AI-assisted ethics? considerations of AI simulation for the ethical assessment and design of assistive technologies

Silke Schicktanz[1,2]*, Johannes Welsch[1], Mark Schweda[3], Andreas Hein[4], Jochem W. Rieger[5] and Thomas Kirste[6]

[1]University Medical Center Göttingen, Department for Medical Ethics and History of Medicine, Göttingen, Germany, [2]Hanse-Wissenschaftskolleg, Institute of Advance Studies, Delmenhorst, Germany, [3]University of Oldenburg, Department of Health Services Research, Division for Ethics in Medicine, Oldenburg, Germany, [4]University of Oldenburg, Department of Health Services Research, Division Assistance Systems and Medical Device Technology, Oldenburg, Germany, [5]University of Oldenburg, Applied Neurocognitive Psychology Lab, Oldenburg, Germany, [6]University of Rostock, Institute for Visual and Analytic Computing, Rostock, Germany

Current ethical debates on the use of artificial intelligence (AI) in healthcare treat AI as a product of technology in three ways. First, by assessing risks and potential benefits of currently developed AI-enabled products with ethical checklists; second, by proposing ex ante lists of ethical values seen as relevant for the design and development of assistive technology, and third, by promoting AI technology to use moral reasoning as part of the automation process. The dominance of these three perspectives in the discourse is demonstrated by a brief summary of the literature. Subsequently, we propose a fourth approach to AI, namely, as a methodological tool to assist ethical reflection. We provide a concept of an AI-simulation informed by three separate elements: 1) stochastic human behavior models based on behavioral data for simulating realistic settings, 2) qualitative empirical data on value statements regarding internal policy, and 3) visualization components that aid in understanding the impact of changes in these variables. The potential of this approach is to inform an interdisciplinary field about anticipated ethical challenges or ethical trade-offs in concrete settings and, hence, to spark a re-evaluation of design and implementation plans. This may be particularly useful for applications that deal with extremely complex values and behavior or with limitations on the communication resources of affected persons (e.g., persons with dementia care or for care of persons with cognitive impairment). Simulation does not replace ethical reflection but does allow for detailed, context-sensitive analysis during the design process and prior to implementation. Finally, we discuss the inherently quantitative methods of analysis afforded by stochastic simulations as well as the potential for ethical discussions and how simulations with AI can improve traditional forms of thought experiments and future-oriented technology assessment.

KEYWORDS

intelligent assistive technology, ethical reflection, simulation, person with dementia, conceptual approach

1 Introduction

In the science fiction movie Dark Star (1974, director John Carpenter), the captain of a starship argues with an artificial intelligence (AI)-controlled bomb about whether it should detonate. The dispute is whether the bomb's decision to detonate itself is based on correct data, namely, a correct order perceived by the bomb's sensory input. The captain's arguments regarding the limits of what the bomb can know about its own existence or intelligence, however, just serve to convert the bomb into a nihilist. Finally, it detonates itself and kills the ship's human crew.

Such movie scenes function like thought experiments, a common methodology in philosophy and science, and can help us anticipate implications or test for the ethical or epistemic coherence (Walsh, 2011) of an argumentation or idea. In the Dark Star case, the risk is that future AI could develop its own morality, with harmful outcomes for humans. However, it is never clear whether the argumentation generated by thought experiments translates to real situations, for example, to AI systems currently under development for dementia care (Schweda et al., 2019). Our goal in this paper is to expand the thought-experiment approach by exploring the opportunities and rationales for using computational simulation as a tool for ethical reflection on human-AI interaction. Our idea is that such algorithmic simulations can augment ethical reflection with empirical, simulated data during the design phase of systems, thereby improving the anticipation of ethical problems in the use of AI technology in various settings.

Following the High Level Expert Group of Artificial Intelligence (2019), we define AI-systems as software systems that analyze their environment and take actions to achieve some goals independently. This general definition does not predefine the type of mathematics and algorithms implemented—e.g., symbolic rule-based or sub-symbolic AI-like neuronal networks—nor does it require specifics on the level of automation (Shneiderman, 2021, *p*. 48).

Ethics in technology development is traditionally guided by general principles that can be employed in thought experiments to test which main principles seem to have what consequences or are more likely to gain public acceptance. The employment of such thought experiments, e.g., trolley-dilemma experiments for exploring ethical aspects of automated vehicles, recently have been criticized as too narrow or abstract (e.g., Goodall, 2019; De Freitas et al., 2021). Furthermore, empirical validity in such thought experiments is often low, and reasoning can be biased by the prejudice of ethicists or technology developers.

Overall, mainstream ethical evaluation approaches regarding new technologies, such as biotech, nanotech or artificial intelligence, tend to conceptualize technology as a mere object of ethical reflection in terms of the "ethics of AI." While this makes sense for biosciences or nanotechnology, it need not be the only way to reflect on AI.

Empirically informed ethical reasoning is a more recently established standard with the potential to significantly reduce bias–including expert bias–and to improve generalizability to real-world situations (Schicktanz et al., 2012; Mertz et al., 2014). Therefore, the social and moral perspectives about, for example, genetic, biological, nano, or AI technology as held by practitioners, stakeholders, and affected persons are collected empirically using qualitative methods. Although this empirically informed approach has some advantages in comparison to traditional, non-empirical methods, it also has some epistemic limitations.

In applied ethics, experts often think of AI as a feature of specific products, be it a feature that analyzes the environment and adapts actions to reach a particular goal or a feature that helps to make (moral) decisions. The dominance of this approach in the discourse is evident from a review of the literature (see below, Section 2). As AI technology often has complex or even hidden outcomes, it has been argued recently that "explainability" and "trust" are essential criteria for ethical evaluation (High Level Expert Group of Artificial Intelligence, 2019; Amann et al., 2020; Coeckelbergh, 2020; Markus et al., 2021; Border and Sarder, 2022). However, explainability and trust focus again on a human-AI interaction, again conceptualizing AI mainly as an end-product and humans as being capable of understanding it. This assumption does not always hold, e.g., when in healthcare and disability settings. Here, one cannot guarantee that the users of technology and the people it affects are able to monitor, interact with, or to understand an AI system's outputs. These individuals cannot "trust" the system because trust requires specific cognitive and emotional features.[1] The aim of our "AI-Assisted Ethics" approach is to anticipate ethical trade-offs and social implications in complex, contextualized settings where criteria such as trust or explainability might not function or are not appropriate. Furthermore, complexity is particularly important in situations where direct anticipation of outcomes and implications is limited, e.g., because the characteristics of people involved are very heterogeneous. In these cases, individuals might interact with the AI very differently, which may turn greatly restrict the generalizability of empirical observations of behavior and values to other individuals.

Our article results from a truly interdisciplinary cooperation between ethicists, social scientists, engineers, and machine learning specialists. It combines insights from various sub-studies that built on each other, with a specific focus on intelligent assistive technologies (IAT) in healthcare settings and especially for the care of persons with dementia. The following proposal has been developed from the comprehensive exchange between these sub-studies. First, we summarize the main strands of the general discussion regarding AI ethics and machines ethics as well as insights regarding the ethics of human-AI interaction in the particular setting of care for older people and persons with dementia (2). As the assessment of impacts of AI-technology in cases such as dementia care is often difficult or risks to neglect the complexity of values and interactions of involved agents, we developed in a next step a conceptual approach to consider AI as a tool (in a simulation) to anticipate ethical and social issues of implementing IAT (3). Here, we focus on the context of dementia care (3.1., see also info box). The concept of such an *in silico* simulation (3.2) considers multiple agents who can interact in various ways with different ethical values. By an 'Ethical Compliance Quantification' evaluation different design alternatives can be quantitatively compared and can inform stakeholder discussions. Hence, results from an exemplary simulation model to test for the developed ethical compliance quantification are presented to illustrate the conceptual approach (3.3). Hereby we construct an example from research in technology-assisted dementia care to discuss the advantages and challenges of this approach. This simulation is informed by value statements drawn from interviews. It utilizes stochastic human behavior models that encompass behavioral data,

---

[1] Therefore, language-based assistive systems to foster a "dialog" between human and machine are neither meaningful nor appropriate.

pre-set values for simulating realistic settings, variables, data sets and setting variables. Finally, we discuss the differences between qualitative and inherently quantitative methods of ethical reasoning and how the simulation approach can enhance ethical reasoning for technology assessment (4.), and provide a short conclusion section (5).

2 AI as a product of technology and as an object of ethical reflection

In the following, we summarize the main strands of discussion regarding AI ethics and ethical machines. We then focus on the ethics of human-AI interaction in a particular setting: "intelligent assistive technology" for the care of older people and persons with dementia.[2]

Various authors have developed catalogs of values and ethical principles to aid in this kind of ethical assessment (Currie et al., 2020; Spiekermann, 2016; Umbrello and van de Poel, 2021; van Wynsberghe, 2013; for an overview, see Schicktanz and Schweda, 2021). Prevalent ethical criteria include self-determination, not harming or actively promoting human welfare, privacy, and sustainability (Hofmann, 2013; Novitzky et al., 2015; Ienca et al., 2018; Vandemeulebroucke et al., 2018). In many of these approaches, some ethical principles are prioritized over others. Although these values are sometimes proposed as guides for design processes, more often they are treated as criteria for assessing existing technologies.[3]

---

2 Weber states in a recent article that "There is no good and generally accepted definition for age-appropriate assistive systems." (Weber, 2021: 29, own translation). Kunze and König (2017):1), Hofmann (2013):390) and the umbrella association of German health insurers (Spitzenverband der Gesetzlichen Krankenversicherungen in Deutschland, 2019:22) report similar findings. In general, according to Ienca and colleagues, the following definition has become accepted: "Assistive technology is the umbrella term used to describe devices or systems which allow to increase, maintain or improve capabilities of individuals with cognitive, physical or communication disabilities" (Ienca et al., 2017:1302; cf. World Health Organization, 2018; Endter, 2021:15; Novitzky et al., 2015:709; Manzeschke et al., 2013:8). Based on this, Ienca and colleagues describe as "intelligent" those assistive systems "with [their] own computation capability and the ability to communicate information through a network" (Ienca et al., 2017:1302). For our purpose here, we assume that the more complex such IAT systems are, the more relevant our considerations regarding the usage of AI for ethical consideration become.

3 In this context, it is important to distinguish between ethics and morality. Ethics is understood as a philosophical reflection about the meaning and justification of various kinds of normative statements, legal practices, or everyday judgements. By contrast, morality is understood as the everyday application of a set of moral principles, e.g., norms and values, in judgment and decision making. Although this underlying set of principles often remains implicit and unarticulated, human agents are usually able to provide a simple explanation of such norms and values upon request (so it is not fully opaque). From these definitions, it follows that if the artificial, automated system cannot reflect and explain its decisions in an appropriate way in varying situations, it should rather be labeled as a moral machine because while it fulfills the criteria of moral decision-making, it does not fulfill the criteria of ethical reflection. By contrast, to describe a machine truly as an ethical machine would, in analogy to human ethical thinking, require that the criteria of "reflection" are fulfilled. This includes at least four components or stages: a) the potential to revise pre-implemented norms, b) the availability of a set of alternative approaches with an understanding of how they differ, c) discussing the pros and cons of revision, and d) providing a final justification of the final conclusion. It is an open question whether the new standard of explainability in AI would satisfy the criteria of ethical reflection or whether it would remain on the level of just making moral criteria comprehensible.

This latter fact might have motivated Ienca and colleagues to call for "a coordinated effort to proactively incorporate ethical considerations early in the design and development of new products" (Ienca et al., 2018:1035). In a recent paper on "embedded ethics," McLennan and colleagues also highlight the need for an "ongoing practice of integrating ethics into the entire development process" (McLennan et al., 2022:3) based on a "truly collaborative, interdisciplinary enterprise" (ibid). This is reminiscent of approaches dating back to the 1990s, when engineers and philosophers started to develop strategies for considering ethical issues and values for the design of human-machine interaction. This has been called computer ethics, social informatics, participatory design, and value-in-design. As Friedemann and Kahn 2007 distinguishes, there exist three main ideas about how values and ethical principles are related to the development of new technologies. In the embodied approach, values are incorporated in technology by the designers. In the exogenous approach, values are determined and imposed by the users after a technology is implemented. Interactional approaches focus on the interaction of designers and users; these include approaches like value-in-design and participatory design. Interestingly, all three approaches can be found in current AI-technology design.

Pertinent questions for this kind of technology assessment are as follows. How should we (or how should we not) use AI/IAT technologies? Are there ethically acceptable risks, do opportunities outweigh risks, or might the use of AI/IAT technologies create conflict with basic human rights and ethical principles such as human dignity, self-determination, or justice? In some fields, such as technologies for the care of older people, it seems that this assessment often takes place after a prototype of the technology has been developed, but not during the design process.

By contrast, the central question in machine ethics is whether AI-technologies that can operate more or less autonomously can and should be constructed to operate in a morally acceptable way. This touches upon ethical questions regarding adequate concepts and standards, as well as on the criteria of morality as such. It encompasses issues of moral agency and responsibility, as well as informatics and engineering questions regarding effective technological implementation through algorithms and "training" (Anderson and Anderson, 2007). This debate differentiates between top-down and bottom-up approaches to the problem of implementing morality-sensitive technology (Wallach et al., 2009). Top-down approaches try to specify moral precepts in a deductive manner by means of the successive specification and application of a set of general moral norms. In this vein, fundamental moral philosophical principles such as the utilitarian principle of utility (maximization of utility) or the Kantian categorical imperative (principle of universalizability of maxims) are operationalized in terms of algorithms that constitute the procedural rules of the autonomous technical system, its "moral modus operandi." For example, van Wynsberghe (2013:411-413) sees a fundamental need to endow care robots (which can be considered a special case of assistive systems) with moral values during the development process. By contrast, bottom-up approaches try to specify moral precepts in an inductive manner by developing moral competences through a series of pertinent moral experiences. An example of this can be seen by the MIT moral machine experiment (Awad et al., 2018) by gathering large data sets of

humans answering online moral dilemma. For AI, this means learning morality. Here, the technical system is not equipped with general rules and a basic moral orientation but is trained rather by repeated confrontations with a variety of pertinent "cases," i.e., moral problems and their solutions, thus emulating the process of human moral development. One might expect such learning processes to result mainly in punishment avoidance in standard learning techniques or to level out practical compromises between different moral opinions. To go beyond this level and reach a coherent ethical framework would require the intellectual capacity to identify new top categories, rules for consistency, and inductive theoretical reflection. This might be beyond the capacities of AI according to some scholars (Brundage, 2014).

Whether moral precepts can be derived through technology and whether deriving moral precepts is a proper and feasible objective of AI has been debated over the last 2 decades (Anderson et al., 2004; Nallur, 2020). Misselhorn (2021) who talks of "algorithm morality" or "artificial morality," favors a "hybrid approach" combing fundamental moral rules (e.g., never harm or kill a human) with AI-based learning of contextual moral rules for interacting with humans (e.g., respecting privacy for person X, and favoring safety issues for person Y). This also allows for the integration of empirical information on actual user preferences. Furthermore, some have proposed to use AI-tools for enhancing human ethical judgment, hence the idea is not to make machines more ethical, but to use AI to improve ethical judgements of humans. For example, Walton (2016) discusses how different methods (Bayesian vs. computational methods) can contribute to testing the plausibility of arguments. This could be also relevant for analytical moral argumentation, even if the author himself does not discuss this option explicitly. Lara and Deckers (2020) provide a theoretical approach to the use of AI as an auxiliary (supportive) system for 'enhancing' human morality. By a Socratic technique, the machine helps the human agent to learn to reason ethically, but the aim is not to delegate decisions to the technical system or to train a system to be compatible with particular values (p. 282). Volkman and Gabriels (2023) build on this idea of 'AI mentors' but stress the need for a 'total' socio-technical system "to operate through a diversity of perspectives challenging one another" (p. 10). Their general idea of support that strives for more complexity and considers many perspectives shows analogies with our still more specific approach. Our approach refers to a specific field of application where we see particular challenges, as described in the following. We focus here on how to improve the process of ethics-by-design by considering the diversity and uncertainty of moral perspectives during the process. This processual focus is consistent with our deliberative participatory ethics background (Schicktanz et al., 2012). We do not claim that it automatically provides better moral outcomes.

A specific field of human-AI interaction in which the human agents involved differ according to 1) their role (e.g., professional vs. informal caregiver), 2) their values regarding care and assistive technology (e.g., privacy over safety), and 3) their cognitive and emotional capacities is tied to technologies for monitoring and assistance of people with physical and mental impairments, e.g., persons with dementia. These technologies are increasingly equipped with different types of AI and therefore also fit under the term IAT (Ienca et al., 2017). As a review by Löbe and AboJabel (2022) revealed, assessments of risks, benefits, and empowerment for persons with dementia often are undertaken when a prototype is introduced in care settings experimentally to test usability,

safety, or social acceptance. Such testing can be understood as *in situ* simulation if the setting is a natural setting or as *in vitro* simulation[4] if conducted in a laboratory that mimics smart homes or care units. *In silico*, noted below, are computational simulations of such settings.

As dementia poses particularly ethical challenges to the use of AI-based monitoring and assistive systems due to limits regarding "classical" informed consent, the possibility of changing values and preferences without clear verbal expression, and the extremely high burden on caregivers, the assessments in this field of application promise to provide highly sensitive insights in fundamental problems regarding the development and use of new technologies in eldercare. In a next step, abstracting the results from dementia care to other, particular sensitive fields of care giving, the approach can be also very fruitful. However, here, dementia is for various reasons (see Section 3.1) a reasonable starting point.

In a previous expert interview study (Welsch, 2022a; Welsch, 2022b; Welsch and Schicktanz, 2022, Abojabal et al., under review) we found that the interviewed experts stressed the fact that providing clear definitions of AI or IAT is difficult. Nevertheless, many interviewees gave specific examples of IATs: reminder systems, orientation systems, smart home applications, and robots. Advanced AI features like machine learning or deep learning is not necessarily a constitutive part of this; existing IAT makes use of traditional algorithms more often. The users and purposes of such IAT have been characterized as quite complex, as these IATs include a wide variety of digital applications which contribute to improving the self-determination, the mobility, the social participation, and, in sum, the quality of life of users. Hence, IAT users are not one homogenous group, but include different, interacting groups–often characterized by having different experiences, values, or preferences–such as people in need of care, family caregivers, other relatives, and professional caregivers. This is an important point to consider for an ethics-by-design approach, as different users may be differently affected and have different moral intuitions about the way IATs should operate. Furthermore, it becomes clear that such technologies can also have multiple goals: self-determination, mobility, quality of life, quality of care, safety, or social participation. This wide range of goals will likely create conflicts during the design phase and in actual use (cf. Schweda et al., 2019; Welsch and Schicktanz, 2022).

While ethics-by-design approaches and, in particular, the participation of future users is often seen as important, however, neurodegenerative diseases—common in old age—pose a major challenge for participatory design approaches, e.g., if people cannot communicate, as in later stages of dementia. This problem is exacerbated by short project durations which prevent the investments of time needed for participation. This points to another serious problem of technology assessment in practice: new technologies are developed, but time and money limitations cut short ethical reflection about their implementation. This can be one motivation for demanding standardized ethical evaluation checklists in technology development. However, the implementation of ethical evaluation checklists and their thorough application appears to be a difficult problem given developers' limited time and the complexity

---

4 See Chandrasekharan et al., 2013 for the differentiation of in-vitro and *in silico* simulations and thought experiments.

of implementation conditions which involve multiple agents with potentially different goals and communication capacities.

Overall, AI is generally thought of as being integral to IAT products, that is, as a feature of the device or system designed with specific end-users in mind. At the same time, a number of practical desiderates and limitations of a classical ethics-by-design approach have become clear. Here we propose using AI as a tool for solving these challenges by presenting an AI-assisted procedural ethics. This problem has also been addressed by Aliman and Kester (2019) who propose to use various socio-technological feedback loops, e.g., by preemptive simulations, to ethically enhance AI technologies.

3 Methodology: model conceptualization

3.1 Premises regarding the need of AI-assisted ethics for supporting IAT development

In our understanding, ethical reflection about modern technologies, including AI, entails taking the following steps: recognition of problems (not only dilemmas); consideration of relevant facts; knowledge of various ethical approaches, principles, and theories to test for alternative conclusions; testing for consistency with accepted norms (does the application of this rule violate uncontroversial norms?); testing for adequacy (can abstract rules be applied to concrete situations without contorting them?); justification of specific decisions (an aspect of explainability), and finally, societal legitimacy of the whole reflective procedure. Such a process can be called "complex" in that it cannot be replaced by a fixed set of values. Most of the above-mentioned approaches start from *a priori* moral intuitions and theoretical generalizations (such as "values").

In the cases where ethical considerations are applied prior to or during the development of a technology, they have to rely on principles that may be too general for concrete design decisions (a limitation of the top-down models noted above). In order to become more relevant for a concrete design decision, ethical issues must rely on analogies from previous situations which are extrapolated to the new situation. This extrapolation is prone to error, but not all errors are evident before product implementation. Obviously, it would be desirable to fill the gap between too general and too specific (but extrapolated) recommendations for ethical design to better adjust to the needs and goals of users, especially when they are vulnerable as, for example, persons with dementia.

Adapting IAT systems to complex settings–characterized by multiple agents with different goals, varying moral intuitions, and different cognitive states and communication skills–during the design phase requires a different approach.[5] The situation could be improved if human ethical reflection would accompany the

design process so that experiments with different designs could be conducted to detect practical moral problems and potential value conflicts. *In situ* experimentation, however, raises other problems. It can be unethical to expose vulnerable people, e.g., those with dementia, to new, prototypical technology. For example, the COACH prompting system intended to assist older adults with dementia with handwashing served only to prompt fear and anger in some cases (Mihailidis et al., 2018). Also the review by Alkadri and Jutai, 2016 concludes that many of such technologies for this target group is weak regarding safety and efficacy (Alkadri and Jutai, 2016). Furthermore, the costs of experimentation can easily exceed available resources. In our field of study, i.e., technology-assisted dementia care, another important challenge needs to be considered: communication between human and AI, now often seen as a solution in which the machine "explains" to humans the criteria used for a decision, is not feasible. In contrast to the scene in Dark Star discussed above, persons with dementia have very varying and limited capacities for effectively communicating with a machine. Nor is this group able to give detailed comments to designers or scholars,[6] hence interactive approaches such as participatory design are limited.

These problems lead us to follow the idea of re-thinking AI as an integral tool of the ethical design process, not just a product of technology. Thus we propose to use *in silico* simulation, which is a computational simulation of the technology in its environment as a proxy for *in situ* experimentation. Ideally, these simulations should encompass multiple human agents, a representation of their goals, an individualized model of their internal decision making (from deterministic to stochastic models), and their environment including the device or procedure under development. Simulations can be run repeatedly at little cost and without unethical exploitation of people. The simulations can serve to assess the effects of a product on agents in a setting while varying inputs. Hence, they would allow reflection on the model-building process (Chandrasekharan et al., 2013:242). Other forms, such as *in situ* experiments or thought experiments, focus on the outcome with questions of ethical acceptability, inefficiency, and safety.

As we suggest using simulation for gaining insight and somewhat oppose it to experiments, it is necessary to briefly reflect on the epistemic advantage of using simulation in our setting. There is a substantial debate on this issue, as there are scholars who significantly question the epistemic benefit of simulations in comparison to experiments and other that take the opposite position (see for instance the positions taken and the sources reviewed in Peck, 2004; Parke, 2014; Di Paolo et al., 2000). The relation between simulation and experiment in general is subject to a multi-faceted discussion (see Winsberg, 2022 for an overview). A simulation of a real-world phenomenon based on a mathematical model of this phenomenon may be considered inferior for two reasons (a) the mathematical model may be deficient or (b) the simulation algorithm may require simplifications (such as discretization) that limit precision, up to

---

5   That this is a complex situation for which the persons involved require training has also been proposed by projects that try to develop simulations of patients with Alzheimer dementia for training facility staff, e.g., "Virtual Patient Simulation Tool for Training Health and Social Care Staff Working with People with Alzheimer's Disease or Related dementia–VIRTUALZ" https://anr.fr/Project-ANR-17-CE19-0028.

6   Such as children, persons with dementia, persons with severe cognitive impairments, or persons with very limited communication skills.

the point of unlimited divergence between simulation results and model content[7]. Considering the first issue, we think that the discussion reflected above does not pertain to the use of simulation we consider. With the simulation model we propose here we do not strive to test existing theories or develop new theories of real-world phenomena. Rather, we suggest to use simulation for analyzing the implications and stepwise construction and explication of a normative model. The normative model we consider consists of ethical value dimensions, the set of events that are being considered relevant with respect to these ethical value dimensions, and mathematical operationalizations of how events are to be quantified (as scores) with respect to values. This mathematical description is the model that is the object of investigation. The method for experimenting with mathematical models indeed is the simulation. This reflects the definition of the term "simulation" already given in Korn and Wait (1978) that a simulation is an experiment performed on a model.

This provides additional clarification to our position as far as this paper is concerned: our main claim is not, that simulation helps to faithfully analyze the real-world effect of an IAT on a set of ethical values. We rather do propose that a simulation is helpful for the iterative development of an ethical value model for IAT design in the first place, for analyzing its implications, and for gaining insight into the consistency or even existence of a value model; we will come back to this in the discussion in Section 4. In addition, we think it is important that the need to provide a mathematical description of the ethical value orientation forces assumptions to be made explicit and thus made accessible to critical review–this is a property that thought experiments do not necessarily have.

Eventually, we also want to develop IAT that provide optimal assistive strategy with respect to a given mathematical model of the values. If one assumes such a value model to exist, this then is conceptually a surprisingly well-defined task, as it can be framed as a standard engineering-level optimization problem. In this paper, we suggest that both tasks can be solved in the same framework. However, as we will see in the example discussed below (Section 3.2.3), strategy optimization may be more sensitive to simulation validity: in strategy optimization, the distribution of events in the state space must adequately reflect the real world in order to correctly identify the optimum. This is the topic of issue (b) identified above. We are confident that the study we discuss below does not fulfill this stronger requirement. However, considering the success of simulation in much more complex situations (see, e.g., Bicher et al., 2021), we are confident that it is possible to build models of adequate validity.

In the following, we give an example of how an AI-assisted simulation can work. The simulation is situated in the field of IAT for dementia care and is a system that guides persons with dementia who have lost their orientation inside a care facility. It illustrates the complexity of the situations that should be considered and what kind of assessment loops are conceivable (see Info Box 1). This example is a conceptual proposition that can be adopted to other

settings. We do not claim that the current model has the optimal structure, parameters, or even sampling strategy.

3.2 A concept for an AI-assisted simulation

Our AI-assisted ethics simulation (Figure 1) comprises the several elements, explained below in turn.

3.2.1 Multiple agents

A multi-agent simulation environment provides a simulated world where simulated agents can interact. "Simulation" means that the state of the world and the state of the agents in this world can be represented by a set of variables in a programming language such that the values of these variables (referred to here as the variable "score"[8]) represent the state of the simulated world at any given time. There might, for instance, be a variable called "location" that contains two scores that indicate the location of an agent in a two-dimensional simulation world. The simulation proceeds in steps, where at each step the set of variables is manipulated according to the rules that define the temporal evolution of this simulated world. There might for instance be a "move" rule that an agent whose "destination" variable contains a position that is not equal to the "location" variable will update its location by an amount of "step length" in direction of "destination." Eventually, these rules are represented as pieces of program code.

The interaction of agents is modeled by rules that depend on (and change) the variables representing the state of two or more agents. In general, simulation environments allow the definition of stochastic rules, whereby the outcome depends on a sampling of some random process. For instance, the step length used in a specific application of the "move" rule may be given by sampling from a normal distribution defined by a mean step length and a certain standard deviation. Specifically, in simulations where agents represent humans, such stochastic rules are important for simulating behavior non-deterministically.

A simulation run is produced by initializing the state variables (e.g., location of the simulated patient and current disorientation state, locations of simulated caregivers, etc.) with pre-defined scores (such as the location coordinates) and then by stepwise advancing simulation until the simulation state fulfills a specified termination condition (such as having reached a certain simulation time point or reaching a certain simulation state). If the simulation uses stochastic rules, different "runs" of the simulation may result from the same initial conditions. Based on many runs, it then becomes possible to analyze the statistical properties of state variables in the simulation and their temporal development by combining the records in the run protocols. For instance, one could estimate the expected number of steps required to reach a given destination from a given starting point by averaging the step counts obtained from multiple run

---

7 Consider the 'Attofox'- problem (Mollison, 1991); but note that this is an illustration of the *opposite* situation: the discretization is more realistic than the continuous model.

8 Note that the term "value" has in our simulation setting two meanings: "variable value" and "ethical value". By "variable value" we mean the quantity stored in a variable of the simulation model. By ethical value we mean normative concepts that have a clear moral connotations and serve for moral orientation, e.g., such as autonomy, freedom, safety, or wellbeing. To avoid confusion, we use the term "score" for variable value although this is not common in simulation modeling.

Info Box 1

Problem statement: IAT system to guide persons with dementia who have lost their orientation inside a care facility.

When residents wander and lose their orientation, it can be a challenge for everyone living and working in care facilities. Hence, an IAT system might help to actively guide patients through buildings. It might lock doors depending on the perceived cognitive state of the patients and on an assessment of safety and privacy. It might also call for human assistance (Landau and Werner, 2012; Ray et al., 2019; Bayat and Mihailidis, 2021; Lancioni et al., 2021).

In nursing homes, residents with limited orientation, for instance due to cognitive decline, often experience a reduced ability to manage their activities autonomously and safely. One obvious problem is getting lost in the nursing home on the way to a destination. Indeed, a substantial amount of care-giver attendance is required for providing guidance to disoriented residents. As a possible IAT for supporting autonomy and safety, one could imagine a "smart bracelet" that detects disorientation, provides orientation cues as appropriate, and calls a caregiver in case the problem persists. Such a system may increase autonomy of residents. It may decrease the amount of caregiver attendance to routine activities and thus free up caregiver resources for socially more salient activities. The benefit of such a system depends on its reliability in detecting an instance where help is required and on the effectiveness of its orientation cues. At the same time, such a system affects different stakeholder values: autonomy and safety for the resident, workload for caregivers, workforce efficiency, nursing quality, and safety regulations for the nursing home operator. It seems reasonable to assume that these values interact with each other. Some may reinforce each other; others may contradict each other. Even this situation can be considered complex, as we have seen in our own empirical research with affected persons. Patients and professionals might differ regarding the criteria of acceptance of such technological guidance (Buhr and Schweda, in prep.; Köhler et al., 2022). Values such as autonomy (of the person with dementia), privacy (of the user but also of other residents), safety, wellbeing, and costs (e.g., professional time) are balanced or prioritized differently across different stakeholders, as an empirical ethics study revealed (Buhr et al., under review; Welsch and Schicktanz, 2022). For example, we identified a group of persons with dementia (calling this type "individual self-determination") for whom disorientation technology should provide directions and guidance but should not inform third persons nor restrict the person's range of mobility. Another type of patient ("relational autonomy") accepts any technology that prevents them from wandering or getting disorientated with the goal of relieving caregivers' burden. They would also consent to having others be tracked or third persons alarmed in cases of disorientation. Thus, we see no empirical justification here for a "one value-profile fits all" approach. Further, the advantages or disadvantages of a technology that gives priority to different values must also be assessed with regard to "realistic" outcomes, potential side-effects, and how free they actually are to select between multiple values (e.g., in light of legal restrictions including liability issues).

---

protocols. The interesting aspect here is that the quantity "expected step count" is not a pre-defined parameter of the simulation but rather a quantity that arises from analysis of multiple simulation runs.

The following thought experiment will help in understanding the usefulness of simulation-based analyses. Assume you are building a new eldercare facility. In the planning process, it would clearly be of interest to see how a quantity such as the expected step count changes in response to modifications of the floor plan or other design elements. The above-described simulation could measure the specific benefit of such changes in terms of any given desirable outcome, such as reducing overall walking times, and can thus provide a quantitative rationale for choosing between corresponding design options in the real world. The use of simulation-based techniques is standard in analyzing the effects of design decisions and exploring what-if-scenarios in a wide range of application domains (such as economic decision making or the analysis of climate change).

Now, focusing on IAT again, we can use this technique also in the specific situation in which the quantities derived from simulations represent the degree of an IAT's compliance with or violation of a set of ethical values. Concerning the aspect of investigating ethics in IAT, this is an interesting shift in perspective from considering how to embed ethical values into IATs to considering how an IAT's actions reflect such values in practice. We call the computation of numerical scores that represent compliance with a set of ethical values ethical compliance quantification (ECQ, see below). As we will discuss below, such an approach is not only interesting because it might provide relevant information for the ethical assessment of an IAT. It also requires all assumptions to be made explicit in order to render them computable and is therefore also an interesting mechanism to discuss and investigate the design and effect of value structures in specific use cases.
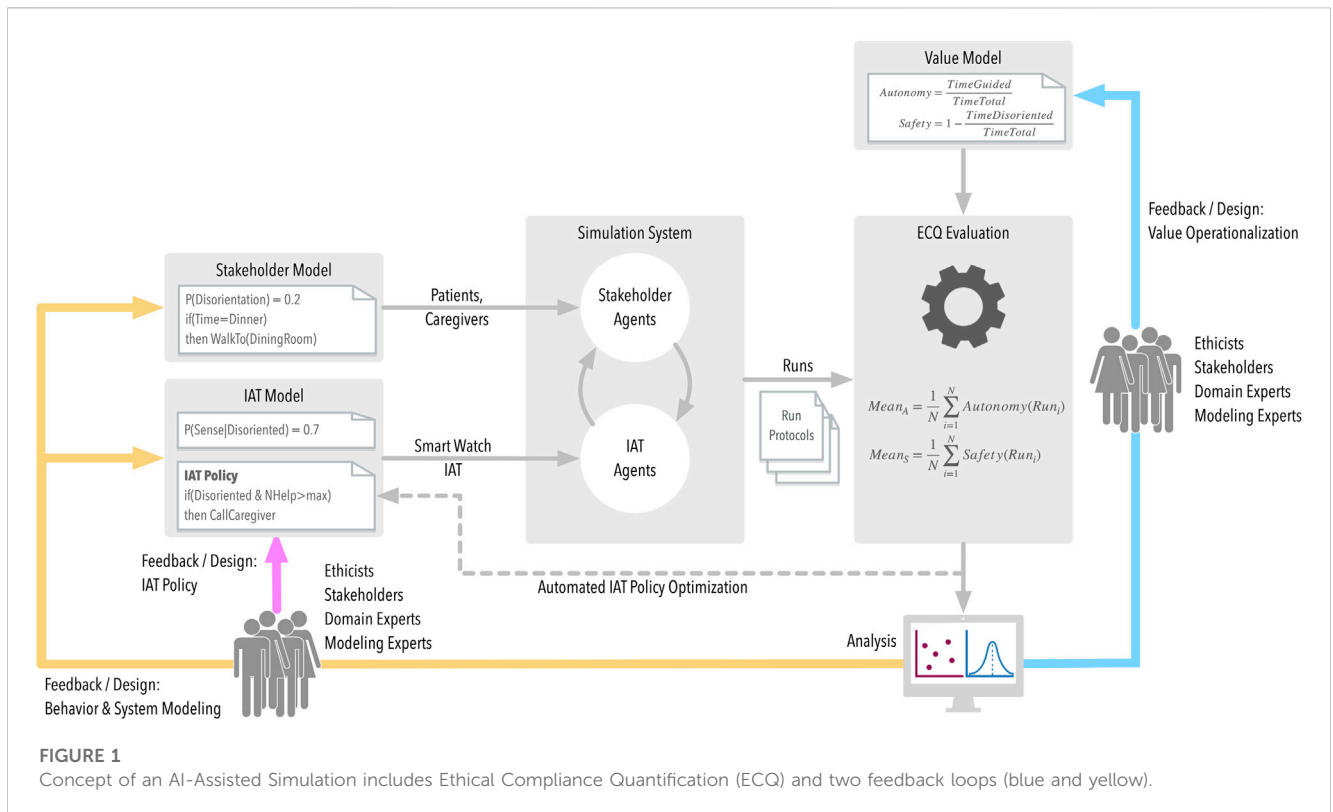
Figure 1 identifies the central components involved in this simulation-based approach and their interplay. The objective of the simulation system component is to simulate the interaction between human stakeholders and IAT in a given environment (such as the interaction between residents, nurses, and a smart-watch-based orientation IAT during nursing home routines, as outlined in our use case below). This means that it is necessary to provide computational rules and stochastic processes that define–in non-deterministic fashion–stakeholder[9] behavior (stakeholder model)[10]. In the process of this definition, it might become necessary to quantify the mental states in stakeholder models–such as the state of a patient's sense of orientation, their likelihood of losing their way altogether, or preferences for certain forms of interaction.[11] It is also necessary to describe the IAT behavior (which is usually comparatively easy because the IAT implementation itself provides the blueprint), as well as the IATs sensor characteristics that define how well it is able to observe the current situation (IAT model). Depending on the application setting, the IAT's sensor reliability may be crucial for being able to make right decisions. A central component in defining IAT behavior is the IAT's "policy," i.e., a set of rules that define how the IAT will chose what assistive action in which situation; it represents the IATs decision-making component. From the viewpoint of the IAT designer, the objective is to define a policy that–within the technical limits of the system environment–achieves optimal results. Such "optimal" results should also be ethically compliant. How to reach this will be discussed in the next step.

---

9   Stakeholder means here all the people whose concerns should be considered in system design.

10   Any kind of knowledge on stakeholder behavior is a reasonable source for model building, empirically or theoretically based. This information then must be transformed into an algorithmic structure a machine can execute. This is the stakeholder model.

11   It should be noted that there exist several cognitive architectures–such as ACT-R, Psi, or SOAR–that provide building blocks for creating a computational model of mental states (Though, the use of such an architecture is not a necessity for setting up a simulation.)

FIGURE 1
Concept of an AI-Assisted Simulation includes Ethical Compliance Quantification (ECQ) and two feedback loops (blue and yellow).

3.2.2 Ethical compliance quantification (ECQ)

The objective of the ECQ evaluation is to provide a quantitative statement on IAT adherence to a score model, based on a set of run protocols generated by the simulation. From an ethicist's viewpoint, the value model is the crucial component, as it provides the translation between the sequence of events in the simulation runs, especially the IAT actions in specific situations, and their ethical assessment. Let us illustrate what defining in such a value model means. Let us consider the care facility floor planning as discussed above. The first step in defining a value model is to identify its values or, better, its value dimensions. Since we will provide scores (numbers) for values, a set of scores–one for each value–defines a point in a space where each dimension corresponds to a value. A very simple ethical value system might ask for "efficiency" and "fairness." The next step in defining a value model is to provide formulae that instruct how to compute a quantitative score as data for the value dimensions of "efficiency" and "fairness" from a simulation run. In our thought experiment world, where stakeholders move between locations, efficiency might for instance be given by the ratio of straight-line distance to distance travelled, while fairness might be given by the quotient of the efficiency scores for different stakeholders (the value "1" representing optimal fairness when all stakeholders experience equal "efficiency"). Then an ECQ setup can be used to compare different floor plans with respect to their rating on the different value dimensions. Even this very simple thought experiment illustrates the core challenge in defining a value model: providing a model that adequately reflects how values are connected to the real world. For instance, consider the–rather trivial–example definition of fairness. One might

rightfully wonder, if it is really fair to compare just efficiency and ignore the physical fitness of stakeholders (e.g., the fitter one is, the longer one can walk). So, stakeholders might rightfully call for a correction factor for the fairness computation that reflects physical fitness.

This simple example illustrates the multilateral nature and the value-sensitive design process required for defining a value model, because it makes value judgements explicit. And by this, it exposes the degrees of freedom that are available in designing the mapping from event sequences to value ratings. Note that simulation-based ECQ also allows the assignment numbers to qualitative value statements: for instance, by counting how often a certain qualitative requirement is observed or violated in a number of simulation runs.

Note that the ECQ-concept provides something impossible in the real world: to evaluate the quantification across different design alternatives for all of the involved models. By varying the IAT policy, it becomes possible to assess the impact of different design alternatives on the compliance to values (policy feedback), possibly with the objective of arriving at an optimal IAT policy. Varying the value model allows assessing the plausibility of the resulting value quantification and thus the plausibility of the value model itself (value operationalization feedback). Finally, by varying stakeholder and IAT model, the sensitivity of the ECQ results to the ecological validity of the simulation model can be assessed.

3.2.3 An example

SimDem (Shaukat et al., 2021) is a simulation system we developed to analyze a smart-watch based IAT in a nursing home for dementia patients. A "smart watch" supports residents by detecting deviations from routes and then prompts the wearer
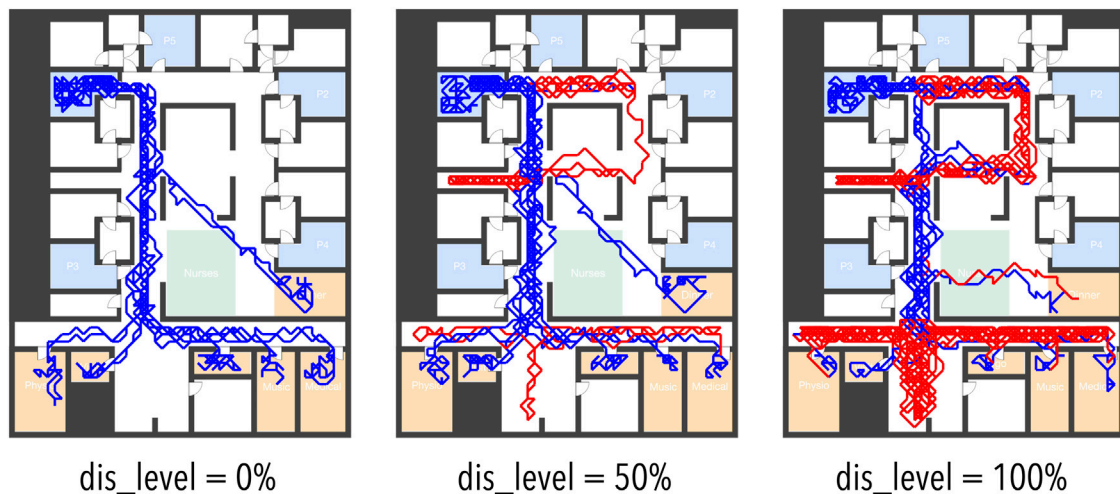
FIGURE 2
Visualization of Trajectories for Simulated Patient Agents with Various Levels of Disorientation Probability (dis_level). Blue = patient state oriented; Red = patient state disoriented.

about which direction to go to reach the destination. One very basic design issue now is the question of how many guidance interventions should trigger the assumption in the system that the wearer is permanently disorientated and thus alert a caregiver. On the basis of previous expert interviews, our own reasoning, and literature research, one might label such values as "safety" and "fairness."

When performing ECQ, the very first step is creating the simulation model. In this case, it is a 2D virtual nursing home (the floor plan based on a real nursing home), where the way-finding behavior for the patient simulation (see Figure 2) has been assessed via observations of real subjects[12].

The simulated IAT–the simulated smart watch–has a certain probability for detecting disorientation and there is a certain probability that a smart watch intervention will help the supported person regain orientation (both these probabilities are design parameters of the simulation model). Based on this setup, it is then possible to perform multiple simulation runs and analyse the quantitative effect of different assistive strategies on values of interest. In figure three, we show the aggregated results from 1809 runs, using different value models. We use this figure to discuss the crucial aspect of value model definition. Concerning the value model, it is first of interest to operationalize "safety." It turns out that there are multiple ways to do this. One might consider the relative amount of time in disorientation as "unsafe" time. This approach produces–as a function of the intervention policy–the reddish colored box plots in Figure 3, labelled "Safety (Original)"

Figure 3 shows that this operationalization is not plausible. The plot shows that, using this operationalization, the resulting score for the strategy of immediately calling a nurse (Nhelp = 0) indicates a higher non-compliance (i.e., a longer time in unsafe state) than the score for the strategy of waiting for five failed smart watch interventions (Nhelp = 5). But, obviously, the more failed interventions we wait for, the longer the disoriented patient will wander unguided. Therefore, this operationalization clearly results in score values that disagree with common sense. The implausibility of this value operationalization design is obvious once the plot provides a visualization of the outcome: as soon as a nurse is accompanying a patient, the situation should be considered as safe by the value operationalization, independent of the patient's disorientation state. Note how the ECQ approach allows discovery of such mistakes in value operationalization through visualizing the value scores across different strategies, as shown in this example.

Providing a more plausible value operationalization now is straightforward: as suggested above, we only consider the time during which a patient is disoriented while not guided by a nurse as "unsafe" time. Using this improved operationalization of "safety" we now see that indeed, immediately calling a nurse is safer than waiting for multiple interventions (see Figure 3, purple box plot, labelled as "Safety (Refinef)"). Note that this plot also reveals that having no IAT at all ("Nurse Only") is the least safe strategy (aside from leaving the patient completely unattended, "No Help") Note that this plot also reveals that having no IAT at all ("Nurse Only") is the least safe strategy (aside from leaving the patient completely unattended, "No Help"). The reason for this is that without a smart watch detecting disorientation, nurses have to actively discover disoriented patients. In this simulation setting, the smart watch therefore always increases safety.

Using the improved operationalization of "Safety," it is now interesting to see how this value is affected by IAT policy in comparison to the "Fairness" value, which reflects the relative amount of time available for forms of caregiving other than route guidance. The rationale behind this is that the more time a caregiver is occupied with route guidance, the less time is available for other, possibly more important tasks, such as social interactions. This

---

12 The probability of selecting a wrong turn is based on data from a study on indoor wayfinding of the University Medicine Rostock. Participants were 8 subjects diagnosed with light to medium dementia (Male/Female = 4/4; Age M = 73.4, SD = 6.3; MMSE M = 22.5, SD = 3.4). Study protocol approved by ethics commitee of University Rostock, approval number A2012-0083.

FIGURE 3
Value scores computed different value models across 1809 simulation runs (see text for details).

information is provided by comparing the box plots for the values "Safety (Refined)" and "Fairness" in Figure 3. We see that there is an obvious value conflict, as decreasing Fairness violations (by waiting longer before calling a nurse) leads to increasing Safety violations. Figure 3 also shows that the safety gain provided by the smart watch IAT in general comes at the price of increased workload, as caregivers are now proactively called for guidance as soon as the IAT gives up on interventions.

Obviously, given the simplicity of the simulation model and its operationalization, this example is of limited significance for the practical design of an orientation-support IAT based on smart watch devices. However, the example does clearly show the potential of ECQ as a means to provide insight into the ethical assessment of IAT, which is the point of interest here. We see that ECQ allows visualization of value trade-offs, and the potential non-linear dependency of score functions on policies. It also shows that ECQ helps in operationalizing values with respect to "real world events" in such a way that the operationalization overlaps with commonly accepted moral precepts. It also shows that human intuition is not guaranteed to provide a plausible operationalization (as illustrated by the first version of the safety operationalization).

4 Discussion

Simulation-based ECQ is a method for exploring the ethics design space, for developing "ethics awareness" in designers, and for informing ethicists about not only outcomes of different scenarios, but how different variables influence the process. Furthermore, as in our case, it allows the simulation or anticipation of complex ethical trade-offs, not only purely hypothetically or very generally (as thought experiments), but as visualized trade-offs regarding human-AI and human-human interaction that cannot be explained or rationalized by the persons involved. In the following discussion, we want to focus on three main challenges.

1) How qualitative values can (or must be) operationalized for such computational simulations and what this requires.
2) In which contexts and for what purposes the advantages of such an AI-assisted simulation outweigh their disadvantages and limits.
3) Why AI-assisted ethics simulations can be compared to thought experiments but provide innovative epistemic dimensions for ethical reasoning.

First, the methodology discussed in this paper is not about a specific value–such as autonomy or safety, but on how to improve the process of ethical reflection for IAT development by considering the diversity of values. Aliman and Kester 2019 argue in favor of a consequentialist approach that predicts the overall utility of a future outcome for a given population. While they thus make a very general argument for AI-assisted simulation with regard to the value of utility, our methodology rather proposes a strategy to understand the impact of an IAT regarding different moral values that can be operationalized. In this sense, our proposal is agnostic regarding the specific values considered during design, but it is not agnostic with regard to the requirement of a participatory and pluralistic approach. Hereby, our methodology is aware of the central challenge of value operationalization. Unless a value is operationalized, it cannot be analyzed by ECQ. While this may be seen as a drawback of the method, we see it rather as an advantage. ECQ poses a *challenge* to value experts to operationalize their value concepts, because ECQ provides the *opportunity* to make use of such an operationalization. A claim that a value cannot be quantified can now be challenged by providing an operationalization, counter-challenging the opponent to show where it violates the value system. In a similar vein, one of the core benefits of using ECQ will be to expose situations where an operationalization indeed cannot be found–or rather cannot be agreed upon. By forcing stakeholders to give an explicit semantics to their value concepts, ECQ exposes conflicts that are indeed fully independent of the question of

"machine ethics," but rather are caused by our own inconsistent or ambiguous opinions concerning moral behavior (Note that behavioral economics has shown that even single persons may make contradictory assessments of situations, depending on whether a situation is experienced or remembered, see also Aliman and Kester, 2019 for this problem.)

We think that the 'values' relevant for machine ethics are far from being sufficiently defined and hinge on theoretical backgrounds, often of conflicting philosophical stances (see also Volkman and Gabriels, 2023). Consider the confusion between "autonomy" as moral self-determination, reflective self-governance or freedom of choice and "automation" understood as "absence of human control", but often seen as model for "autonomous decision of a machine". For example, the popular "levels of automation" model discussed in Sheridan (1980) provides a quantitative theory of "autonomous decisions". This model identifies eight automation levels, level 1 being no automation, level 8 removing any human involvement. While in a simulation, the levels of automation can be tested and quantified, their moral assessment–which level of interaction are better or morally more acceptable–cannot be quantified or answered. ECQ provides a methodology to experiment with different operationalizations of values to analyze which of them coincide with intuitive ethical judgement. The stakeholder feedback-loop external to the simulation furthermore permits a critical reflection on the consequences of the selection of values and operationalizations and allows for radical revision.

Second, another objection concerning the simulation approach might be: Why not simply ask the user for the level of support she would like? This objection is also highlighted by Aliman and Kester (2019) who discuss the possibility of predicting the utility of an AI-based device by the potential users. As they aptly remark, "predicted utility is subject to diverse considerable cognitive biases and often crucially differs from instant utility." (ibid.: 28) It should be emphasized that this applies even more to persons with dementia whose value preferences can change or become unpredictable with progressing dementia. With respect to people with dementia, another obvious reply is that some of them will not be able to express a well-considered preference. But, on a more general level, this is an aspect that holds for all stakeholders. It is in general difficult to assess the consequences of a rather abstract decision ("How many times should the smart watch provide navigation hints before calling a caregiver?") with respect to the impact on the personal experience. Moreover, empirical or participatory approaches that, for example, interview stakeholders also have limitations. Sample sizes are often small, the situations that can be morally assessed are anecdotal, the expectations of future technologies are biased by the experience with current technologies, the experiences and values might be biased by the individual perspectives, they provide only limited and biased reflections of reality, more complex technological features that are opaque to the individual are not considered, and the information gained is static and again requires thought experiments to consider novel what-if-scenarios. All these restrictions limit generalizations of ethical design. This is because sufficient experience to judge the decision impact for a novel technology does not exist in the rule. The simulation approach allows stakeholders (e.g., ethicists, engineers, healthcare providers, patient advocates) to see what her decisions would mean in "practice." This makes it particularly helpful in the context of new technologies that have not been implemented yet. Of course, we do not suggest *not* to ask

the stakeholders, but rather to provide sufficient information before asking and hence, to have a more informed and reflected discussion about potential outcomes and ethical trade-offs. In this sense, the simulation approach does not aim to surpass or replace but to complement participatory approaches.

Third, we started above with the role of thought experiments and their importance as a tool for reflecting on new technologies (or new ideas in general). However, thought experiments have their limitations: They are fictional, and often neglect physical, biological, or social conditions since they are usually primarily aimed to test for logical implications and conceptual premises and therefore tend to operationalize critical variables categorically even when they are continuous. This can become problematic when ethical implications for single agents are analyzed as categorization can introduce errors and contradictory results when only few cases and few variables are considered. An ethical technology assessment, on the other hand, is understood as the exploration and evaluation of more or less likely (or plausible) future scenarios. Even if possible future developments are anticipated under uncertainty or ignorance, relevant and reliable physical, biological, and social knowledge must be considered. The simulation approach can support this anticipation beyond a thought experiment by systematically running through a whole range of possible baseline conditions and their respective outcomes. It can introduce statistical variations and be re-run multiple times to produce a population of outcomes on the variables of interest. In principle, well developed methods for model checking and model evaluation can be applied to better understand functional relations between components and variables in the simulations. However, similar to thought experiments, simulations require that variables are adequately operationalized. In addition, other important variables of the virtual world in which the simulation unfolds need to be sufficiently realistically implemented. In our example, physical variables reflecting the state of the simulation, like space and time, must be part of the simulations and appropriately implemented. Even such "technical" variables can have an influence on the outcome of the simulation and can produce biases if implemented inappropriately. For example, modelling a discrete quantity with a continuous number can lead to meaningless results of an "in between state" on a quantity that is categorical of can only take discrete values. The final ethical evaluation–i.e., whether respective developments or at least individual consequences are considered desirable, undesirable, or even unacceptable from a moral point of view–must, however, be made through human reflection by the observers of such simulation. Therefore, the use of AI in this context does not mean a replacement of human ethical reflection through ethical machines, or that machines can make moral decisions. Instead, our proposed model can be subsumed as a form of "human-centered AI" (Shneiderman, 2021) in the field of ethics which strives to support humans in reasoning about complex systems by means of computational simulations (again, just like in the case of the world climate). In the ethical context, also usually more than one stakeholder is involved, and the values of different stakeholders can be in conflict (e.g., of caretakers and patients). In that case, a compromise needs to be found. Empirically analyzing the effects of different compromises for IAT policy in practice is obviously not possible and would also be morally problematic. Instead, a simulation as ECQ can systematically explore different alternatives without intervening with the actual practice. In this sense, ethical reasoning can directly benefit from the simulation system and our approach could be considered as a form of

"intelligent augmentation" of ethical reasoning. Indeed, both have in common that they stress two aspects (Shneiderman, 2021, 9). First, their design method builds on user experience, stakeholder engagement, and iterative refinement. Second, they are designed as a "supertool" (Shneiderman) to amplify, augment, and empower human performance, but emphasize human control. An additional challenge to such simulation as we have proposed is the significance of space and time for the results. Therefore, we emphasize that we do not advocate to use simulation as full replacement for real world experiments. Rather, our objective is to help stakeholders do develop insight into the conceptual validity of their models of how to quantify the effect of assistive system actions on ethical values. In our example, we have modeled these two dimensions as a continuum, i.e., moral evaluations and of the simulated users with dementia were stable over time. This is not to be expected for the reality: Real users' values and preferences will change over time due to the progress of dementia, to the experiences made with IAT or other endogene and exogene factors. Hence, an AI-based simulation cannot eliminate the need for an ex post assessment via experiments and corresponding individualization. Nevertheless, it contributes to an ex ante ethical alignment of the new technologies for vulnerable groups, e.g., persons with dementia.

5 Conclusion and outlook

AI-assisted simulations can address shortcomings of the current gold standard of empirically informed ethical reasoning, as well as of traditional approaches such as thought experiments and forecasting methods. They could help in the exploration of numerous complex what-if scenarios with great flexibility and provide objective observations that can be visualized and analyzed processually. The process of visualization seems especially relevant as it helps to manifest trade-offs and observations.

In particular, our contribution considers how empirical data about the scope of stakeholders' value preferences and potential ways of behavior can inform a "supertool" to permute the range of ethically relevant baseline parameters and thus simulate different possible outcomes. In this vein, ethically motivated empirical research and AI-assisted simulation strategies are combined and complement each other. In this sense, what we propose here is neither traditional ethics of technology nor machine ethics, but AI-assisted ethics as a new, innovative methodology for empirically informed ethical reflection. However, as an interdisciplinary working group, we also realized that time for collaborative learning is needed to achieve a productive combination of theoretical and methodological perspectives. Of course, in our case, the object of such ethical reflection is also AI-technology. While this does not necessarily have to be the only conceivable use case for AI-assisted ethics, the approach proves to be particularly suited to this still young field of technology development with its comparatively low degree of practical implementation and actual empirical experience. In future the AI model simulations, the inner states (e.g., values, emotions) and behavior of the simulated agents could be included, tracked, and related via AI methods to the interventions (e.g., Francilett et al., 2020). Such a prediction model could produce information about the combined effects of the intervention on various inner states and make this information available for ethical analysis. The results of simulations with such models could provide data similar to results generated by empirical

interviews for use in ethical reasoning. Agents may be simulated using symbolic or sub-symbolic AI techniques. Both have a tradition in cognitive psychology and gaming. In the end, the approach can lead to better-informed ethical reasoning by providing data on how humans are affected by a AI-based system and may help to identify critical factors that lead to problematic situations and support the investigation of ways to mitigate them. It can do this in complex, realistic situations with multiple actors and technical components interacting with each other. This opens new perspectives for the systematic ethical reflection of technological futures in the middle ground between dystopia and utopia.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

All authors have contributed to an early stage of conceptual design of this article during a series of internal workshops. SiS and TK have drafted and structured the first manuscript version, together with JW who drafted main parts of subchapter 2. MS has helped in drafting subchapter 2, too, and made substantial contributions for the discussion part. JR has co-edited subchapter 1, 2, and 3. AH has helped to design subchapter 4 and also to conceptualize Figure 1. All authors contributed to the article and approved the submitted version.

Funding

Acknowledgments

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abojabal, H., Welsch, J., and Schickanz, S. (under review). Cross-cultural perspectives on intelligent assistive technology in dementia care: Comparing Israeli and German experts' attitudes

Aliman, N., and Kester, L. J. H. M. (2019). *Transformative AI governance and AI-empowered ethical enhancement through preemptive simulations*. Delphi 1/2019. doi:10.21552/DELPHI/2019/1/6

Alkadri, J., and Jutai, J. (2016). Cognitive impairment and assistive devices: Outcomes and adverse effects. *J. Rehabil. Assist. Technol. Eng.* 3, 2055668316668146. doi:10.1177/2055668316668146

Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V. I.Precise4Q consortium (2020). Explainability for artificial intelligence in healthcare. A multidisciplinary perspective. *BMC Med. Infrom. Decis. Mak.* 20 (1), 310. doi:10.1186/s12911-020-01332-6

Anderson, M., and Anderson, S. L. (2007). Machine ethics. Creating an ethical intelligent agent. *Ai Mag.* 28 (4), 15–26. doi:10.1609/aimag.v28i4.2065

Anderson, M., Anderson, S. L., and Armen, C. (2004). *Towards machine ethics*. Conference paper. Available at: https://www.aaai.org/Papers/Workshops/2004/WS-04-02/WS04-02-008.pdf (Accessed August 31, 2022).

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature* 563, 59–64. doi:10.1038/s41586-018-0637-6

Bayat, S., and Mihailidis, A. (2021). Outdoor life in dementia. How predictable are people with dementia in their mobility? *Alzheimers. Dement. (Amst.).* 13 (1), e12187. doi:10.1002/dad2.12187

Bicher, M., Rippinger, C., Urach, Ch., Brunmeir, D., Siebert, U., and Popper, N. (2021). Evaluation of contact-tracing policies against the spread of SARS-CoV-2 in Austria: An agent-based simulation. *Med. Decis. Mak.* 41 (8), 1017–1032. doi:10.1177/0272989X211013306

Border, S. P., and Sarder, P. (2022). From what to Why, the growing need for a focus shift toward explainability of AI in digital pathology. *Front. Physiol.* 12, 821217. doi:10.3389/fphys.2021.821217

Brundage, M. (2014). Limitations and risks of machine ethics. *J. Exp. Theor. Artif. Intell.* 26 (3), 355–372. doi:10.1080/0952813X.2014.895108

Buhr, E., and Schweda, M. (in prep). *Moral issues of assistive technologies in dementia care: An ethical analysis of views and attitudes of affected people in Germany.*

Buhr, E., Welsch, J., and Shaukat, M. S. (under review). *Value preference profiles. A source for ethics by design in technology-assisted dementia care.*

Chandrasekharan, A., Nersessian, N. J., and Subramanian, V. (2013). "Computational modeling: Is this the end of thought experimenting in science?," in *Thought experiments in philosophy, science and the arts*. Editors J. Brown, M. Frappier, and L. Meynell (London: Routledge), 239–260.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Sci. Eng. Ethics.* 26 (4), 2051–2068. doi:10.1007/s11948-019-00146-8

Currie, G., Hawk, K. E., and Rohren, E. M. (2020). Ethical principles for the application of artificial intelligence (AI) in nuclear medicine. *Eur. J. Nuc. Med. Mol. Imaging.* 47, 748–752. doi:10.1007/s00259-020-04678-1

De Freitas, J., Cenis, A., Smith, B. W., and Frazzolio, E. (2021). From driverless dilemmas to more practical commonsense tests for automated vehicles. *Proc. Natl. Acad. Sci. U S A* 118 (11), e2010202118. doi:10.1073/pnas.2010202118

Di Paolo, E. A., Noble, J., and Bullock, S. (2000). *Simulation models as opaque thought experiments*. Cambridge, MA: MIT Press, 497–506.

Endter, C. (2021). *Assistiert Altern. Die Entwicklung digitaler Technologien für und mit älteren Menschen*. Wiesbaden: Springer VS.

Francillette, Y., Boucher, E., Bier, N., Lussier, M., Bouchard, K., Belchior, P., et al. (2020). Modeling the behavior of persons with mild cognitive impairment or Alzheimer's for intelligent environment simulation. *User Model User-Adapt. Interact.* 30 (5), 895–947. doi:10.1007/s11257-020-09266-4

Friedmann, B., and Kahn, P. H., Jr. (2007). "Human values, ethics, and design," in *The human-computer interaction handbook*. Editors J. A. Jacko and A. Sears (Boca Raton, FL: CRC Press), 1267–1292.

Goodall, N. J. (2019). More than trolleys: Plausible, ethically ambiguous scenarios likely to Be encountered by automated vehicles. *Transf. Interdiscipl. J. Mobili. Stud.* 9 (2), 45–58. doi:10.3167/TRANS.2019.090204

High Level Expert Group on Artificial Intelligence (2019). *Ecthics guidelines for trustworthy AI*. Available at: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf (Accessed August 31, 2022).

Hofmann, B. (2013). Ethical challenges with welfare technology. A review of the literature. *Sci. Eng. Ethics.* 19, 389–406. doi:10.1007/s11948-011-9348-1

Ienca, M., Jotterand, F., Elger, B., Caon, M., Pappagallo, A., Kressig, R. W., et al. (2017). Intelligent assistive technology for Alzheimer's disease and other dementias. A systematic review. *J. Alzheimer. Dis.* 56 (4), 1301–1340. doi:10.3233/jad-161037

Ienca, M., Wangmo, T., Jotterand, F., Kressig, R. W., and Elger, B. (2018). Ethical design of intelligent assistive technologies for dementia: A descriptive review. *A Descr. Rev. Sci. Eng. Ethics.* 24, 1035–1055. doi:10.1007/s11948-017-9976-1

Köhler, S., Görß, D., Kowe, A., and Teipel, S. (2022). Matching values to technology: A value sensitive design approach to identify values and use cases of an assistive system for people with dementia in institutional care. *Ethics. Inf. Technol.* 24, 27. doi:10.1007/s10676-022-09656-9

Korn, G. A., and Wait, J. V. (1978) *Digital continous-systems simulation*, Prentice-Hall, Englewood Cliffs, N.J.

Kunze, C., and König, P. (2017). "Systematisierung technischer Unterstützungssysteme in den Bereichen Pflege, Teilhabeunterstützung und aktives Leben im Alter," in *Umgebungsunterstütes leben. Beiträge zum usability day XV*. Editors I. Hämmerle and G. Kempter (Lengerich: Pabst), 15–22.

Lancioni, G. E., Desideri, L., Singh, N., O'Reilly, M. F., Sigafoos, J., De Caro, M. F., et al. (2021). Use of technology to sustain mobility in older people with cognitive impairment and dementia. A scoping review. *Disabil. Rehabil. Assist. Technol.* 222, 1–15. doi:10.1080/17483107.2021.1900935

Landau, R., and Werner, S. (2012). Ethical aspects of using GPS for tracking people with dementia. Recommendations for practice. *Int. Psychogeriatr.* 24 (3), 358–366. doi:10.1017/S1041610211001888

Lara, F., and Deckers, J. (2020). Artificial intelligence as a socratic assistant for moral enhancement. *Neuroethics* 13, 275–287. doi:10.1007/s12152-019-09401-y

Löbe, C., and AboJabal, H. (2022). The role of intelligent assistive technology for empowering people with dementia. A scoping review. *Arch. Geronto.* Epub. doi:10.1016/j.archger.2022.104699

Manzeschke, A., Weber, K., Rother, E., and Fangerau, H. (2013). *Ethische Fragen im Bereich Altersgerechter Assistenzsysteme. Ergebnisse der Studie.* VDI/VDE Innovation + Technik.

Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care. A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inf.* 113, 103655. doi:10.1016/j.jbi.2020.103655

McLennan, S., Fiske, A., Tigard, D., Müller, R., Haddadin, S., and Buyx, A. (2022). Embedded ethics. A proposal for integrating ethics into the development of medical AI. *BMC Med. Ethics.* 23, 6. doi:10.1186/s12910-022-00746-3

Mertz, M., Inthorn, J., Renz, G., Rothenberger, L. G., Salloch, S., Schildmann, J., et al. (2014). Research across the disciplines: A road map for quality criteria in empirical ethics research. *BMC Med. Ethics.* 15, 17. doi:10.1186/1472-6939-15-17

Mihailidis, A., Boger, J. N., Craig, T., and Hoey, J. (2008). The COACH prompting system to assist older adults with dementia through handwashing. An efficacy study. *BMC Geriatr.* 8, 28. doi:10.1186/1471-2318-8-28

Misselhorn, C. (2021). Artificial systems with moral capacities? A research design and its implementation in a geriatric care system. *Artif. Intell.* 278, 103179. doi:10.1016/j.artint.2019.103179

Mollison, D. (1991). Dependence of epidemic and population velocities on basic parameters. *Math. Biosci.* 107 (2), 255–287. doi:10.1016/0025-5564(91)90009-8

Nallur, V. (2020). Landscape of machine implemented ethics. *Sci. Eng. Ethics.* 26, 2381–2399. doi:10.1007/s11948-020-00236-y

Novitzky, P., Smeaton, A. F., Chen, C., Irving, K., Jacquemard, T., O'Brolcháin, F., et al. (2015). A review of contemporary work on the ethics of ambient assisted living technologies for people with dementia. *Sci. Eng. Ethics.* 21, 707–765. doi:10.1007/s11948-014-9552-x

Parke, E. C. (2014). Experiments, simulations, and epistemic privilege. *Philos. Sci.* 81 (4), 516–536. doi:10.1086/677956

Peck, St. L. (2004). Simulation as experiment: A philosophical reassessment for biological modeling. *Trends Ecol. Evol.* 19 (10), 530–534. doi:10.1016/j.tree.2004.07.019

Ray, P. P., Dash, D., and De, D. (2019). A systematic review and implementation of IoT-based pervasive sensor-enabled tracking system for dementia patients. *J. Med. Syst.* 43, 287. doi:10.1007/s10916-019-1417-z

Schicktanz, S., and Schweda, M. (2021). Aging 4.0? Rethinking the ethical framing of technology-assisted eldercare. *Hist. Philos. Life Sci.* 43, 93. doi:10.1007/s40656-021-00447-x

Schicktanz, S., Schweda, M., and Wynne, B. (2012). The ethics of public understanding of ethics -Why and how bioethics expertise should include public and patients voices. *Med. Health Care Philos.* 15 (2), 129–139. doi:10.1007/s11019-011-9321-4

Schweda, M., Kirste, T., Hein, A., Teipel, A., and Schicktanz, S. (2019). The emergence of co-intelligent monitoring and assistive technologies in dementia care - an outline of technological trends and ethical aspects. *Bioethica Forum* 12 (1/2), 29–37.

Shaukat, M. S., Hiller, B. J., Bader, S., and Kirste, T. (2021). SimDem A multi-agent simulation environment to model persons with dementia and their assistance. *arXiv.* doi:10.48550/arXiv.2107.05346

Sheridan, T. B. (1980). Computer control and human alienation. *Technol. Rev.* 83 (1), 60–73.

Shneiderman, B. (2021). *Human centered AI.* Oxford: University Press.

Spiekermann, S. (2016). *Ethical IT innovation. A value-based system design approach.* Boca Raton, FL: CRC Press.

Spitzenverband der Gesetzlichen Krankenversicherungen in Deutschland (2019). *Digitalisierung und Pflegebedürftigkeit. Nutzen und Potenziale von Assistenztechnologien.* Hürth.

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

van Wynsberghe, A. (2013). Designing robots for care. Care centered value-sensitive design. *Sci. Eng. Ethics.* 19 (2), 407–433. doi:10.1007/s11948-011-9343-6

Vandemeulebroucke, T., Dierckx de Casterlé, B., and Gastmans, C. (2018). The use of care robots in aged care. A systematic review of argument-based ethics literature. *Arch. Gerontol. Geriatr.* 74, 15–25. doi:10.1016/j.archger.2017.08.014

Volkman, R., and Gabriels, K., (2023): AI moral enhancement. Upgrading the socio-technical system of moral engagement. *Sci. Eng. Ethics* 29, 11. doi:10.1007/s11948-023-00428-2

Wallach, W., Franklin, S., and Allen, C. (2009). A conceptual and computational model of moral decision making in human and artificial agents. *Top. Cogn. Sci.* 2, 454–485. doi:10.1111/j.1756-8765.2010.01095.x

Walsh, A. (2011). A moderate defence of the use of thought experiments in applied ethics. *Prac* 14, 467–481. doi:10.1007/s10677-010-9254-7

Walton, D. (2016). Some artificial intelligence tools for Argument evaluation: An introduction. *Argumentation* 30, 317–340. doi:10.1007/s10503-015-9387-x

Weber, K. (2021). "Altersgerechte Assistenzsysteme: Ein Überblick," in *Gute Technik für ein gutes Leben im Alter? Akzeptanz, Chancen und Herausforderungen altersgerechter Assistenzsysteme.* Editors K. Weber, S. Haug, D. Frommeld, and U. Scorna (Bielefeld: transcript Verlag), 27–62. doi:10.14361/9783839454695-002

Welsch, J., and Schicktanz, S. (2022). "Developing value preference profiles as tool for value-oriented technology design for living with dementia? Practical and methodological considerations," in Presentation given at the World Congress of Bioethics 2022 (Switzerland: University of Basel). Available at: https://organizers-congress.org/frontend/index.php#.

Welsch, J. (2022a). "Empowerment and Technology. An ethical-empirical exploration of technology-assisted dementia care," in Poster presentation, International Symposium The Future of Assistive Technologies in Dementia Care (Delmenhorst).

Welsch, J. (2022b). "Digital-assisted care in the digital desert? Experts' perspectives on the preconditions of modern monitoring and assistive technology in dementia care," in Presentation given at the Jahrestagung der Akademie für Ethik in der Medizin e.V.

Winsberg, E. (2022). "Computer simulations in science," in *The stanford encyclopedia of philosophy (winter 2022 edition).* Editors E. N. Zalta and U. Nodelman URL: https://plato.stanford.edu/archives/win2022/entries/simulations-science/.

WHO - World Health Organization (2018). *Assistive technology.* Key facts. Available at: https://www.who.int/news-room/fact-sheets/detail/assistive-technology (Accessed August 31, 2022).