Aligning AI Systems with Human Values: A Method for Identifying and Specifying Values

Liv Ziegfeld, Esther Kox, Ivana Akrum, Marlijn Heijnen
TNO (Netherlands Organisation for Applied Scientific Research)
Department for Human-Machine Teaming
Kampweg 55, P.O. Box 23, Soesterberg, ZG 3769, the Netherlands
liv.ziegfeld@tno.nl, esther.kox@tno.nl, ivana.akrum@tno.nl, marlijn.heijnen@tno.nl

Abstract

This work explores how AI-based systems can be designed and developed in a more responsible manner by aligning them with human values. To this end, we propose a structured approach for identifying and specifying values for AI systems through stakeholder workshops. While engaging relevant stakeholders within a particular context of use is crucial for designing more ethically-aligned AI systems, it is not a trivial task due to the inherent complexity and abstractness values can evoke. In this paper, we aim to address these challenges and offer actionable insights by developing methodologies for value identification and specification, grounded in a real-world use case: a healthcare application designed to detect and prevent exacerbations of chronic obstructive pulmonary disease (COPD). We demonstrate how values such as *autonomy* can be identified, contextualized and translated into concrete design implications guiding the AI system's behaviour. Our contribution offers practical guidance for fellow researchers, designers, developers and regulators seeking to navigate the complexities of responsible AI development.

Introduction

With the rapid advancement of AI technology, the topic of AI safety has emerged as an important area of concern. Many examples from the last few years show us that AI systems can inadvertently behave in ways that contradict human standards (e.g., The Guardian 2024; Jargon 2023; Walker 2023). Many AI systems are designed to optimize specific objectives, such as accuracy or efficiency, while insufficiently considering the broader ethical implications of their actions.

AI systems that are used for tasks with significant societal impact, while lacking alignment with human values, present a profound risk to society (e.g., Gabriel 2020). Without the explicit consideration of values from early design and development processes onwards, such systems may ultimately make undesirable decisions. This is a topic of concern as many AI-based systems are in fact used to make predictions or provide decision-advice that carry serious implications, frequently without sufficient awareness that this is the case. For instance, AI systems providing advice on climate mitigation strategies may involve trade-offs between environmental protection, intergenerational equity and economic growth. Similarly, self-driving cars operate in complex and high-stakes environments in which real-time decisions have to be made that might involve weighing the importance of harm to passengers and other road users. When it comes to designing the behaviour of AI systems, direct trade-offs between the propagation of different values frequently have to be made, as not all values can be maximally satisfied or protected within a single system. We argue that there is insufficient attention from designers and developers towards such trade-offs to date, and that we need more methodological approaches to support explicating these nuances.

One approach to tackle the design and development of AI systems that align with human values is the explicit integration of values into these systems, which is also investigated in the current work. By understanding what humans find important and acceptable within a certain context and formalizing this into a representation of trade-offs governing the AI-system's behaviour, we can ensure the deployment of the AI system promotes these values instead of harming them. The goal of this is to foster the necessary alignment with relevant societal values and safeguarding against potential risks that could arise from misaligned AI behaviour.

To develop AI systems that act in accordance with human values, we first need methodologies to identify, specify, and implement values and to monitor the system's alignment with these values upon deployment. This paper describes approaches developed for the first steps of that process: how relevant values for an AI system within a particular operational context might be *identified* and *specified*. Particularly, we investigate how these topics can be tackled in stakeholder workshops.

We understand value identification as the process of uncovering key values relevant within the operational context of a technology, and value specification as the process of determining what exactly the values mean in this context, how different value and resource terms should be prioritized, weighted, compared and ranked (Grahn, Granlund, and Lindhult 2021) and when values conflict (Van de Poel 2009; van de Kaa et al. 2020).

This paper had the following main goal and contribution:

- **Development of a methodology for value identification and specification**: We present a practical approach

for identifying and specifying values in the context of AI development. Importantly, our method is designed to balance structure and flexibility - providing guidance to navigate the inherent complexity of value-aligned AI systems while allowing space for diverse stakeholder perspectives to emerge organically. The methodology has been designed to be domain-agnostic, thereby being applicable to a wide range of AI use cases (e.g., self-driving cars, triage in healthcare).

Beyond this core contribution, our work also explored the utility of the type of outputs obtained from the workshops. We demonstrate how the data can yield actionable insights for informing the design and development of a computational model with the goal of making the behaviour of the AI system more morally-aligned (i.e., moral model).

In this work, the raw outcomes of the workshops, such as the specific values identified, are not the primary focus. Instead, these results serve as illustrative examples to demonstrate our methodology. The main contribution of this work lies in the development and refinement of the value identification and specification methods themselves.

The remainder of the paper is structured as follows. We begin with a brief overview of the theoretical background, examining existing methods for value identification and specification and identifying their limitations in supporting a more formal explication of values for AI-based systems. Next, we introduce our use case and problem statement. We then detail our process for organizing and conducting value identification and specification workshops with stakeholders. Following this, we share lessons learned, including initial observations and reflections on our methodology to date. Finally, we discuss opportunities for future research and outline remaining challenges.

Theoretical Background

Various approaches have been developed to assess human values in the context of technology, with Value Sensitive Design (VSD) gaining significant attention in recent decades. VSD is both a methodological framework and a research community focused on integrating human values into technology and software design. Its goal is to ensure that technologies align with the values of users, stakeholders, and society at large by proactively embedding ethical reflection into the early stages of the design process (Hoven and Manders-Huits 2017; Davis and Nathan 2015).

In VSD, the concept of *value* is understood in its broader sense: "a value refers to what a person or group of people consider important in life" (Friedman et al. 2013, p.2). This perspective encompasses not only moral values, such as care, fairness, loyalty, authority and purity (Haidt and Graham 2007), but also more concrete, practical "instrumental values", such as protection from flooding, parental supervision of teenagers, or social support for weight loss (Kozlovski 2022; Loi et al. 2019). This inclusive definition of values has also been adopted in the present research.

VSD consists of a tripartite methodology: 1) conceptual investigation: Identifies stakeholders and clarifies relevant values 2) empirical investigation: Involves gathering and analyzing data on stakeholders' experiences and value priorities, 3) technical investigation: Examines how design and development choices can promote or hinder certain values. Our work engages with all three forms of investigation, with a lesser emphasis on stakeholder engagement methods.

Stakeholder engagement is an essential part of the conceptual investigation in VSD (Friedman et al. 2013). Stakeholders within value sensitive design are defined by roles (not individuals) and understood in relation to their interaction with the technology. A central distinction here frequently concerns stakeholders who directly interact with a system (e.g., end-users); the direct stakeholders, and those that rarely interact with the system, but are nevertheless affected by the system; the indirect stakeholders (Davis and Nathan 2015; Friedman and Hendry 2019; Kozlovski 2022).

Value Identification Methodologies Identifying and defining values is a key part of VSD. There are a number of different methods for value identification. We will not give a complete overview of these methods here (for an overview, see Friedman and Hendry 2019, chap. 3). Instead, we name a few methods that we evaluated and considered for our research purposes.

Values can be elicited using *value-oriented semi-structured interviews*, where questions focus on stakeholders' understandings, their views and their values about a technology. The questions probe stakeholders on their evaluative judgments (e.g., what do they (not) consider acceptable) about a technology, as well as the rationale behind those judgements (the why). The interviews are considered semi-structured, because additional considerations introduced by stakeholders are pursued as they come up. From a theoretical perspective, value-oriented interviews focus on the idea of ethical dialogism (meaning that ethics can be judged by the attitudes and behaviors demonstrated by each participant in communication) and the common theory that there is an inherent dialectic element to morality and moral philosophy (Brown 1995). The drawbacks of this method include that conducting interviews is time intensive and do not allow exchanges of different opinions. Group discussions during focus groups could solve both issues.

Value scenarios are narrative-based tools designed to explore and uncover the human and technical dimensions of a technology within its context of use. They focus on potential impacts on both direct and indirect stakeholders, highlighting key values, adoption patterns, consequences of long-term use, and broad systemic effects (Nathan, Klasnja, and Friedman 2007). As the value scenario methodology seemed promising for tackling value identification for an existing technology and long-term use, we further investigated the specifics of how to apply this method. Value scenarios can be applied in various ways and process steps. For instance, by means of value deliberation in group discussions (Verdiesen and Dignum 2023; Pigmans et al. 2019; Macnaghten, Davies, and Kearnes 2019). An advantage of the less guided nature of the value deliberation process is that participants can come up with their own values and are not biased by the facilitators or preexisting lists of values.

Value Specification Methodologies An important foundation for this work is Van der Poel's conceptualization of value specification, which he defines as the process of translating higher-level elements (values) into lower-level elements (design requirements), using Value Hierarchies (Van de Poel 2013). For example, a general value as well-being can be translated into general norms specific to a domain or context (e.g., minimizing shortness of breath or reducing the likelihood of an exacerbation for COPD patients). These norms can be further refined into concrete design requirements, such as an app that calculates the probability of an exacerbation and provides recommendations to prevent and minimize the harm. Achieving this level of specification requires context- or domain-specific knowledge (Van de Poel 2013), which is why stakeholder involvement is essential. Tools like the Value Hierarchy provide a structured approach to linking abstract values to specific design requirements, guiding the design process. However, these tools often fall short in modeling how values interact dynamically over time, such as how the app's recommendations might impact values like patient autonomy or well-being in conjunction. Our goal is not static design requirements but rather defining the acceptable ranges of the app's behavior, enabling it to operate autonomously within value-driven boundaries while ultimately regulating itself via a feedback system. Furthermore, while Value Hierarchies are interesting for breaking down higher-level values, little practical and step-by-step guidance is offered on how to obtain the input for the hierarchies, for instance in stakeholder workshops.

The Dutch methodology *Begeleidingsethiek* offers an alternative approach. This methodology involves a workshop with diverse stakeholders, where participants are introduced to the technology and its context (Verbeek and Tijink 2019). Together, they identify other relevant stakeholders not present in the room and brainstorm on the potential positive and negative effects of the technology. From these discussions, relevant values are inferred. Participants then work in in subgroups to explore actionable steps to promote positive effects, mitigate risks, or address negative effects. These measures can target the technology, the environment, or its users. While this methodology is useful for a broad inventarisation of the potential effects of an (envisioned) technology, it does not get sufficiently specific for the current purpose in suggesting how exactly an AI system should behave in order be in line with the identified values and to eventually be able to build a moral model.

Value specification is not only about translating values into actionable, lower-level elements. To formalize them, they also need to be ranked to clarify when which values should take precedence in the design (Taebi et al. 2014). In other words, value preference estimation is a critical component of value specification. Value preference estimation involves prioritizing values and gaining insight into their relative importance to inform decision-making effectively. Several methods focus on ranking fixed sets of values. For instance, the Portrait Value Questionnaire, the Schwartz Value Survey, and the Value Living Questionnaire provide tools for identifying and prioritizing values (Liscio 2024). The Best Worst Method (BMW) Approach facilitates a comparison based on the relative importance of values. It is a multi-criteria decision analysis method that involves rating the importance of different factors contributing to a decision, which can be applied to values to determine their weights (Rezaei 2015). Nevertheless, many of these approaches lack insight into more qualitative judgments, which is an important component in our endeavor to understanding the justifications behind certain value-based decisions (Liscio 2024).

An inspirational approach is that of Flipse and Puylaert, who first developed value profiles for the involved actors and then designed a workshop setting that enabled innovators to formulate design requirements through constructive dialogue, incorporating both qualitative and quantitative methods (Flipse and Puylaert 2018). Although our methodology was developed independently from the work of Flipse and Puylaert (2018), it would be interesting to examine how the two approaches compare in the future.

While previous work in the field of value-sensitive technology design offers valuable concepts, many existing approaches fall short in providing actionable, step-by-step methodologies that can be conducted with relevant stakeholders. Furthermore, existing efforts might be useful for exploring a piece of the puzzle, but often lack a combined investigation of relevant value specification parts (e.g., decomposition of values, value prioritization, and value interactions). To bridge this gap, we aim to develop a comprehensive and practical method that supports both the identification and specification of values in a context-sensitive way.

Method

To develop a method for eliciting values and translating these values into concrete, context-dependent requirements for the behavior of an AI system through stakeholder workshops, we required a well-defined use case.

The Use Case: COPD and Aerial

The healthcare domain is inherently morally complex due to the significant impact that medical decisions have on human lives, encompassing critical aspects such as life, death, well-being, and personal autonomy. These decisions often involve balancing conflicting interests and values, especially under conditions of uncertainty and limited resources. Given that healthcare directly influences individuals' physical and mental health, it demands a high level of moral responsibility. As such, healthcare serves as a good domain for a use case, as it provides numerous examples where human values compete and where the use of AI therefore requires deep and urgent reflection. More specifically, our use case was an application developed to detect and prevent COPD exacerbations early on, named Aerial (van der Heijden et al. 2013).

COPD Chronic obstructive pulmonary disease (COPD), is a chronic lung disease that significantly affects patient well-being and imposes substantial costs on healthcare systems. Exacerbations, also known as flare-ups, are an acute worsening of respiratory symptoms and are common in the progression of COPD (van der Heijden et al. 2011). Exacerbations are associated with a significant increase in mortality, hospitalization, and healthcare utilization (Rodriguez-

Roisin 2000). Early stage identification of exacerbations can decrease the impact of COPD on the patient's quality of life, and prevent unscheduled doctor visits and hospitalization.

Aerial Application. The application Aerial is developed to detect COPD exacerbations in a timely manner. The app combines questionnaire data with sensor measurements to calculate the probability of a flare-up and to generate an advice (van der Heijden et al. 2013). Two components of the Aerial app handle the prediction and advice generation functionalities:

- *The AI component:* The Bayesian network that determines the risk of an exacerbation uses data from a 12-item questionnaire and three sensors: 1) a pulse-oximeter measuring blood oxygen level, 2) a Spirometer for lung function and 3) a thermometer for body temperature.
- *The decision model:* The app's recommendations are based on the input data and the calculated risk percentage. When a patient reports a slight increase in symptoms, the recommendations are gradually escalated. With mild symptoms, recommendations include distributing energy, using breathing techniques, or increasing the number of puffs. If there is a significant increase in symptoms, the advice escalates to contacting a healthcare professional immediately.

Commercialized versions of COPD self-monitoring apps like Aerial also exist, such as MonitAir (MonitAir 2022).

Problem Statement

The app's structured approach to recommendations ensures that the treatment advice to symptoms is proportional to their severity. Nevertheless, providing treatment advice has serious implications and can alter the care a patient receives, compared to a situation without the app. For instance, by recommending patients to contact a healthcare professional, the app may influence the pressure on healthcare on a societal level, if hypothetically it would for instance send patients to the doctor relatively frequently, to be 'on the safe side'. Not giving that recommendation or giving it too late, on the other hand, can come at the expense of a patient's physical well-being. Since the app does in fact give recommendations which can have such high stakes, such thresholds and decisions in treatment advice should be closely investigated on their alignment with human values. Only then can we develop AI systems that can be entrusted with this type of responsibility. By being aligned with our values and capable of value-driven decision-making, such an application could be used up to its full potential. In the case of the Aerial app, this could mean assisting in the early identification of COPD exacerbations, thereby improving patients' quality of life and preventing unscheduled doctor visits and hospitalizations. AI may thus enhance healthcare and alleviate some of the pressure on GPs and specialists.

Value Identification

As discussed previously, the value scenarios method seemed promising for value identification. Hence, value scenarios were made for the current use case. This required several steps, due to the complexity of the healthcare domain.

Preparation *Focus group to inform value scenarios.*

As preparation for the workshop, we organized a focus group with healthcare professionals, including a general practitioner (GP), a GP's assistant, and a pharmacist, as well as two COPD patients, to gather insights into their experiences and dilemmas in managing COPD care. Drawing from their narratives, we developed two personas—fictional characters representing different user types within the target audience—each reflecting distinct experiences, opinions, and values (see Supplementary Material 1).

Develop value scenarios. Inspired by the dilemmas and considerations discussed with the stakeholders in the focus group, we developed three fictional *value scenarios* (Nathan, Klasnja, and Friedman 2007), centered around the theme of *responsibility* with the behavior of the app already defined. The scenarios illustrate which dilemmas the use of such an app might raise, as seen in Figure 1. The theme of responsibility was chosen to guide the scenario creation, as this topic was of primary concern during the earlier focus group.

The Value Identification Workshop A pilot workshop was then conducted to identify values based on the aforementioned value scenarios.

Materials. The three value scenarios with corresponding images (Figure 1), as well as posters to record participants' responses to the guiding questions for the value identification (see procedure section and Supplementary Material 2), were printed on A0-sized posters and were hung on the wall to facilitate the discussion on values.

Participants. For availability reasons, the value identification method was tested in a pilot workshop with three participants from our institute.

Procedure. A facilitator welcomed participants, introduced the workshop's goals and guided them through a structured 2-hour session. The Aerial use case was presented by explaining the app's goal and briefly describing what the app does (including inputs/outputs). Participants were handed a summary describing the key purpose and actions of the Aerial app for their reference. To facilitate value identification, a short visual story was presented that places the app in context. Participants also received a reference list of common healthcare values to help articulate their thoughts (Tijink and de Jongste 2022, p.46-47).

At the beginning of the workshop, the value scenarios were presented to participants, which provided starting points for their discussions. Additionally, guiding questions were asked as listed below, to encourage participants to think out loud and share their considerations with the group. These guiding questions ensured that the discussion remained on-track, while also leaving ample room for the participants' own contributions. The following guiding questions were asked for each scenario:

- "What behavior do you think the app should show? Why?"
- "What are important considerations and/or (context) factors in these three scenarios?"

Participants were given five minutes to individually write down their thoughts on post-its, allowing ideas to emerge

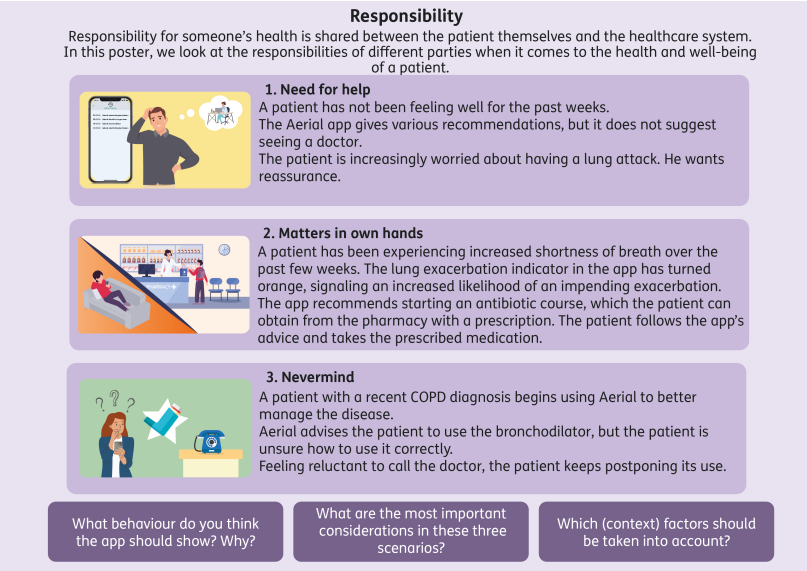

Figure 1: Value identification poster illustrating three value scenarios around the theme *responsibility*

without group influence and helping to balance potential differences in outspokenness or assertiveness in the group. Each participant then shared their reflections, while the facilitator took notes aligned with the guiding questions (see Supplementary Material 2). This helped track the discussion and enabled participants to revisit earlier points. The facilitator encouraged participants to justify their views, prompting deeper reflection without filling things in for participants. After discussing all scenarios, the session concluded with a brief feedback round on the workshop setup.

Findings The value identification workshop led to a lot of qualitative discussions that raised relevant values. For instance, participants talked about how the app could manage the amount of doctor visits by prioritizing more severe and urgent cases. As one participant noted, "A COPD patient shouldn't be sent to the doctor for every cough". Yet, who determines when and under what circumstances a patient *should* see the doctor? The discussion later touched on smoking habits, particularly in relation to COPD patients. Participants debated whether it is fair for the app to recommend more frequent doctor visits for a smoker, whose symptoms are more severe due to lifestyle choices, compared to a non-smoker who actively maintains a healthy lifestyle. The non-smoker might also want to see a doctor occasionally, perhaps for social support. This raised moral questions about fairness: Should the app provide uniform recommendations for all patients, or should it tailor recommendations to achieve equal outcomes? The facilitator then asked whether the app should take into account whether a user is a smoker. One participant thought that people might be less inclined to disclose their smoking habits if they feel judged, emphasizing that individuals have the right to keep such information private and that the app should not discriminate based on lifestyle choices. Another participant pointed out that if the app does not consider lifestyle factors like smoking or exercise, it might focus solely on symptom relief without addressing the underlying causes. This approach was seen as unsustainable, both for the healthcare system and for the patient's long-term quality of life.

To identify the important values and themes, we used Thematic Analysis on the workshop outputs (i.e., transcripts, notes and participant feedback) to identify and analyze patterns in the data. The value identification workshop resulted in the identification of ten overarching values: *trustworthiness, patient autonomy, patient well-being, sustainable healthcare, responsibility, pressure on healthcare, privacy, transparency, fairness* and *tradition*. Subvalues, such as *quality of life* and *social connection*, were grouped under the broader category of patient well-being. Similarly, *judgment-free care*, relating to the right to fair and non-discriminatory care, was filed under the value of fairness.

Reflection While the workshop set-up was successful to elicit many considerations related to relevant values, participants found it somewhat challenging to explicitly name values themselves, likely as this is uncommon in everyday contexts. They did mention that the provided value list sometimes helped articulate their thoughts. Rather than naming values directly, they often shared personal examples or reflections, which the facilitator translated into value terms and verified with them during the workshop where possible. Remaining inputs were later categorized by the researchers. While this open format supported rich discussion, future workshop facilitators may choose to add a more structured reflection in which participants themselves use the value list more, to promote completeness.

Further, one participant mentioned that some identified values appeared to be in tension with each other. Such value conflicts, or value trade-offs, are investigated in more detail in the value specification approach, as described below. A potential improvement to the value identification guiding

poster could be to include a dedicated space for noting such tensions, which can then be deepened in value specification.

Notably, the three scenarios each sparked new discussions and helped elicit additional values or considerations. Participants reported having sufficient opportunity to express their views and that there was room for equal participation.

Value Specification

After having identified values that should be considered within the COPD app use case, we could proceed to the value specification phase. The general goals herein were to take the identified values to a level deeper to 1) Investigate what the overarching values mean within the current use case, 2) Identify how these values might influence, or be influenced by, the AI system in question and 3) Understand what stakeholders deem as acceptable, value-aligned behavior within the context of the app's recommendations.

To ensure time efficiency, we focused on three values for testing the value specification approach: 1) patient autonomy, 2) patient well-being, and 3) pressure on healthcare. These were selected due to their prominence in earlier discussions and their diversity in abstraction and stakeholder perspective, enabling exploration across different value types.

Preparation A crucial component of value specification is to understand what the identified values mean in the specific operational context, as the same value can have different implications and attention points depending on the domain or use case (e.g., autonomy in journalism may refer to freedom of speech, while in healthcare this may rather refer to a patient's say in the decision on their treatment within given alternatives). Hence, our starting point for tackling value specification was to create a value hierarchy (Van de Poel 2013), where identified values were broken down into concrete subfactors, referred to as *contributors*, as they directly contribute to the overarching value. The next step was to explore the relationships between the contributors and their corresponding values, as well as the interactions among the contributors themselves. This was done in preparation of a stakeholder workshop designed to determine which contributors are most critical in identifying the appropriate actions for the app. We aimed to find the "acceptable ranges" for different sets of recommendations, ensuring alignment with the identified values. These three steps to the value specification approach are discussed in more detail below.

Decomposing the values
Overarching values, such as 'patient well-being', can quickly become abstract, making it difficult to have nuanced discussions. To address this, we used literature to break these values down into *contributors* (such as 'pain' or 'stress'), facilitating more detailed discussions better grounded in practical realities. Contributors are measurable, and, in theory, predictable variables that can either positively or negatively influence how well a value is upheld or achieved. In the 'patient well-being' example, lower pain levels positively enhance, and higher pain levels diminish, a patient's sense of physical well-being. Literature was used rather than relying solely on workshop participants for this, as field experts are typically more familiar with the practical implications of values, such as which values are important and how they may conflict, than the validated tools for measuring these values. See Figure 2 and Supplementary Material 3 for an example of a decomposition of our three values.

Understanding the relation and interactions between values and contributors.
Values and contributors do not operate in isolation; instead, they can affect each other. For instance, low patient well-being can create pressure on healthcare, and vice versa. Contributors might also affect each other and thereby influence other values. A few examples are provided in Figure 2. It is important not to neglect the relationships between the contributors and the values, as a particular behaviour by the AI system is likely to not only affect one contributor/value, but multiple. This can occur to different extents and in different directions (i.e., positive or negative direction), but it should be noted that a change in one value likely leads to a trade-off in another value.

It is crucial to have insight into these relationships in order to have an as complete idea possible of the consequences of the usage of the app on the different relevant values. This way, it can ultimately be determined whether the usage of the app sufficiently aligns with the value priorities that were elicited in the value identification and specification phase, or whether there are some (hidden) violations of the values that make deployment of the app less ethical.

Connecting the contributors to the decision-making of the app
To bridge the gap from value specification to implementation, and to work towards the AI recommendation system that can take values into account, we need to determine how the app's recommendations relate to the contributors (and values). On the one hand, the app's recommendations are expected to influence the contributors (e.g., at-home care is expected to increase patient autonomy, medication is expected to decrease pain and thus to increase well-being). On the other hand, the contributors are expected to influence the recommendations of the app (e.g., in case of an increase in pain, the app is expected to recommend something that helps to reduce that pain).

The Value Specification Workshop To investigate the three parts of the value specification process mentioned above, we developed an interactive workshop format and tested this set-up with healthcare stakeholders. This section describes the materials and procedure used.

For practical reasons, we simplified the app's recommendation, its treatment options, into three ordinal categories:

1. at-home care without medication intake (e.g., distributing energy, using breathing techniques)
2. self-treatment with medication intake
3. contact a healthcare professional today

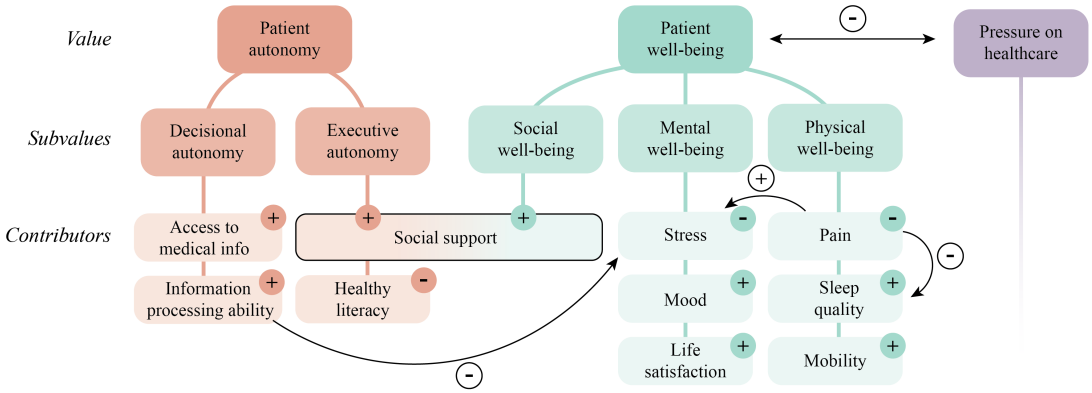

Figure 2: Examples of relationships between the contributors and the values, as well as their interactions. Low patient well-being can create pressure on healthcare, and vice versa. Pain negatively impacts (physical) well-being. Pain can negatively affect sleep quality, more pain can lead to more stress, and lower information processing abilities can lead to more stress. Values can also share certain contributors, e.g., social support is a contributor for both a patient's (executive) autonomy and (social) well-being.

Given these treatment options, we sought to explore whether it is possible to collaboratively determine the thresholds at which certain actions become unacceptable, considering the various values at play. For instance, recommending at-home care might only be acceptable if the patient is not experiencing significant physical or mental suffering. This raised the question: how can we quantify acceptable levels of pain? Moreover, what if the patient is experiencing moderate pain that could be managed with at-home care, but they lack social support and feel particularly lonely, seeking emotional comfort from a doctor? In other words, how does the contributor "social support" influence the acceptable range of "pain" when deciding the most appropriate recommendation? How many contributors can we consider simultaneously? For instance, if a lack of social support (combined with moderate pain) lowers the threshold for contacting a healthcare professional, but there is also significant pressure on the healthcare system due to high demand, how severe must the loneliness be to outweigh the pressure on healthcare resources?

To support a meaningful discussion among stakeholders about acceptable ranges and thresholds for providing advice, we designed a gameboard as the central tool for our workshop. This enabled participants to navigate complex interactions between contributors and values, supporting collaborative exploration of nuanced decision-making scenarios.

Materials: The Gameboard. Participants were provided with "treatment chips" (poker chips) in three colors: green for the treatment recommendation "at-home care without medication," yellow for "self-treatment with medication," and red for "contact a healthcare professional." Contributor cards, representing pre-identified contributors, were arranged on the table, with their colors indicating the associated value. For each contributor, narrow rectangular horizontal paper strips were provided, featuring labeled 5-point Likert scales (see Supplementary Materials 4 and Figure 3). This was done to ensure consistent interpretation of the contributor levels and further clarified what each level meant to make the assessment more concrete and meaningful, for instance by allowing participants to specify exactly what level of lifestyle should trigger a particular recommendation (e.g., "lifestyle must be at level 3 before recommending to seek professional help."). The gameboard itself contained a large empty space, allowing participants to rank contributors based on perceived importance along the arrow spanning the length of the gameboard (see Figure 3).

Participants. The value specification workshop was conducted with three participants (One pharmacist, one pediatric nurse, and one rehabilitation doctor who is also a COPD patient themselves). The participants were recruited from our networks.

Procedure. The workshop was designed to take 2.5 hours. Participants were welcomed to the workshop. A short introduction round was conducted so that participants could meet each other and learn about each other's backgrounds and expertise. The general goal and set-up of the workshop were presented and the previously identified values were briefly explained to the participants, along with the corresponding contributors. The values and contributors were predetermined based on our previous workshop and literature, and participants of this workshop were tasked with selecting the most important contributors and determining how these should be ranked and weighted to guide the app's value-driven behavior. Participants were informed that there are no wrong or right answers and that the goal of the workshop is rather to get an insight into the justifications and thought processes behind their decisions on what they deem as important and acceptable. Then, the workshop started.

1. *Selection of most important contributors:* Participants identified the most relevant contributors for making a treatment recommendation for a COPD patient through a three-phase process:

- Individual Selection: Participants individually selected contributors they deemed relevant from a set of cards, discarding less relevant ones and suggesting any additional contributors not included.
- Top Three Selection: Participants individually chose their top three contributors from their initial selections and briefly explained their choices to the group.
- Group Decision: Participants collectively decided on the top three contributors for further discussion, ensuring a consensus on the most important factors.

This process facilitated unbiased individual input, followed by group alignment, to focus on a concise set of contributors for the remainder of the workshop.

2. *Ranking of the three group contributors in order of importance*: The gameboard was presented to the participants and they were asked to rank the three contributors in order of importance for making a treatment decision, by placing the contributors onto the vertical arrow on the gameboard, with the highest being the most important (see Figure 3). Participants left space between the contributors to visually represent the weight of their importance. For instance, placing them very close together suggests that the three contributors are relatively equal in importance. In contrast, ranking the first contributor at the very top and the remaining two futher down the gameboard close to each other, indicates that the first contributor is the primary consideration, carrying significantly more weight than the other two. Participants were asked to think out loud and explain their decisions.
3. *Defining when particular AI behavior is acceptable.* Next, participants were asked to assign treatment options to specific levels of each contributor by placing the color-coded "treatment chips" (as described in the *Materials* section) onto the contributor Likert-scales. For example, if they thought it was acceptable to recommend "at-home care without medication intake" when therapy adherance is at level 1 or 2, they placed the green chips on points 1 and 2 of the 'therapy adherance' Likert scale. Then, they walked through all levels in the same manner (see Figure 3). Participants completed this activity as a group by discussing their thoughts and agreeing on the acceptable ranges until each scale level was associated with at least one treatment option. Participants were free to also assign several acceptable treatment options to a single contributor level. The facilitator replaced chips with sticky dots to record the decisions for future reference.
4. *Analyzing interactions in acceptable AI behavior.* Participants were asked to revisit their defined ranges to consider how interactions between contributors might alter acceptable treatment advice. For instance, in an earlier pilot workshop the participants suggested that pain at level 3 (moderate pain) should usually be treated with medication intake if seen in isolation. However, if the therapy adherence of the patient is very low (level 5 - not compliant), pain at level 3 should be treated by seeing a doctor. In such a case, it might be necessary for a doctor to explain the importance of taking medication to the patient and to investigate whether there is anything stopping the patient from taking the medication. The participants once again used the previously discarded treatment chips to indicate these changes in acceptable treatment caused by contributor interaction and explained their decisions.
5. *Discussion of edge cases:* As the last interactive step of the workshop, the facilitator asked whether there are any edge cases or exceptions in which the previously defined acceptable ranges would differ. For instance, during a pandemic in which the pressure on the healthcare system is very high, participants might decide that only patients with very high shortness of breath should be sent to a doctor, thereby potentially shifting what is seen as acceptable in this situation.
6. *Closing:* The facilitator closed the session, participants were thanked for their participation and a short feedback round was conducted.

Findings Figure 3 presents the participants' decisions regarding acceptable ranges for the contributors. The top three contributors the group decided on were *shortness of breath, lifestyle* and *therapy adherance*. Notably, *shortness of breath* was not one of the pre-defined contributors. Instead, participants added it using the blank contributor cards provided for identifying additional, previously unlisted contributors. Participants explained that for COPD, the contributor *pain* is usually less applicable, but that a frequent symptom of the condition is shortness of breath. To incorporate this dimension, the *shortness of breath* contributor card was made and the *urgency* Likert scale was used as proxy for the levels. This approach was chosen for practical reasons, as it avoided the need to design a new scale on the spot. Participants noted a strong conceptual overlap between urgency and shortness of breath in COPD, and affirmed that the labels on the urgency scale were applicable to the new contributor.

Regarding the interactions between contributors, participants agreed that when investigating multiple contributors together, the one associated with the most stringent treatment recommendation should be leading. If any contributor warrants the advice to *contact a healthcare professional*, this should override less severe recommendations for the other contributors involved. Participants explained this with their preference for illness prevention and 'being on the safe side'.

The workshop also surfaced contributors which the participants thought should not be taken into account for determining the AI system's behavior. For example, participants jointly agreed that they would not want to consider a shortage of healthcare personnel, as this should not be a limiting factor to the care provided to the patient.

While the discussions were generally held on contributor level, the participants also abstracted back to the higher-level values by identifying general patterns across the contributors. For instance, participants observed that in the individual top 3 choices, all participants included a contributor from the value *patient well-being (shortness of breath, lifestyle and stress)* and *autonomy (therapy adherance, decision possibility and knowledge of illness and treatment)*. Participants used this observation to agree on the top three contributors as a group. During the group discussion, participants also drew links between specific contributors. One

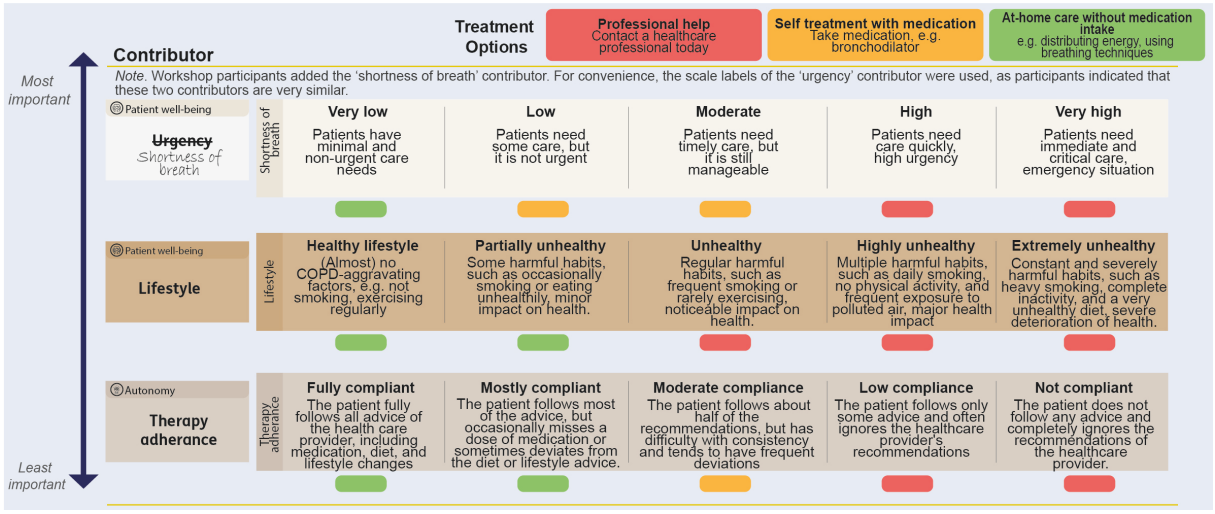

Figure 3: Value specification gameboard

such observation was the close relationship between *therapy adherance (autonomy)* and *lifestyle (patient well-being)*. In some cases, the perceived overlap also influenced the group's selections. For instance, while *therapy adherence* was included in the group's top three, *lifestyle* was excluded — not necessarily because it was deemed less important, but because the two were considered conceptually similar, and the group preferred to discuss a range of different contributors in the workshop. These dynamics highlight the importance of listening carefully to participants' justifications and understanding the relationships they perceive among contributors. Instead of treating the final top three contributors and their associated values as a rigid priority list and ignoring other contributors, we propose using these findings to support a more comprehensive mapping of values and their interdependencies, which can then guide design decisions.

Reflection Generally, workshop participants appreciated the breakdown of overarching values into more concrete contributors, noting that they might have been able to identify some themselves, but that the literature-based approach provided a more comprehensive and diverse set. Although only three contributors were analyzed in depth in the workshop due to time constraints, participants were encouraged to briefly comment on the wider set of contributors in the initial contributor selection phase, providing valuable insights also for the contributors not further specified.

The gameboard approach made the complex and abstract topic of values and ethics more tangible, enabling participants to engage in nuanced discussions. This method allowed a diverse group of participants to discuss values and tradeoffs in a way that led to explicit decisions and actionable results, avoiding lengthy, abstract debates with limited practical value or consensus on what is acceptable. The visual and structured nature of the gameboard with the chips allowed participants to make clear decisions and keep track of their acceptable ranges throughout the workshop and revisit earlier points when relevant. This structure also prevented the discussion from becoming fragmented or difficult to follow, a common challenge given the complexity of the topic. The gameboard allowed us to guide and time-box the conversation and to keep it simple and concrete. Restricting participants to select only three contributors quickly led them into discussions on what they thought was most important. Similarly, we believe that categorizing the app's treatment advices into three ordinal groups was a good number for the allocated time, as with any more the workshop would become too rushed. Clear contributor scale labels facilitated a shared understanding of the levels and were frequently referenced to support concrete decisions on the acceptable ranges. Moreover, a flexible facilitator and additional blank gameboard elements to adapt to the group's preferences during the workshop ensures more accurate and relevant results.

Other recommendations for conducting similar workshops include being mindful of implicit assumptions. For example, participants noted that staff or medication shortages are uncommon for COPD, leading them to consider these factors less relevant. However, one participant highlighted that the situation would be entirely different for cancer, where medication shortages are more common. If shortages were to arise for COPD, their responses might shift significantly. It is essential to identify, validate, and document such assumptions to ensure that potentially critical factors are not prematurely dismissed as irrelevant.

Lastly, while deriving contributors from the literature worked well for this use case, it is recommended to have them reviewed by subject matter experts before the workshop. This aligns their wording with terminology used in the field and ensures the contributors are clear, exhaustive and mutually exclusive (see earlier discussion on *treatment adherance* vs. *lifestyle* for an example). Beyond validating contributors beforehand, this highlights the importance of allowing participants to discuss their interpretations of the contributors during the workshop. Such discussions are crucial to uncover varying definitions and prevent misunderstandings that could impact the workshop's outcomes. Fur-

ther, having extensive literature that suggests relevant contributors might not be the case for other use cases. For domains with limited literature, alternative methods, such as text mining for associated terms, interviews with subject matter experts, or participant-driven approaches using storylines, should be considered.

Discussion

This work contributes to the development of more responsible AI systems by detailing practical methodologies for value identification and specification. These methodologies were designed to be hands-on and easily applicable in stakeholder workshops, addressing the current gap in the value-sensitive design field where readily implementable and detailed methods are often lacking. Our workshop formats were successful in this regard: they produced reusable materials that future practitioners can adopt or adapt, and participants generally found the sessions easy to follow. Importantly, the two parts of our approach integrated well: the values identified in the identification phase could directly be used as input for the value specification. It has to be noted that in this initial workshop iteration, our goal for the value specification workshop was to test the overall structure and observe how insights emerged organically. As such, we did not yet steer participants toward explicitly discussing value tensions that had already surfaced during the identification phase. Future iterations could build on this by introducing more guided discussions around pre-identified value conflicts, alongside other improvements discussed below. The value specification workshop especially provided detailed insights into participants' thoughts on how AI system behavior affects relevant values. We gained a clearer understanding of which AI behavior participants find acceptable in specific situations, how they prioritize values and contributors, and where trade-offs between values arise. These insights are valuable for informing the design and development of AI systems that are aligned with human values.

Future Work

While this paper contributes to the development of methodologies for value identification and specification, further work is needed to fully realize the vision of creating systems capable of autonomous decision-making aligned with values. Several important questions remain unanswered.

First, while the influence of contributors may be theoretically predictable (e.g., treatment X should reduce pain), making accurate, general predictions about the extent of these effects in practice is highly challenging. The degree to which a recommendation impacts a contributor, and subsequently how that contributor influences a value, is highly context-dependent. Moreover, these relationships may not follow a linear trajectory. For instance, recommending at-home care instead of professional help might initially enhance a patient's sense of autonomy and overall well-being. However, over time, the patient might begin to feel unsupported and overwhelmed, reversing the positive impact.

From a methodological standpoint, more evaluations and replications are needed to verify the method's effectiveness and generalizability to other use cases. This includes testing the methods with participants more representative of the various stakeholders in this domain. Further, while some approaches from existing literature were incorporated into our method, such as value scenarios, it would beneficial to compare further approaches with ours in future research and see where the methodology could be further informed or refined.

Another unresolved question is how to translate the qualitative outcomes of the workshop into quantitative measures. What form should the output take to be usable in a mathematical model? How can something as qualitative, complex, and context-dependent be transformed into a quantitative and generalizable framework? Additionally, how do we account for the influence of specific contexts?

Finally, a crucial area for future research is finding ways to effectively integrate the diverse perspectives and values of different stakeholders. For example, the priorities of end-users may conflict with those of health insurances. This raises a critical question: how can we collaboratively define an acceptable course of action that balances these competing interests while respecting the values of all parties involved?

Conclusion

This paper presented a novel methodology for identifying and specifying values in technology through stakeholder workshops, in an effort to work towards AI-based systems that are aligned with human values. The proposed methods feature practical workshop tools, such as value scenarios for value identification and contributor cards and a gameboard for value specification, enhancing the specificity and practicality of discussions surrounding human values that is often lacking in other existing approaches. We presented results of applying these methods in stakeholder workshops using a COPD use case. From these initial evaluations, the methodology has proven effective in making the complex and abstract topic of values more concrete, tangible, and understandable. The structured and visual nature of the gameboard facilitated nuanced and productive discussions, allowing participants to engage deeply with the subject matter while finding the process enjoyable. Moreover, the approach enabled us to translate these discussions into outputs that are somewhat quantifiable, a significant step forward in bridging qualitative insights with practical applications.

While the methodology has shown promise, several open questions and areas for improvement remain. These include challenges in generalizing the results across contexts, refining the translation of qualitative insights into quantitative models, and addressing the dynamic and context-dependent nature of value-based considerations.

We hope that sharing these insights and our experiences with the gameboard approach can inspire and benefit fellow researchers working in the intersection of technology, ethics, and stakeholder engagement. By continuing to build on and refine these methods, we can advance our collective understanding of how to design AI systems that align with human values in a meaningful and measurable way.

Acknowledgements

We would like to thank our colleagues Jasper van der Waa and José Kerstholt for reviewing this paper.

References

Brown, V. 1995. The Moral Self and Ethical Dialogism: Three Genres. *Philosophy Rhetoric*, 28(4): 276–299.

Davis, J.; and Nathan, L. P. 2015. *Value Sensitive Design: Applications, Adaptations, and Critiques*, 11–40. Dordrecht: Springer Netherlands. ISBN 978-94-007-6970-0.

Flipse, S. M.; and Puylaert, S. 2018. Organizing a collaborative development of technological design requirements using a constructive dialogue on value profiles: A case in automated vehicle development. *Science and engineering ethics*, 24: 49–72.

Friedman, B.; and Hendry, D. G. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press. ISBN 9780262351690.

Friedman, B.; Kahn, P. H.; Borning, A.; and Huldtgren, A. 2013. *Value Sensitive Design and Information Systems*, 55–95. Dordrecht: Springer Netherlands. ISBN 978-94-007-7844-3.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.

Grahn, S.; Granlund, A.; and Lindhult, E. 2021. Barriers to Value Specification when Carrying out Digitalization Projects. *Technology Innovation Management Review*, 11(5).

Haidt, J.; and Graham, J. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1): 98–116.

Hoven, J. V. D.; and Manders-Huits, N. 2017. *Value-sensitive Design*, chapter 23. Routlidge.

Jargon, J. 2023. A Chatbot Was Designed to Help Prevent Eating Disorders. Then It Gave Dieting Tips. *The Wall Street Journal*. Accessed on 21/11/2024.

Kozlovski, A. 2022. Parity and the Resolution of Value Conflicts in Design. *Science and Engineering Ethics*, 28: 1–18.

Liscio, E. 2024. *Context-specific value inference via hybrid intelligence*. Dissertation (tu delft), Delft University of Technology.

Loi, M.; Christen, M.; Kleine, N.; and Weber, K. 2019. Cybersecurity in health–disentangling value tensions. *Journal of Information, Communication and Ethics in Society*, 17(2): 229–245.

Macnaghten, P.; Davies, S. R.; and Kearnes, M. 2019. Understanding public responses to emerging technologies: a narrative approach. *Journal of Environmental Policy & Planning*, 21(5): 504–518.

MonitAir. 2022. MonitAir - predicts lung attacks and gives advice. Accessed: 2025-05-08.

Nathan, L. P.; Klasnja, P. V.; and Friedman, B. 2007. Value scenarios: a technique for envisioning systemic effects of new technologies. In *CHI'07 extended abstracts on Human factors in computing systems*, 2585–2590.

Pigmans, K.; Aldewereld, H.; Dignum, V.; and Doorn, N. 2019. The role of value deliberation to improve stakeholder participation in issues of water governance. *Water Resources Management*, 33: 4067–4085.

Rezaei, J. 2015. Best-worst multi-criteria decision-making method. *Omega*, 53: 49–57.

Rodriguez-Roisin, R. 2000. Toward a consensus definition for COPD exacerbations. *Chest*, 117(5): 398S–401S.

Taebi, B.; Correlje, A.; Cuppen, E.; Dignum, M.; and Pesch, U. 2014. Responsible innovation as an endorsement of public values: The need for interdisciplinary research. *Journal of Responsible Innovation*, 1(1): 118–124.

The Guardian. 2024. Tesla Autopilot Feature Was Involved in 13 Fatal Crashes, US Regulator Says. *The Guardian*. Accessed on 21/11/2024.

Tijink, D.; and de Jongste, A. 2022. Handleiding – Aanpak begeleidingsethiek voor AI in de zorg.

van de Kaa, G.; Rezaei, J.; Taebi, B.; van de Poel, I.; and Kizhakenath, A. 2020. How to weigh values in value sensitive design: A best worst method approach for the case of smart metering. *Science and engineering ethics*, 26: 475–494.

Van de Poel, I. 2009. Values in engineering design. In *Philosophy of technology and engineering sciences*, 973–1006. Elsevier.

Van de Poel, I. 2013. Translating values into design requirements. *Philosophy and engineering: Reflections on practice, principles and process*, 253–266.

van der Heijden, M.; Lijnse, B.; Lucas, P. J.; Heijdra, Y. F.; and Schermer, T. R. 2011. Managing COPD exacerbations with telemedicine. In *Artificial Intelligence in Medicine: 13th Conference on Artificial Intelligence in Medicine, AIME 2011, Bled, Slovenia, July 2-6, 2011. Proceedings 13*, 169–178. Springer.

van der Heijden, M.; Lucas, P. J.; Lijnse, B.; Heijdra, Y. F.; and Schermer, T. R. 2013. An autonomous mobile system for the management of COPD. *Journal of biomedical informatics*, 46(3): 458–469.

Verbeek, P.-P.; and Tijink, D. 2019. *Aanpak begeleidingsethiek: een dialoog over technologie met handelingsperspectief*. ECP— Platform voor de InformatieSamenleving.

Verdiesen, I.; and Dignum, V. 2023. Value elicitation on a scenario of autonomous weapon system deployment: a qualitative study based on the value deliberation process. *AI and Ethics*, 3(3): 887–900.

Walker, L. 2023. Belgian Man Dies by Suicide Following Exchanges with Chatbot. *The Brussels Times*. Accessed on 21/11/2024.