Beyond ideals: why the (medical) AI industry needs to motivate behavioural change in line with fairness and transparency values, and how it can do it

AI & SOCIETY
https://doi.org/10.1007/s00146-023-01684-3

MAIN PAPER

Beyond ideals: why the (medical) AI industry needs to motivate behavioural change in line with fairness and transparency values, and how it can do it

Alice Liefgreen[1,2] · Netta Weinstein[2] · Sandra Wachter[3] · Brent Mittelstadt[3]

Received: 15 July 2022 / Accepted: 21 April 2023

Abstract
Artificial intelligence (AI) is increasingly relied upon by clinicians for making diagnostic and treatment decisions, playing an important role in imaging, diagnosis, risk analysis, lifestyle monitoring, and health information management. While research has identified biases in healthcare AI systems and proposed technical solutions to address these, we argue that effective solutions require human engagement. Furthermore, there is a lack of research on how to motivate the adoption of these solutions and promote investment in designing AI systems that align with values such as transparency and fairness from the outset. Drawing on insights from psychological theories, we assert the need to understand the values that underlie decisions made by individuals involved in creating and deploying AI systems. We describe how this understanding can be leveraged to increase engagement with de-biasing and fairness-enhancing practices within the AI healthcare industry, ultimately leading to sustained behavioral change via autonomy-supportive communication strategies rooted in motivational and social psychology theories. In developing these pathways to engagement, we consider the norms and needs that govern the AI healthcare domain, and we evaluate incentives for maintaining the status quo against economic, legal, and social incentives for behavior change in line with transparency and fairness values.

Keywords Artificial intelligence · Healthcare · Medicine · Fairness · Bias · Motivation · Behaviour change

1 Introduction

Forms of artificial intelligence (AI) have been adopted across the healthcare sector to augment the accuracy, efficiency, and quality of information feeding into healthcare decision-making (Haleem et al. 2019). AI is increasingly a trusted resource used by clinicians to make diagnostic and treatment decisions—assisting in imaging, diagnosis, risk analysis, lifestyle monitoring and health information management (Davenport and Kalakota 2019). This paper explores conclusions drawn within the AI healthcare literature that, given the trust healthcare practitioners place in AI, extensive investment is required to align AI technologies with certain values and ethical principles, to ensure they meet their end goal of improving medical care. These ethical principles resemble classical principles of bioethics and comprise, amongst others, justice (encompassing fairness) and transparency (Gabriel 2020; Royakkers et al. 2018).[1]

Despite their recognition, there is a lack of research investigating how to practically implement and integrate these principles within the AI development cycle—a process that requires trade-offs and overcoming a multitude of challenges (Ayling and Chapman 2021; Morley et al. 2021; Vakkuri et al. 2019; Whittlestone et al. 2019). While value alignment[2]

✉ Alice Liefgreen
alice.liefgreen@swansea.ac.uk

[1] Hillary Rodham Clinton School of Law, University of Swansea, Swansea SA2 8PP, UK

[2] School of Psychology and Clinical Language Sciences, University of Reading, Whiteknights Road, Reading RG6 6AL, UK

[3] Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford OX1 3JS, UK

[1] Other principles are beneficence, non-maleficence, autonomy, and explicability (see Floridi and Cowls, 2019).

[2] Value alignment in AI refers to the process of ensuring that the values and ethical principles of the AI system is aligned with those of society, to ensure the outcomes of the system align with widely accepted ethical and moral principles and do not create negative outcomes for the individuals utilizing the systems, or society more broadly.

Published online: 20 May 2023

has gained increasing attention, it is primarily approached from a normative and technical standpoint (Gabriel 2020; Stray et al. 2021; Sutrop 2020), and has largely neglected the need to foster alignment among the values and ethical principles of the developers, stakeholders, and agents involved in the creation of the AI system with shared values, such as transparency and fairness. In this paper, we highlight the importance of developing strategies to motivate individuals and organizations to instil these values and ethical principles into their AI systems from the outset (Mittelstadt 2019; Morley et al. 2020; Weinstein et al. 2013; Whittlestone et al. 2019) and argue that this is a crucial step to creating long-lasting behavioral change that results in new bias mitigation technologies being used, not just made available, across the healthcare sector. Identifying weaknesses of AI systems and agreeing on principles these systems should reflect and abide by is not sufficient to ensure the technology used in healthcare domains (and beyond) is trustworthy or ethical (Mittelstadt 2019). Addressing these challenges necessitates extensive human investment in solving the problem of bias in technology. Critically questioning the effectiveness of current and future technologies to understand their limitations owing to bias and taking layered measures to change the way that technology is designed and implemented requires sustained human effort and the willingness to risk significant reputational, project management, and economic costs. In addition, we argue that the effort required to truly understand and correct for bias in AI will be afforded only when driven by people's ethics and values—their deeply-held worldviews regarding what is the important and right thing to do.

To identify means of addressing these challenges, we conducted a review of the literature on topics including the identification and mitigation of biases in AI systems, behavior change interventions, value systems theory, self-determination theory and AI and ethics. We brought together multiple perspectives from specialists in the fields of AI law and ethics, data ethics and philosophy, and behavioral and motivational psychology. We then synthesized these perspectives in the context of existing literature and frameworks from these interrelated fields. This enabled us to develop a thorough and nuanced understanding of the challenges and opportunities associated with translating ethical principles of fairness and transparency into practice and leverage this understanding to present possible solutions to these challenges.

This paper is structured as follows. We begin by examining some of the most common biases that permeate AI systems, such as data-driven bias, and discuss some of the solutions that have been implemented to address them, including data curation processes and auditing practices. Next, we turn our attention to means of promoting fair AI through increased transparency in the design and auditing processes of AI development and deployment—arguing that this requires a deeper understanding and leveraging of the values that drive decisions in AI development. Subsequently, we delve into the field of motivational psychology—which focuses on understanding how social environments influence behavior and developing strategies for promoting behavioral change (e.g., Deci and Ryan 2008)—to promote trustworthy AI technology. More specifically, we examine the potential of self-determination theory, autonomy-supportive strategies, value salience with identity, and the approach-avoidance framework to inspire behavior change and bridge the gap between values and behavior regarding trustworthy AI. Finally, we place these strategies in context by considering the enablers and barriers to behavior change, such as the role of social norms and the economic advantages of maintaining the status quo, as well as the financial and regulatory incentives of developing effective AI governance solutions to promote fairness in AI technology.

By bringing together insights from the fields of psychology, AI, philosophy, and computer science, we hope to provide a holistic understanding of the challenges and potential solutions for promoting trustworthy AI in healthcare and beyond.

2 Bias in healthcare technology

Benefits of healthcare technology are wide-ranging and far-reaching—including increased speed and accuracy in imaging, diagnostics for patients suffering from complex diseases, more accurate predictive screening and prognosis, and overall increased efficiency (Davenport and Kalakota 2019; Esmaeilzadeh 2020; Yoon and Lee 2021). For example, IBM developed the *Watson Care Manager System* to improve cost efficiency, design individually tailored healthcare plans, and aid in the effective use of managerial resources (Sun and Medaglia 2019). Elsewhere in clinical care, the *Ultranomics system* uses AI to analyze echocardiography scans that help detect patterns in heartbeats and diagnose coronary heart disease (Dinakaran and Anitha 2018). AI technology has also been adopted within radiology, with the development of systems able to automatically detect and quantify, as well assess, the growth and classification of abnormalities in CT scans. These systems promise greater accuracy and efficiency in prediction and prognosis compared to human radiologists using the naked eye (Jalal et al. 2021).

Despite these benefits, similarly to other domains in which AI has been in use, bias and discrimination have been flagged as key risks facing AI applications in healthcare.[3]

[3] Excellent overviews of biases in healthcare AI have been provided in recent years (Cho 2021; Gerke et al. 2020; Lysaght et al. 2019; Norori et al. 2021; Sargent 2021).

Groups which have historically faced healthcare inequalities, including minority racial and ethnic groups, women, lower-income patients, and other underrepresented or disadvantaged groups, face similar challenges with medical AI tools (Chen et al. 2022; Wachter et al. 2021b). Within this domain, biases can result from long-standing societal prejudices and inequalities amplified by problematic datasets and AI models that learn from them (Parikh et al. 2019a, b; Wachter et al. 2021a, b). There are a wide range of instances in which algorithms make spurious associations between protected factors, such as race, and disease outcome, when the underlying causal factor stems from social determinants of health (e.g., lack of access to care, delayed screening, insurance type) and not the class of a protected factor itself (Chen et al. 2022; Obermeyer et al. 2019; Vyas et al. 2020). The medical field has numerous examples where racial, gender or age disparities affect clinical decision-making, quality of treatment, and diagnosis (Norori et al. 2021; Webster et al. 2022). For example, it is recognized that Black patients have lower survival rates than white patients for different cancer categories, as well as cardiovascular disease (Lam et al. 2018; Norori et al. 2021; Tajeu et al. 2020). Nevertheless, AI systems used to predict the risk of cardiovascular disease are trained on datasets with a majority of (white) male cases, thereby displaying less accuracy for female patient groups or people of color (Antoniades and Oikonomou 2021; Norori et al. 2021).

Another example of data-driven bias in healthcare involves polygenic risk scores that use data from genome-wide association (GWAS) studies to calculate a person's inherited proneness to disease (Norori et al. 2021). Although a polygenic risk score has excellent potential as a predictive biomarker, 81% of GWAS studies have been conducted in individuals of European ancestry (Popejoy and Fullerton 2016), affecting the score's generalizability across different populations, and ultimately resulting in biased predictions that perpetuate inequalities in health outcomes. Relatedly, Larrazabal et al. (2020) recently provided empirical evidence supported by a large-scale study of a consistent decrease in algorithm performance for underrepresented genders when diagnosing various thoracic diseases under different gender imbalance conditions. When the software was reprogrammed with more gender-balanced data, it performed better by detecting thoracic diseases more accurately on X-rays, proving that more diverse datasets can improve the software's clinical performance in a diverse patient population.

2.1 Technical approaches to mitigating biases

Effective governance of AI systems, which ensures the delivery of equitable healthcare and mitigates the presence of bias, requires extensive effort on the part of those involved in the development and deployment of AI systems. Part of these efforts should be focused on tackling unfairness and discrimination which stems from training algorithms on skewed datasets, or datasets reflecting socio-economic and historical inequalities in our world.

One approach to mitigate these effects is to find ways to train models using datasets representing larger and more diverse patient populations (Food and Drugs Administration 2019). Training datasets should ideally be *diverse* along three dimensions: (i) *individual*, considering different biological factors, such as age, sex, and race; (ii) *population*, reflecting diverse disease prevalence, access to healthcare, and cultural factors; and (iii) *technical*, containing data originating from different types of medical machinery, using various acquisition or reconstruction parameters (Barclay 2021). A dataset should not only be diverse but also *balanced*, meaning it has an even distribution of a set of relevant features across the dataset. Despite the need for fair representation, clinicians often rely on publicly available datasets which are not representative of different sub-groups (Norori et al. 2021). This is because historically, vulnerable groups have been omitted from, or misrepresented in, medical datasets –meaning that AI systems trained on historical data have limited predictive ability when it comes to these groups (Parikh et al. 2019a, b). Obtaining raw medical data that is diverse and balanced is a major challenge for various reasons, including data protection regulations which tend to be particularly restrictive for health-related data (for discussion see Murdoch 2021).

Data curation and enrichment is another area which needs extensive human investment to ensure it does not contribute to training biased algorithms. AI typically needs large amounts of data to train and prepare for real-world applications. Before use, training datasets requires human curation and enrichment processes including filtering, cleaning, and labelling (Gerke et al. 2020). These processes can be very time-consuming but are essential to ensure the dataset accurately reflects a "ground truth" and is thus a valid foundation to train the model. In the healthcare domain, this can mean validating the output as a true positive or negative when the algorithm is in the learning phase (e.g., accurately labelling a tumor on a given X-ray). This labelling phase is carried out by experts within the domain and remains subjective and open to interpretation. For example, in radiology, where experts label scans used to train AI algorithms, individual differences arise in labelling practices. Experienced pathologists will often disagree about histopathology and diagnosis, particularly early-stage lesions. This level of human bias, which introduces subjectivity to the dataset's "ground truth," can lead to biased models and outcomes (Willemink et al. 2020).

2.2 Auditing practices to mitigating biases

AI technology is prone to bias and yet is trusted by medical practitioners to assist their practice (Juravle et al. 2020). Recognizing this problem, AI auditing frameworks have been developed to advocate for transparency in recounting an AI system's development and deployment. This includes reporting all collected artefacts, datasets used, test results, risk mitigation plans, and final decisions made, as well as any changes made to the AI system following an audit (Liu et al. 2022; Oala et al. 2021). Recognizing bias and the need for greater awareness, *toolkits* based on these frameworks have been put forth to evaluate the fairness or bias of machine learning models and automated decision-making/prediction systems. These toolkits monitor bias and unfairness issues throughout a model's lifecycle (from early production to implementation), thus facilitating the appraisal of an algorithm's performance as well as its failings (Morley et al. 2021). Ultimately, they are designed to help developers, policymakers, and laypeople obtain a greater understanding of the limitations and functionality of AI systems and move toward the development of fairer algorithms that do not perpetuate discriminatory behavior (Stevens et al. 2018, 2020; Zehlike et al. 2017).

Recently, auditing frameworks tailored to medical AI systems have also been advanced. These aim to guide the auditor (i.e., a developer or user, such as a medical practitioner) through a systematic process of considering potential algorithmic errors—defined by Liu et al. (2022) as any outputs of the AI system which are inaccurate, including those inconsistent with the expected performance of the system—in the context of a given clinical task, mapping the factors that might contribute to the occurrence of errors, and anticipating their potential consequences (Liu et al. 2022; Oala et al. 2021; Panigutti et al. 2021). These frameworks additionally outline approaches for testing algorithmic errors via, for example, sub-group testing or exploratory error analyses (Liu et al. 2022). The aim is to take into consideration a dynamic set of technical, clinical, and regulatory considerations found in law and outlined in regulatory bodies when evaluating medical AI systems. This approach highlights the need to go beyond providing solely quantitative performance measures to tackle issues of bias, and also offer qualitative ones (Oala et al. 2021). This comprehensive practice can help auditors and users identify ways to mitigate the impact of weaknesses in their technology at various levels, such as modifying the artificial intelligence model, modifying the model threshold, or modifying the instructions for its use (Liu et al. 2022; Oala et al. 2021).

2.2.1 Transparent communications in auditing practices

Transparency is a key value that, if activated in developers of the technology, can drive increased engagement in creating fair systems. So far, there are mixed findings on the effect of reporting measures of uncertainty, or limitations in performance, directly to users of AI systems. The majority of this research, however, has focused on the influence of this information on healthcare practitioners' levels of *trust*—finding in some cases that information on a prediction's uncertainty fosters trust in users, and in others reporting it hinders the development of trust (Cai et al. 2019a; b; Glass et al. 2008; Ha et al. 2020; Kim and Song 2022; Papenmeier et al. 2019; Robinette et al. 2017; Vorm 2018).

However, the possible benefits of transparently communicating auditing results were recently empirically demonstrated in a study carried out by Raji and Buolamwini (2019). The authors selected target companies from the original Gender Shades Study[4] (Buolamwini and Gebru 2018), as well as companies not targeted within the initial audit, and conducted a follow-up audit on these companies to test whether there had been improvement in model performance across identified subgroups which fared poorly in the original study. The findings of this follow-up audit revealed universal improvement across intersectional subgroups in the systems utilized by all targeted companies. Although post-audit performance for the minority class (darker-skinned females) was still the worst relative to other classes, the gap between this subgroup and the best performing subgroup (lighter-skinned males) reduced significantly after companies released updates following the initial audit (Raji and Buolamwini 2019). Additionally, the authors found that minimizing subgroup performance disparities did not jeopardize overall model performance but rather improved it, highlighting the alignment of fairness objectives to the commercial incentive of improved performance and accuracy (Raji and Buolamwini 2019).

While further research is needed on this matter, knowledge of risks and a system's performance is arguably an important prerequisite to achieving the goals of transparency, increasing autonomy and control for users of AI systems, and facilitating accountability practices. Research

[4] The 'Gender Shades Study' is a renowned public audit conducted by Buolamwini and Gebru (2018) which evaluated bias present in automated facial analysis algorithms and datasets concerning phenotypic subgroups. When reviewing three commercial gender classification systems, (Buolamwini and Gebru 2018) found that the datasets used by these systems were overwhelmingly composed of lighter-skinned subjects. More problematically, even when the system used new facial analysis balanced datasets, the authors found that the systems had significantly higher misclassification rates for darker-skinned faces (darker-skinned females in particular).

should also focus on identifying the factors that drive organizations to address algorithmic bias proactively and engage in auditing practices in the first place. This will facilitate the development of frameworks to improve engagement and awareness, improve the efficacy of algorithmic auditing practices, and formalize procedures for the transparent communication of a system's development, performance, and evaluation.

3 How to promote investment in fair and transparent technology

Advocates of technical solutions (including toolkits) to obtain fair and trustworthy AI technology, have increasingly recognized that only through disclosure and transparency in the design and auditing process will AI solutions improve over time. Yet, there is a notable gap in our knowledge base. Whereas plenty of attention has been placed on identifying where biases lie within AI systems, and on developing technical solutions to mitigate their presence, comparatively less attention has been placed on evaluating the effectiveness of these existing approaches and developing strategies to incentivize individuals and organizations to instill fairness into their AI systems and development processes from the outset (Burr et al. 2020; Crawford 2016; Jobin et al. 2019).

Companies developing and using AI have put forth a variety of information campaigns. These primarily communicate the presence of bias in AI. They also describe where issues lie, and which aspects of the technology development and implementation process are most vulnerable to bias. Some benefits of this approach can be recognized in the positive effect of public audits, partly attributed to increased corporate and public awareness of the problematic discriminatory consequences of the algorithms which incited the companies to speedily release product updates (Raji and Buolamwini 2019). In addition, literature in computer science has previously cited this notion of promoting fairness through user awareness and education (e.g., Hamilton et al. 2014). The approach of recognizing the dangers of AI systems and communicating these to relevant stakeholders undoubtedly has merit. However, research demonstrating that raising awareness leads to behavior change in this domain remains scarce.

Although important, knowledge of an issue alone is often not enough to lead to behavior change (Arlinghaus and Johnston 2018). This is a perception that is not new in the field of psychology, with much research showing that motivating behavior change requires more than providing knowledge and awareness of the behavior that needs to be changed (Arlinghaus and Johnston 2018; Corace and Garber 2014; Feldman and Sills 2013; Hussein et al. 2021). Rather, prior research has shown that to ensure sustained behavior change, one needs to engage with the target subjects in ways that elicit their own goals, interests, and plans so that they can develop their value for, and interest in, the positive behavior (Connell et al. 2019; Kullgren et al. 2016; Patrick and Williams 2012; Teixeira et al. 2011).

As a result of the wave of research identifying issues relating to bias within AI systems and identifying the legal and ethical principles which AI systems should follow, there is now widespread agreement around the basic principles that ethical and fair AI should meet in healthcare domains and beyond. These include beneficence, non-maleficence, autonomy fairness, explainability, accountability, and understandability (see Jobin et al. 2019; Royakkers et al. 2018). To promote behavior change in line with creating AI systems that abide by these principles and are representative of human values, we argue for a need to understand, and subsequently leverage, the values that underlie decisions made by individuals involved in crafting and deploying these systems.

Developers play a vital role in the pipeline of an AI system, as they are involved in data pre-processing, parameter and model selection, and are responsible for making model architecture changes to fit a particular use. Communicating to developers about the different biases that creep into the AI pipeline is useful to raise their awareness of these issues and enable them to address fairness limitations. However, the effectiveness of these communications in translating to behavioral change will depend on their content and on contextual and individual factors relating to, for example, values and beliefs. Understanding these factors and leveraging them when issuing communications about bias in AI and developing de-biasing strategies could be an essential means to effectively translate increased awareness of the issue into behavior change. This entails viewing investing in fair AI as a value-expressive behavior, building on previous work that sees certain behaviors as giving expression to particular values (Bardi and Schwartz 2003).

3.1 Behavior change and values

Values are stable characteristics that describe what people hold most *important* and guide attitudes and behaviors, the latter of which are responsive to contexts and changing situations (Feather 1990; Maio and Olson 1994; Rokeach 1973). Even among large commercial organizations, individuals' values drive decision-making that ultimately produces value-congruent or value-incongruent products. In the context of healthcare technology, developers are the ones to instantiate values in AI systems via, for example, dataset curation and enrichment (Gerdes 2022; Sanderson et al. 2022). Relatedly, the values of an organization influence how discrimination is defined within the organization and what solutions are developed to tackle its presence. Therefore, understanding the values and norms of developers and members of AI organizations is key to promoting fairness by design—the practice

of designing and deploying AI solutions that reflect and promote the values and ethical principles shared by society from the outset. This approach moves away from adopting reactive approaches to address issues of fairness and bias after an AI system is completed, and closer to a proactive method oriented toward infusing the systems with these shared values from the outset.

Given that technologies inevitably embody values (Nissenbaum 2001), pre-emptive attention to values during the design stage can ensure that the process and the final product (e.g., AI system) reflect these values in the long run (Felzmann et al. 2020). This is a key principle behind ethical AI design, including approaches like safe-by-design (SBD) (Baum 2016), value-sensitive design (VSD) (Friedman et al. 2013; Umbrello 2019; Umbrello and van de Poel 2021; van den Hoven et al. 2015a, b), and value alignment solutions (Gabriel and Ghazavi 2021a, b; Yudkowsky 2011)—methodologies that hold the premise that technologies are value-laden and that human values are implemented during and after their design (Friedman et al. 2013).

Social and moral psychology theories such as value systems theory, also stress the importance of focusing on the values and principles of those involved in creating and deploying AI systems. By linking ethical principles to personal and organizational values through values theory, organizations can inspire their employees to act in accordance with these principles, leading to a culture of ethics and accountability that supports the consistent implementation of transparency and fairness in AI systems. Literature in organizational psychology has evidenced that the values and motivations of individuals are crucial to reducing bias and discrimination within organizations—and that corporate efforts at reducing discrimination are ineffective without motivated individuals (Dobbin and Kalev 2018; Sullivan et al. 2001). Similarly, individuals are more likely to adopt changes in policy or practice if they perceive an organizational shift in line with their values (e.g., see Angehrn 2005). These notions remain essential when considering ways of eradicating discrimination from AI systems and creating and rolling out solutions that promote fairer and less biased AI systems.

A practical dimension of values from psychology contrasts between self-transcendent values, those that reflect care and concern for others and the world outside of oneself, and self-enhancing values, those that represent a focus on oneself, for example, in terms of increasing one's success or sense of power (Schwartz 1992, 2012). Taxonomies of specific values distinguished in terms of where they lie on this model are highly consistent across many cultures (Schwartz 1992). The distinction between the two categories of values—self-transcendent values that comprise universalism (caring for others equally) and self-enhancing—helps to predict whether value-congruent behavior will be undertaken (Nordlund and Garvill 2002; Schoenefeld and McCauley 2016).

Most who have been queried in cross-cultural samples endorse the self-transcendent values—universalism (caring for others equally) and benevolence (value of enhancing the welfare of the community; Schwartz 2012)—needed to invest in trustworthy and fair technology (Bardi and Schwartz 2003). In principle, it should be easy to get people to do the 'right' thing regarding trustworthy technology. So, why are there still such prominent and unacknowledged biases in technology? Studies show a discrepancy between people's stated values and their behaviors, and small to moderate relations have been reported linking values and corresponding behaviors (Bardi and Schwartz 2003; Cieciuch 2017; Schwartz et al. 2017; Schwartz and Butenko 2014). This value-behavior gap is an elusive foe. Despite efforts to identify sources for eliminating it taken over many years, it has historically escaped researchers' actions in several value-driven behavior domains, such as in environmental behaviors (Kollmuss and Agyeman 2002; Linder et al. 2022), prosocial behavior (Dovidio et al. 2017) and prejudice reduction (Paluck et al. 2021). The difficulty arises in part because values exist in the abstract. In contrast, specific situations have concrete needs and challenges that may prevent them from being consciously associated with their corresponding values or that may create barriers to change. Put differently, whereas an ideal situation would be to follow one's values consistently, one may fail to do so because they fail to recognize how their values could best be expressed in a specific situation or because they feel blocked in expressing their values because of the current demands of the situation (Maio 2010).

Further, to change one's existing behavior patterns to align these with values, one must change habits (Legate et al. 2021; Legate and Weinstein 2021, 2022). The ways that people apply their values to their day-to-day work and life become habitual through repetition. Certain decisions and priorities are familiar through their recurrence in the work environment; certain ways of talking about and evaluating priorities become normative. Those behaviors that may not be well-aligned to one's values have been reinforced in conditions where contradictory outcomes are appreciated at the institutional level (Legate and Weinstein 2022). Therefore, these habits must be re-examined, and new information integrated to change the output. To do so, we must consider the wider professional environment in which people operate, and the barriers it might raise to engaging in value-consistent behavior.

It is now recognized that strategies for changing behavior must be sensitive to the context of the behavior (for example, recognizing existing workplace cultures or practical challenges of undertaking behavior; Hutchison 2019; Miner and Costa 2018; Pless and Maak 2004), consider employing

multiple strategies at once to produce change (Coleman and Pasternak 2012), and ultimately test sustained behavioral change (Volpp and Loewenstein 2020). A recent study by Winecoff and Watkins (2022) emphasized that the discordant values of start-up entrepreneurs (and developers) and their funders often hindered the product's (AI system) ability to reflect their values. Findings emphasized that while entrepreneurs aimed to preserve their values—especially those connected to scientific integrity and organizational autonomy within their organization—the demands of technology entrepreneurship often ran counter to these values by focusing predominantly on rapid innovation and financial gain. In this case, the need to operate within the fast-moving technology industry constrained how entrepreneurs and developers fully transformed their ethical values into substantive practices. Whittlestone et al. (2019) recently argued for the need to focus research attention not only on identifying principles and values that AI should follow, but also on identifying the tensions that inevitably arise when implementing these principles in practice—such as tensions between the goals of an AI system and the risks it introduces to other values, or the tensions between the interests and values of an organization, a user, and a developer. Understanding and leveraging the values of developers and the environment in which the AI system is being developed is an important step to developing effective solutions to the problem of biased AI.

4 What motivational psychology can tell us about behavior change

Tackling these various social challenges, in motivational psychology, strategies have been devised that reduce the gap between values and behavior. Motivational approaches try to package or frame information in ways that increase the effectiveness of information being delivered to encourage sustained engagement. These can be applied to inspiring behavior change in line with fairness and transparency values in the service of trustworthy AI technology.

One approach to framing information to motivate change comes from self-determination theory (SDT; Deci and Ryan 1985). SDT posits that value-consistent behavior, such as, in our case, decisions to promote fairness or transparent communications, is undertaken when they are energized by different and distinct motives that reflect reasons which are autonomous and self-driven (i.e., internalized and influenced by one's interests and values) or controlled and other-driven (i.e., external and influenced by expectations and pressures from the environment; Ryan and Deci 2017). Inspired by SDT theory, there has been significant interest in identifying autonomy-supportive strategies for conveying information. Much of this information has been in the context of healthcare (Altendorf et al. 2019; Legate et al. 2021; Moon and Woo 2021; Moon et al. 2021) sports, and education (Reeve 2016; Vansteenkiste et al. 2004a, b; Vansteenkiste et al. 2004a, b). These survey and experimental studies have shown that promoting internalized, autonomous motivation through autonomy-supportive strategies—providing a rationale to convey the importance of the behavior being motivated, creating a sense that one has a choice about how to behave, and avoiding using pressure and shame—are more effective at sustained behavioral change (see Legate and Weinstein 2021, 2022). This is because they give those being motivated the psychological space to consider their own (rather than the required) commitment to their values and whether their behaviors align with those values, while simultaneously inspiring personally meaningful reasons to build their commitment in the first place.

One of the ways of creating autonomous motivation for action is by pairing value salience with identity. Indeed, previous research shows that values are linked to behavior insofar as they are central to identity (Verplanken and Holland 2002). In the context of environmental behaviors, identity is seen as a core driver for environmentally friendly 'green behaviors', especially in cases where those behaviors come at a cost or challenge (Jaspal et al. 2014). In healthcare contexts, researchers have also proposed that small shifts in identity open the door to seeing the possibility of change, reinforcing this change, and subsequently supporting even more behavior change (Kearney and O'Sullivan 2003). Much like is the case for values, individuals who engage in certain behaviors can distance their behaviors from their identity. For example, someone can hold implicitly prejudiced beliefs but not self-identify as racist (Archakis et al. 2018). In healthcare, similar observations have been made regarding addiction—addictive behavior can be compartmentalized from people's identities as addicts (e.g., Choi et al. 2010). From an SDT perspective, having an identity consistent with one's values helps the behavior to be fully autonomous, in a form called integrated motivation. Behaviors motivated by integration reflect the core aspects of the self—they are aligned with the internal drivers that energize behaviors (Deci and Ryan 2008).

Other theories in the fields of social psychology support this point. For example, cognitive-dissonance theory suggests that people will try to reduce discomfort caused by holding conflicting beliefs (Festinger 1957). In the context of building AI systems, developers may have to reconcile their personal values with the ethical considerations of designing a given system, to minimise cognitive dissonance. Relatedly, the approach-avoidance framework outlines that individuals are motivated to approach stimuli or situations that are associated with positive outcomes and to avoid stimuli or situations that are associated with negative outcomes (Elliot and Thrash 2002). These areas of research suggest that an effective motivational strategy could involve highlighting the

alignment of ethical principles of fairness and transparency with positive outcomes, such as increased trust in the AI system, enhanced reputation of the organization, and improved stakeholder satisfaction.

Recent literature evaluating the challenges of implementing ethical principles in practice (e.g., see Goirand et al. 2021 for a discussion relating to the healthcare industry and Mökander and Sheth, (2023) to the biopharmaceutical industry) has identified a range of difficulties, including a lack of clear governance structures, limited understanding of AI, ethical dilemmas, difficulty in aligning governance principles with business goals, balancing competing interests, legal and regulatory challenges, lack of standardization and best practices, and limited engagement with external stakeholders. Strategies based on motivational psychology theories, such as self-determination theory and value systems theory, might offer potential solutions to some of these challenges. Self-determination theory can enhance intrinsic motivation by aligning personal and organizational values, promoting a collaborative culture, and providing employees with a sense of control and autonomy. This can lead to improved ethical decision-making and more effective implementation of AI governance principles. Relatedly, other strategies stemming from motivational psychology principles, such as involving stakeholders in the design and implementation process, providing transparency and control over data use, and fostering a sense of community and collaboration—can help align AI-based applications with ethical principles such as transparency, privacy, and fairness.

5 Barriers to behaviour change

It is useful to draw on these principles based on social psychology to develop effective AI governance solutions and motivate individuals to invest in them. However, these principles do not exist in a vacuum. To translate them into practice, we must consider the norms and needs that govern our context of interest, and which might act as barriers to change. This will enable us to operationalize value-activation interventions aimed at incentivizing the adoption of tools that promote fairness in AI technology.

5.1 Norms

In the values literature, one of the notable drivers of inaction is that even when people hold self-transcendent values that should drive their positive behaviors, they perceive others to lack those same values, contributing to a perception of helplessness to produce change (Sanderson and Dawe 2019). Thus, social norms and perceived social norms are crucial elements to consider when developing behavior change interventions, especially if the goal is to achieve sustained behavior change on a large scale. Though more than one definition exists, broadly, social norms can be thought of as standards for behavior that people utilize to guide their behavior and evaluate that of others (Smith 2020). Individuals often have norms that define the kind of behavior they should or should not engage in. Similarly, organizations have norms that will dictate how members of the organization should behave and how the organization operates. Even more broadly, societies have shared beliefs about prescribed or banned behaviors.

These norms can be highly influential in guiding behavior at each level, via various channels. The *informational* influence of norms occurs when people conform to a behavior, such as wearing a mask, because actions of those in their social groups are indicative of these actions being the correct behavior. For example, they are convinced it is the right thing to do to reduce transmission risk of an infectious disease. *Normative* influence occurs when people conform to be accepted by their social group. For example, they wear a face mask to not stand out negatively (Neville et al. 2021). When discussing social norms, it is worth noting that these have both injunctive elements (defining what should be done) and descriptive elements (defining what is done). *Injunctive* norms refer to beliefs about what behavior is approved or disapproved of, and motivate behavior adoption using social rewards or punishments associated with the behavior. *Descriptive* norms describe the status-quo behavior and boost consensus to adopt it primarily by providing information about what behavior is likely to be effective and adaptive in a specific context (Neville et al. 2021).

Research on social norms and behavior change has shown that messages reinforcing positive norms can effectively change behavior, especially when evoking a shared identity (Reynolds et al. 2015). Descriptive normative communications centered around alerting users and organizations about the behavior of other similar individuals and organizations seem particularly effective in doing so in a range of contexts spanning from adopting pro-environmental behaviors (De Groot et al. 2013; Göckeritz et al. 2010; Goldstein et al. 2008; Nolan et al. 2008) to tax compliance behavior (Behavioural Insights Team 2012). In contrast to coercive approaches, manipulating norms in communications aims at producing bottom-up changes that will ultimately be adopted at a societal level (Bicchieri 2016). This approach seems valuable when considering solutions to promote fair AI and motivate behavioral change within the AI industry in favor of fairness by design.

5.2 Economic advantages

The task of re-examining longstanding habits and changing one's practice is challenging because normative decisions, even those resulting in biased technology, sometimes serve

worthwhile functions—such as offering economic advantages. For example, Google's combination of search and advertisement activities (e.g., advertising networks), has in the past decade come under scrutiny for resulting in conflicts of interest and incentives to bias search result rankings. Rieder and Sire (2014), argue that the company's role as a search provider and a content provider strongly motivates it to organize search results in a biased and self-serving way. This, combined with its recognized dominance on the market, compounded in Google being repeatedly fined for abusing its dominance as a search engine and (i) imposing restrictive clauses in contracts with third-party websites which prevent rivals from placing their search adverts on these websites (European Commission 2019), and (ii) giving an unfair advantage to their comparison-shopping services (European Commission 2017).

Another example of when *not* addressing bias in AI systems can be an economic advantage relates to adverts for Science, Technology, Engineering and Mathematics (STEM) jobs. A research study found that women see fewer advertisements about entering STEM professions than men (Lambrecht and Tucker 2019).This is not due to companies purposefully targeting men in a disproportionate fashion, but rather because of economic ad sales. For advertisers, it is more expensive to get female views than male views on digital ads. This results in the creation of ad algorithms which, designed to save costs, end up targeting the cheaper male viewers (Lambrecht and Tucker 2019; Maron 2018). In this example, the primary issue seems to arise from having specific goals and objectives which are at odds with ensuring fair outcomes. The auction-based ad allocation system attempts to save advertisers money in their ad targeting; this might be a meaningful goal, but it places no value on fairness and gender equality.

Examples of this issue can also be found within the academic research domain. Big data obtained from social platforms such as Twitter have become increasingly popular for studying human behavior. This is not because it is a representative dataset (often this is not the case), but because it is the cheapest, fastest, and easiest to access for researchers. Researchers affected by time and funding constraints make use of Twitter as it is readily available and voluminous. In this case, questions about fairness and representativeness might take a backseat to other needs such as publication and funding.

5.3 Financial costs of fairness tools

The financial costs of implementing fairness metrics in an algorithm or system are another critical inhibitor to consider when devising effective behavior change strategies. AI fairness metrics typically increase the prediction performance for the sensitive or protected group (Hardt et al. 2016; Zietlow et al. 2022)—this can mean lowering it for the non-protected group, which often represents the more extensive user base of a system (von Zahn et al. 2021). Hence, one could expect that implementing fairness metrics results in some financial costs. This has been recently empirically recorded in e-commerce (von Zahn et al. 2021). AI development still occurs largely within the commercial sector, which is guided by financial principles. Therefore, it is unsurprising that, in part, biased algorithms and biased datasets (e.g., unbalanced) are not a company's primary concern unless perhaps they lead to financial deficits. Imagine that a company's new vision AI system fails to recognize the imagery of certain cultural practices. If the communities that engage the most in these cultural practices are not its primary customer base, the company has little incentive to address those biases as doing so would not result in any economic benefits.

Similarly, implementing solutions such as auditing toolkits can incur both financial and administrative costs including investing time and resources into preparing for audits, as well as the costs of implementing any changes to the AI system, or its use, that come to light following the audit (Mökander and Floridi 2022). Recently, the European Commission estimated that certification for an AI system that abides by the EU AI Act could cost on average EUR 16,800–23,000, equating to approximately 10–14% of the development cost (Moritz et al. 2021; Norori et al. 2021). Others have argued this to be a conservative estimate (Haataja and Bryson 2021; Mueller 2021).

6 Enablers of behavioral change

Nonetheless, there are also clear incentives to design and implement fair AI systems. These include improving various metrics such as data security, talent acquisition, reputational management, process optimization, economic advantage, and regulatory preparedness (Holweg et al. 2022; Mökander and Floridi 2022; The Economist Intelligence Unit 2020). Here, we consider two possible incentives that can be leveraged as enablers of behavioral change in line with designing AI systems imbued with transparency and fairness values: economic incentives and regulatory incentives.

6.1 Economic(s) incentives

Adopting fair AI solutions can be enriching because it aligns with positive societal, organizational, and personal norms and values, but it can also be financially advantageous. The previously mentioned costs of implementing fairness metrics and auditing practices are smaller than the costs of addressing system limitations later in the development process. Implementing solutions at the design phase, compared to

the testing phase or deployment stage, can ultimately cut costs by a significant amount (Dawson et al. 2010).

Considering another example, banks earn money by approving loans to reasonable credit risks, not by turning away performing loans. Biased algorithms that disadvantage a group of candidates and lock them out of the financial system, ultimately cap the revenue that banks could be accruing. This principle led the credit bureau Experian to create 'Boost' (Henry and Morris 2018)—a program that allows users with limited credit history to increase their scores with information about managing money. This allowed more than half of the users initially labelled as having a "poor" credit rating (simply due to lack of information), to be re-labelled as having a "fair" credit rating—giving them access to previously denied loans and allowing lenders to extend more credit to profitable customers.

DataRobot (2019)—in collaboration with the World Economic Forum—surveyed more than 350 U.S. and U.K.-based technology leaders to understand how organizations identify and mitigate bias in AI. Survey respondents included CIOs, IT directors, IT managers, data scientists and development leads involved in the design and deployment of AI systems. Crucially, 36% of respondents declared that their organizations suffered from AI bias. Damage was reported to be in the form of lost revenue by 62% of companies, lost customers in 61% of cases, lost employees due to AI bias in 43% of cases, and incurred legal fees due to a lawsuit or legal action relating to bias in 35% of cases (DataRobot 2019).

In addition to being "the right thing to do," minimizing bias can therefore also be portrayed as an economic imperative—making organizations more profitable and productive (Wachter 2021) and enabling them to appear more alluring and trustworthy to customers/users who prefer to use products of companies that reflect their values (Zeno Group 2020).

6.2 Laws and regulations

Another incentive to implement effective AI governance solutions is to improve business metrics such as regulatory preparedness. Laws and regulations can be powerful drivers of behavioral change. The increased discussion of the biased nature of AI systems has mobilized organizations and governments around the world to regulate how these technologies are developed and how they are utilized in decision-making contexts. Many existing and new regulatory frameworks create obligations to look for bias in AI due to their scope or by explicitly addressing the technology or concept. In the U.S, local, state, and federal regulations have already been enforced and members of Congress are introducing bills like the Algorithmic Accountability Act (2019) and the Algorithmic Fairness Act (2020), to promote ethical AI decision-making. While the federal laws do not explicitly target AI, the Federal Trade Commission (FTC) has, for example, issued a regulation noting that biased AI violates the Fair Credit Reporting Act (FCRA) given that the Equal Credit Opportunity Act (ECOA) renders it illegal for a company to utilize biased algorithms resulting in credit discrimination based on sensitive attributes such as national origin, race, or sex (Federal Trade Commission 2020, 2021).

In Europe, similar regulatory efforts can be observed. A recent regulation is the draft Artificial Intelligence Act or AI Act (2021). This is Europe's first attempt to regulate AI comprehensively. Whilst this is still a draft, a clear regulatory trajectory can already be seen. The regulation classifies the different AI applications according to their risk, ranging from (i) low or minimal risk, (ii) limited risk, (iii) high risk, and (iv) unacceptable risk. The bulk of the regulation primarily aims at self-assessment procedures for operators of high-risk AI systems. Providers of high-risk systems (e.g., employment, critical infrastructure, education) need to undergo an ex-ante conformity assessment to test compliance with the regulation and the harmonized standards (Ebers et al. 2021). The draft requires that operators examine possible bias and ensure that the training, validation, and testing datasets are relevant, representative, free of errors, and complete (Art 10 (2) f and (3)), keep records (Art 12) and guarantee human oversight (Art 14). At this stage, it is still too early to describe the exact procedures and obligations the regulation will require of AI providers, but there will likely be several legal duties to test for and mitigate biases in high-risk systems.

There are, however, already frameworks in place that impact AI systems. In 2018 the General Data Protection Regulation (GDPR) came into force. Whilst this framework is primarily aimed at protecting privacy and regulating the use of data, some provisions have an impact on automated decision-making and AI. Art 22 of the GDPR stipulates that fully automated decisions that do have legal or similarly significant effects on data subjects warrant special legal safeguards. If such a decision is rendered, for example, a loan or employment decision, the data subject has a right to demand human intervention, express their point of view and contest the decision. In addition, Articles 13–15 GDPR grant individuals the right to obtain "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" (Wachter et al. 2017, p.5). While these provisions do not directly address questions of bias, the heightened transparency requirements aim to increase individuals' understanding of how and why automated decisions are made.

Moreover, the European Non-Discrimination Directives, even though not designed to regulate AI, are highly likely to have an increased impact on automated decision-making and the deployment of AI. These directives prohibit direct and indirect discrimination. Direct discrimination prohibits

the use of protected attributes such as gender, ethnicity, ability, or sexual orientation to make decisions in a protected sector, such as, for example, employment, or when offering goods and services (e.g., healthcare). Indirect discrimination is also prohibited, which refers to the use of an "apparently neutral provision, criterion, or practice [that] would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons". If such practices lead to unequal outcomes (e.g., unequal access to health services), the disparity would need to be legally justified (Hacker 2018; Zuiderveen Borgesius 2020).

The US has a similar system in place (Barocas and Selbst 2016). One of the crucial differences between these two systems is that the European system does not necessitate intent in direct discrimination cases. In other words, in the EU a person can be liable under law for direct (or indirect discrimination), even if they did not know about the discriminating nature of the system. This provision aims to encourage providers to investigate bias and discrimination as widely as possible, both before and after the deployment of an AI system, as ignorance about biased performance or outcomes will not free them from liability (Wachter et al. 2020).

6.2.1 Psychology and legal compliance

Similarly to the aforementioned notions that making people aware of bias in technology is not enough to ensure people act to reduce this bias and creating technical solutions to mitigate bias is not enough to ensure people adopt these solutions—drawing up laws and regulations to regulate AI (although an important feat) does not in itself necessarily guarantee that people will comply with them, and it does not mean they will serve their intended purpose. Research within psychology and cognitive science has shown that compliance with laws is not only a cost–benefit calculation but is shaped by various factors such as social norms, social identity, imitation of others, and ethics (see Baier 2016; Bradford et al. 2015; Licht 2008; Nadler 2017)). Laws that reflect the affected community's values and norms, enhance their ability to gain compliance (Acemoglu and Jackson 2017). As such, the processes mentioned so far, relating to values and norms, also play a role in moderating compliance with laws and legislation. This is a bi-directional relationship, with values and attitudes towards a certain behavior influencing how laws regulating that behavior are perceived—and laws being able to alter people's attitudes towards the regulated behavior.

Research has found that laws can influence people's behavior through various means (see Bilz and Nadler 2014 for an overview), such as by leveraging their motivation to maintain the esteem of others. The most promising domains for changing behavior seem to be those in which people can be prompted to take the path of least resistance (Shenhav et al. 2017). This is in line with theories of diffusion of technology and innovation, which state (and find empirical support towards) that 'ease of use' and 'workflow integration' are crucial factors that predict attitudes towards technological change (Vale Martins et al. 2021; Zhou et al. 2019). Overall, what psychological research on legal compliance suggests is that even though conventional law and economics can explain differences in people's willingness to engage in regulated behavior up to a certain point, people's attitudes and beliefs about the behavior can also be altered (Bilz and Nadler 2014; Cialdini 2007; Roy 2021). If legal regulation can transmute the social meaning of behavior, and can change people's perceptions regarding the desirability of this behavior, perhaps deep-rooted and widespread behavior change can be achieved. These notions should be considered when devising regulations and solutions to mitigate bias in AI, to maximize their effectiveness and maximise their adoption.

7 Conclusions and suggestions

There is no shortage of research identifying biases in healthcare AI products and developing solutions—typically supported by ethical principles—to address these. However, there is a shortage of research investigating how to best motivate the widespread adoption of these solutions and promote fairness and transparency values amongst key developers and stakeholders of the technology. In this paper, we outlined how we can draw on theories stemming from social and motivational psychology to increase engagement with de-biasing and fairness-enhancing practices within the AI (healthcare) industry to promote sustained behavioral change. In doing so, we framed investment in fair AI as a value-expressive behavior, building on previous work that sees certain behaviors give expression to certain personally held values.

Ensuring that AI systems are designed and used legally, ethically, and safely requires organizations to not only have the right values and tools in place but also be able to effectively communicate these to individuals involved in the development and distribution of the systems. In this paper, we outlined how we can draw from motivational, and more broadly social, psychology when considering how to design effective communications. Researchers should empirically investigate the use of autonomy-supportive communication strategies as means of facilitating long-term behavior change (going beyond top-down directives) within the AI healthcare industry, in line with transparency and fairness values. Motivating individuals to care about fairness by design can ultimately act as a catalyst for internal and societal change. However, for these motivational strategies to be maximally effective, we argue they should also consider

and incorporate the forces that are working for, and against, the desired behavior change. As such, they need to consider the existing norms and habits of a given environment and any known inhibitors and incentives of behavior change. Values are not activated in a vacuum but within an organizational, and broader societal, ecosystem. Research on optimizing value-activation within the healthcare AI industry to increase uptake of fairness by design practices should also inform research on best practices for creating legal mandates and regulations, and maximizing compliance with these.

Ultimately, developing real-world translational research requires extensive investment, and without care and time, researchers can draw hasty conclusions about real-world effects that, over broader periods, do not facilitate actual change. This type of research requires a process spanning multiple stems. The field must begin from robust basic research that is internally valid and reproducible and develop this work into field experiments and studies that extend scientific conclusions to real-world settings. The process requires commitment from academics and practitioners collaborating to draw theory-rich but applicable tests of research questions and hypotheses. It also requires academics from different disciplines to work together to produce broader research that considers context alongside individual differences.

Acknowledgements We would like to thank Lizzie Barclay for her insightful discussions which are at the core of this paper. We would additionally like to thank our colleagues on the Governance of Emerging Technologies programme; Johann Laux, Chris Russell, Prashan Madumal and Rory Gillis, for their feedback and support.

Funding This work has been supported through research funding provided by the Wellcome Trust (Grant nr 223765/Z/21/Z), Sloan Foundation (Grant nr G-2021-16779), the Department of Health and Social Care (via the AI Lab at NHSx), a British Academy Postdoctoral Fellowships (Grant nr PF2\180114), Luminate Group, and the Miami Foundation.

Data availability Data sharing is not applicable to this article as no datasets were generated or analysed for the current article.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

References

Acemoglu D, Jackson MO (2017) Social norms and the enforcement of laws. J Eur Econ Assoc 15(2):245–295. https://doi.org/10.1093/jeea/jvw006

Altendorf MB, van Weert JCM, Hoving C, Smit ES (2019) Should or could? Testing the use of autonomy-supportive language and the provision of choice in online computer-tailored alcohol reduction communication. Digit Health 5:2055207619832767. https://doi.org/10.1177/2055207619832767

Angehrn AA (2005) Learning to manage innovation and change through organizational and people dynamics simulations. In: Proceedings of the international simulation & gaming association conference (ISAGA 05)

Antoniades C, Oikonomou EK (2021) Artificial intelligence in cardiovascular imaging—principles, expectations, and limitations. Eur Heart J. https://doi.org/10.1093/eurheartj/ehab678

Archakis A, Lampropoulou S, Tsakona V (2018) "I'm not racist but I expect linguistic assimilation": the concealing power of humor in an anti-racist campaign. Discourse Context Media 23:53–61. https://doi.org/10.1016/j.dcm.2017.03.005

Arlinghaus KR, Johnston CA (2018) Advocating for behavior change with education. Am J Lifestyle Med 12(2):113–116. https://doi.org/10.1177/1559827617745479

Ayling J, Chapman A (2021) Putting AI ethics to work: are the tools fit for purpose? AI Ethics. https://doi.org/10.1007/s43681-021-00084-x

Baier M (2016) Social and legal norms: towards a socio-legal understanding of normativity. Routledge

Barclay L (2021) Bias in medical imaging AI: checkpoints and mitigation. Aidence. https://www.aidence.com/articles/bias-in-medical-imaging-ai/

Bardi A, Schwartz SH (2003) Values and behavior: strength and structure of relations. Pers Soc Psychol Bull 29(10):1207–1220. https://doi.org/10.1177/0146167203254602

Barocas S, Selbst AD (2016) Big data's disparate impact essay. Calif Law Rev 104(3):671–732

Behavioural Insights Team (2012) Applying behavioural insights to reduce fraud, error and debt. 38.

Bicchieri C (2016) Norms in the wild: how to diagnose, measure, and change social norms. Oxford University Press, Oxford

Bilz K, Nadler J (2014) Law, moral attitudes, and behavioral change. In: Zamir E, Teichman D (eds) The Oxford handbook of behavioral economics and the law, pp 240–267. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199945474.013.0010

Bradford B, Hohl K, Jackson J, MacQueen S (2015) Obeying the rules of the road: procedural justice, social identity, and normative compliance. J Contemp Crim Justice 31(2):171–191. https://doi.org/10.1177/1043986214568833

Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st conference on fairness, accountability and transparency, pp 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Burr C, Taddeo M, Floridi L (2020) The ethics of digital well-being: a thematic review. Sci Eng Ethics 26(4):2313–2343. https://doi.org/10.1007/s11948-020-00175-8

Cai CJ, Jongejan J, Holbrook J (2019a) The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th international conference on intelligent user interfaces, pp 258–262. https://doi.org/10.1145/3301275.3302289

Cai CJ, Winter S, Steiner D, Wilcox L, Terry M (2019b) 'Hello AI': uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. In: Proceedings of the ACM on human-computer interaction, 3(CSCW), pp 1–24. https://doi.org/10.1145/3359206

Chen RJ, Chen TY, Lipkova J, Wang JJ, Williamson DFK, Lu MY, Sahai S, Mahmood F (2022) Algorithm Fairness in AI for Medicine and Healthcare (arXiv:2110.00603). arXiv. http://arxiv.org/abs/2110.00603

Cho MK (2021) Rising to the challenge of bias in health care AI. Nat Med 27(12):12. https://doi.org/10.1038/s41591-021-01577-2

Choi Y, Choi SM, Rifon N (2010) "I Smoke but I Am Not a Smoker": phantom smokers and the discrepancy between self-identity and behavior. J Am Coll Health 59(2):117–125. https://doi.org/10.1080/07448481.2010.483704

Cialdini R (2007) Descriptive social norms as underappreciated sources of social control. Psychometrika 72:263–268. https://doi.org/10.1007/s11336-006-1560-6

Cieciuch J (2017) Exploring the complicated relationship between values and behaviour. In: Roccas S, Sagiv L (eds) Cieciuch, Jan (2017). Exploring the complicated relationship between values and behaviour. In: Roccas, Sonia; Sagiv, Lilach. Values and behavior. Springer, Cham, pp 237–247. https://doi.org/10.1007/978-3-319-56352-7_11

Coleman MT, Pasternak RH (2012) Effective strategies for behavior change. Prim Care Clin off Pract 39(2):281–305. https://doi.org/10.1016/j.pop.2012.03.004

European Commission (2017) Antitrust: Google fined €1.49 billion for online advertising abuse [Text]. European Commission—European Commission. https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1770

European Commission (2019) Antitrust: Commission fines Google €2.42 billion [Text]. https://ec.europa.eu/commission/presscorner/detail/en/IP_17_1784

Connell LE, Carey RN, de Bruin M, Rothman AJ, Johnston M, Kelly MP, Michie S (2019) Links between behavior change techniques and mechanisms of action: an expert consensus study. Ann Behav Med 53(8):708–720. https://doi.org/10.1093/abm/kay082

Corace K, Garber G (2014) When knowledge is not enough: changing behavior to change vaccination results. Hum Vaccin Immunother 10(9):2623–2624. https://doi.org/10.4161/21645515.2014.970076

Crawford K (2016) Can an algorithm be agonistic? Ten scenes from life in calculated publics. Sci Technol Hum Values 41(1):77–92. https://doi.org/10.1177/0162243915589635

DataRobot (2019) The state of AI bias in 2019. DataRobot AI Cloud. https://www.datarobot.com/lp/the-state-of-ai-bias-in-2019/

Davenport T, Kalakota R (2019) The potential for artificial intelligence in healthcare. Future Healthc J 6(2):94–98. https://doi.org/10.7861/futurehosp.6-2-94

Dawson M, Burrell DN, Rahim E, Brewster S (2010) Integrating software assurance into the software development life cycle (SDLC). J Inf Syst Technol Plan 3(6):7

De Groot JIM, Abrahamse W, Jones K (2013) Persuasive normative messages: the influence of injunctive and personal norms on using free plastic bags. Sustainability 5(5):5. https://doi.org/10.3390/su5051829

Deci EL, Ryan RM (1985) The general causality orientations scale: self-determination in personality. J Res Pers 19(2):109–134. https://doi.org/10.1016/0092-6566(85)90023-6

Deci EL, Ryan RM (2008) Facilitating optimal motivation and psychological well-being across life's domains. Can Psychol Psychol Can 49(1):14–23. https://doi.org/10.1037/0708-5591.49.1.14

Dinakaran S, Anitha P (2018) A review and study on AI in health care issues. Int J Sci Res Comput Sci Eng Inf Technol. https://doi.org/10.32628/CSEIT183886

Dobbin F, Kalev A (2018) Why doesn't diversity training work? The challenge for industry and academia. Anthropol Now 10(2):48–55. https://doi.org/10.1080/19428200.2018.1493182

Dovidio J, Piliavin J, Schroeder D, Penner L (2017) The social psychology of prosocial behavior. Psychol Press. https://doi.org/10.4324/9781315085241

Ebers M, Hoch VRS, Rosenkranz F, Ruschemeier H, Steinrötter B (2021) The European Commission's proposal for an artificial intelligence act—a critical assessment by members of the robotics and AI law society (RAILS). J 4(4):Article 4. https://doi.org/10.3390/j4040043

Elliot AJ, Thrash TM (2002) Approach-avoidance motivation in personality: approach and avoidance temperaments and goals. J Pers Soc Psychol 82:804–818. https://doi.org/10.1037/0022-3514.82.5.804

Esmaeilzadeh P (2020) Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Making 20(1):NA–NA

Feather NT (1990) Reactions to equal reward allocations: effects of situation, gender and values. Br J Soc Psychol 29(4):315–329. https://doi.org/10.1111/j.2044-8309.1990.tb00913.x

Federal Trade Commission (2020) Using Artificial Intelligence and Algorithms. Federal Trade Commission. http://www.ftc.gov/business-guidance/blog/2020/04/using-artificial-intelligence-and-algorithms

Federal Trade Commission (2021) Aiming for truth, fairness, and equity in your company's use of AI. Federal Trade Commission. http://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai

Feldman DB, Sills JR (2013) Hope and cardiovascular health-promoting behaviour: education alone is not enough. Psychol Health 28(7):727–745. https://doi.org/10.1080/08870446.2012.754025

Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A (2020) Towards transparency by design for artificial intelligence. Sci Eng Ethics 26(6):3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Festinger L (1957) A theory of cognitive dissonance. Row, Peterson

Food and Drugs Administration (2019) Proposed regulatory framework for modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD). https://apo.org.au/node/228371

Friedman Jr BPHK, Borning A (2013) Value sensitive design and information systems. 34

Gabriel I (2020) Artificial intelligence, values, and alignment. Mind Mach 30(3):411–437. https://doi.org/10.1007/s11023-020-09539-2

Gabriel I, Ghazavi V (2021a) The challenge of value alignment: from fairer algorithms to AI safety. ArXiv:2101.06060 *[Cs]*. http://arxiv.org/abs/2101.06060

Gabriel I, Ghazavi V (2021b) The challenge of value alignment: from fairer algorithms to AI safety (arXiv:2101.06060). arXiv. https://doi.org/10.48550/arXiv.2101.06060

Gerdes A (2022) A participatory data-centric approach to AI ethics by design. Appl Artif Intell 36(1):2009222. https://doi.org/10.1080/08839514.2021.2009222

Gerke S, Minssen T, Cohen G (2020) Ethical and legal challenges of artificial intelligence-driven healthcare. Artif Intell Healthc. https://doi.org/10.1016/B978-0-12-818438-7.00012-5

Glass A, McGuinness DL, Wolverton M (2008) Toward establishing trust in adaptive agents. In: Proceedings of the 13th

international conference on intelligent user interfaces—IUI '08, 227. https://doi.org/10.1145/1378773.1378804
Göckeritz S, Schultz PW, Rendón T, Cialdini RB, Goldstein NJ, Griskevicius V (2010) Descriptive normative beliefs and conservation behavior: the moderating roles of personal involvement and injunctive normative beliefs. Eur J Soc Psychol 40(3):514–523. https://doi.org/10.1002/ejsp.643
Goirand M, Austin E, Clay-Williams R (2021) Implementing ethics in healthcare AI-based applications: a scoping review. Sci Eng Ethics 27(5):61. https://doi.org/10.1007/s11948-021-00336-3
Goldstein NJ, Cialdini RB, Griskevicius V (2008) A room with a viewpoint: using social norms to motivate environmental conservation in hotels. J Consum Res 35(3):472–482. https://doi.org/10.1086/586910
Ha T, Kim S, Seo D, Lee S (2020) Effects of explanation types and perceived risk on trust in autonomous vehicles. Transport Res Part F Traffic Psychol Behav 73:271–280. https://doi.org/10.1016/j.trf.2020.06.021
Haataja M, Bryson JJ (2021) What costs should we expect from the EU's AI Act? SocArXiv. https://doi.org/10.31235/osf.io/8nzb4
Hacker P (2018) Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. Common Market Law Rev 55(Issue 4):1143–1185. https://doi.org/10.54648/COLA2018095
Haleem A, Javaid M, Khan IH (2019) Current status and applications of Artificial Intelligence (AI) in medical field: an overview. Curr Med Res Pract 9(6):231–237. https://doi.org/10.1016/j.cmrp.2019.11.005
Hamilton K, Karahalios K, Sandvig C, Eslami M (2014) A path to understanding the effects of algorithm awareness. In: CHI '14 extended abstracts on human factors in computing systems, pp 631–642. https://doi.org/10.1145/2559206.2578883
Hardt M, Price E, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Adv Neural Inf Process Syst 29. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html
Henry N, Morris J (2018) Scaling up affordable lending: inclusive credit scoring
Holweg, Younger R, Wen Y (2022) The reputational risks of AI. California Management Review. https://cmr.berkeley.edu/2022/01/the-reputational-risks-of-ai/
van den Hoven J, Vermaas PE, van de Poel I (2015a) Design for values: an introduction. In: van den Hoven J, Vermaas PE, van de Poel I (eds) Handbook of ethics, values, and technological design: sources, theory, values and application domains, pp 1–7. Springer Netherlands. https://doi.org/10.1007/978-94-007-6970-0_40
van den Hoven M, Vermaas P, van de Poel I (2015b) Design for values: an introduction. In: van den Hoven J, Vermaas P, van de Poel I (eds) Handbook of ethics, values, and technological design: sources, theory, values and application domains, pp 1–7. Springer Science+Business Media. https://doi.org/10.1007/978-94-007-6970-0_1
Hussein R, Whaley CRJ, Lin ECJ, Grindrod K (2021) Identifying barriers, facilitators and behaviour change techniques to the adoption of the full scope of pharmacy practice among pharmacy professionals: using the theoretical domains framework. Res Social Adm Pharm 17(8):1396–1406. https://doi.org/10.1016/j.sapharm.2020.10.003
Hutchison ED (2019) Dimensions of human behavior: person and environment, 6th edn. SAGE
Jalal S, Parker W, Ferguson D, Nicolaou S (2021) Exploring the role of artificial intelligence in an emergency and trauma radiology department. Can Assoc Radiol J 72(1):167–174. https://doi.org/10.1177/0846537120918338
Jaspal R, Nerlich B, Cinnirella M (2014) Human responses to climate change: social representation, identity and socio-psychological action. Environ Commun 8(1):110–130. https://doi.org/10.1080/17524032.2013.846270
Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):9. https://doi.org/10.1038/s42256-019-0088-2
Juravle G, Boudouraki A, Terziyska M, Rezlescu C (2020) Trust in artificial intelligence for medical diagnoses. In: Progress in brain research, vol 253, pp 263–282. Elsevier. https://doi.org/10.1016/bs.pbr.2020.06.006
Kearney MH, O'Sullivan J (2003) Identity shifts as turning points in health behavior change. West J Nurs Res 25(2):134–152. https://doi.org/10.1177/0193945902250032
Kim T, Song H (2022) Communicating the limitations of AI: the effect of message framing and ownership on trust in artificial intelligence. Int J Hum Comput Interact. https://doi.org/10.1080/10447318.2022.2049134
Kollmuss A, Agyeman J (2002) Mind the Gap: why do people act environmentally and what are the barriers to pro-environmental behavior? Environ Educ Res 8(3):239–260. https://doi.org/10.1080/13504620220145401
Kullgren JT, Williams GC, Resnicow K, An LC, Rothberg A, Volpp KG, Heisler M (2016) The promise of tailoring incentives for healthy behaviors. Int J Workplace Health Manag 9(1):2–16. https://doi.org/10.1108/IJWHM-12-2014-0060
Lam C, Cronin K, Ballard R, Mariotto A (2018) Differences in cancer survival among white and black cancer patients by presence of diabetes mellitus: estimations based on SEER-Medicare-linked data resource. Cancer Med 7(7):3434–3444. https://doi.org/10.1002/cam4.1554
Lambrecht A, Tucker C (2019) Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manage Sci 65(7):2966–2981. https://doi.org/10.1287/mnsc.2018.3093
Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E (2020) Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci 117(23):12592–12594. https://doi.org/10.1073/pnas.1919012117
Legate N, Weinstein N (2022) How to motivate people to care about prejudice reduction in the workplace. In:Handbook of self-determination theory. Oxford University Press, Oxford
Legate N, Nguyen TT, Weinstein N, Moller A, Legault L, Adamkovic M, Adetula GA, Agesin BB, Ahlgren L, Akkas H, Almeida I, Anjum G, Antoniadi M, Arinze AI, Arvanitis A, Rana K, Badalyan V, Becker M, Bernardo O (2021) A global experiment on motivating social distancing during the COVID-19 pandemi. https://doi.org/10.31234/osf.io/n3dyf
Legate N, Weinstein N (2021) Can we communicate autonomy support and a mandate? How motivating messages relate to motivation for staying at home across time during the COVID-19 pandemic. Health Commun. https://doi.org/10.1080/10410236.2021.1921907
Licht AN (2008) Social norms and the law: why peoples obey the law. Rev Law Econ 4(3):715–750. https://doi.org/10.2202/1555-5879.1232
Linder N, Giusti M, Samuelsson K, Barthel S (2022) Pro-environmental habits: an underexplored research agenda in sustainability science. Ambio 51(3):546–556. https://doi.org/10.1007/s13280-021-01619-6
Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L (2022) The medical algorithmic audit. Lancet Digit Health 4(5):e384–e397. https://doi.org/10.1016/S2589-7500(22)00003-6
Lysaght T, Lim HY, Xafis V, Ngiam KY (2019) AI-assisted decision-making in healthcare. Asian Bioethics Rev 16:299–314

Maio GR, Olson JM (1994) Value—attitude-behaviour relations: the moderating role of attitude functions. Br J Soc Psychol 33(3):301–312. https://doi.org/10.1111/j.2044-8309.1994.tb01027.x

Maio GR (2010) Chapter 1—Mental representations of social values. In: Advances in experimental social psychology, vol 42, pp 1–43. Academic Press. https://doi.org/10.1016/S0065-2601(10)42001-8

Maron DF (2018) Science career ads are disproportionately seen by men. Scientific American. https://www.scientificamerican.com/article/science-career-ads-are-disproportionately-seen-by-men/

Miner K, Costa P (2018) Ambient workplace heterosexism: implications for sexual minority and heterosexual employees. Stress Health. https://doi.org/10.1002/smi.2817

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1(11):11. https://doi.org/10.1038/s42256-019-0114-4

Mökander J, Sheth M (2023) Challenges and best practices in corporate AI governance: lessons from the biopharmaceutical industry

Mökander J, Floridi L (2022) Operationalising AI governance through ethics-based auditing: an industry case study. AI Ethics. https://doi.org/10.1007/s43681-022-00171-7

Moon H, Woo K (2021) An integrative review on mothers' experiences of online breastfeeding peer support: motivations, attributes and effects. Maternal Child Nutr 17(3):e13200. https://doi.org/10.1111/mcn.13200

Moon K, Riege A, Gourdon-Kanhukamwe A, Vallée-Tourangeau G (2021) The moderating effect of autonomy on promotional health messages encouraging healthcare professionals' to get the influenza vaccine. J Exp Psychol Appl 27(2):187. https://doi.org/10.1037/xap0000348

Moritz L, Renda A, Yeung T (2021) Clarifying the costs for the EU's AI Act. *CEPS*. https://www.ceps.eu/clarifying-the-costs-for-the-eus-ai-act/

Morley J, Floridi L, Kinsey L, Elhalal A (2020) From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 26(4):2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L (2021) Ethics as a service: a pragmatic operationalisation of AI ethics. Mind Mach 31(2):239–256. https://doi.org/10.1007/s11023-021-09563-w

Mueller B (2021) Artificial Intelligence Act. 16

Murdoch B (2021) Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 22(1):122. https://doi.org/10.1186/s12910-021-00687-3

Nadler J (2017) Expressive law, social norms, and social groups. Law Soc Inq 42(1):60–75. https://doi.org/10.1111/lsi.12279

Neville FG, Templeton A, Smith JR, Louis WR (2021) Social norms, social identities and the COVID-19 pandemic: theory and recommendations. Soc Personal Psychol Compass 15(5):e12596. https://doi.org/10.1111/spc3.12596

Nissenbaum H (2001) How computer systems embody values. Computer 34(3):120–119. https://doi.org/10.1109/2.910905

Nolan JM, Schultz PW, Cialdini RB, Goldstein NJ, Griskevicius V (2008) Normative social influence is underdetected. Pers Soc Psychol Bull 34(7):913–923. https://doi.org/10.1177/0146167208316691

Nordlund AM, Garvill J (2002) Value structures behind proenvironmental behavior. Environ Behav 34(6):740–756. https://doi.org/10.1177/001391602237244

Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A (2021) Addressing bias in big data and AI for health care: a call for open science. Patterns 2(10):100347. https://doi.org/10.1016/j.patter.2021.100347

Oala L, Murchison AG, Balachandran P, Choudhary S, Fehr J, Leite AW, Goldschmidt PG, Johner C, Schörverth EDM, Nakasi R, Meyer M, Cabitza F, Baird P, Prabhu C, Weicken E, Liu X, Wenzel M, Vogler S, Akogo D, Wiegand T (2021) Machine learning for health: algorithm auditing & quality control. J Med Syst 45(12):105. https://doi.org/10.1007/s10916-021-01783-y

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453. https://doi.org/10.1126/science.aax2342

Paluck E, Porat R, Clark C, Donald G (2021) Prejudice reduction: progress and challenges. Annu Rev Psychol. https://doi.org/10.1146/annurev-psych-071620-030619

Panigutti C, Perotti A, Panisson A, Bajardi P, Pedreschi D (2021) FairLens: auditing black-box clinical decision support systems. Inf Process Manag 58(5):102657. https://doi.org/10.1016/j.ipm.2021.102657

Papenmeier A, Englebienne G, Seifert C (2019) How model accuracy and explanation fidelity influence user trust (arXiv:1907.12652). arXiv. http://arxiv.org/abs/1907.12652

Parikh RB, Gdowski A, Patt DA, Hertler A, Mermel C, Bekelman JE (2019a) Using big data and predictive analytics to determine patient risk in oncology. Am Soc Clin Oncol Educ Book 39:e53–e58. https://doi.org/10.1200/EDBK_238891

Parikh RB, Teeple S, Navathe AS (2019b) Addressing bias in artificial intelligence in health care. JAMA 322(24):2377. https://doi.org/10.1001/jama.2019.18058

Patrick H, Williams GC (2012) Self-determination theory: its application to health behavior and complementarity with motivational interviewing. Int J Behav Nutr Phys Act 9(1):18. https://doi.org/10.1186/1479-5868-9-18

Pless N, Maak T (2004) Building an inclusive diversity culture: principles, processes and practice, vol 54. University of St. Gallen. https://doi.org/10.1007/s10551-004-9465-8

Popejoy AB, Fullerton SM (2016) Genomics is failing on diversity. Nature 538(7624):161–164. https://doi.org/10.1038/538161a

Raji ID, Buolamwini J (2019) Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 429–435. https://doi.org/10.1145/3306618.3314244

Reeve J (2016) Autonomy-supportive teaching: what it is, how to do it. In: Liu WC, Wang JCK, Ryan RM (eds) Building autonomous learners: perspectives from research and practice using self-determination theory, pp 129–152. Springer. https://doi.org/10.1007/978-981-287-630-0_7

Reynolds KJ, Subašić E, Tindall K (2015) The problem of behaviour change: from social norms to an ingroup focus: norms and behaviour change. Soc Pers Psychol Compass 9(1):45–56. https://doi.org/10.1111/spc3.12155

Rieder B, Sire G (2014) Conflicts of interest and incentives to bias: a microeconomic critique of Google's tangled position on the Web. New Media Soc 16(2):195–211. https://doi.org/10.1177/1461444813481195

Robinette P, Howard AM, Wagner AR (2017) Effect of robot performance on human-robot trust in time-critical situations. IEEE Trans Hum Mach Syst 47(4):425–436. https://doi.org/10.1109/THMS.2017.2648849

Rokeach M (1973) The nature of human values, pp x, 438. Free Press

Roy S (2021) Theory of social proof and legal compliance: a sociocognitive explanation for regulatory (non) compliance. German Law J 22(2):238–255. https://doi.org/10.1017/glj.2021.5

Royakkers L, Timmer J, Kool L, van Est R (2018) Societal and ethical issues of digitization. Ethics Inf Technol 20(2):127–142. https://doi.org/10.1007/s10676-018-9452-x

Ryan RM, Deci EL (2017) Self-determination theory: basic psychological needs in motivation, development, and wellness, pp xii, 756. The Guilford Press. https://doi.org/10.1521/978.14625/28806

Sanderson K, Dawe J (2019) Perspectives: getting to the heart of workforce wellbeing in health and social care: from personal practice to organisational change. J Res Nurs JRN 24(8):729–733. https://doi.org/10.1177/1744987119890922

Sanderson C, Douglas D, Lu Q, Schleiger E, Whittle J, Lacey J, Newnham G, Hajkowicz S, Robinson C, Hansen D (2022) AI ethics principles in practice: perspectives of designers and developers (arXiv:2112.07467). arXiv. http://arxiv.org/abs/2112.07467

Sargent SL (2021) AI bias in healthcare: using ImpactPro as a case study for healthcare practitioners' duties to engage in anti-bias measures. Can J Bioethics 4(1):112–116. https://doi.org/10.7202/1077639ar

Schoenefeld JJ, McCauley MR (2016) Local is not always better: the impact of climate information on values, behavior and policy support. J Environ Stud Sci 6(4):724–732. https://doi.org/10.1007/s13412-015-0288-y

Schwartz SH, Butenko T (2014) Values and behavior: validating the refined value theory in Russia. Eur J Soc Psychol 44(7):799–813. https://doi.org/10.1002/ejsp.2053

Schwartz SH, Cieciuch J, Vecchione M, Torres C, Dirilen-Gumus O, Butenko T (2017) Value tradeoffs propel and inhibit behavior: validating the 19 refined values in four countries. Eur J Soc Psychol 47(3):241–258. https://doi.org/10.1002/ejsp.2228

Schwartz SH (1992) Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. Adv Exp Soc Psychol

Schwartz SH (2012) Toward refining the theory of basic human values. In: Methods, theories, and empirical applications in the social sciences, pp 39–46. https://doi.org/10.1007/978-3-531-18898-0_6

Shenhav A, Rand DG, Greene JD (2017) The relationship between intertemporal choice and following the path of least resistance across choices, preferences, and beliefs. Judgm Decis Mak 12(1):18

Smith JR (2020) Group norms. In: Oxford research encyclopedia of psychology. https://doi.org/10.1093/acrefore/9780190236557.001.0001/acrefore-9780190236557-e-453

Stevens C, Liu CH, Chen JA (2018) Racial/ethnic disparities in US college students' experience: discrimination as an impediment to academic performance. J Am Coll Health 66(7):665–673. https://doi.org/10.1080/07448481.2018.1452745

Stevens A, Deruyck P, Veldhoven ZV, Vanthienen J (2020) Explainability and fairness in machine learning: improve fair end-to-end lending for kiva. IEEE Symp Ser Comput Intell SSCI 2020:1241–1248. https://doi.org/10.1109/SSCI47803.2020.9308371

Stray J, Vendrov I, Nixon J, Adler S, Hadfield-Menell D (2021) What are you optimizing for? Aligning Recommender Systems with Human Values (arXiv:2107.10939). arXiv. https://doi.org/10.48550/arXiv.2107.10939

Sullivan W, Sullivan R, Buffton B (2001) Aligning individual and organisational values to support change. J Chang Manag 2:247–254. https://doi.org/10.1080/738552750

Sun TQ, Medaglia R (2019) Mapping the challenges of Artificial Intelligence in the public sector: evidence from public healthcare. Gov Inf Q 36(2):368–383. https://doi.org/10.1016/j.giq.2018.09.008

Sutrop M (2020) Challenges of aligning artificial intelligence with human values. Acta Baltica Historiae Et Philosophiae Scientiarum 8(2):54–72

Tajeu GS, Safford MM, Howard G, Howard VJ, Chen L, Long DL, Tanner RM, Muntner P (2020) Black-white differences in cardiovascular disease mortality: a prospective US study, 2003–2017. Am J Public Health 110(5):696–703. https://doi.org/10.2105/AJPH.2019.305543

Teixeira PJ, Patrick H, Mata J (2011) Why we eat what we eat: the role of autonomous motivation in eating behaviour regulation. Nutr Bull 36(1):102–107. https://doi.org/10.1111/j.1467-3010.2010.01876.x

The Economist Intelligence Unit (2020) *Staying ahead of the curve The business case for responsible AI* (p. 78). The Economist. https://pages.eiu.com/rs/753-RIQ-438/images/EIUStayingAheadOfTheCurve.pdf

Umbrello S, van de Poel I (2021) Mapping value sensitive design onto AI for social good principles. AI Ethics 1(3):283–296. https://doi.org/10.1007/s43681-021-00038-3

Umbrello S (2019) Beneficial artificial intelligence coordination by means of a value sensitive design approach. 3(5). https://doi.org/10.3390/bdcc3010005

Vakkuri V, Kemell K-K, Kultanen J, Siponen M, Abrahamsson P (2019) Ethically aligned design of autonomous systems: industry viewpoint and an empirical study (arXiv:1906.07946). arXiv. https://doi.org/10.48550/arXiv.1906.07946

do Vale Martins R, Alturas B, Alexandre I (2021) Perspective for the use of adoption theories in artificial intelligence. In: 2021 16th Iberian conference on information systems and technologies (CISTI), pp 1–4. https://doi.org/10.23919/CISTI52073.2021.9476340

Vansteenkiste M, Simons J, Lens W, Sheldon KM, Deci EL (2004a) Motivating learning, performance, and persistence: the synergistic effects of intrinsic goal contents and autonomy-supportive contexts. J Pers Soc Psychol 87(2):246–260. https://doi.org/10.1037/0022-3514.87.2.246

Vansteenkiste M, Simons J, Soenens B, Lens W (2004b) How to become a persevering exerciser? Providing a clear, future intrinsic goal in an autonomy-supportive way. J Sport Exerc Psychol 26(2):232–249

Verplanken B, Holland R (2002) Motivated decision making: effects of activation and self-centrality of values on choices and behavior. J Pers Soc Psychol 82:434–447. https://doi.org/10.1037/0022-3514.82.3.434

Volpp KG, Loewenstein G (2020) What is a habit? Diverse mechanisms that can produce sustained behavior change. Organ Behav Hum Decis Process 161:36–38. https://doi.org/10.1016/j.obhdp.2020.10.002

von Zahn M, Feuerriegel S, Kuehl N (2021) The cost of fairness in AI: evidence from E-commerce. Bus Inf Syst Eng. https://doi.org/10.1007/s12599-021-00716-w

Vorm ES (2018) Assessing demand for transparency in intelligent systems using machine learning. Innov Intell Syst Appl INISTA 2018:1–7. https://doi.org/10.1109/INISTA.2018.8466328

Vyas DA, Eisenstein LG, Jones DS (2020) Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N Engl J Med 383(9):874–882. https://doi.org/10.1056/NEJMms2004740

Wachter S (2021) How fair AI can make us richer. Eur Data Prot Law Rev EDPL 7(3):367–372

Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Privacy Law 7(2):76–99. https://doi.org/10.1093/idpl/ipx005

Wachter S, Mittelstadt B, Russell C (2020) Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. West Virginia Law Rev 123(3):735–790

Wachter S, Mittelstadt B, Russell C (2021a) Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. SSRN Electron J. https://doi.org/10.2139/ssrn.3792772

Wachter S, Mittelstadt B, Russell C (2021b) Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. Comput Law Secur Rev 41:105567. https://doi.org/10.1016/j.clsr.2021.105567

Webster CS, Taylor S, Thomas C, Weller JM (2022) Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. BJA Educ 22(4):131–137. https://doi.org/10.1016/j.bjae.2021.11.011

Weinstein N, Ryan RM, Deci EL (2013) Motivation, meaning, and wellness: a self-determination perspective on the creation and internalization of personal meanings and life goal. In: The human quest for meaning, pp 81–106. Taylor and Francis. https://doi.org/10.4324/9780203146286

Whittlestone J, Nyrup R, Alexandrova A, Cave S (2019) The role and limits of principles in AI ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 195–200. https://doi.org/10.1145/3306618.3314289

Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, Folio LR, Summers RM, Rubin DL, Lungren MP (2020) Preparing medical imaging data for machine learning. Radiology 295(1):4–15. https://doi.org/10.1148/radiol.2020192224

Winecoff AA, Watkins EA (2022) Artificial concepts of artificial intelligence: institutional compliance and resistance in AI startups. https://doi.org/10.1145/3514094.3534138

Yoon N, Lee H-K (2021) AI recommendation service acceptance: assessing the effects of perceived empathy and need for cognition. J Theor Appl Electron Commerce Res 16(5):5. https://doi.org/10.3390/jtaer16050107

Yudkowsky E (2011) Complex value systems in friendly AI. In: Schmidhuber J, Thórisson KR, Looks M (eds) Artificial general intelligence, pp 388–393. Springer. https://doi.org/10.1007/978-3-642-22887-2_48

Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R (2017) FA*IR: a fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 1569–1578. https://doi.org/10.1145/3132847.3132938

Zeno Group (2020) 2020 Zeno Strength of Purpose Study. https://drive.google.com/file/d/1ni3dl4jAEWn7d0KxD_-rB05p2ZoBJJlC/view?usp=sharing&usp=embed_facebook

Zhou E, Li D, Madden A, Chen Y, Ding Y, Kang Q, Su H (2019) Modeling adoption behavior for innovation diffusion. In: 14th International conference, iConference 2019, Washington, DC, USA, March 31–April 3, 2019, Proceedings, pp 339–349. https://doi.org/10.1007/978-3-030-15742-5_33

Zietlow D, Lohaus M, Balakrishnan G, Kleindessner M, Locatello F, Schölkopf B, Russell C (2022) Leveling down in computer vision: pareto inefficiencies in fair deep classifiers (arXiv:2203.04913). arXiv. https://doi.org/10.48550/arXiv.2203.04913

Zuiderveen Borgesius FJ (2020) Strengthening legal protection against discrimination by algorithms and artificial intelligence. Int J Hum Rights 24(10):1572–1593. https://doi.org/10.1080/13642987.2020.1743976

Floridi L, Cowls J (2019) A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review. https://doi.org/10.1162/99608f92.8cd550d1