University of Reading

Values tensions and values tradeoffs in the development of healthcare artificial intelligence technology: a conceptual model of decisions to create trustworthy technology

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Tovmasyan, A. ORCID: https://orcid.org/0000-0002-9297-0084, Weinstein, N. ORCID: https://orcid.org/0000-0003-2200-6617 and Mittelstadt, B. (2025) Values tensions and values tradeoffs in the development of healthcare artificial intelligence technology: a conceptual model of decisions to create trustworthy technology. Social Influence, 20 (1). 2478940. ISSN 1553-4529 doi: 10.1080/15534510.2025.2478940 Available at https://centaur.reading.ac.uk/122096/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1080/15534510.2025.2478940

Publisher: Taylor and Francis

All outputs in CentAUR are protected by Intellectual Property Rights law,

University of Reading

including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Social Influence

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/psif20

Values tensions and values tradeoffs in the development of healthcare artificial intelligence technology: a conceptual model of decisions to create trustworthy technology

Anna Tovmasyan, Netta Weinstein & Brent Mittelstadt

To cite this article: Anna Tovmasyan, Netta Weinstein & Brent Mittelstadt (2025) Values tensions and values tradeoffs in the development of healthcare artificial intelligence technology: a conceptual model of decisions to create trustworthy technology, Social Influence, 20:1, 2478940, DOI: 10.1080/15534510.2025.2478940

To link to this article: https://doi.org/10.1080/15534510.2025.2478940

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 17 Mar 2025.

Submit your article to this journal

Article views: 23

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=psif20

SOCIAL INFLUENCE
2025, VOL. 20, NO. 1, 2478940
https://doi.org/10.1080/15534510.2025.2478940

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Values tensions and values tradeoffs in the development of healthcare artificial intelligence technology: a conceptual model of decisions to create trustworthy technology

Anna Tovmasyan[a,b], Netta Weinstein[a,b] and Brent Mittelstadt[b]

[a]School of Psychology and Clinical Language Sciences, University of Reading, Reading, UK; [b]Oxford Internet Institute, University of Oxford, Oxford, UK

ABSTRACT

Value-action gaps may exist when values are endorsed due to external pressures, from low perceived behavioral control, or when values conflict with community norms or one another. In the context of developing healthcare artificial intelligence (AI), prosocial values like fairness and transparency are particularly important, yet they may clash with an instrumental value for profit. We recruited 185 healthcare AI developers (66.49% male, 58.92% White, Mage = 27.56) to examine facilitators and barriers of value-action congruence. Value endorsement, internalization, attitudes, norms, and message frequency were positively linked to behavioral intention to act in line with one's values, while low perceived control and value tension were negatively associated with it. Findings highlight the need for organizational cultures that emphasize prosocial values.

ARTICLE HISTORY
Received 8 January 2024
Accepted 5 March 2025

KEYWORDS
Artificial intelligence; values; values trade-off; transparency; inclusion

Research into the interplay between values, attitudes, and behavior underscores the significant role of social influence in shaping individuals' decision-making processes. For example, although holding certain values and attitudes may play a central role in guiding people's behavior (Homer & Kahle, 1988; Vermeir & Verbeke, 2008), values are often surprisingly poor predictors of real-world value expression (e.g., Barr, 2006; Flynn et al., 2009). Two psychological theories could help explain this. First, the Theory of Planned Behavior (TPB; Ajzen, 1991) suggests that an individual's intention to engage in a behavior is influenced by three main factors: *attitudes*, *perceived social norms*, and *perceived behavioral control*. These factors interact to predict behavioral intentions, which, in turn, guide actual behaviors. Separately, Self-Determination Theory (SDT; Ryan & Deci, 2000) gives reason to believe that the extent to which certain values are *internalized* will be associated with the willingness to act in line with the values (ibid.). The two approaches provide well-evidenced perspectives on how value-laden decisions might be made, but there is little work integrating them or applying them to values and value tradeoffs; indeed, little attention has been given to value tradeoffs despite their potential for explaining the value-action gap (Ajzen & Fishbein, 1975; LaPiere, 1934).

CONTACT Anna Tovmasyan a.tovmasyan@reading.ac.uk

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

The current study uses healthcare artificial intelligence (AI) as a proof of concept to address these gaps and test the extent to which attitudes, perceived social norms, and values are associated with healthcare AI developers' tradeoff decisions, specifically their intentions to transparently communicate about bias in their technology or pursue competing values.

Attitudes, Values, and Behavior

Values (e.g., fairness, transparency, success) can be understood as abstract ideals that provide important guiding principles (Rokeach, 1968; Schwartz, 1992, 1994; Schwartz et al., 1999) and remain relatively stable throughout the course of human life (Rokeach, 1973; Stern, 2000). In contrast, attitudes represent individuals' likes and dislikes toward various entities, ranging from concrete objects to abstract concepts (Hanel et al., 2021), though ultimately, attitudes relate back to values. Specifically, the value-attitude-behavior hierarchy (Homer & Kahle, 1988) posits that holding certain values is related to holding the value's corresponding attitudes (Hanel et al., 2021) and, ultimately, intending to behave (Vermeir & Verbeke, 2008) and behaving in a way consistent with that value (Homer & Kahle, 1988). Supporting these views, much research has shown that both values and attitudes can predict a range of behaviors, including pro-environmental (Dunlap et al., 1983; Gatersleben et al., 2014; Sharma & Jha, 2017; Urien & Kilbourne, 2011), prosocial (Hackett, 2014; Lönnqvist et al., 2013; Schwartz, 2010), entrepreneurial (Kirkley, 2016), romantic (G. Maio et al., 2023), risk-taking (Iversen, 2004; Sleczka et al., 2018), and health-related choices, such as substance use (Lins de Holanda Coelho et al., 2018; Morell-Gomis et al., 2018). Further, values and attitudes may drive not only the behaviors of individuals, but also those of organizations (e.g., NHS; Smith & Malcolm, 2010) and entire communities (Mydland & Grahn, 2012).

Applying those literatures and the value-attitude-behavior hierarchy (Vermeir & Verbeke, 2008) to the context of healthcare AI, attitudes shaped by values of ethical responsibility and patient well-being may lead to intentions to transparently communicate limitations in AI technology. However, research has shown that while holding certain values and attitudes may play a central role in guiding behavior (Homer & Kahle, 1988; Vermeir & Verbeke, 2008), they are often surprisingly poor predictors of real-world value expression (e.g., Barr, 2006; Flynn et al., 2009). This phenomenon is referred to in the literature as the 'value-action gap' (Ajzen & Fishbein, 1975; LaPiere, 1934). For example, people may have the value of preserving the environment yet take little to no action that gives expression to this value (Barr, 2006; Flynn et al., 2009). Sometimes this may be due to factors such as lack of access to recycling facilities or competing demands on their time and resources (Gifford, 2011), but such barriers do not provide a comprehensive account for this gap.

Another reason for the values-action gap (Ajzen & Fishbein, 1975; LaPiere, 1934) may be that in consequential real-life behaviors, multiple co-existing values act on any given behavior – as one value inspires a corresponding action, another value acts as a deterrent to the same action. Individuals may hold abstract ideals that do not fully translate into concrete action, particularly if those conflict with other important values, resulting in values tension (Hitlin & Piliavin, 2004). For example, a company leader may hold values of achieving success and protecting the environment and must decide between

maximizing profits and minimizing environmental impact. In such cases, a decision on how to act may depend on the trade-off between competing values (Le Grand, 1990).

Values when Developing Healthcare Technology

One of the contexts where value-behavior alignment is becoming increasingly important is in the development of healthcare AI. Healthcare providers rely more and more on these technologies to make diagnostic and treatment decisions, as well as allocation and patient care management (Asan et al., 2020). Yet there is evidence that such systems may perpetuate bias regarding gender, ethnicity, and income (Nelson, 2019). Further, some patients and healthcare providers report resistance to using AI algorithms because of a lack of understanding how the systems may be making decisions (Cadario et al., 2021; Richardson et al., 2021; Shinners et al., 2020). Therefore, it is crucial that their developers behave in line with the values of fairness and transparency to ensure that the technology they produce makes accurate and equal diagnostic and treatment decisions that provide high-quality care without discrimination (Ghassemi & Mohamed, 2022). Further, it is important to encourage full disclosure of technology limitations when these standards are not met (Walmsley, 2021). Specifically, the fairness value is particularly important in ensuring that technologies are developed and deployed in a way that does not discriminate against individuals or groups based on their race, gender, age, religion, or other characteristics. Following this value in practice involves costly decisions that most often require additional heavy investment (Barocas & Selbst, 2016; Wachter et al., 2020), yet fairness values provide reason to accept these costs – that is, investing resources can enable those developers who value it to design and train AI algorithms to do so in a way that does not perpetuate existing biases or inequalities.

Alongside fairness, transparency is increasingly recognized as essential in the development of healthcare AI technologies (Walmsley, 2021). The principle of transparency refers to the ability of individuals and organizations to access information about how AI algorithms are designed, trained, and deployed, as well as making customers and society aware of the possibility of bias in, and limitations of, the technology (Burrell, 2016; Mittelstadt, 2021). By promoting transparency, AI developers can build trust and accountability among their users, regulators, and other stakeholders (Veale & Binns, 2017). Transparency can also help ensure that AI systems are being used ethically and that potential biases or errors are being addressed in a timely and effective manner (Mittelstadt et al., 2016). Thus, for those developers who value transparency, there are multiple reasons to apply this self-transcendence value to the AI workplace.

Yet, the implementation of transparency can present challenges in terms of protecting intellectual property and trade secrets, as well as potentially harming sales of the product (Wachter et al., 2017). For example, being transparent about an error in the algorithm that may result in technology being less accurate in making diagnostic decisions for certain groups of people may result in having to do additional training for the algorithms, which could be expensive and time-consuming, as well as potentially discouraging consumers from buying the products. Here, the self-transcendence values of transparency and fairness are important (Walmsley, 2021), yet they may be in tension with a self-enhancement value of most organizational contexts: achieving profit (Schwartz, 2003).

Therefore, the question arises as to which value healthcare technology developers are more likely to prioritize and act in accordance with.

Overall, healthcare stands out as an ideal setting to explore the complex interplay between ethical considerations and AI design. In particular, the notion of fairness takes center stage within healthcare AI applications. The need for unbiased decision-making processes is paramount in healthcare settings, where the consequences of unfair or biased outcomes can have profound implications for individuals' health and well-being. By focusing on healthcare AI, we aim to delve into the tensions that may arise between different values and ethical principles in AI design, with a particular emphasis on fairness considerations.

It is important to note that the values of profit, fairness, and transparency may not necessarily be in tension with one another, and one may, in many cases, make decisions that are in line with all three of these values. Often, the profit is inherently linked to the other two values. For example, when a customer complains that a product does not work well on a segment of the population, this lack of fairness could translate to reputation loss and loss of revenue; in these cases, decisions can express all three values at once. Yet, in situations where one value is placed in tension with another, the incongruity may pose a significant decision dilemma to an AI developer, and thus it is important to explore the processes that may underlie decision-making in these contexts.

Motivation for Values as an Explanation for How Trade-Off Decisions are Made

Given that technology developers in the healthcare industry may simultaneously hold fairness, transparency, and profit values that inform their behaviors, how do they choose to respond when having to act on one over another value? The answer may lie in the extent to which each value is internalized or perceived as being generated from within the self and self-identity, as opposed to being externally informed. This view is in line with Self-Determination Theory, a motivational theory which posits people are more likely to act on their intentions when those intentions are internalized in a manner that is consistent with other personally held beliefs and identity, and they feel truly important, rather than beng primarily driven by feelings of guilt, shame, or external pressure (Deci & Ryan, 2008). *Value internalization* refers to the process by which individuals transform externally imposed values into personally endorsed values that become integrated into their self-identity (ibid.). The more an individual internalizes a value, the more likely they are to feel a sense of personal ownership and commitment to that value, which, in turn, increases their motivation to act in line with that value (Vansteenkiste et al., 2004; Williams et al., 2004). Given the impacts of internalized values on behavior, it is plausible that when multiple values come into tension with one another, the value that is most deeply held (i.e., most fully internalized) will exert the most influence on behavior. In this case, when deciding which value to act upon, individuals will choose to behave in accordance with the one higher in internalization. To our knowledge, no studies to date have tested this assumption by integrating the values and motivation literatures.

Perceived Social Norms as an Explanation for How Trade-Off Decisions are Made

Further, values – however well-internalized – may compete with other influences that shape behavior in a workplace environment, such as social norms, which can override an individual's values or beliefs and limit their ability to act on them (Gollwitzer & Sheeran, 2006), resulting in a value-action gap (Ajzen & Fishbein, 1975; LaPiere, 1934). Perceived social norms reflect the perceived social pressures and influences that impact behavior (Terry et al., 1999). On the one hand, values influence who individuals consider as important referents and what societal norms they prioritize (Stern et al., 1999), but on the other hand, as posited by the Theory of Planned Behavior (TPB), perceived social norms may sometimes play a more determinant role in behavior, for example, if a certain value is not internalized to a large extent (Deci & Ryan, 2008) or is in tension with another value that is more consistent with the social norms (Hitlin & Piliavin, 2004). Further, social pressures can play a more detrimental role via perceived behavioral control, where developers may believe they cannot act in line with certain values due to outside influences (Fennis & Aarts, 2012). In this study, we will examine competing explanations for predicting behavioral intentions, namely, that workplace norms (Treviño et al., 2014) and the frequency of messages (Podsakoff et al., 1996) regarding fairness, transparency, and profitability of healthcare AI systems, because both may lead individuals to prioritize conformity over their personal values (Cialdini & Goldstein, 2004).

The planned study

The value-action gap observed in previous research may be due to value tensions that exist between multiple values acting together on one action (Kennedy et al., 2009). Such tensions, and the trade-off decisions that are made to express one value over another, offer a fascinating example of where real-world decision-making is guided by complex factors acting together on it in meaningful ways. These decisions are also consequential. For example, when developers invest in building inclusive technology that represents patients from many different backgrounds (e.g., sex, race, or standing on other protected characteristics), or choose to communicate transparently about the limitations of technology, patient care is better for more of the population. However, such value-based decisions come with certain costs, such as increasing the production time of the product or reducing an organization's profit obtained from developing the technology. This study will examine how the values (self-transcendence values of fairness and transparency versus self-enhancement value of profit) of healthcare AI systems developers, their attitudes, subjective norms (measured by norms at the workplace, and frequency of messages about fairness and transparency), and perceived behavioral control correspond to their intention to communicate transparently about the limitations and to increase the accuracy and fairness of their technology. By integrating the TPB, SDT, and values literature, this study utilizes a bold approach, aiming to contribute to the refinement and expansion of these frameworks concerned with social influence, with healthcare AI being a proof-of-concept for these broader principles. Further, the study has implications

Figure 1. Summary of the hypotheses.

beyond the area of healthcare AI, contributing to understanding how decision-making processes are made when competing influences are in place.

We tested four *a-priori* directional hypotheses to build conceptual understanding of how value trade-offs are made, and specifically, the role of motivation in making them (see Figure 1):

Hypothesis 1: Based on values literature, we hypothesized that developers would intend to engage in behavior that is in line with their values. We anticipated that if the data do not support this hypothesis, such findings would lead us to conclude there is no evidence that the values tested link with behavioral intentions in the context of healthcare AI developers as hypothesized. This would indicate that other factors, not accounted for in our model, may play a more substantive role in guiding behavioral intentions. Such a result would highlight the need for a more nuanced understanding of how specific contexts or additional variables (e.g., immediate incentives, organizational culture) affect value–behavior relationships.

Hypothesis 2: Based on values and SDT literatures, we hypothesized that the more a value is internalized, the more likely healthcare AI technology developers would be to

intend to engage in behavior in line with this value. We anticipated that a non-significant finding here would suggest that internalization may not be as critical a mediator in the value–behavior relationship as posited by SDT. This could imply that, in organizational contexts, even deeply internalized values may not lead to corresponding behavioral intentions.

Hypothesis 3: Based on TPB literature, and accounting for values and SDT literatures, we hypothesized that when controlling for values internalization, attitudes, workplace norms, frequency of messages, and perceived behavioral control will all predict behavioral intention. We anticipated that non-significant results for this hypothesis would challenge the comprehensive applicability of the TPB in organizational context. It might suggest that the combined effects of these factors are less predictive of behavioral intentions among healthcare AI developers than anticipated.

Hypothesis 4: Based on values and SDT literature, we hypothesized that when healthcare AI technology developers endorse multiple values that are in tension (fairness and transparency versus profit), value internalization would interact with value tension, so that the value that is more internalized drives behavior. If this hypothesis yields non-significant findings, it would imply that value internalization does not necessarily resolve tensions between conflicting values in the expected manner. This outcome would suggest that other mechanisms, such as external influences or situational constraints, might play a more significant role in resolving such value conflicts. It would suggest the need for a reexamination of how value tensions are navigated, and the potential need to incorporate a broader range of psychological or situational factors into our understanding of these dynamics.

Method

Ethical Approval

The research complied with APA and BPS ethical regulations. Ethical approval for this study was attained by the Psychology Department's Ethics Committee at the University of Reading (22–064-NW) before undertaking study procedures. Written informed consent was obtained from all participants of the study before starting, and they were debriefed on study goals and reminded of data management processes at the end of the study. Participants were able to withdraw from the study at any point.

Open Research Practices

In line with best practices, materials, analysis code, and data can be found on Open Science Framework (https://osf.io/3fyqx/).

Methodological Approach

The concept of values is regarded as a complex one that poses a challenge in experimental manipulation both because of its nuance and because it largely reflects deep-seated and long-lasting tendencies to care for some outcomes over others (Maio, 2016). Consequently, the majority of studies investigating the relationship between values and behavior have utilized correlational designs to examine the relationship between the two variables (e.g., Hansen, 2008; Mackenbach, 2014; Souchon et al., 2017). In this study, we adopted a similar methodological approach to examine the relationship between values and behavior in a workplace setting, using a sample of AI software developers. We directly compared competing interpretations of behavioral intentions to determine the best explanations of the behavioral intentions that developers would demonstrate.

Participants

We recruited healthcare AI developers for this study, i.e., professionals who specialize in creating AI technologies for applications within the healthcare industry. We specifically aimed to recruit those whose primary focus is on designing AI systems and software that can analyze, interpret, and make decisions based on healthcare data. Participants were recruited through e-mail, Twitter advertisements, and by contacting relevant organizations. Example organizations that were targeted by recruitment include Aidence, One HealthTech, and Open Medical. We also recruited on Prolific, a website for recruiting participants for academic studies. To recruit via Prolific, we used the following screeners: Work – 'Information Technology', Education – 'Computer Science', Industry – 'Medical/healthcare' and 'Software'. We required that participants are at least 18 years old, spoke fluent English, and met the above criteria regarding their affiliation to healthcare AI development. We had no other exclusion criteria.

Sample Size

Power was calculated for effects at Level 1, with three observations per person (one per value, with three values tested per person). To achieve .95 power to detect a medium effect size of $f = .15$ at the standard .05 alpha error probability we identified that $N = 107$ was determined to test the first hypothesis (two predictors), $N = 107$ to test the second hypothesis (two predictors), $N = 146$ to test the third hypothesis (six predictors), and $N = 107$ to test the fourth hypothesis (two predictors). Given the novel nature of investigation, we lack a more precise means to anticipate the effect size that will be observable in the final sample. However, because we were also interested in exploratory analyses, we opened the door to a larger sample size than is required for detecting main and moderation conditions and interaction effects. We set a stopping rule precisely 1 month after we achieved $N = 146$.

The final sample size consisted of 185 participants. Of these, 66.49% identified as male, 25.41% identified as female, 7.57% identified as non-binary or third gender, and 0.54% refused to declare their gender. In addition, 58.92% participants identified as White,

20.54% as Black, 9.73% as other, 5.95% as mixed or multiple ethnic groups, and 4.87% as Asian. Mean age was 27.56 years, $SD = 7.36$, range = 19–72 years. On average, participants spent 2.28 years in a developer role ($SD = 2.28$, range = 0–20 years).

Procedure

Participants signed a consent form and were asked if they wish to be contacted about participation in similar future studies. Participants were then asked to create their identifier word, which was used to identify their responses in case of participation in future studies or if they wish to withdraw their data from this study. Participants were asked to complete a value endorsement, workplace norms, behavioral intention, values internalization, and demographics questionnaires, and were fully debriefed at the end of the study.

Materials

All materials can be found on the Open Science Framework: https://osf.io/3fyqx/. To avoid order effects, questions were presented to participants randomly.

Demographics and Background Measures.

We asked participants to describe their gender (from a dropdown), ethnicity (from a dropdown), age (from a dropdown), years in the role (from a dropdown), and job role (open-ended).

Value-Relevant and SDT-Relevant Measures

Values Endorsement

We were interested in examining three values that relate to developers' work: profit, fairness, and transparency. We asked participants to complete a three-item questionnaire adapted from the Schwartz Value Survey (SVS; Schwartz, 1992, 2006), which represented the values of fairness (referred to as 'universalism' in SVS, assessed with the item: 'They think it is important to produce technology which is equally accurate across different populations') and profit (assessed with the item: 'Being very successful at work is important to them'). We also developed similar questions to assess the value of transparency (assessed with the item: 'It is important to them to be open and transparent at the workplace'), which is not represented in SVS. Responses for each item were presented on a scale from 0 ('*not at all like me*') to 5 ('*very much like me*'). Out of 1665 data points, there were 123 outliers in the dataset.

Values Tension

We computed three scores to assess the tensions between the three value endorsement scores: 1) profit being in tension with transparency and fairness; 2) transparency being in tension with profit and fairness; 3) fairness being in tension with profit and transparency. Out of 1665 data points, there were 66 outliers in the dataset. The scores were computed

using the following formula, and then reversed (so that higher scores indicate more discrepancy between the variables, and thus more tension):

Profit value tension = reversed(profit value endorsement – (transparency value endorsement + fairness value endorsement)/2)

Transparency value tension = reversed(transparency value endorsement – (profit value endorsement + fairness value endorsement)/2)

Fairness value tension = reversed (fairness value endorsement – (profit value endorsement + transparency value endorsement)/2)

Values Internalization

To assess values internalization, we examined the spectrum of motivation that ranges from least to most internalized as described by the SDT (Ryan & Deci, 2000). The corresponding scale follows practices by Neyrinck et al. (2006), Pelletier et al. (1998), and others in relation to values (see review in Howard et al., 2017 for these scales across a broader spectrum of behavior) and assesses different motivations and levels of value internalizations (integrated [fully internalized and connected to self-identity], identified [understood to be personally important and meaningful], introjected [valued to avoid shame and guilt], external [held to avoid punishment and judgment], amotivation [not valued]). We examined the motivations for each of the three values (transparency, fairness, and profit) by asking participants to identify how much they have each value for the following reasons. Out of 1665 data points, there were 15 outliers in the dataset. To compute the internalization score, we used the following formula for each of the three values, resulting in three separate internalization scores:

Internalization = [(integrated motivation + identified motivation)] – [(introjected motivation + external motivation + amotivation)].

TPB-Relevant Measures

Workplace Norms

Workplace norms encompass the beliefs about the prevailing attitudes and behaviors within the work environment, including the expectations of colleagues, supervisors, and organizational culture. Participants responded to six statements, which were adapted from a measure of workplace norms by Dixon et al. (2015) to examine to what extent participants' workplace acts in line with profit, fairness, and transparency values (two items per value). Example questions are 'My company expects me to communicate transparently', 'My colleagues take responsibility for transparently communicating the issues that may be limiting the technology' (on a scale from 0 ['*not at all*'] to 100 ['*extremely*']). The correlation between each pair of items within each subscale was r = .53 for fairness, r =.53 for transparency, and r = .67 for profit. Therefore, following preregistered plans to seek relations above .70, we analyzed each of two items per subscale separately. Out of 1665 data points, there were 33 outliers within the first item and 30 outliers within the second item.

Frequency of Messages

Frequency of messages was measured in terms of exposure to workplace communication frequency related to the value in question. Participants were asked to estimate how frequently ('*never*' = 0, '*once a month*' = 1, '*a few times a month*' = 2, '*weekly*' = 3, '*multiple times a week*' = 4, '*every day*' = 5) they hear (through conversation) or see (e.g., on social media) others talking about the following topics: transparency in communications, fairness or bias in technology, and performance metrics. There were no outliers for this measure.

Attitudes

Participants were asked to rate their attitudes on profit, fairness, and transparency in healthcare AI using a 9-point semantic differential scale (−4 to + 4), with the following adjectives in each end: negative/positive, unpleasant/pleasant, bad/good, and undesirable/desirable. This is a measure widely used in the literature measure, and has demonstrated a Cronbach's α of > .95 in previous research (see Lins de Holanda Coelho et al., 2018; Maio & Olson, 1994). We averaged items to produce one score per value. For our sample, α for fairness = .91, 95% CI [.89, .93]; α for transparency = .91, 95% CI [.89, .93], α for profit = .95, 95% CI [.94, .96]. There were no outliers within the attitudes measure.

Perceived Behavioral Control

Participants were asked to assess their capacity for acting in line with the values by rating their agreement with six statements (two regarding profit, two regarding fairness, and two regarding transparency) on a 7-point scale (from 1 – *completely disagree* to 7 – *completely agree*). An example statement is 'I am confident that I am able to produce fair technology'. This measure was adapted from Ajzen (2006), which showed a high reliability score (α = .95, Liao et al., 2022). The correlation between the items was r = .57 for fairness, r = .42 for transparency, r = .63 for profit. Therefore, we included items separately in the analysis following pre-registered plans. There were no outliers within the measures.

Outcome Measures

Behavioral Intention

Participants were presented with scenarios of workplace value-congruent behavior dilemmas and given options outlining ways they would address the problem introduced within each scenario. Participants were asked to think about their particular line of work when responding. Each response option represents one of the values (profit, fairness, and transparency). For each option, participants rated on a scale from 0 to 100 how likely they are to act this way. The scenarios and outcomes were written by a professional working in a radiology field with an aim to have realistic representation of situations at the workplace and have been validated with professionals working in the field (N = 13). During the validation, we asked healthcare professionals which of three values they would attribute to each course of action. This allowed us to examine whether participants are intending to act in line with the values of profit, fairness, and transparency. There were no outliers within the measure.

An example scenario is as follows:

You've received feedback from a customer who found evidence that your algorithm doesn't perform as well on a subgroup of their local population (e.g., an ethnic minority group). How likely are you to spend the month after . . .

Making other customers aware of the problem in case their local populations are similarly affected.

Pause the deployment of the software and take the time to find evidence that speaks to the technology's performance on local populations (e.g., ethnic minorities compared to other groups), since it is important for better understanding whether the product produces fair results.

Design a marketing strategy that highlights the strengths of the product.

Values Trade-Off Decision Task

For each scenario measuring behavioral intention to act in line with corresponding value (described above), participants were asked to select the option, out of those presented, that they see to be the most beneficial. Effectively, they were choosing between actions that express value of profit versus value of fairness versus value of transparency. This was done by asking participants 'What is the most beneficial course of action?' from options presented to them in scenarios. The trade-off decision allowed us to examine how values are prioritized when having to select between competing behaviors.

Analysis Plan

A correlation table was created in JASP (Version 0.17.1). All other analyses were conducted in R (Version 1 March 1056), using Hierarchical Linear Modelling (package lme4). Missing data was imputed via multiple imputation using 'mice' package in R. All predictors were centered. Outliers were kept in the dataset. We interpreted two measures to be significantly related when $p < .05$, but in the Discussion section we review conclusions drawn with a more conservative Bonferroni correction and an adjusted benchmark of $p < .01$.

Planned Primary Analyses

Hypotheses were tested by using hierarchical linear modeling with data nested within individuals. First, we examined measures' reliability without nesting. Then, we restructured the data. Individual-level data were modeled at Level 2, with three value-relevant constructs (namely, for fairness, transparency, and profit values) nested within individuals at Level 1. This data structure yielded 1665 data points. Our hypotheses concern these Level 1 variables, which link value endorsement, internalization, tension, norms, and message frequency (as five predictors tested separately or in combination according to the a-priori hypothesis) to corresponding value behavioral intention defined as an outcome variable in each of our proposed models. We did not have a-priori hypotheses that concern Level 2 predictors and therefore did not plan to define them.

To test **Hypothesis 1**, which suggested that developers intend to engage in behavior that is in line with their values, we used the following analysis code:

model_h1 <- lmer(behavior ~ value + attitudes + (1 | developer_id), data = data_scaled)

We planned that if attitudes and endorsement significantly relate to behavioral intention, we would consider Hypothesis 1 to be supported. We also planned that of only one of the tested predictor variables significantly linked to behavioral intention, we would consider Hypothesis 1 to be partially supported.

To test **Hypothesis 2**, which suggests that the more a value is internalized, the more likely healthcare AI technology developers are to intend to engage in behavior in line with this value, we used the following analysis code:

model _h2 <- lmer(behavior ~ internalization + (1 | developer_id), data = data_scaled)

We anticipated that if internalization significantly associates with behavioral intention, we would consider Hypothesis 2 to be supported.

To test **Hypothesis 3**, which suggests that when holding one another constant and controlling for the effects of values internalization, attitudes, workplace norms, frequency of messages, and perceived behavioral control would all predict behave intention, we used the following analysis code:

model _h3 <- lmer(behavior ~ attitudes + workplace_norms + frequency_of_messages + perceived_behavioral_control + internalization + (1 | developer_id), data = data_scaled)

We anticipated that if value internalization, attitudes, workplace norms, frequency of messages, and perceived behavioral control relate to behavioral intention, we would consider Hypothesis 3 to be supported.

To test **Hypothesis 4**, which expected that when healthcare AI technology developers endorse multiple values that are in tension (fairness and transparency versus profit), value internalization would interact with value tension, so that the value that is more internalized drives behavior, we used the following analysis code:

*model_h4 <- lmer(most_beneficial_course_of_action ~ internalization*tension + (1 | developer_id), data = data_scaled)*

We planned that if internalization associates with the most value-consistent course of action chosen, over and above value tension, this would support Hypothesis 4.

Exploratory Analyses

Exploratory analysis used hierarchical regression to examine the relationship between norms, frequency of messages, perceived behavioral control, and behavioral intention when controlling for whether participants were employed, number of years in the role, and their level of seniority.

Results

See Table 1 for correlations between the variables.

Table 1. Pearson's correlations among study variables.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Value endorsement	—													
2. Attitudes	.266***	—												
3. Norms (item 1)	.194***	.250***	—											
4. Norms (item 2)	.209***	.280***	.635***	—										
5. Frequency of Messages	.087***	−.037	.097***	.060*	—									
6. Behavioral Control (item 1)	−.058*	−.055*	−.128***	−.158***	.069**	—								
7. Behavioral Control (item 2)	−.072**	−.055*	−.116***	−.162***	.059*	.888***	—							
8. Value internalization	.081***	.076**	.080**	.114***	.031	−.014	−.030	—						
9. Value Tension	−.643***	−.145***	.026	.034	.026	.007	−.009	.003	—					
10. Behavior	.190***	.200***	.256***	.254***	.059*	−.100***	−.142***	.107***	−.035	—				
11. Values trade-off (Yes/No)	.014	.121***	−.098***	−.078**	−.045	.004	.016	.047	−.045	.090***	—			
12. Seniority	.070**	.090***	.149***	.174***	−.201***	−.125***	−.181***	.068**	−.001	.193***	−.051*	—		
13. Current employment	−.052*	−.138***	−.172***	−.215***	.105***	.307***	.377***	−.017	.003	−.225***	.099***	−.608***	—	
14. Years in the Role	.144***	.087**	.074**	.180***	.098***	.092***	.112***	.009	.000	.045	−.019	−.002	−.075**	—

*$p < .05$, **$p < .01$, ***$p < .001$.
For the values trade-off variable, 0 represents not choosing the corresponding value, 1 represents choosing the corresponding value.
For the current employment variable, 1 represents being employed, 2 represents not being employed.

Table 2. Summary of regression models.

Predictor	b	SE	t	df	p
Model 1					
Intercept	68.43	1.04	65.94	175.38	<.001
Endorsement	2.54	0.75	3.39	1194.75	<.001
Attitudes	1.73	0.39	4.47	1478.45	<.001
Model 2					
Intercept	68.43	1.12	61.08	183.42	<.001
Internalization	0.73	0.24	3.04	1346.84	.002
Model 3					
Intercept	68.43	0.94	73.18	161.15	<.001
Attitudes	1.79	0.38	4.66	1236.39	<.001
Norms Item 1	0.10	0.03	3.17	1647.17	.002
Norms Item 2	0.06	0.03	1.81	1624.86	.070
Frequency of messages	0.98	0.45	2.16	972.97	.031
Behavioral control Item 1	0.65	0.58	1.13	1282.53	.260
Behavioral control Item 2	−1.73	0.53	−3.27	1387.10	.001
Internalization	0.63	0.23	2.76	1038.53	.006
Model 4					
Intercept	0.43	0.02	23.42	182.60	<.001
Internalization	0.01	0.004	2.59	1149.00	.010
Value tension	−0.02	0.01	−2.00	1481.00	.045
Internalization x Value tension	−0.00	0.004	−0.02	1660.00	.984

Endorsement and internalization refer to value endorsement and internalization scales. Specific items for norms and for behavioral control were analyzed separately because they correlated below the threshold for our pre-registered plans ($r < .70$). Internalization x Value tension reflects the interaction term between internalization and value tension.

Planned Primary Tests

Model examining how values and attitudes are associated with value-consistent behavioral intention showed that both values ($b = 2.55$, $SE = .75$, $t(1194.75) = 3.39$, $p < .001$) and attitudes ($b = 1.73$, $SE = .39$, $t(1478.46) = 4.47$, $p < .001$) predicted behavioral intention to act in line with corresponding values; thus, Hypothesis 1 was supported. The model examining the association between value internalization and value-consistent behavioral intention showed that the more the value was internalized, the more developers intended to act in line with this value ($b = .73$, $SE = .24$, $t(1346.84) = 3.04$, $p = .002$), thus supporting Hypothesis 2. The model examining the association between TPB variables and value-consistent behavioral intention showed that when controlling for internalization, attitudes ($b = 1.79$, $SE = .38$, $t(1236.39) = 4.69$, $p < .001$) and norms ($b = .10$, $SE = .03$, $t(1647.17) = 3.17$, $p = .002$) positively predicted behavioral intention, while perceived behavioral control was negatively associated with it ($b = -1.74$, $SE = .53$, $t(1387.10) = -3.27$, $p = .001$). Therefore, Hypothesis 3 was partially supported. The model examining the interaction between value tension and value internalization showed that, while higher internalization ($b = .01$, $SE = .00$, $t(1149) = 2.59$, $p = .010$) and lower value tension ($b = -.02$, $SE = .01$, $t(1481) = -2.00$, $p = .045$) were associated with intention to act in line with a corresponding value, they did not interact; thus, Hypothesis 4 was not supported. See Table 2 for the summary of confirmatory results.

Exploratory Tests

Exploratory analysis indicated that when controlling for whether participants are employed, years in the role, and level of seniority, perceived behavioral control and

Table 3. Results of exploratory analysis.

Predictor	*b*	*SE*	*t*	df	*p*
Intercept	63.21	5.60	11.28	133.43	<.001
Attitudes	1.76	0.37	4.72	1045.36	<.001
Seniority	0.26	0.98	0.26	131.79	.795
Employment Status	6.60	4.42	1.49	131.51	.138
Years in the Role	0.09	0.44	0.21	132.87	.831
Norms Item 1	0.11	0.04	2.91	1299.82	.004
Norms Item 2	0.12	0.04	2.99	1293.49	.003
Frequency of Messages	0.79	0.51	1.54	745.93	.125
Behavioral Control Item 1	0.69	0.70	0.98	1202.97	.330
Behavioral Control Item 2	−1.14	0.67	−1.70	1286.63	.090
Internalization	0.55	0.26	2.16	536.22	.031

frequency of messages were not significant predictors anymore. For the summary of exploratory analysis, see Table 3.

Discussion

The current study aimed to use values, Self-Determination Theory (Ryan & Deci, 2000) and the Theory of Planned Behavior (Ajzen, 1991) literatures to explore which factors may be associated with behavioral intention of developers to engage in value-based actions of transparency (i.e., valuing honestyabout limitations of technology), fairness (valuing technology that is equally effective at diagnosing and treating people from different backgrounds and demographics), and profit (valuing earning potential). These values drive key decisions made by develops as they invest resources in building ethical AI for use in healthcare (Hendrix et al., 2022; Solanki et al., 2023).

Based on values literature (Rokeach, 1973; Schwartz, 1992), we hypothesized that developers would intend to engage in behavior that is in line with their values (H1). This hypothesis was supported. Based on Self-Determination Theory literature (Ryan & Deci, 2000), we hypothesized that the more a value is internalized and thus felt to be personally meaningful and part of one's identity, the more likely healthcare AI technology developers would intend to engage in behavior in line with this value (H2). This hypothesis was also supported. Next, based on the Theory of Planned Behavior literature (Ajzen, 1991), and accounting for values and SDT literatures, we hypothesized that when controlling for values internalization, attitudes, workplace norms, frequency of messages, and perceived behavioral control would all predict behavior intention (H3). This hypothesis was partially supported – although attitudes, frequency of messages, and workplace norms emerged as positive predictors of value-consistent behavioral intention, perceived behavioral control was inversely associated with behavioral intention. Finally, based on values and SDT literature, we hypothesized that when healthcare AI technology developers endorse multiple values that are in tension (fairness and transparency versus profit), value internalization would interact with value tension, so that the value that is more internalized drives behavior (H4). We anticipated this would be the case because given the choice to make between two competing values one holds, the value that is more personally meaningful and tied to identity should drive decision-making. This hypothesis was not supported – there was no evidence that value internalization helped to drive value-expressing behavioral intentions when two values were in direct tension with one another.

In our simultaneous models, the more the value was endorsed, and the more participants held positive attitudes toward a value-consistent behavior, the more likely they were to intend to engage in the value-congruent behavior. These findings support both the Theory of Planned Behavior (Ajzen, 1991) and values (Rokeach, 1973; Schwartz, 1992) literatures, highlighting the central role of merely holding a value or a positive attitude in shaping behavior. Further, consistent with values-attitudes-behavior hierarchy (Homer & Kahle, 1988), we observed significant positive correlations between values, attitudes, and behavior. This finding suggested that cultivating values and positive attitudes toward relevant behaviors is needed to cultivate these behaviors. In the context of the values we studied, this may include framing concrete and actionable fairness- and transparency-related behaviors in the context of employees' existing values for those basic principles (Albarracin & Shavitt, 2018). Additionally, implementing workplace interventions, such as emotional regulation strategies like cognitive reappraisal, can further promote these behaviors (ibid.). Indeed, cognitive reappraisal has been shown to reduce automatic defensive reactions and facilitate more deliberate, value-driven decision-making (Sheppes et al., 2014). Organizations could implement training programs that help employees reframe fairness-related challenges in ways that align with their core values. By integrating these strategies, workplaces may be better positioned to cultivate a workplace climate that encourages fairness and transparency in daily decision-making.

Self-Determination Theory-based expectations were also supported, highlighting the importance that value internalization, which reflects the transformation from external regulations into personally endorsed values and self-regulation (Ryan & Deci, 2000) played in driving value-congruent behavior. Importantly, the benefits of value internalization were independent from *having* the value and had additive effects – developers acted in line with their values when they both had a value and had internalized motivation from the value.

Internalization reflects a more autonomous, self-driven motivation, which can be cultivated by providing choice, support, and learning opportunities (Kenny et al., 2024) to help explore and identify the personal significance of a value. Within workplace environments, providing options to participate in courses related to fairness and transparency in emerging technologies and other workplace environments may help to cultivate this form of motivation and facilitate subsequent behaviors that are consistent with one's values.

It is possible that adopting a transformational leadership style facilitates value internalization (Sun & Henderson, 2017). Transformational leaders are characterized by their ability to inspire, motivate, and lead by example, which could bring a shared vision to a company and therefore promote higher levels of value alignment among employees (Pandey et al., 2016). Such leaders are also likely to use autonomy-supportive strategies such as articulating concrete goals, fostering intellectual stimulation, and addressing the unique needs of individuals (Hannah et al., 2016). By reinforcing workplace values, transformational leaders can create an environment where employees are more likely to internalize these values and engage in value-corresponding behaviors.

Our expectations in line with the Theory of Planned Behavior, that attitudes, workplace norms, frequency of messages, and perceived behavioral control would all be positively associated with behavioral intention when controlling for values

internalization, were not supported. While attitudes, norms, and frequency of messages were positively related to intention to behave in line with one's values, frequency of messages was not a significant predictor. Further, we were surprised to see that perceived behavioral control had an inverse relationship with behavioral control, in direct contrast to what was expected. Although perceived behavioral control is typically associated with greater behavioral intention, such unexpected findings have been identified in the past. For example, Farah (2017) found that as bank customers felt more in control of their banking choices, they were less, not more, likely to switch banks. Though this example is quite different from our own study, a possible explanation for effects across both investigations is that when people feel like they have sufficient control over their current situation, they may feel less inclined to change or engage in behaviors that would require effort or introduce uncertainty. Effects may also be explained through the mechanism of a status quo bias, a cognitive bias that describes individuals preferring situations to remain constant rather than change (Samuelson & Zeckhauser, 1988). It may have been that employees who believed they have control had become complacent, feeling that changing behavior or adopting new values was unnecessary or irrelevant.

Although we speculate potential mechanisms, it is worth noting that other studies, including Kashif et al. (2018), have shown that perceived behavioral control is typically associated with increased behavioral intention in organizational contexts. More generally, in organizational settings, perceived control is often influenced by hierarchical position and the autonomy associated with one's role (Parker et al., 2001). In our study, employees in senior positions may perceive greater control over their work environment and decisions, while those in subordinate roles might feel constrained. This variation in perceived control across organizational levels could explain the contradictory findings. In fact, exploratory analyses in this study showed that the effect of perceived behavioral control on behavioral intention diminished after controlling for seniority. Additionally, when adjusting for multiple comparisons, the frequency of messages was no longer a significant predictor of behavioral intention. Taken together, these findings challenge the comprehensive applicability of this component of the theory of planned behavior in organizational settings, particularly among healthcare AI developers. They suggest that the combined effects of attitudes, norms, message frequency, and perceived behavioral control may be less predictive of behavioral intentions than previously anticipated. Future research could examine how organizational hierarchy and seniority influence perceived behavioral control and its subsequent impact on behavioral intentions.

Integrating values and Self-Determination Theory literatures, we sought to test a novel interaction – that value-tension costs to behavior would be mitigated when having higher internalization for a value-consistent behavior. Although both higher internalization and lower value tension were both independently predictive of behavioral intention as we describe above, they did not interact with one another. Said another way, there was no evidence that value internalization, in isolation, resolved tensions between conflicting values in the expected manner. Internalization of values alone may not be sufficient to reduce the cognitive dissonance or internal conflict associated with value tension, challenging some previous models that posit a direct link between internalization and value conflict resolution (Rokeach, 1973; Schwartz, 1992). Further, the effect sizes for both variables were small, and for value tension, the

effect was not significant at .01 significance level. This questions the robustness of these predictors in organizational contexts, where complex external factors may be at play. Future research could examine whether barriers for behaving in a certain way, social pressure (Müller-Hansen et al., 2017) and external constraints, such as audits or organizational consequences and culture (Sonjaya, 2024) may be more effective in guiding behavior in the face of value tension.

Findings from this study underscore the critical role of organizational climate in shaping employees' values of fairness, inclusion, and associated behaviors; ultimately, there may be broader implications for workplace performance (Italiani et al., 2022; Syarief et al., 2022). A positive organizational climate, defined by trust, transparency, and respect, enhances employees' intrinsic motivation to adopt and maintain these values (Men & Stacks, 2014). Key practices such as providing regular feedback, recognizing behaviors aligned with organizational values, establishing formal codes of conduct, setting informal norms, leading by example, and considering individual differences among employees all contribute to shaping attitudes and behaviors that reflect these core values (Besio & Pronzini, 2014; Grojean et al., 2004). The concept of 'ethical leadership,' as highlighted in the literature, emphasizes that such an environment promotes not only ethical behavior and alignment with organizational values but also a sense of belonging and commitment among staff (Avey et al., 2012). In this way, a supportive organizational climate serves as a foundation for nurturing fairness, inclusion, and transparency, driving behaviors that align with the organization's mission and objectives, ultimately enhancing success, sustainability, and, in the context of healthcare, improving patient outcomes.

The results of the study should be considered in light of several limitations. First, the items used to measure norms and perceived behavioral control were poorly correlated, leading us to analyze them separately. While the direction of the effects was consistent across items, this issue suggests potential measurement limitations. It is possible that the constructs were not well captured by the specific items used or that participants interpreted them in different ways (Podsakoff et al., 2012). Future research should refine these measures to ensure stronger internal consistency and construct validity. Second, we measured behavioral intention rather than actual behavior, which limits the extent to which our findings can be directly applied to real-world actions. Indeed, the intention-behavior gap (Sheeran & Webb, 2016) remains a well-documented challenge in psychology, with numerous situational and psychological barriers influencing whether intentions translate into action (Ajzen, 2011; Gollwitzer, 1999). Future studies should incorporate longitudinal or experimental designs to assess whether the relationships we observed hold when examining real behavior over time. Additionally, the inclusion of behavioral measures, such as self-reports of enacted behaviors or observational data, could provide more robust insights into how values and attitudes shape actual decision-making.

In all, the results of the study indicated that value endorsement, value internalization, and attitudes are important predictors of intention to behave in line with values in the organizational context of healthcare AI development. Workplace norms and frequency of messages were also found to be associated with increased behavioral intention, whereas perceived behavioral control had an inverse relationship with behavior. The findings highlight the importance of initiatives to promote relevant values at the workplace.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Alfred P. Sloan Foundation [G-2021-16779]; Department of Health and Social Care; Wellcome Trust [223765/Z/21/Z].

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Ajzen, I. (2006, January). *Constructing a theory of planned behavior questionnaire*.

Ajzen, I. (2011). The theory of planned behaviour: Reactions and reflections. *Psychology & Health*, *26*(9), 1113–1127. https://doi.org/10.1080/08870446.2011.613995

Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin*, *82*(2), 261–277. https://doi.org/10.1037/h0076477

Albarracin, D., & Shavitt, S. (2018). Attitudes and attitude change. *Annual Review of Psychology*, *69* (1), 299–327. https://doi.org/10.1146/annurev-psych-122216-011911

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research*, *22*(6), e15154. https://doi.org/10.2196/15154

Avey, J. B., Wernsing, T. S., & Palanski, M. E. (2012). Exploring the process of ethical leadership: The mediating role of employee voice and psychological ownership. *Journal of Business Ethics*, *107*(1), 21–34. https://doi.org/10.1007/s10551-012-1298-2

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

Barr, S. (2006). Environmental action in the home: Investigating the 'value-action' gap. *Geography*, *91*(1), 43–54. https://doi.org/10.1080/00167487.2006.12094149

Besio, C., & Pronzini, A. (2014). Morality, ethics, and values outside and inside organizations: An example of the discourse on climate change. *Journal of Business Ethics*, *119*(3), 287–300. https://doi.org/10.1007/s10551-013-1641-2

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, *5*(12), 1636–1642. https://doi.org/10.1038/s41562-021-01146-0

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology/Psychologie Canadienne*, *49*(3), 182. https://doi.org/10.1037/a0012801

Dixon, G. N., Deline, M. B., McComas, K., Chambliss, L., & Hoffmann, M. (2015). Saving energy at the workplace: The salience of behavioral antecedents and sense of community. *Energy Research and Social Science*, *6*, 121–127. https://doi.org/10.1016/j.erss.2015.01.004

Dunlap, R. E., Grieneeks, J. K., & Rokeach, M. (1983). Human values and pro-environmental behavior. In W. D. Conn (Ed.), *Energy and material resources: Attitudes, values, and public policy* (pp. 145–168). Westview Press.

Farah, M. F. (2017). Application of the theory of planned behavior to customer switching intentions in the context of bank consolidations. *International Journal of Bank Marketing*, *35* (1), 147–172. https://doi.org/10.1108/IJBM-01-2016-0003

Fennis, B. M., & Aarts, H. (2012). Revisiting the agentic shift: Weakening personal control increases susceptibility to social influence. *European Journal of Social Psychology*, *42*(7), 824–831. https://doi.org/10.1002/ejsp.1887

Flynn, R., Bellaby, P., & Ricci, M. (2009). The 'value-action gap' in public attitudes towards sustainable energy: The case of hydrogen energy. *Sociological Review*, *57*(2_suppl), 159–180. https://doi.org/10.1111/j.1467-954X.2010.01891.x

Gatersleben, B., Murtagh, N., & Abrahamse, W. (2014). Values, identity and pro-environmental behaviour. *Contemporary Social Science*, *9*(4), 374–392. https://doi.org/10.1080/21582041.2012.682086

Ghassemi, M., & Mohamed, S. (2022). Machine learning and health need better values. *NPJ Digital Medicine*, *5*(1), 51. https://doi.org/10.1038/s41746-022-00595-9

Gifford, R. (2011). The dragons of inaction: Psychological barriers that limit climate change mitigation and adaptation. *The American Psychologist*, *66*(4), 290. https://doi.org/10.1037/a0023566

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *The American Psychologist*, *54*(7), 493. https://doi.org/10.1037/0003-066X.54.7.493

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 38, pp.69–119). Elsevier Academic Press.

Grojean, M. W., Resick, C. J., Dickson, M. W., & Smith, D. B. (2004). Leaders, values, and organizational climate: Examining leadership strategies for establishing an organizational climate regarding ethics. *Journal of Business Ethics*, *55*(3), 223–241. https://doi.org/10.1007/s10551-004-1275-5

Hackett, J. D. (2014). Values anchoring: Strengthening the link between values and activist behaviors. *Social Influence*, *9*(2), 99–115. https://doi.org/10.1080/15534510.2013.787117

Hanel, P. H., Foad, C., & Maio, G. R. (2021). Attitudes and Values. In M. Hogg (Ed.), *Oxford research encyclopedia of psychology*. https://doi.org/10.1093/acrefore/9780190236557.013.248

Hannah, S. T., Schaubroeck, J. M., & Peng, A. C. (2016). Transforming followers' value internalization and role self-efficacy: Dual processes promoting performance and peer norm-enforcement. *Journal of Applied Psychology*, *101*(2), 252. https://doi.org/10.1037/apl0000038

Hansen, T. (2008). Consumer values, the theory of planned behaviour and online grocery shopping. *International Journal of Consumer Studies*, *32*(2), 128–137. https://doi.org/10.1111/j.1470-6431.2007.00655.x

Hendrix, N., Veenstra, D. L., Cheng, M., Anderson, N. C., & Verguet, S. (2022). Assessing the economic value of clinical artificial intelligence: Challenges and opportunities. *Value in Health*, *25*(3), 331–339. https://doi.org/10.1016/j.jval.2021.08.015

Hitlin, S., & Piliavin, J. A. (2004). Values: Reviving a dormant concept. *Annual Review of Sociology*, *30*(1), 359–393. https://doi.org/10.1146/annurev.soc.30.012703.110640

Homer, P. M., & Kahle, L. R. (1988). A structural equation test of the value-attitude-behavior hierarchy. *Journal of Personality & Social Psychology*, *54*(4), 638. https://doi.org/10.1037/0022-3514.54.4.638

Howard, J. L., Gagné, M., & Bureau, J. S. (2017). Testing a continuum structure of self-determined motivation: A meta-analysis. *Psychological Bulletin*, *143*(12), 1346. https://doi.org/10.1037/bul0000125

Italiani, N., Musmuliadi, M., & Diju, A. (2022). The influence of leadership, organizational climate, and work motivation on employee's performance. *Interdisciplinary Social Studies*, *1*(12). https://doi.org/10.55324/iss.v1i12.285

Iversen, H. (2004). Risk-taking attitudes and risky driving behaviour. *Transportation Research: Part F, Traffic Psychology and Behaviour*, *7*(3), 135–150. https://doi.org/10.1016/j.trf.2003.11.003

Kashif, M., Zarkada, A., & Ramayah, T. (2018). The impact of attitude, subjective norms, and perceived behavioural control on managers' intentions to behave ethically. *Total Quality Management & Business Excellence*, *29*(5–6), 481–501. https://doi.org/10.1080/14783363.2016.1209970

Kennedy, E. H., Beckley, T. M., McFarlane, B. L., & Nadeau, S. (2009). Why we don't" walk the talk": Understanding the environmental values/behaviour gap in Canada. *Human Ecology Review*, *16*(2), 151–160. http://www.jstor.org/stable/24707539

Kenny, M. E., Medvide, M. B., & Gordon, P. (2024). Cultivating purpose and internalized motivation through workplace learning. *Gifted Education International*, *40*(3), 295–311. https://doi.org/10.1177/02614294241264398

Kirkley, W. W. (2016). Entrepreneurial behaviour: The role of values. *International Journal of Entrepreneurial Behavior and Research*, *22*(3), 290–328. https://doi.org/10.1108/IJEBR-02-2015-0042

LaPiere, R. T. (1934). Attitudes vs. actions. *Social Forces*, *13*(2), 230–237. https://doi.org/10.2307/2570339

Le Grand, J. (1990). Equity versus efficiency: The elusive trade-off. *Ethics*, *100*(3), 554–568. https://doi.org/10.1086/293210

Liao, T., Tang, S., & Shim, Y. (2022). The development of a model to predict sports participation among college students in Central China. *International Journal of Environmental Research and Public Health*, *19*(3), 1806. https://doi.org/10.3390/ijerph19031806

Lins de Holanda Coelho, G., Hp Hanel, P., Vilar, R., P Monteiro, R., Gouveia, V. V., & R Maio, G. (2018). Need for affect and attitudes toward drugs: The mediating role of values. *Substance Use & Misuse*, *53*(13), 2232–2239. https://doi.org/10.1080/10826084.2018.1467454

Lönnqvist, J. E., Verkasalo, M., Wichardt, P. C., & Walkowitz, G. (2013). Personal values and prosocial behaviour in strategic interactions: Distinguishing value-expressive from value-ambivalent behaviours. *European Journal of Social Psychology*, *43*(6), 554–569. https://doi.org/10.1002/ejsp.1976

Mackenbach, J. P. (2014). Cultural values and population health: A quantitative analysis of variations in cultural values, health behaviours and health outcomes among 42 European countries. *Health & Place*, *28*, 116–132. https://doi.org/10.1016/j.healthplace.2014.04.004

Maio, G., Litzellachner, L. F., Karremans, J. C., Buiter, N., Breukel, J., & Maio, G. R. (2023). Values in romantic relationships. *Personality & Social Psychology Bulletin*, *50*(7), 1066–1079. https://doi.org/10.1177/01461672231156975

Maio, G. R. (2016). *The psychology of human values*. Psychology Press.

Maio, G. R., & Olson, J. M. (1994). Value—attitude-behaviour relations: The moderating role of attitude functions. *British Journal of Social Psychology*, *33*(3), 301–312. https://doi.org/10.1111/j.2044-8309.1994.tb01027.x

Men, L. R., & Stacks, D. (2014). The effects of authentic leadership on strategic internal communication and employee-organization relationships. *Journal of Public Relations Research*, *26*(4), 301–324. https://doi.org/10.1080/1062726X.2014.908720

Mittelstadt, B. (2021, November 10). Interpretability and transparency in artificial intelligence. In C. Véliz (Ed.), *The Oxford Handbook of Digital Ethics* (online ed.). Oxford Academic. 378–410. doi:https://doi.org/10.1093/oxfordhb/9780198857815.013.20

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Morell-Gomis, R., Lloret Irles, D., Moriano, J. A., Edú-Valsania, S., & Laguía González, A. (2018). Predicting cannabis use among adolescents in four European countries: Combining personal values and the theory of planned behaviour. *Addiction Research & Theory*, *26*(6), 498–506. https://doi.org/10.1080/16066359.2018.1443214

Müller-Hansen, F., Schlüter, M., Mäs, M., Donges, J. F., Kolb, J. J., Thonicke, K., & Heitzig, J. (2017). Towards representing human behavior and decision making in Earth system models – an overview of techniques and approaches. *Earth System Dynamics*, *8*(4), 977–1007. https://doi.org/10.5194/esd-8-977-2017

Mydland, L., & Grahn, W. (2012). Identifying heritage values in local communities. *International Journal of Heritage Studies*, *18*(6), 564–587. https://doi.org/10.1080/13527258.2011.619554

Nelson, G. S. (2019). Bias in artificial intelligence. *North Carolina Medical Journal*, *80*(4), 220–222. https://doi.org/10.18043/ncm.80.4.220

Neyrinck, B., Vansteenkiste, M., Lens, W., Duriez, B., & Hutsebaut, D. (2006). Cognitive, affective and behavioral correlates of internalization of regulations for religious activities. *Motivation and Emotion*, *30*(4), 321–332. https://doi.org/10.1007/s11031-006-9048-3

Pandey, S. K., Davis, R. S., Pandey, S., & Peng, S. (2016). Transformational leadership and the use of normative public values: Can employees be inspired to serve larger public purposes? *Public Administration*, *94*(1), 204–222. https://doi.org/10.1111/padm.12214

Parker, S. K., Wall, T. D., & Cordery, J. L. (2001). Future work design research and practice: Towards an elaborated model of work design. *Journal of Occupational & Organizational Psychology*, *74*(4), 413–440. https://doi.org/10.1348/096317901167460

Pelletier, L. G., Tuson, K. M., Green-Demers, I., Noels, K., & Beaton, A. M. (1998). Why are you doing things for the environment? The motivation toward the environment scale (mtes) 1. *Journal of Applied Social Psychology*, *28*(5), 437–468. https://doi.org/10.1111/j.1559-1816.1998.tb01714.x

Podsakoff, P. M., MacKenzie, S. B., & Bommer, W. H. (1996). Transformational leader behaviors and substitutes for leadership as determinants of employee satisfaction, commitment, trust, and organizational citizenship behaviors. *Journal of Management*, *22*(2), 259–298. https://doi.org/10.1177/014920639602200204

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*(1), 539–569. https://doi.org/10.1146/annurev-psych-120710-100452

Richardson, J. P., Smith, C., Curtis, S., Watson, S., Zhu, X., Barry, B., & Sharp, R. R. (2021). Patient apprehensions about the use of artificial intelligence in healthcare. *NPJ Digital Medicine*, *4*(1), 140. https://doi.org/10.1038/s41746-021-00509-1

Rokeach, M. (1968). Beliefs, attitudes and values: A theory of organization and change.

Rokeach, M. (1973). *The nature of human values*. Free Press.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, *55*(1), 68. https://doi.org/10.1037/0003-066X.55.1.68

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59. https://doi.org/10.1007/BF00055564

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, *25*(1), 1–65.

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *The Journal of Social Issues*, *50*(4), 19–45. https://doi.org/10.1111/j.1540-4560.1994.tb01196.x

Schwartz, S. H. (2003). A proposal for measuring value orientations across nations. *Questionnaire Package of the European Social Survey*, *259*(290), 261.

Schwartz, S. H. (2006). A theory of cultural value orientations: Explication and applications. *Comparative Sociology*, *5*(2–3), 137–182. https://doi.org/10.1163/156913306778667357

Schwartz, S. H. (2010). Basic values: How they motivate and inhibit prosocial behavior. In M. Mikulincer & P. R. Shaver (Eds.), *Prosocial motives, emotions, and behavior: The better angels of our nature* (pp. 221–241). American Psychological Association. https://doi.org/10.1037/12061-012

Schwartz, S. H., Lehmann, A., & Roccas, S. (1999). Multimethod probes of basic human values. In J. Adamopoulos & Y. Kashima (Eds.), *Social psychology and cultural context: Essays in honor of Harry C. Triandis* (pp. 107–123). Sage.

Sharma, R., & Jha, M. (2017). Values influencing sustainable consumption behaviour: Exploring the contextual relationship. *Journal of Business Research*, *76*, 77–88. https://doi.org/10.1016/j.jbusres.2017.03.010

Sheeran, P., & Webb, T. L. (2016). The intention–behavior gap. *Social and Personality Psychology Compass*, *10*(9), 503–518. https://doi.org/10.1111/spc3.12265

Sheppes, G. (2014). Emotion regulation choice: Theory and findings. *Handbook of Emotion Regulation*, *2*, 126–139.

Shinners, L., Aggar, C., Grace, S., & Smith, S. (2020). Exploring healthcare professionals' understanding and experiences of artificial intelligence technology use in the delivery of healthcare: An integrative review. *Health Informatics Journal*, *26*(2), 1225–1236. https://doi.org/10.1177/1460458219874641

Sleczka, P., Braun, B., Grüne, B., Bühringer, G., & Kraus, L. (2018). Family functioning and gambling problems in young adulthood: The role of the concordance of values. *Addiction Research & Theory*, *26*(6), 447–456. https://doi.org/10.1080/16066359.2017.1393531

Smith, J., & Malcolm, A. (2010). Spirituality, leadership and values in the NHS. *International Journal of Leadership in Public Services*, *6*(2), 39–53. https://doi.org/10.5042/ijlps.2010.0353

Solanki, P., Grundy, J., & Hussain, W. (2023). Operationalising ethics in artificial intelligence for healthcare: A framework for AI developers. *AI and Ethics*, *3*(1), 223–240. https://doi.org/10.1007/s43681-022-00195-z

Sonjaya, Y. (2024). The influence of corporate culture on audit practices and ethics. *Golden Ratio of Auditing Research*, *4*(2), 107–124.

Souchon, N., Maio, G. R., Hanel, P. H., & Bardin, B. (2017). Does spontaneous favorability to power (vs. universalism) values predict spontaneous prejudice and discrimination? *Journal of Personality*, *85*(5), 658–674. https://doi.org/10.1111/jopy.12269

Stern, P. C. (2000). New environmental theories: Toward a coherent theory of environmentally significant behavior. *The Journal of Social Issues*, *56*(3), 407–424. https://doi.org/10.1111/0022-4537.00175

Stern, P. C., Dietz, T., Abel, T., Guagnano, G. A., & Kalof, L. (1999). A value-belief-norm theory of support for social movements: The case of environmentalism. *Human Ecology Review*, *6*(2), 81–97. https://www.jstor.org/stable/24707060

Sun, R., & Henderson, A. C. (2017). Transformational leadership and organizational processes: Influencing public performance. *Public Administration Review*, *77*(4), 554–565. https://doi.org/10.1111/puar.12654

Syarief, A. S. I., Iskandar, N. I., & Muhajir, M. N. A. (2022). The effect of organizational climate and work motivation on employee performance at Sawerigading Hospital Palopo. *Jurnal Economic Resource*, *5*(2), 279–285. https://doi.org/10.57178/jer.v5i2.366

Terry, D. J., Hogg, M. A., & White, K. M. (1999). Attitude-behavior relations: Social identity and group membership. In D. J. Terry & M. A. Hogg (Eds.), *Attitudes, behavior, and social context* (pp. 67–93). Psychology Press.

Treviño, L. K., Weaver, G. R., & Reynolds, S. J. (2014). Behavioral ethics in organizations: A review. *Journal of Management*, *40*(1), 123–152.

Urien, B., & Kilbourne, W. (2011). Generativity and self-enhancement values in eco-friendly behavioral intentions and environmentally responsible consumption behavior. *Psychology & Marketing*, *28*(1), 69–90. https://doi.org/10.1002/mar.20381

Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality & Social Psychology*, *87*(2), 246. https://doi.org/10.1037/0022-3514.87.2.246

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 2053951717743530. https://doi.org/10.1177/2053951717743530

Vermeir, I., & Verbeke, W. (2008). Sustainable food consumption among young adults in Belgium: Theory of planned behaviour and the role of confidence and values. *Ecological Economics*, *64*(3), 542–553. https://doi.org/10.1016/j.ecolecon.2007.03.007

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *SSRN Electronic Journal*, *31*, 841. https://doi.org/10.2139/ssrn.3063289

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *SSRN Electronic Journal*, *123*, 735. https://doi.org/10.2139/ssrn.3792772

Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & Society*, *36*(2), 585–595. https://doi.org/10.1007/s00146-020-01066-z

Williams, G. C., McGregor, H. A., Zeldman, A., Freedman, Z. R., & Deci, E. L. (2004). Testing a self-determination theory process model for promoting glycemic control through diabetes self-management. *Health Psychology*, *23*(1), 58. https://doi.org/10.1037/0278-6133.23.1.58