. *arXiv [Preprint] arXiv:*1812.11276.

Zhuang, Y. T., Wu, F., Chen, C., and Pan, Y. H. (2017). Challenges and opportunities: from big data to knowledge in AI 2.0. *Front. Information Technol. Electronic Eng.* 18, 3–14. doi: 10.1631/FITEE.1601883

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership