AI and Ethics
https://doi.org/10.1007/s43681-024-00647-8

ORIGINAL RESEARCH

Measuring meaningful human control in human–AI teaming: effects of team design in AI-assisted pandemic triage

Karel Van den Bosch[1] · Jurriaan Van Diggelen[1] · Sabine Verdult[1] · Tjalling Haije[1] · Jasper Van der Waa[1]

Received: 28 December 2023 / Accepted: 2 December 2024

Abstract
AI will increasingly be used to collaborate with humans on ethical tasks. This study contributes to the need for practical methods to assess whether a human–AI system is under Meaningful Human Control (MHC). We propose three measurable components of MHC: subjective-, normative-, and moral control. To empirically evaluate the qualities of the MHC measuring approach, we developed an experiment in which a human–AI team performs triage during a pandemic outbreak. Participants performed the role of physician. Moral pressure was induced by a rapid influx of patients and limited resources. Three designs of human–AI collaboration were tested as a repeated within-subjects factor: (A) agent provides information and decision advice; (B) human assigns some patients to agent for triage; (C) human instructs agent to autonomously conduct triage on all patients. The measures were sufficiently sensitive to detect effects of the three human–AI team design on MHC: When advised by an agent (A), or when issuing tasks to an agent (B), participants felt more engaged, were able to exercise more control, and were more compliant with ethical guidelines. When the agent performed triage autonomously (C), participants reported a lower moral load, and judged the collaboration as less believable. Subjective, normative, and moral control can serve as a practical approach for assessing MHC.

Keywords Human-agent teaming · Artificial intelligence · Joint decision making · Meaningful human control · Moral decision making · Triage · Pandemics

1 Introduction

The progress of Artificial Intelligence (AI) allows the deployment of intelligent agents in increasingly complex tasks [1]. As AI becomes more competent and ubiquitous, it is crucial that humans maintain control over it. This is considered especially important when a Human-Agent Team (HAT) is faced with making decisions with moral and legal implications. Examples of morally sensitive tasks can be found in healthcare [2, 3], autonomous driving [4], AI-based defense systems [5], and in many other societal domains [6]. Although some propose that moral decision making by artificially intelligent machines should in principle be possible [7], others argue that incorporating moral values in decision making requires virtue or moral character [e.g., 8, 9], a capacity that is expected to remain beyond reach for AI in the foreseeable future. According to this latter view, moral decision making is regarded as a uniquely human competence. When intelligent technology is used in morally salient tasks, it is essential to maintain *Meaningful Human Control* (MHC) [10]. This means that the technology is demonstrably and verifiably responsive to the human's moral reasons relevant to the circumstances [11]. Human's trust in the ethical capabilities of the AI is crucial for accepting the AI as partner [12, 13]. The level of a human's trust in AI should be calibrated; that is: in line with the capabilities of the AI [14, 15]. Inappropriate trust in autonomous AI systems has

✉ Karel Van den Bosch
karel.vandenbosch@tno.nl

Jurriaan Van Diggelen
jurriaan.vandiggelen@tno.nl

Sabine Verdult
sabine.verdult@tno.nl

Tjalling Haije
tjalling.haije@tno.nl

Jasper Van der Waa
jasper.vanderwaa@tno.nl

[1] Human-Machine Teaming, TNO, PO Box 23, 3769 ZG Soesterberg, The Netherlands

Published online: 28 January 2025

often led to violations of safe, morally just and responsible behavior [e.g., 16, 17].

Thus, to maintain MHC, roles and responsibilities should be appropriately assigned within the human–AI team. This control does not necessarily mean that the AI must always be continuously monitored or that a human should invariably approve all its decisions before the AI performs an action. More efficient and indirect forms of exercising control are possible, provided that the human is capable of averting violations of moral standards, and accepts full responsibility and legal accountability of the human–AI team decisions [10].

Many endorse the notion of MHC in human–AI systems and some argue that it should be legally enforced [e.g., 5]. International AI regulations have been developed that emphasize the need to maintain human oversight and keep the human in control over the team's behavior and actions (e.g., NATO,[1] EU AI-act[2]).

Various definitions and theoretical properties of MHC have been discussed [e.g., 5, 11, 18]. However, these discussions have so far not resulted in a commonly shared definition, nor in concrete and testable requirements for researchers, designers, and engineers. As a result, it remains unclear how to assess whether an AI system is under MHC. In response to the need for tangible instruments to assess MHC, several authors [e.g., 19, 20, 8] have proposed actionable properties that can be used to test whether an AI system is under MHC. We aim to advance these efforts by proposing a practical method for assessment of MHC and test it in a semi-realistic context. In this paper we propose three components of MHC: *subjective* control, *normative* control, and *moral* control (see Sect. 2.1). Furthermore, we introduce a set of measures for these components, which we apply in a human–AI team collaboration testbed [21] to empirically evaluate the measuring approach. Thus, the main objective of this study is to empirically evaluate the value of the proposed method for assessing MHC in a human–AI collaboration task.[3] With this work we aim to contribute to an operational approach for founded and better assessment of MHC. Furthermore, we believe that experiences emerging from empirical evaluation of MHC will also enhance our understanding of MHC.

In a testbed that simulates the domain of pandemic triage, we implemented three different designs of human–AI collaboration. The human–AI team has to decide which patients are given access to scarce hospital beds. The triage decisions have moral implications, thus posing challenges for the human to achieve MHC over the performance of the team. The three human–AI team designs differed in terms of task allocation and decision authority assigned to the human and AI. It is expected that the different designs will bring about dissimilar outcomes on the distinguished components of MHC, enabling us to assess whether our proposed measures of MHC are *feasible*, *sensitive*, and *complementary*. To enable measurement of MHC in a practical context, measures should be relatively easy to collect, and they should be sensitive enough to reveal potential differences between designs of human–AI teams that can be expected to yield different effects on MHC. Furthermore, the set of measures should be complementary, in the sense that they address different aspects or manifestations of MHC.

In addition to evaluating our measuring approach, we also have the objective to explore the impact of the distinguished human–AI team designs on MHC, as the selected designs reflect often proposed options for assigning responsibilities and permissions to team members [e.g., 22]. We believe that the challenges introduced in our testbed resemble those that would arise when AI-support would be introduced in real medical triage. Under this assumption outcomes are indicative for what to expect. However, as the study presents a simplification of the pandemic-triage task, results cannot be used to draw final conclusions on how to incorporate intelligent technology into real-world medical applications.

Thus, the primary goal of the study is to develop and evaluate measures for founded and better assessment of MHC. The secondary goal is to explore the effects of different human–AI team designs on MHC.

2 Meaningful human control and its measurement

Various scholars emphasize the importance of having MHC over intelligent autonomous technology. The discussion on what exactly constitutes and defines MHC [e.g., 19] has resulted in a number of requirements [5, 20, 23]:

- Human operators are making informed, conscious decisions about the deployment of technology.
- Human operators have sufficient information to evaluate and verify actions taken by the system, taking the context into account.
- The technology is extensively tested before deployment, and human operators are properly trained for the tasks.

[1] NATO principles of responsible AI (2021). https://www.nato.int/cps/en/natohq/official_texts_187617.htm.

[2] EU AI Act: first regulation on artificial intelligence (2023), https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[3] In this paper we use the term AI also for intelligent agents that can reason and decide (semi-)autonomously. Hence, the terms human-agent team, human–machine team, and human–AI team are regarded as synonyms.

- Explicit links exist between the actions of the technology and of humans who are aware of their moral responsibility.

MHC should not be viewed as a fixed property of a human–AI system, but as a property that emerges from the interactions between multiple humans and technology over a longer period of time [8]. In reality, systems function in dynamically changing contexts, with a wide variety of stakeholders [6, 23]. MHC is more than enabling the human to intervene at the moment when the technology is actually deployed [11]. MHC should instead emerge from all levels of a system's life, including from decisions taken by: the initiating organization, designers, developers, operators, subjects, and the society at large [6, 11, 20]. If, for example, during system deployment task conditions occur that make direct supervisory control by the human infeasible (e.g., due to rapidly evolving events that require instantaneous decision making; or when there is no network connection between technology and human), then some level of MHC may still be achieved by defining in advance under what conditions the autonomous technology is allowed to act, and in which way [8]. In that manner, human control can be executed before the morally sensitive situation occurs.

Some scholars point out that the terms being used to define MHC lack precision [24, 25]. It is argued that this imprecision of concepts, in combination with the complexity of the application domains, will lead to unceasing debates as to what constitutes 'informed decisions'; 'sufficient information'; 'proper training', and so on. The point here is that a thorough appreciation of the task, the team, and the task domain in question is needed to assess whether a human has meaningful control over a system. This means that studying MHC cannot take place at the conceptual level only; empirical and pragmatic research in situational contexts is also needed to bring the field forward.

In the next section we propose three components of MHC and discuss how these components could be measured when testing for MHC in concrete human–AI systems.

2.1 Three components of MHC

The concept of 'control' is a crucial construct in psychology [e.g., 26]; control enables people to act purposefully, and to achieve goals. In the context of humans collaborating with AI-driven technology, human control implies having power over intelligent partners that have the capacity to act intelligently in the task space and are thus dynamically influencing the course of action. Control by the human can be seen as a process of achieving goal-directed results in the face of disturbances (changing circumstances) that would otherwise prevent achievement of these goals [27, 28]. As stated earlier, our point is that establishing human control within human–AI teams is not something that can be arranged at the level of the human operator only; it also needs appropriate measures at all levels of the team's organizational context [8, 11]. However, in this paper we focus on MHC in the operational context of the team. That is, the human participating as a leader of the human–AI team.

According to Perceptual Control Theory [28], when a person acts or reacts in the task environment, they do so to achieve goals that are in accordance with their subjective intents. Humans that lead a team with machine partners have multiple intents. A fundamental intent is to control the team's behavior to make sure that decisions and actions by the team do not violate the values that the human aims to secure. This concerns experienced, or subjective control. A related intent is to control the behavior of the team in such a fashion that it meets the requirements imposed by an external authority. This refers to the control that is needed to make the team comply with norms issued by an external authority, often conveyed in the forms of guidelines. Again another intent is the natural desire of humans to act morally [29]. This requires a human to have control over the team to make it act in alignment with the human's personal moral values. This refers to having moral control.

Subjective-, normative-, and moral control should not be regarded as independent goals of a human. Instead, they interactively and simultaneously influence the human's thinking and decision making in a complicated manner. The above proposed components may not cover all aspects of meaningful human control. However, we deem these as important and necessary, and being able to measure these components will support assessing MHC in concrete situations. In this paper we investigate the three distinguished intents as components of MHC and propose associated measures for these components (see Fig. 1).

In the sections below we discuss the components of MHC and propose how they may be assessed when humans and agents collaboratively carry out tasks. The oval in the upper-left of Fig. 1 represents the human-agent team. The nature of the human-agent teamwork is dependent on the team design, that dictates how tasks are assigned within the team and how humans and agents collaborate (see Sect. 3).

2.1.1 Subjective control

The degree of control that people experience over a task or team is affected by many factors. For example, if a decision needs to be made quickly, then the human may have little or no time to evaluate a proposed action by the agent. This likely hampers the human's opportunity to assess the origins of the agent's action or advice which will likely affect the human's trust in its partner [30]. Likewise, if the

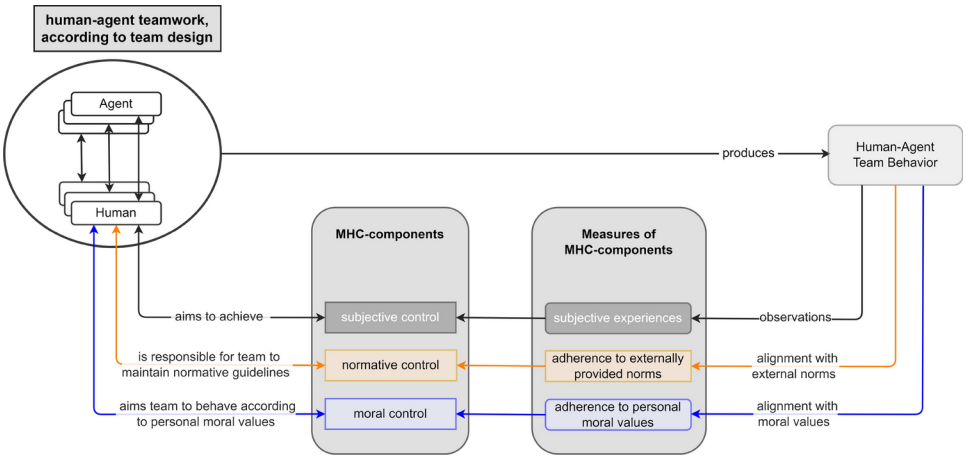

Fig. 1 Components of meaningful human control and associated measures

agents behave in a manner that is completely different from what the human thinks was instructed to them, the human may also feel out of control. In contrast, when the human has the time and skills to monitor the agent's behavior, and when it understands and endorses the suggestions that the agent gives, then this enhances the feeling of control by the human.

There is a difference between subjectively experienced control and actual control. There is evidence that individual differences exist in the way people attribute opportunities to exercise control over machines (i.e., their locus of control), which influences their trust, and also their self-assessed control over machines in actual cases [31]. Questions based upon simulated recall (prompted by e.g., questions, photos, video) can be used to obtain insight into how a person experienced an interaction process [32]. By using this method, subjective control can be assessed by afterwards presenting critical task events to the human team leader and to subsequently ask what was going on in the human-agent team (e.g., was there sufficient time for monitoring before decisions were made?), and to evaluate how well they could make sense of the processes within the team (e.g., was the relevant information available?; does the human leader consider him or herself to be sufficiently skilled and trained to make sense of the processes?). And also: is the human content with the decisions made by the team, given the circumstances and information available at the time? Measuring data about these questions can provide evidence for subjective control (see grey lines in Fig. 1).

2.1.2 Normative control

Human-agent teams always operate in the context of a wider normative system [6] and are therefore required, or even obliged, to act in accordance with the values of that system, often formally represented in the form of normative guidelines by an external authority (e.g., a medical doctor ensuring compliance with medical guidelines; a military commander respecting the rules of engagement). In practice, the human is assigned with the responsibility to ensure that the team adheres to the guidelines [33]. Why allocate that responsibility to humans, and not to agents? This has to do with the inherent imprecision of terms in guidelines [24], and the need to hold humans accountable [34]. For example, in the complexity of a real military situation, when can be concluded that an opponent's act qualifies as hostile? And when exactly is a medical treatment proportional to the risks? Humans are believed, unlike AI, to be capable of judging the context, and to determine whether a considered action meets, or does not meet, the guidelines. Furthermore, humans have legal personhood, but robots do not. So by assigning responsibility to humans, a human can be held accountable for the team's actions.

Whether a human has normative control can be assessed by collecting data that reveal whether the decisions and actions taken by the team are in alignment with guidelines, as assessed by a representative of the issuing authority, e.g., by a supervisor, or domain expert(s). If they do align, then this provides evidence for normative control (see orange lines in Fig. 1). However, this evidence is not perfect. The system might coincidentally adhere to the norms without deliberate human control. Therefore, *adherence to externally provided norms* should be regarded as a weak measure, correlating with, but not perfectly indicating normative control.

2.1.3 Moral control

In their seminal paper, Santoni de Sio and Van den Hoven [10, p. 7], argued that for a decision maker to have MHC, "*the system should demonstrably and verifiably be responsive to the human moral reasons relevant to the circumstances [..]*". The capacity of a human to enforce a system to do so, refers to having moral control (see the blue lines in Fig. 1). Moral control complements normative control. Although the objective of normative guidelines is to provide unambiguous directions for decision making, this is seldom

achieved. In the real world there inevitably arise situations that cannot properly be solved by the guidelines, clearing the way for different interpretations that may vary according to situational and personal factors [35]. It is believed that humans are better than machines at determining whether or not guidelines apply in the actual context at hand. However, this sometimes proves also for humans to be very difficult. For example, suppose that guidelines for medical triage state that younger patients are more eligible to medical care than older patients, and that fitter patients are more eligible than unfit patients. How then to decide between a fit older patient, and an unfit younger patient? Obviously, the difficulties increase when more factors need to be taken into account. In such complex situations guidelines fail to provide clear-cut directions, and humans have to rely on their own personal moral values and to decide accordingly [36]. If a human successfully accomplishes to make the system behave in alignment with the human's personal moral values (e.g. by providing clear instructions in advance), then this should be considered as having moral control.

3 Design of human-agent teams and MHC

Intelligent agents may be used to improve and alleviate decision making efforts in complex tasks [37]. It requires effective collaboration, based upon common ground, predictability, and understanding [38]. Coordinated team collaboration can be accomplished by smartly designing the interactions within human-agent teams.

Various methods exist to characterize interactions of human-agent teamwork. The sociotechnical systems approach [39] acknowledges that social and technical elements of work are interrelated and cannot be decoupled. When intelligent technology is implemented, it affects the entire system by altering the interactions among work system elements, and these effects should be considered to achieve an effective and robust work flow.

For the present study we adopted the method of Team Design Patterns (TDPs) [22, 40, 41] to develop different designs for human–AI teams. TDPs describe essential elements of teamwork in a formalized and generalizable way, for example:

- How tasks are divided among the team members.
- How tasks are monitored by different team members.
- How interdependence of tasks in teamwork is handled.
- How different phases of teamwork follow up on each other.

Typically, different team design patterns are possible for the same task. A TDP may, for example, assign no or very little autonomy to an agent, demanding human supervision and approval for all decision-making performed by an agent. Such a TDP supports MHC by allocating all decision making to the human. A downside is that, -when available time is limited-, the human is at risk to be overloaded with work, which could, in the end, result in losing MHC.

Another design solution is a TDP in which the human provides extensive instructions to the agent beforehand on how to decide and act in morally sensitive situations. Subsequently, the agent acts without human guidance, and attempts to follow the given moral instructions. This TDP may prevent excessive workload for the human. However, the domain may be so complex that it is impossible to capture the nature of human decision making into numerical values that robots need to understand [42]. A potential risk of this design of teamwork is therefore that the agent, when confronted with situations for which the instructions do not yield an unequivocal decision, needs to improvise, leading to decisions that misalign with the human's moral values (hence losing MHC).

Yet another design solution would be to let the human and agent divide the work: easy and straightforward tasks are performed by the agent, and all difficult (potentially morally sensitive) tasks are performed by the human. This TDP may offer a best of both worlds' solution, but discriminating in advance between easy and difficult tasks may be more difficult than it seems.

In conclusion, multiple solutions to organizing team behavior exist, and they all have their potential strengths and vulnerabilities from the perspective of achieving and maintaining MHC. The question which team design solution benefits MHC the most has no universal answer: the best solution depends on the task context, constraints, workload, and the values at play. Therefore, the influence of team design on MHC should be assessed and empirically studied within a particular use case, which is the topic of the next section.

4 A case study: MHC in AI-assisted pandemic triage

The conceptualization and measurement of MHC in a human-agent team (see Sect. 2.1) is tested in an experimental use case of medical triage during a viral pandemic crisis. A computer simulation of a medical emergency unit was developed, and participants were required to conduct triage on the incoming patients. They first performed triage alone, then several times in various collaboration conditions with an AI-based team partner. We do not claim that our simulation is a valid representation of pandemic triage, would such conditions occur in real life. However, triage

under crisis conditions was selected as use case because: it involves moral implications of decision making; humans may benefit from assistance by AI-based teammates; and it allows various solutions for designing human-agent collaboration. For example, agents could help by giving decision advice, or by taking over some or all of the decisions from the human. Reviews have shown that the medical domain is a fruitful application area of AI [e.g., 43] and medical experts consider triage to be a plausible and suitable task for studying MHC in human–AI collaboration [21]. Furthermore, achieving human control over technology is especially important in this domain, as it has been suggested that insufficient human oversight can potentially result in people being unable to prevent biases from affecting medical decisions [44].

The experiment was conducted in February 2021, in the midst of the second wave of the COVID-19 pandemic. At that time, it was feared that the available medical resources would soon no longer be sufficient to provide new patients with care as normal, in the news denoted as 'code black'. In anticipation of this, the government installed a protocol for conducting triage based on non-medical considerations [45]. This initiated a widely held public discussion on the ethical justifications and implications of the protocol. It is assumed that the heated public debate at the time supported the immersion and ethical involvement of participants that we were looking for in this study. As we now know, the intensity of the COVID-19 crisis decreased a few months after the experiment, so 'code black' never had to be implemented.

The purpose of the use case is to present human participants with a complex and immersive task in which they have to take decisions with ethical implications. Participants take decisions by themselves, but also with support from, and in collaboration with, intelligent agents. Thus, we used a task simulation that requires participants to interact with intelligent agent partners, fostering that participants feel the deliberations involved in complex decision making, and to experience how it feels to have or lack control over the team's performance. Such an interactive method is more appropriate for studying MHC than passive research methods (e.g., requesting participants to evaluate or judge a presented narrative about human-agent collaboration) [13].

We developed three designs for human-agent cooperation (see Sect. 4.1.2) and measured the effects of team design on MHC (see Fig. 1), and also on teamwork and on team performance. The following questions were investigated:

- RQ1: How does team design affect MHC over making triage decisions, as expressed by measures of subjective-, normative-, and moral control?
- RQ2: How does team design affect team performance?

To investigate whether or not participants' triage decisions comply with given guidelines, a reference is needed for what in a concrete decision situation constitutes an ethically compliant decision, and what not. That is because guidelines are precisely what the word suggests: they provide guidance, no clear-cut decision recipes. For example, not assigning a patient with severe symptoms to the intensive care unit may be regarded as a violation of the guidelines as that is unjustly withholding essential care. Similarly, referring a patient with moderate symptoms to the intensive care unit may also be regarded as a violation of the guidelines as this might mean that a seriously ill patient who comes in later has to be denied essential care. Thus, whether or not a particular triage decision complies with the guidelines is partly a subjective matter and its judgment requires expertise on the content matter. In this study we presented the materials to an expert in medical triage, and asked for each case which triage decisions should be considered compliant with ethical guidelines, and which are not-compliant (see Appendix A—Medical expert's judgments on triage decisions).

4.1 Methods

4.1.1 Participants

Twenty-one participants (11 female, 10 male) took part on a voluntary basis. Inclusion criteria were a good understanding of written English language, a higher education level, and affinity with technology. Being involved in the medical field (as student or in profession) was used as exclusion criterion. These criteria were to obtain a homogeneous sample of subjects. The study was approved by the ethics committee of the authors' affiliation.[4] Written informed consent was obtained from all participants. Participants were reimbursed with 15 euros per hour. The age characteristics of the obtained sample of participants were: range: 19–47; mean: 27.1; sd: 8.7. Subjects rated their technical skill level on average as 4.2 on a 5 point Likert-scale (sd = 0.89).

4.1.2 Design

A within-subjects design was used, with *Type of Human-Agent Team* as the within-subjects factor with four levels: baseline, TDP-1, TDP-2, and TDP-3. All participants conducted the medical triage task four times under different conditions:

- Baseline: Human makes Decisions—no agent support.
- TDP-1: Data-Driven Decision Support—agent provides support and advice - human makes decisions.

[4] issued on September 27th, 2020, with registration number 2020-082.

- TDP-2: Dynamic Task Allocation—human and agent divide patients for doing triage.
- TDP-3: Supervised Autonomy—human pre-instructs agent - agent autonomously does triage on all patients.

Every participant first received the baseline condition, the three TDP-conditions were subsequently administered in random order to eliminate the possibility of order effects. Four sets of patients were created for each of the within-subjects conditions (see Sect. 4.1.4.3).

4.1.3 Task

The task is conducting triage on a series of patients during a pandemic virus outbreak. Code black has been enforced in the scenario, implying circumstances in which medical care is not unlimitedly available to all patients. Participants perform the role of physician in attendance of a hospital's emergency unit. They have to evaluate and triage each patient for treatment at either: (a) *intensive care*; (b) *hospital ward*; or (c) *home treatment by a general physician*. The available resources are limited: there are three intensive care beds, and six beds on the hospital's ward. Capacity for home treatment is unlimited. The urgency of the situation is simulated by a rapid influx of patients, and by presenting more patients with severe symptoms than can be accommodated with the sparse available intensive care beds. Participants are presented with normative medical and ethical guidelines for triage (see Sect. 4.1.4.5).

4.1.4 Materials

4.1.4.1 Briefing of participants Participants were told that the task was inspired by the COVID-19 pandemic, but that all presented patients and situations were fictitious, and not related to any real patients or events. Participants were informed that they would be asked to play the role of a physician, but that neither medical education nor medical expertise was required. It was emphasized that the purpose of the study was not about the process and quality of medical decision making, but rather to investigate how people experience making difficult decisions, and whether and how technology could help people to conduct complex decision tasks.

4.1.4.2 Implementation of the task environment A computer simulation of a medical emergency unit was developed (see Fig. 2). The photos of the fictitious patients are artificially constructed faces of non-existing people.[5]

4.1.4.3 Composing series of patients The simulator generated fictitious patients by assigning values for fitness,

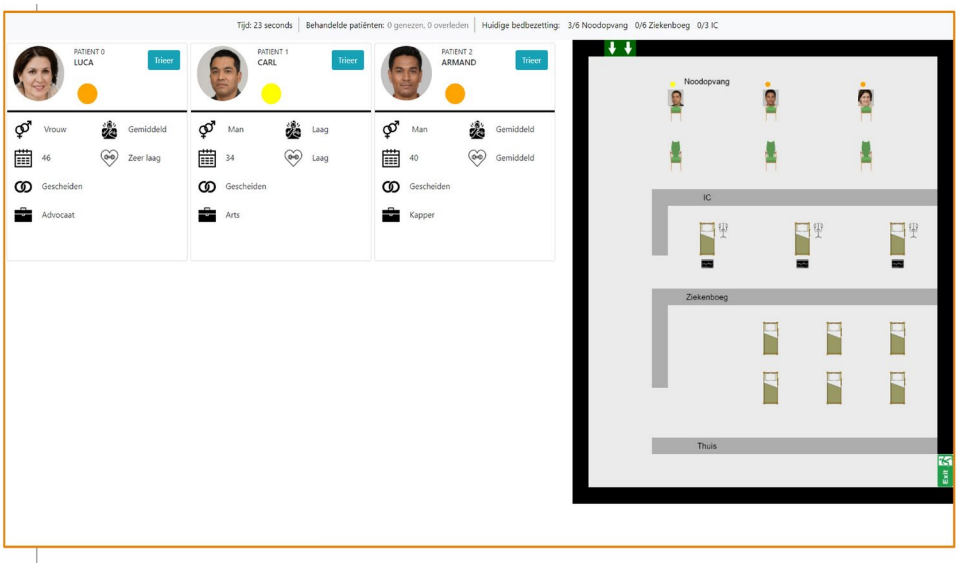

Fig. 2 Layout of the simulation of the task environment. The left panel shows the cards with information about the patients currently waiting to be triaged in the emergency shelter (in this example: four patients). Patient information concerned medical condition (severity of symptoms; the patient's fitness) and social factors (gender, age, marital status, profession). The colored dot indicates severity of symptoms: green=mild; orange=average; red=severe. The right panel shows the emergency shelter (upper section); the Intensive Care (second section); the Ward (third section); and home treatment (lower section). In total, there are 3 IC beds and 6 hospital beds available. The capacity of home treatment is unlimited. The screen's top bar shows summary information

[5] see https://www.newscientist.com/article/2308312-fake-faces-created-by-ai-look-more-trustworthy-than-real-people/.

severity of symptoms and social variables in such a manner that a coherent and believable pattern of properties emerged (e.g., no 18 year aged female patient with four children). The social variables were: age; gender; marital status; and profession. Four similar series of 16 patients were constructed for the purpose of this experiment; one series per *Type of Human-Agent Team* (see Sect. 4.1.2).

In order to make potential ethical dilemmas in pandemic triage more immersive and palpable, a short narrative was composed, giving the patient an identity and background. See Fig. 3 for an example of a patient narrative.

4.1.4.4 Patient health model To simulate the health condition of a patient over time, a patient-health model was developed. This model used as input: the severity of symptoms, the properties of the patient (e.g., age, fitness), and the assigned medical treatment (i.e., intensive care; ward; or home treatment). The impact these parameters on the patient's health were simulated in a plausible manner, but we do not claim medical validity in any way. The purpose of the model was to generate plausible feedback to the participant, and to generate standardized and uniform outcome measures.

4.1.4.5 Ethical Guidelines The guidelines used in this experiment are a simplified version of the official guidelines that have been prepared for use in hospitals in the Netherlands [45]. The objective of the official guidelines is to achieve triage decisions in accordance with ethical safeguards as established by the federation of medical specialists. They specify how in crisis conditions the scare medical resources should be assigned to patients, taking both medical and non-medical parameters into account. An example of a non-medical consideration is that a patient with a healthcare-profession who is contaminated on duty, is more eligible for medical care than a similar patient with a different profession, or with the same profession but not contaminated on duty. We prepared a simplified version of these guidelines for this study (see Appendix B—Ethical guidelines).

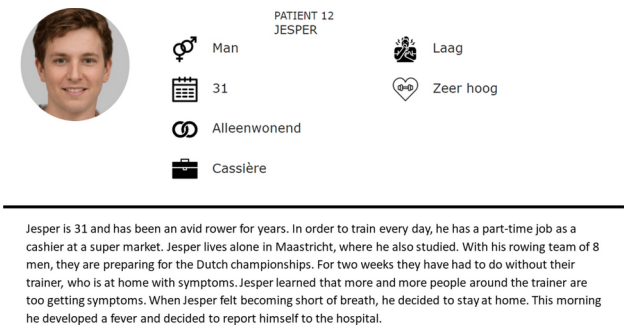

Fig. 3 Example of patient card (translated from Dutch)

4.1.4.6 Instructions to participants The participants received an instruction video, accompanied with handouts. They were told that under normal circumstances, the physician decides with only the interests of the individual patient in mind. But that there is now a crisis situation, requiring the physician to look beyond the interest of the individual patient, and take decisions that benefit the entire group of patients most. Participants were informed about the guidelines to be used for allocating the scarce medical resources in such a manner that 'good is done for as many people as possible'. See Appendix C—Instruction to participants for the full instruction.

4.1.4.7 Eliciting participants' personal moral values To investigate whether the participant has moral control over triage decisions by the team, we elicited the participant's personal moral values. This moral value elicitation was conducted after the participant had completed the base-line condition, and before any of the TDP-sessions. This was to ensure that the participant had acquired some knowledge on the role of patient properties on triage within this experimental task. For each of the four social factors, the participant was asked to evaluate two statements. The first concerned the attributed relevance of the social factor in general when taking triage decisions. For example, the questions concerning age were: 'with scarce beds, certain patients may be given priority on the basis of age' (agree/ not agree). The second question concerned in what direction the social factor should affect a triage decision. For example, if a participant chose 'agree', they were then asked to evaluate the following statement: 'in general, people under 60 are more/ less eligible for a bed with scarce beds than people over 60' (more/less). Appendix D—Moral value elicitation shows the full set of statements that were used to determine a participant's personal moral values.

4.1.5 Design of supporting intelligent agents

Three designs for collaborating with Intelligent Decision Support Technology have been worked out for this experiment (see Sect. 4.1.2). See [21] for an extensive discussion on these designs for human-teaming.

The agents were provided with algorithms to apply ethical guidelines for triage decisions (see Appendix E—Calculating triage score), as well as with a model of the individual participant's personal moral values, as elicited in

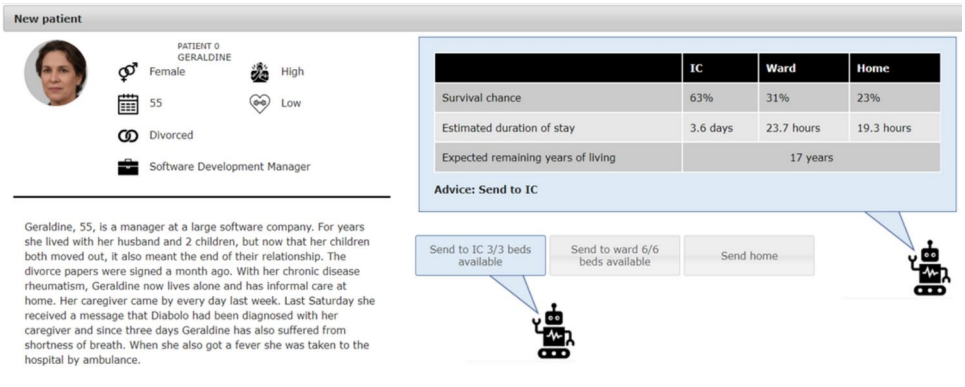

Fig. 4 Illustration of intelligent agent support under conditions of TDP-1. The upper right section shows that the agent provides statistical information obtained from data on other patients; below that: the agent gives decision advice

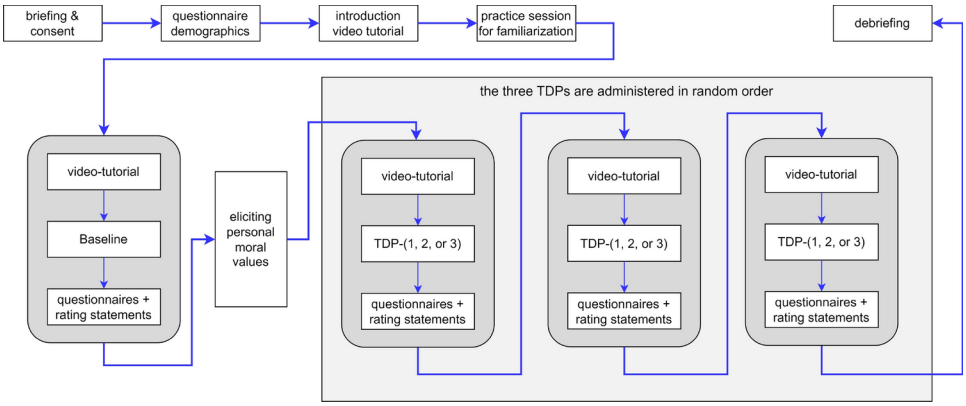

Fig. 5 Flowchart of the procedure of the experiment

the beginning of the experiment. Details are explained in Sect. 4.1.4.7.

Figure 4 illustrates how agent support in TDP-1 worked out in the experiment. It was emphasized to the participants that the agent's information and advice was based upon preliminary data on other patients, while the pandemic was still in its early phase. Thus, although they should consider data and figures as valuable information, participants were warned not to regard it as the undisputable truth, as the given advice may contain errors and inconsistencies.

4.1.6 Procedure

Figure 5 shows a flowchart of the experiment. The experiment was conducted in a quiet room, three participants per session. Participants sat in front of a laptop on which instruction videos and the task was presented. After briefing and the signing of the informed consent, questionnaires were administered for collecting demographic data. Then the experimenter introduced the experiment to the participants; explained the triage task to be conducted and how the guidelines should be applied. Participants also received a printed version of the guidelines (see Sect. 4.1.4.6), which they were allowed to use at any time during the experiment.

Then participants were presented a tutorial video on how to conduct the triage task (approx. 7 min). After watching the video, participants had the opportunity to ask questions to the experimenter. Then participants first received a practice session to make them familiar with the interface and task instructions. This involved, for example, how to select a patient for examination, to be familiar with overview information shown in the upper line of the screen, how to read a patient's information card, how to refer a patient to IC, ward, or home treatment, and so on. Conducting triage was not part of the familiarization.

All participants first conducted the baseline condition. After completion, a questionnaire was administered (see Appendix F). Then the procedure for eliciting the participant's personal moral values for medical triage was started. The obtained responses were used to instantly and automatically define an algorithm that reflected the individual participant's values, and the algorithm was implemented into the support agent for the design conditions TDP-2 and TDP-3. Thus, every participant had its personal intelligent agent in these conditions.

The design conditions TDP-1, TDP-2, and TDP-3 were presented in random order for each participant. Before the participant commenced a TDP-session, a short tutorial video was shown, explaining how the collaboration with the intelligent agent was organized. After each TDP-session, the same questionnaire as following the baseline-condition was administered. After completing all tasks, participants were extensively debriefed about the experiment.

4.1.7 Measures

The complete version of the used questionnaires can be found in Appendix F—Questionnaires..

4.1.7.1 Demographic properties Prior to the experiment proper, the following demographic data were collected: the participant's age, gender, education level, and computer experience.

4.1.7.2 Participants' perception of the task The objective was to present a plausible and immersive environment that evokes in the participant the experience of making decisions in morally sensitive situations. We asked participants to reflect upon the triage sessions, and to evaluate questions on likeability, difficulty, moral load (on 5-point scales), and believability (3-point scale) of the task.

4.1.7.3 Subjective control Subjective control was assessed by a questionnaire on self-assessed control, and by recalling on screen the decisions the participant made for a particular patient, and to ask the participant to state its contentment with the earlier made decision. For reasons of time, this was not done for all 16 patients, but for four randomly selected patients.

4.1.7.4 Normative control For each triage decision it was assessed whether or not it complied with the guidelines, as determined by the medical domain expert (see Appendix A—Medical expert's judgments on triage decisions). The proportion of decisions that were compliant with the guidelines was calculated.

4.1.7.5 Moral control By using the obtained values that participants in advance assigned to social factors (i.e.,: age; gender; marital status; and profession), we were able to determine for each triage decision whether it was either consistent with the participant's moral values (i.e., a score of +1), inconsistent (i.e., a score of − 1), or unrelated to these social factors (i.e., a score of 0). If, for example, the value elicitation revealed that a participant assigns priority to patients that are married, compared with patients who are single, then a decision to give intensive care treatment to a married patient was interpreted as 'consistent with personal moral values'; a decision to refer the patient to a hospital bed was interpreted as 'unrelated to personal moral values', and a decision to refer for home treatment was considered as 'inconsistent with personal moral values'. The compliance with the participant's moral value was determined for each triage decision, and the **mean moral-compliance** was calculated over all 16 triage decisions of a condition.

4.1.7.6 Performance *Number of surviving patients*: the patient health model (see Sect. 4.1.4.4) was used to determine how many patients survived infection and treatment after one (simulated) day.

Health of surviving patients: the patient health model was used to assess the health of the surviving patients, one fictitious day after being triaged.

4.2 Results

All participants confirmed to understand the purpose of the experiment, and the nature of the fictitious situation and the task presented to them. They all expressed that they considered themselves to be able to do the triage task with this perspective in mind.

4.2.1 Participants' perception of the task

Figure 6 shows results on participants' evaluation of likeability, difficulty, and believability of the task.

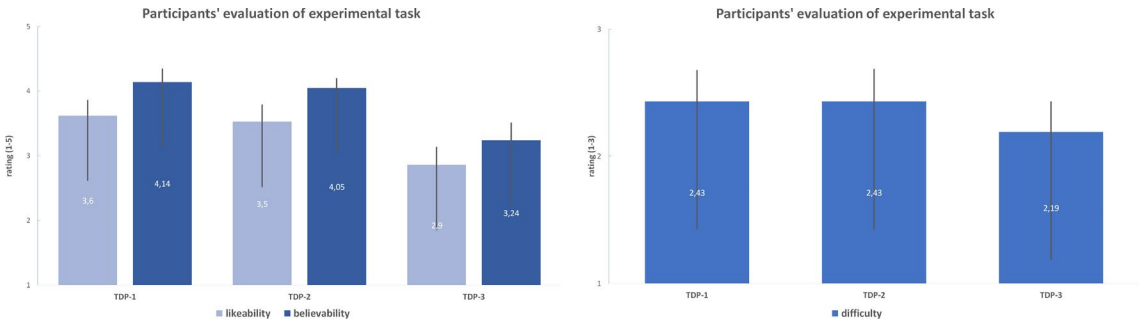

Fig. 6 Participants' evaluation of the experimental task with respect to likeability and believability (left pane, 5-point scale), and difficulty (right panel, 3-point scale), split by type of team

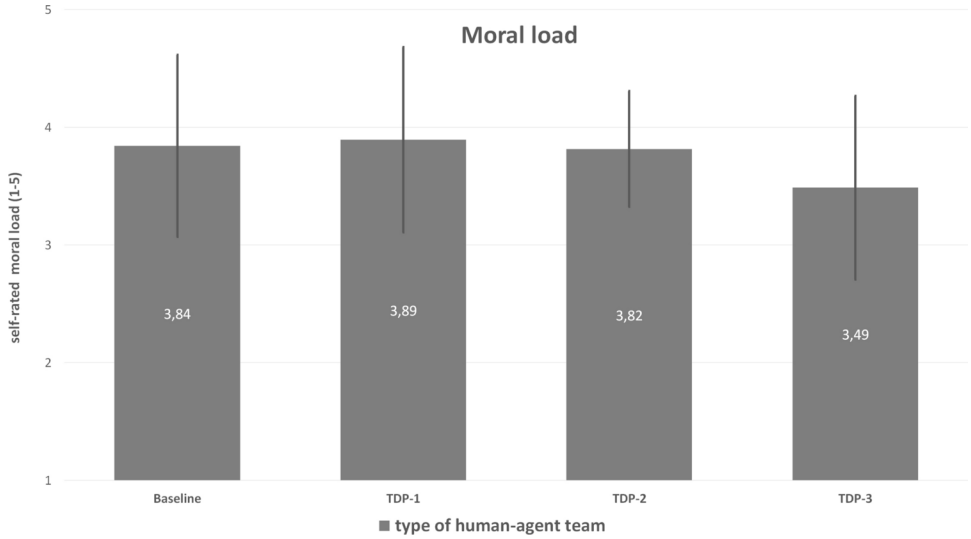

Fig. 7 Self-rated moral load experienced during performing triage, split by team design pattern

Table 1 Statistics of pairwise comparisons between conditions on moral load (* = statistically significant difference)

	Baseline	TDP-1	TDP-2	TDP-3
	Human makes decisions—no agent support	Human makes decisions—agent provides support	Human and agent divide patients	Human pre-instructs agent —agent decides autonomously
Baseline		Z = −.82, n.s	Z = −.46, n.s	Z = − 2,.48, p <.05*
TDP-1			Z = −.77, n.s	Z = − 2.86, p <.01*
TDP-2				Z = − 2.91, p <.01*

Friedman tests were performed with *Type of Human-Agent Team* as within-subjects factor, and with likeability, difficulty, and believability as dependent measures. Results reveal that participants did not like the task equally under the different conditions, as indicated by an effect of *likeability*: χ^2 (2, N = 21) = 10.3, p <.001 with the lowest score given for the *supervised autonomy condition* (TDP3). The experiment leader reported that several participants indicated to feel less involved in the task under TDP3 than in the other two collaboration conditions.

With respect to believability, participants evaluated the believability of the task differently for the three TDPs (χ^2 (2, N = 21) = 9.46, p <.001), with TDP3 receiving the lowest believability score. Subsequent Wilcoxon signed ranks tests showed that participants evaluate an artificially intelligent agent that autonomously makes triage decisions based upon pre-instructed medical and moral guidelines (TDP3) as less believable than agents that provide information and advice (TDP1) (Z = − 2.3, p <.01, ES = 0.41), and also less believable than agents that conduct triage on patients assigned by the human (TDP2) (Z = − 2.4, p <.05, ES = 0.32). The experienced difficulty of the task was not affected by type of human-agent collaboration.

For this study into MHC we selected a medical triage task because decisions potentially involve morality. We deliberately arranged a high influx of patients, scarce resources, and guidelines that are likely to not cover all decision problems. This resulted in situations in which participants could not avoid taking moral principles into account when making decisions. It was believed that this would evoke *moral load*, i.e., the level of effort dedicated to considering and applying moral principles. Comments from participants after the experiment provide evidence that this did happen: several participants indicated that they experienced some of the triage decisions as oppressive. To evaluate the experienced moral load, we asked participants after completing each round of patients, to rate statements on moral load on a 5-point Likert scale. For example, one statement was "*I had to make choices that I would have been reluctant to make in real life*". The mean of participant's ratings was calculated. Figure 7 shows the results.

A Friedman test was run on moral load with *Type of Human-Agent Team* as within-subjects factor χ^2(3, N = 19) = 9.44, p < *.05*), showing that the nature of collaboration between human and agent affected the experienced moral load, as rated afterwards by participants. Table 1 shows the pairwise comparisons between conditions, showing that participants experienced less mental load when their AI-partner took the decisions on behalf of them; less than in any of the other conditions.

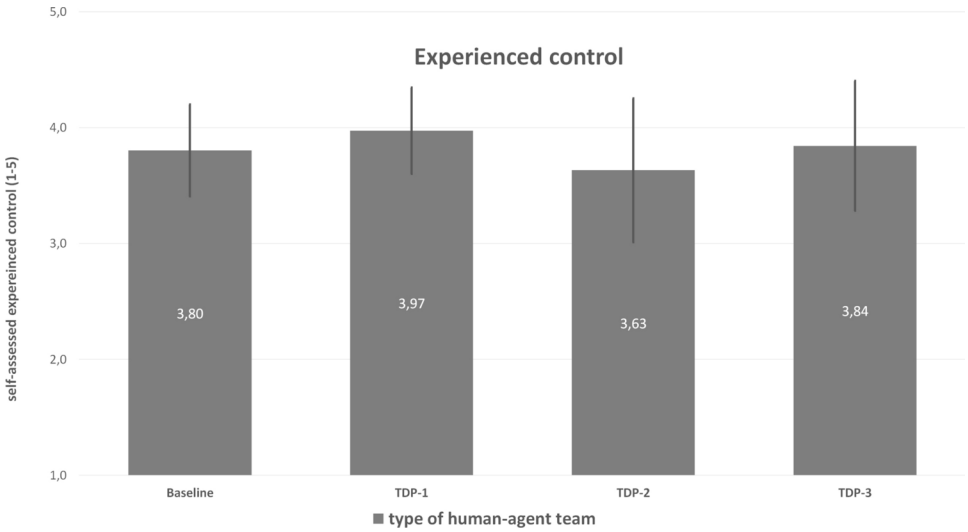

Fig. 8 Self-rated subjective control

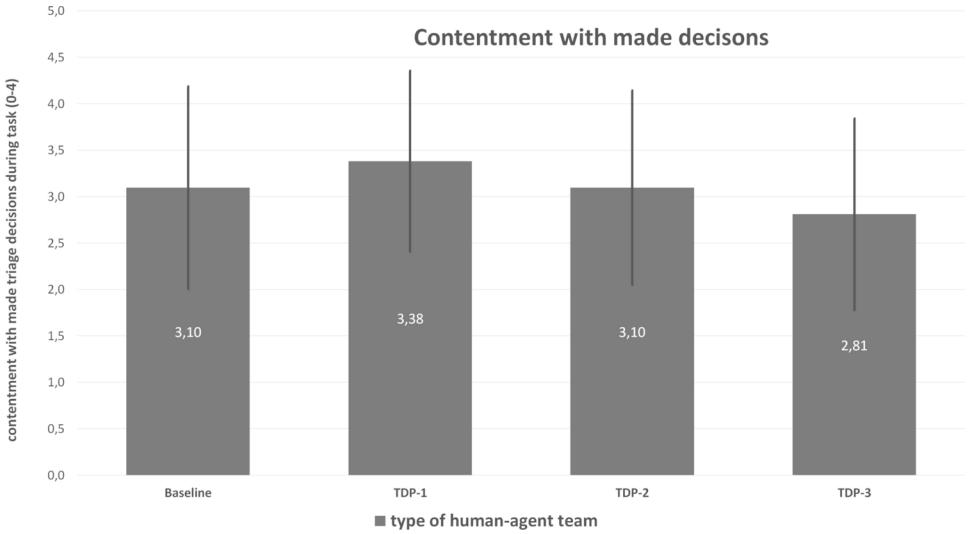

Fig. 9 Participants' contentment afterwards with decisions made during task

4.2.2 Subjective control

Self-assessed control: Fig. 8 shows the results of self-assessed control.

A Friedman test revealed no effect of Type of Human-Agent Team $\chi^2(3, N = 19) = 6.17$, *n.s.*, showing that the team design did not affect the extent to which participants experienced control over the decision making process.

Contentment with made triage decisions: Fig. 9 shows the results.

The data show that participants were afterwards, overall, relatively content with the decisions that the team made while in operation. The average contentment score was 3.1, which suggests that participants felt in control over the decision of the team. The lowest average contentment over decisions was found when the intelligent agents made the actual decisions (TDP-3), although a Friedman test showed no difference between conditions $\chi^2(3, N = 21) = 3.74$, *n.s.*.

4.2.3 Normative control

Results are shown in Fig. 10.

Participants were over-all able to adhere to the ethical guidelines in only about half of the decisions (54%). An analysis of variance reveals that the type of team design affected adherence to ethical guidelines ($F(3,60) = 18.79$, $p < 0.05$). The lowest adherence was found in the baseline condition (participant conducted triage alone, without assistance of an AI-partner). Pairwise comparisons show significant differences between all TDPs, except for the baseline and TDP-3 comparison (see Table 2).

The collaboration condition in which the participants divided doing triage on patients among themselves and the AI-partner (TDP-2) resulted in the highest proportion of ethically compliant decisions.

Fig. 10 Proportion of decisions in accordance with ethical guidelines

Table 2 Statistics of pairwise comparisons between conditions on compliance with ethical guidelines (* = statistically significant difference)

	Baseline	TDP-1	TDP-2	TDP-3
	Human makes decisions—no agent support	Human makes decisions—agent provides support	Human and agent divide patients	Human pre-instructs agent— agent decides autonomously
Baseline		t(20) = 3.15, $p < 0.05$*	t(20) = 5.58, $p < 0.05$*	t(20) = 1.57, $p > 0.05$
TDP-1			t(20) = 3.31, $p < 0.05$*	t(20) = 2.76, $p < 0.05$*
TDP-2				t(20) = 8.37, $p < 0.05$*

Table 3 Mean compliance of triage decisions with participant's personal moral values (sd in parenthesis)

Baseline	TDP-1	TDP-2	TDP-3
Human makes decisions - no agent support	Human makes decisions - agent provides support	Human and agent divide patients	Human pre-instructs agent - agent decides autonomously
− 0.02 (0.55)	0.12 (0.98)	− 0.22 (0.40)	− 0.14 (0.11)

4.2.4 Moral control

Table 3 shows the mean moral compliance score, split by condition.

Results show that participants' scores on compliance with personal moral values are all close to zero, for all conditions. Furthermore, standard errors are high. An analysis of variance reveals that the intercept of mean moral compliance is not different from zero ($F(1,20) = 0.45$, *n.s.*). This demonstrates that the measured personal moral values do not influence triage decisions. Furthermore, no differences between conditions were found ($F(3,18) = 2.49$, $p = .09$).

4.2.5 Performance

Figure 11 shows the results on number of surviving patients.

A repeated measures variance of analysis was performed with *Type of Human-Agent Team* as within-subjects factor, and with the number of surviving patients as dependent measure. Results show an effect of *Type of Human-Agent Team* ($F(3, 18) = 33.7$, $p < 0.01$). Table 4 shows the statistics for pairwise comparisons.

Figure 12 shows the results on the health of surviving patients, one fictious day after being triaged.

Results show that, according to the patient health model, the health of those patients that survive is high. Those patients that survive seem to profit from the assigned treatment. An analysis of variance shows an effect of *Type of Human-Agent Team* ($F(3, 18) = 3.24$, $p < 0.05$). The pairwise comparisons show significant differences between the Baseline and TDP-3 ($t(20) = 2.82$, $p < .05$), and between TDP-2 and TDP-3 ($t(20) = 2{,}14$, $p < .05$).

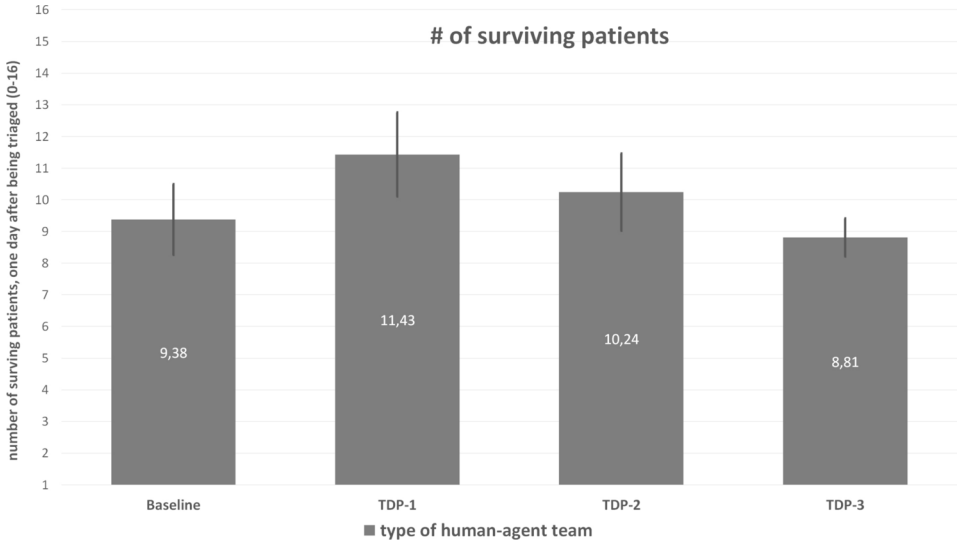

Fig. 11 Number of surviving patients, one fictious day after being triaged

Table 4 Statistics of pairwise comparisons between conditions for number of surviving patients (*=statistically significant difference)

	Baseline	TDP-1	TDP-2	TDP-3
	Human makes decisions-no agent support	Human makes decisions-agent provides support	Human and agent divide patients	Human pre-instructs agent-agent decides autonomously
Baseline		t(20) = 7.1, $p < 0.01$*	t(20) = 2.26, $p < 0.05$*	t(20) = 2.03, $p = 0.06$
TDP-1			t(20) = 2.81, $p < 0.05$*	t(20) = 8.60, $p < 0.01$*
TDP-2				t(20) = 4.94, $p < 0.01$*

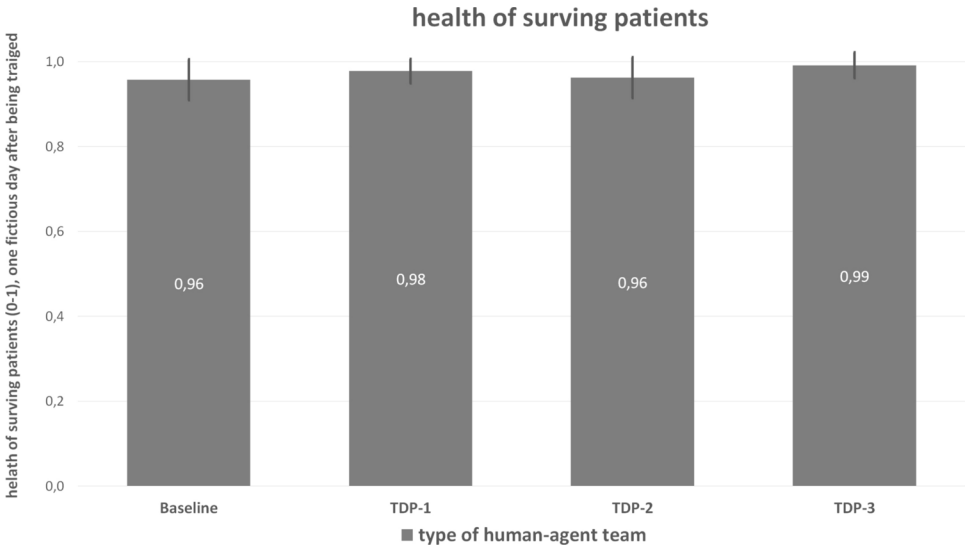

Fig. 12 Mean health of surviving patients, one fictious day after being triaged

5 Discussion

The advancements in AI offer new possibilities for its deployment in complex tasks, but it also raises the question of whether and how humans can preserve meaningful control over the technology's behavior. This is especially crucial for tasks that involve moral and ethical principles. To assess MHC in a human–AI collaboration task, we propose three measurable components: subjective-, normative-, and moral control. The primary goal of this study is to empirically evaluate the value of the proposed method and measurements for assessing MHC in a human–AI collaboration task. As a use case we developed a simulated environment

for conducting medical triage under difficult circumstances (i.e., 'code black' during a pandemic) requiring participants to make decisions with moral implications. A secondary goal of this study was to apply the proposed method for exploring the effects of different human–AI team designs on the distinguished components of MHC. It is argued in Sect. 3 that MHC is dependent upon many factors, such as how tasks are orchestrated within a team, the conditions of the specific context, the interactions taking place, and what authorities are granted to humans and agents. The human–AI team designs used in our study differed with respect to the delegation of tasks, the agent's autonomy, and its decision-making permissions. They reflect common human-agent team designs in the field.

5.1 Subjective control

5.1.1 Participants' opinions and experiences

Results on measures of believability and moral load, as well as comments that participants made to the experimenter, indicate that conducting triage in the task environment generated the immersion and ethical involvement required for appreciating the moral consequences of decisions [36]. Participants experienced the task over-all as likeable (i.e., engaging to do). It should be noted that the participants had no medical expertise, so their opinions reflect the projection of themselves into the role of physician. The administered questions may evoke different responses from medically-schooled people, although a study by [21] reported similar opinions obtained from medical experts when asked for the plausibility of physician-AI collaborations in the future. Asking (potential) stakeholders for their opinions and expectations about the potential deployment of intelligent systems is important to shape future use of technology. A limitation of asking participants is that they may not all have the same frame of reference and may adopt different sources of information and interpretations to respond. Though participants' opinions and judgments are relevant, we do not consider their feedback as a final result. But their evaluation can be valuable for subsequent specialized research.

5.1.2 Measuring subjective control

In response to our questions addressing their judgments, participants indicated to have felt in control over the triage process (see Fig. 8). They claimed to have had a relatively good overview of admitted patients and of the available resources and said not to have felt too much time pressure upon their decision making. This is surprising, as they also perceived the task as difficult (see Sect. 4.2.1). Furthermore, participants felt to have made an appropriate triage decision on most of the patients. When participants were afterwards asked to reflect upon their triage decisions during the task while taking the available information at the time into account, they responded to feel content with the made triage decisions (see Fig. 8). One question administered to assess subjective control was: '*I believe that my decisions have resulted in a good distribution of care among all patients.*'. Though intended to measure subjective control, high scores to this question may not necessarily imply that the participant experienced a high level of control. A participant may have felt no to little control over the agent's decisions, but still judged that the available care to be well distributed. It may also be that the participant understood the question as referring to the decisions that they personally made, excluding the ones that the agent made.[6] Hence, the question may not be a good indicator for subjective control.

Taken together, the picture emerging from these results suggest participants feeling in control. This is not fully in conjunction with the observations of the experiment leader who reported that many participants confessed to find the task difficult and expressed to being aware of the decision dilemmas. Some shared their concerns with the experiment leader about whether they had done the right thing.

When investigating the question whether a human feels control over an AI-based agent, the human's *trust* in the AI-agent is very important. It may therefore be useful to include questions measuring the human's trust in the technology, when measuring a human's control over AI-agent(s) in the team. However, trust and control are distinct concepts. Recently, [12] addressed control with different notions of trust. Research in the context of human-agent collaboration often approaches human's trust as a set of beliefs and expectations, formed by experience and interactions with the AI. A more formal notion defines trust as "*the willingness of a party (e.g., the human) to be vulnerable to the actions of another party (e.g., the AI) based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party*" [46, 47]. This notion considers trust as a 'contract' between humans and AI whereby the human requires the AI to behave in a specific way or to perform a specific action in specific circumstances [48]. An AI-based agent is trustworthy to the human if it shows to be capable of maintaining the contract. This latter notion of trust better fits the concept of MHC. Including measures of this notion of trust in future studies may improve our understanding of experienced control.

[6] we thank an anonymous reviewer for its remarks regarding this question.

5.1.3 Effects of team design on subjective control

When participants had to supervise agents that conducted triage autonomously (as in TDP3), they felt less moral load, liked the task less and indicated to be less involved than in the other conditions. Apparently, working with a fully autonomous AI-based agent makes the task less attractive, perhaps because such an agent provides few cues about its behavior to the human [49]. There is evidence that when AI-partners provide (visual) cues of their actions, the participant is more likely to consider the agents as a teammate. Furthermore, agents that promote transparency through cueing contribute positively to the team's performance [49]. It is possible that, due to a lack of cueing, the agents and participants in condition TDP3 failed to establish a team connection. When using a team design involving highly autonomous agents, it is recommended to implement measures that promote connection between human and agent, such as supporting transparency or by providing explanations [50].

5.2 Normative control

5.2.1 Measuring normative control

The alignment of a participant's decision with the guidelines provides evidence for normative control. A problem with many real world tasks, especially when decisions also require moral considerations (e.g., medical decision making, military command; fire-fighting, et cetera), is that inevitably situations will arise for which the guidelines provide no unambiguous solution. Autonomous and automatic assessment of whether decisions align with guidelines is therefore seldom possible. This was also the case in our pandemic triage task. We decided to involve an experienced medical expert to evaluate the patient cases we presented to the participants. This facilitated the face-validity of norm-decision's. However, in complex tasks such as medical triage, humans may dynamically change their goals to adapt to actual or anticipated changes in the task environment. In our study the agents lacked the capabilities to align their advice and actions with possible goal changes of the human. Recent research proposed a method to enable agents interpreting changes in human intent and goal prioritization [51]. It requires further research whether the development and implementation of individualized models of human intent in agents will enable elevated levels of guidelines compliance. When team decisions align with the guidelines this provides evidence for normative control. There is of course always the possibility that when a participant has no to little normative control, the team nevertheless (coincidentally or otherwise) adheres to the guidelines. However, systematic and sustained alignment with norms is generally a demonstration of normative control. especially when this is supported by other evidence (e.g., explanations by the participant).

Our measuring approach showed that participants were able to adhere to the ethical guidelines in about half of the decisions (54%). This seems a low figure, but it is obviously caused by the bottleneck conditions ('code black') that we intentionally created. The rapid flow of patients coming to the emergency unit in combination with the scarce available beds made it often impossible to act in accordance with the guidelines, simply because there were too few beds available to assign the medical care that the patient should receive. Thus, all participants were presented with situations that forced them to give in with respect to the guidelines. However, many participants developed smart strategies to limit the overall number of guideline violations. Consider, for example, a patient that is severely ill but is nevertheless relatively fit. The guidelines (see Appendix B—Ethical guidelines) prescribe to assign such a patient to intensive care. However, suppose that the participant expects that this patient may also survive without IC-treatment. A participant that in anticipation of an intensifying crisis decides to assign the severely ill patient to ward rather than to IC, saves an available IC-bed for future use. This bed may prove valuable later, when new patients with even worse symptoms arrive. Our participants discovered that strictly applying guidelines for each individual patient would create bottlenecks later on. They therefore 'looked ahead' and took into account of what was likely to come. By intentionally relaxing the ethical guidelines for individual patients, in the end they managed to achieve a higher compliance with ethical guidelines for the entire group of patients, thus being better at achieving to do good for as many people as possible.

5.2.2 Effects of team design on normative control

Best compliance with guidelines is achieved by teams in which the human has the final word over triage decisions, either directly (as in TDP1) or indirectly, through dividing responsibilities between human and AI-partner (as in TDP2). In teams where the agent decides autonomously (as in TDP3), the human had no opportunity to exert influence on the decision making process, eventually leading to a lower level of compliance with the guidelines.

In our study, teams designed as in TDP1 and TDP2 have the highest number of surviving patients (see Fig. 11). That shows that these teams not only better complied with the guidelines; they also succeeded in doing good for as many people as possible.

5.3 Moral control

5.3.1 Measuring moral control

Participants were instructed to follow the ethical guidelines as much as possible when conducting triage. However, just as in real life, situations arose for which the guidelines provided no clear-cut answer. To nevertheless force a decision, participants were told to follow their own moral norms in such cases. In our study we wanted to measure whether participants' personal moral values affect their triage decisions. We asked participants prior to the experimental task about their view on what role a patient's *age*, *gender*, *marital status*, and *profession* should play when to choose between patients that, medically speaking, can be regarded as equal. At the time of the experiment the COVID-19 crisis reached its peak, and it was feared by many that *code black* had to be implemented in hospitals. In the media there was a heated debate whether social factors should be taken into account when assigning medical care. By asking participants to rate statements we developed a personal profile of moral values on these factors. We conjectured that if a participant assigns a higher level of medical care (e.g., an IC-bed) to patients with social characteristics that the participants values, and assigns a lower level of medical care (e.g., home treatment) to patients with social characteristics that the participant considers less important, then this can be regarded as a manifestation of moral control. It would show that the participant accomplished to implement its personal moral values into triage decisions. However, we found no relationship between triage decisions and moral values. One possibility is that the participants had no moral beliefs regarding the included attributes of patients (e.g., marital status, profession, age) with respect to making triage decisions. It may also be that the abstract way of presenting patients (see Fig. 3) does not sufficiently trigger the affect in participants that is possibly required to invoke moral considerations. Based on our experiences, we instead believe that our method for eliciting moral values was not suitable for this study. A drawback of our approach is that we simply *asked* our participants how they valued a particular social characteristic, rather than presenting situations to them in which they could express their preferences through behavior and decisions. It is likely that simply asking participants is not sufficient to initiate a deep reflection upon the role of the proposed social factors. To obtain contemplated opinions from participants, more depth on moral beliefs is needed. Rather than asking straightforward questions, interactive dialogue-based explorations may be better for measuring moral beliefs [e.g., 52]. The importance of emotions for moral judgment have been emphasized by many in the field [e.g., 53, 54, 8]. Although different methods for disclosing people's moral values have been proposed (see for review of methods: [55]), and some of them have been designed for practical applications in organizations [e.g., 56], these methods do not explicitly incorporate the emotional component of moral values. In our study, it would have been desirable to use a method for value elicitation that embeds the moral issues in immersive experiences, to evoke the emotions associated with decision forthcoming from the values.

Other studies provide directions for future research into the relationships between morality and ethics, trust, and control in human-agent teams. For example, [13] explored the relationship between ethics and trust in a human–AI team and found that even when agents violate common ethical standards, humans preserved some levels of trust in the AI. Although humans disagreed with the agent's recommendation, they nevertheless trusted the AI as a teammate. The authors conjecture this to be a result of 'automation bias', the tendency for humans to favor suggestions from automated decision-making systems and to ignore contradictory information. The finding that some participants sustained trust in AI after it recommended unethical decisions may be the result of rationalization. Apparently, the effects of an agent's (non-)compliance with moral values on human trust is subtle, and is not determined solely by the agent's performance. Textor and colleagues used measures of trust reflecting the human's assessment of the agent's intrinsic trustworthiness. It would be interesting to follow up on this by using measures reflecting the human's trust in the agent's compliance with their 'contract', i.e., requiring the agent to behave or perform in a specific established manner. This latter notion of trust reflects more closely the proposed notion of meaningful human control, as it concerns the self-assessed (or experienced) capacity of humans to regulate the behavior and performance of human-agent teams.

5.4 Limitations

The findings are an important first step for measurements of MHC in human–AI teaming, but the study also has limitations. To enforce the emergence of moral dilemmas, we arranged a rapid influx of patients with severe symptoms, while limiting the available intensive care beds. An example of a moral dilemma is that with only one IC bed available, the participant must choose between two or more patients that, medically speaking, are both eligible for the remaining IC-bed. Who to choose, and on what grounds? The participant was free to choose the order in which they admitted patients for examination and triage. This is realistic and conform the official guidelines. However, by making a triage decision the participant also shapes the context for subsequent patients, as a decision influences the remaining available medical resources. Thus, despite that all participants

received the same set of patients in a particular team design condition, the circumstances for the triage decisions were not equal for all participants. Our ambition to create a natural and plausible decision-making task for our study conflicted with executing control on moral decision making.

Another limitation related to this is the presentation of patients using patient cards, showing a picture of the fictious patient, a narrative with anamnesis, and a listing of medical and social variables. Although some participants afterwards confided to have felt moral tension to the experiment leader, it may be that the abstract presentation of patient information is not sufficient for eliciting the feelings, emotions and affect in participants to make them acknowledge the moral relevance of the various patient characteristics. Movie clips, or even live-acting role players may be suitable to establish a more immersive and engaging environment for experimentation.

Thirdly, the study used a sample of twenty-one participants, all with a higher education background. The sample size is not very large. Furthermore, the participants had no medical background. Due to this unfamiliarity with the task, participants probably find it hard to manage the difficult task of triage as it is, let alone collaborating with, and exerting control over, an unknown artificially intelligent agent while doing so. It may be that medically schooled participants with triage experience would be able to exert better control than participants of the present study. This obliges a careful interpretation of outcomes. Results on MHC, and also on how MHC is affected by features of the human–AI team design, should be regarded with care. Its validity for practical applications in the medical field cannot be taken for granted.

Fourthly, we argue that normative control is needed to achieve that decisions comply with ethical guidelines. However, we also argue that in complex real-world settings, it is often difficult to determine whether decisions do or do not comply with ethical guidelines. In our simulation we intentionally created opaque ethical conditions, in which the 'right' answer is often unknowable and ambiguous. How then can compliance, and thus normative control, be assessed in such situations? We solved this by requesting a medical expert to decide beforehand for each patient, for each possible triage decision, whether that decision complies with the guidelines or not. Thus, this expert judgment was made on a one-by-one basis of patients individually, not by taking the history and context of the actual decision context into account. As we suggested in Sect. 5.2, our participants may have occasionally violated the rules deliberately on individual patients with the goal to achieve a higher overall compliance with ethical guidelines. The design of our study does not allow determining whether participants actually used this strategy, and whether it indeed produced the suggested outcome. It would have been possible when we had asked medical experts to evaluate triage decisions in context, including all relevant conditions. Note that this solution implies a more labor-intensive procedure than the one we used. In addition, for logistic reasons we consulted only one experienced medical expert. It is fair to assume that the quality of evaluating compliance could be improved by including multiple experts.

Fifthly, Likert scales are often used because they enable easy and rapid collection of input from participants. However, we believe in hindsight that Likert-scales have drawbacks for measuring how participants experienced situations, and for assessing how they retrospectively evaluate their state of mind when doing the task during the experiment. One crucial drawback is that the participant is requested not to rate an actual experience, but to rate an experience in the past, by thinking back how the decision problem felt at the time, and whether moral deliberations came to mind. People may have difficulty at recalling a vivid image of the experience, which requires bringing back the then felt values, affects, and emotions to the surface. A question for experimenters is what alternatives exist. It would be best if constructs relevant to MHC, such as control, commitment, and moral load, could be measured in real time in the context. However, as administering measurements during the task also intervenes with the task itself, this may not be a viable alternative. A more viable option would be to carefully reconstruct the context, and to request the participant to re-enact the task (i.e., taking the triage-decision again), and to administer a question-guided open interview [cf. 57, 58].

6 Conclusion

Despite significant work on the theoretical properties of MHC and efforts to provide operational definitions, there is limited practical experience with measuring MHC. As it is expected that AI-systems will increasingly be used to collaborate with humans on tasks that involve moral considerations and choices, it is important that practical methods become available to assess whether a human–AI system is under MHC. This paper aims to contribute to that need. We propose three intuitive components of MHC: *subjective*, *normative*, and *moral* control. Furthermore, we propose measurements for these in a specific use case of human–AI collaboration: medical triage in a pandemic crisis. To be of use in practical settings, collecting measures should be feasible; measures should be sensitive enough to detect differences in MHC; and the set of measures should cover the relevant aspects or manifestations of MHC.

6.1 Feasibility

The results of our experiment show that the measurements on subjective control are relatively easy to collect. As subjective control refers to experiences of the user, these measures mainly involve questionnaires where participants rate a series of propositions. In our study they indicated on Likert-scales how well (or not) a proposition matched their own experience or opinion. These questionnaires were easy to administer and yielded useful information about the control that participants felt over the collaboration and decision-making. However, in the paper we express drawbacks on the use of Likert-scales, primarily because they request the participant not to rate an actual experience, but to rate a previously experienced state, feeling or judgment.

Collecting measures to assess whether participants' decisions correspond to guidelines (normative control) is not so easy as it perhaps may seem. In many real world tasks that require appraisal of different values there inevitably arise situations for which the guidelines are ambiguous. In the use case of our study we deliberately introduced that characteristic in the task. Determining whether a decision in a particular context aligns or not with guidelines requires taking the specific circumstances into account. For that reason it is hard or impossible to build in automatic assessment of normative control. In our study we involved an experienced medical expert to collect decision judgments, which solved the needs for our study. However, in complex real world tasks humans adapt to actual or anticipated changes in the task environment. Achieving normative control over AI-based agents in such circumstances requires agents with the capabilities to align their behavior to adaptive changes of the human. Determining whether mutually adaptive human–AI partners collaborate in accordance with normative guidelines will likely require a more resourceful measuring approach.

When thinking about measures intended to assess whether someone exerts moral control over an AI-based system, it is necessary to identify the moral values of the person. The presence of moral values in the brain of a person should be taken as a figure of speech rather than representing physical structures in the brain [59]. Moral values represent dynamic and interlinked thoughts, emotions and sensations [60]. We can invite people to reflect on these, and use the data to construct an explicit approximation of the person's moral values. In our study we asked people directly whether they felt that patients with particular social characteristics should receive more care than other patients. This is a straightforward and relatively easy approach to infer a person's moral values. Yet in hindsight we conclude that this method does not initiate the deep reflection that is needed to obtain the data for valid measurement of moral control [61].

6.2 Sensitivity

The results of our experiment show that the measurements on subjective control are sensitive enough to detect differences in MHC as a function of human–AI team design. A design in which the human exerts control over decisions prior to the task only (as in TDP3, where the normative and moral model of the human is used to instruct the agent in advance how to make decisions) is judged as less believable, less engaging, and less morally charging. Furthermore, people are more likely to feel discontent with triage decisions (made by the agent) in such a design. In a different study using medical experts [21], it was found that people tend to dissociate themselves from decisions taken by the agent, despite the instruction to maintain a critical stance [cf. 62]. This suggests that prior control over agents contributes little to MHC. This could be because people learn from experience that agents do not have a sophisticated appreciation of the context, hence make decisions that people do not agree with. Our study confirms this notion, as we found that when people have direct control over the team's actions, they take decisions that are better aligned with ethical guidelines, and that lead to better outcomes than people who have to rely on decisions taken by pre-instructed agents.

Measurements of normative control show that when the human has close control over the team during the task (as in TDPs 1 and 2) the decisions were better aligned with guidelines.

The measurements used in this study to assess moral control were not sensitive enough, unfortunately. Explanations and suggestions have been discussed above and in Sect. 5.3.

As argued in Sect. 5.4, the size and composition of the participants sample, and the simplification of medical triage in the experiment obliges a careful interpretation of outcomes. The distinguished designs of human–AI teams have been found to have different effects on MHC, but the questions whether these results will also apply in real-world applications demands further research in a more realistic setting.

6.3 Complementarity

The results of our experiment show that the three components can be used to measure MHC from different angles. Though complementary, the measures are not complete. Furthermore, the specification and implementation of measures for the distinguished MHC-components require domain-specific interpretation. For example, measuring subjective control in an automated driving task requires questions tuned to that specific task. Simply carrying over questions from a medical triage task is evidently not possible. Moreover, formulating questions that address the intended

measure in an unambiguous manner is not a straightforward issue, as also became clear during this study (see Sect. 5.1).

6.4 Final comments

The question how roles, tasks and responsibilities should be allocated in human-agent teams of the future needs careful consideration of a variety of factors [39, 63]. As technology develops and teaming between humans and autonomous systems becomes more common in the future, more research will be needed to better comprehend the nature of the complicated interactions between people and these systems. The presented findings are an initial step in an ongoing endeavor to measure MHC, highlighting the need for sustained efforts in fostering a synergetic relationship between humans and intelligent machines of the future. The presented findings do not provide definitive answers, but rather form a starting point for further exploration and understanding in this crucial area.

Appendix A: Medical expert's judgments on triage decisions

We asked an experienced medical domain expert to reflect upon the patient-cases that we presented to our participants. Our domain expert has over 20 years of experience as an ambulance nurse, a function in which making assessments on the nature and urgency of medical treatment is almost daily routine.

It was explained to the medical expert that code black was enforced in the scenario, implying that new patients come in at a rapid pace, and the available sources are insufficient to apply all patients with the care they need. Choices have to be made, and that guidelines have been issued to ensure that ethical safeguards are maintained while doing so. We familiarized our medical expert with the same set of guidelines as we had presented to our participants. The simulated hospital context was not provided, therefore, the number of available beds was not a factor in the decision making process. One by one, the 64 patient cards (see Fig. 3) were presented to the medical expert. For each patient, the expert indicated for each possible triage decision whether or not he considered it in compliance with the guidelines. Thus, for example, the expert may judge that for patient X an *IC-treatment* is in line with the guidelines, but *ward-treatment* and *home-treatment* are not. The expert was asked to explain his judgments. The session took, including three short breaks, 3 h to complete.

Appendix B: Ethical guidelines

The participants were instructed to use the following guidelines when conducting triage. These are a simplified version of the official guidelines that have been prepared for use in hospitals in the Netherlands [45].

See Fig. 13.

Fig. 13 Ethical guidelines; derived from the official guidelines that have been prepared for use in hospitals in the Netherlands

Medical guidelines:

- Nature of **symptoms** — **assign to:**
  - severe & high symptoms: → IC
  - moderate symptoms: → Ward
  - mild & very mild symptoms: → Home treatment
- Level of **fitness** required for treatment type:
  - IC: high fitness required
  - Ward: moderate or high fitness required
  - Home treatment: no fitness requirements

Social factors guidelines

- Profession
  - patients with a health care profession receive priority if their infection occurred on duty
- Age
  - patients of younger age groups have priority over patients from older age groups
    - age groups are: 0-20; 21-40; 41-60; 61-80; 80+
    - when deciding between patients *within the same age group*, age should be discarded in the decision making

Appendix C: Instruction to participants

The participants received an instruction video, accompanied with handouts. Below are summarized fragments of the information being conveyed to the participant.

"The scenario of this experiment is staged during a pandemic. The society has to cope with large number of victims and few available medical resources. The Ministry of Health proclaimed 'code black'. This means that it is accepted as a fact that care can no longer be provided as under normal circumstances. It implies that physicians on duty need to make choices which patients are eligible to receive intensive medical care, and which patients will be given less care than otherwise possible. This is called medical triage. The ultimate purpose of triage is to save as many lives as possible.

Code black means a fundamental change in the perspective of the physician. Because under normal circumstances, the physician decides with only the interests of the individual patient in mind. But under conditions of code black, the physician has to look beyond the interest of the individual patient, and take a decision that benefits the entire group of patients most. The purpose of medical triage under conditions of code black is to allocate the scarce medical resources in such a manner that 'good is done for as many people as possible'.

You will be playing the role of a physician on duty, conducting triage. All patients admitted are infected with the virus. You may select patients in any order for examination. By clicking on a patient, you will see a short anamnesis and summary information about the patient's medical status and social circumstances (see Fig. 3). The patient's fitness is important for assessing whether the patient will be able to survive intensive medical treatment. As the physician you refer a patient to either: treatment at home; hospital ward; or intensive care. Remember to keep in mind that there are no more than six hospital beds, and three IC-beds available. It is important that your decisions do good for as many patients as possible. You decide by making use of the information available about the patient and by following the guidelines as closely as possible [*the participant is explained how the guidelines should be used, see Appendix B*]. If you conclude that the guidelines provide insufficient support for making a decision on a patient, you are free to decide what you think is best. Be aware that patients come in sick, and their illness may deteriorate while waiting to be triaged. There is a chance that a seriously ill patient dies before you can make your triage decision.

In the experiment you will be faced with making difficult choices. Furthermore, new patients will be admitted at an irregular pace. While you are conducting triage, new patients will almost certainly be admitted. How many patients will be brought in during your duty is uncertain, but in general this adds up somewhere between 12 and 20 [*in fact, each condition consisted of 16 patients, but this was not told to the participants*]. On your first duty you will do the task by yourself. After that, you will perform additional duties, helped by intelligent technology. We will explain that later. Any questions?"

Appendix D: Moral value elicitation

To assess whether participants assign importance to social factors in the triage of patients, we asked participants to evaluate two statements for each of the four distinguished factors: age; gender; marital status; and profession. The first question concerned whether the participant assigned relevance of the social factor when taking triage decisions; the second questions was administered only when the participant responded 'yes'. The second question concerned the direction of how the social factor should affect a triage decision (i.e., receiving more or less priority to medical care).

See Table 5.

Table 5 Elicitation of value statements for four distinguished social factors: age; gender; marital status; & profession

Social factor	Question	Options
Age	When IC-beds are scarce, the patient's age should be involved when making a triage decision	Yes/no
	People under 60 are more/less eligible for an IC-bed than people over 60	More/less
Gender	When IC-beds are scarce, the patient's gender should be involved when making a triage decision	Yes/no
	Men are more/less eligible for an IC-bed than women	More/less
Marital status	When IC-beds are scarce, the patient's marital status (i.e., being married with children) should be involved when making a triage decision	Yes/no
	Married people with children living at home are more/less eligible for an IC-bed than people without children living at home	More/less
Profession	When IC-beds are scarce, the patient's profession should be involved when making a triage decision. (people working in healthcare professions are more at risk of being contaminated with the virus)	Yes/no
	People working in a healthcare profession are more/less eligible for an IC-bed than people with other professions	More/less

Appendix E: Calculating triage score

The agents calculate on the fly a triage score for each patient, by taking into account the available IC- and ward-beds, as well as the characteristics of the patients waiting in the emergency room to be triaged.

Basic triage score

First the agent calculates a *basic triage score* by using the *severity of symptoms* parameter:

Severity of symptoms: mild=1; moderate=2; severe=3

Thus, the basic triage score for a particular patient varies between 1 and +3

Priority score

Then the agent calculates a *priority score* for a patient, based on the parameters *fitness*, *age*, *gender*, *marital status*, *profession*.

If the participant has expressed not to regard *age* and/or *profession* to be taken into account when assigning care to patients, then the values of *the ethical guidelines* were used on these parameters.

If, however, the participant has expressed *age* and/or *profession* as relevant properties when determining care for the patient, then the *personal moral values* on these parameters were used in the calculation.

If the participant has expressed *no moral preference* on the parameter:

- Profession: healthcare and contaminated on duty = + 1; others = 0
- Age: 0–20 = + 1; 21–40 = +.5; 41–60 = 0; 61–80 = −.5; 80–100 = − 1

If the participant has expressed *a moral preference* to involve the parameter:

- Profession: +1 if coinciding with participant's preference; − 1 if opposite
- Age: (see Table 6)

The priority score for the parameters *fitness*, *gender* and *marital status* were calculated as follows:

Table 6 moral preference score for the factor 'age'

Age	Prioritize young	Prioritize old
80+	− 1	1
61–80	−.5	.5
41–60	0	0
21–40	+.5	−.5
0–20	+ 1	− 1

- Fitness: very low = − 1; low =-.5; average = 0; high = +.5; very high = + 1
- Gender: +1 if coinciding; 0 if neutral; − 1 if opposite
- Marital status: + 1 if coinciding; 0 if neutral; − 1 if opposite

The raw priority score was calculated as the sum of priorities on each of the five patient characteristics. The maximum priority score is therefore + 5. The patient's final priority score was calculated as a proportion of the maximum: raw priority score/maximum = final priority score. Thus a patient's final priority score has a value between − 1 and + 1.

Triage score

The patient's triage score was calculated as the sum of the basic triage score and the priority score. The range of the triage score was between 0 and + 4.

Agent's triage decisions

For each new patient entering the emergency room, the agent calculates a triage score, and uses the following decision rules:

- triage score $\geq$ 3: IC (if no IC-bed available, then assign ward-bed).
- triage score $\geq$ 2 AND <3: Ward (if none available, then assign home treatment).
- triage score < 2: home treatment.

In TDP-2 (dynamic task allocation) the human and agent divide patients for triage. If circumstance force an agent to 'downscale' the preferred care (e.g., assigning a ward-bed because all IC-beds are occupied), then the agent attends the human to this and allows the human to redirect this patient to himself. If the agent has multiple patients with similar triage scores (a difference of.5 or less), then it redirects these patients to the human for conducting triage.

Examples

A participant has indicated to have a moral preference for assigning scarce care to: older patients rather than younger patients; to females rather than males; to married patients rather than single patients, and feels that patients with a care-profession should be treated with priority compared to patients having other professions.

Suppose that a patient due for triage is a 28-years old male patient, who is single, and is a librarian. He has moderate symptoms and has low fitness. Then the participant's basic triage score for that patient would be 2. The raw priority score would be age (-.5) + profession (− 1) + fitness (−.5) + gender (− 1) + marital status (− 1) = − 4. The final priority score would be − 4/5 = −.8. The triage score would then be 2 + −.8 = 1.2.

In contrast, suppose that another patient is also due for triage. She is a 62-years old female patient, who is married,

and is working as a nurse in a nursing home. She also has moderate symptoms and has low fitness. Then the participant's basic triage score for that patient would also be 2. The raw priority score would be age (.5) + profession (+ 1) + fitness (−.5) + gender (+ 1) + marital status (+1) = + 3. The final priority score would be 3/5 =.6. The triage score would then be 2 +.6 = 2.6.

Appendix F: Questionnaires

Participant's perception of the task

- I found the task (not at all to very) likable to do. (5-point scale)
- I found the task (not at all to very) difficult to do. (5-point scale)
- I found the task (not at all to very) believable. (3-point scale)

Experienced moral load

The experienced moral load was assessed by the following questions (5-point scale):

- I felt responsible for the well-being of the patients.
- I had to make choices where medical and ethical guidelines were inconclusive.
- I had to make choices that I was reluctant to make in real life.
- I felt uncomfortable with some of my choices.

Subjective control

Participants rated how applicable they found the following statements (5-point scale):

- I found it difficult to keep an overview of the patients and the available resources.
- I felt time pressure when making my choices.
- I think I made a decision in most patients that was appropriate at the time with a good distribution of care.
- I believe that my decisions have resulted in a good distribution of care among all patients.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

1. Curtis, C., Gillespie, N., Lockey, S.: AI-deploying organizations are key to addressing 'perfect storm' of AI risks. AI Ethi. **3**(1), 145–153 (2023). https://doi.org/10.1007/s43681-022-00163-7
2. Hille, E.M., Hummel, P., Braun, M.: Meaningful human control over AI for health? A review. J. Med. Eth. (2023). https://doi.org/10.1136/jme-2023-109095
3. Wang, W., Siau, K.: Ethical and moral issues with AI: a case study on healthcare robots. In: Twenty-Fourth Americas Conference on Information Systems (2019). Accessed 16 July 2018
4. Calvert, S.C., Heikoop, D.D., Mecacci, G., Van Arem, B.: A human centric framework for the analysis of automated driving systems based on meaningful human control. Theor. Issues Ergon. Sci. **21**(4), 478–506 (2020). https://doi.org/10.1080/1463922X.2019.1697390
5. Scharre, P., Horowitz, M.C.: Meaningful human control in weapon systems: a primer. Center New Am. Secur. 16 (2015)
6. Peeters, M.M.M., van Diggelen, J., van den Bosch, K., Bronkhorst, A., Neerincx, M.A., Schraagen, J.M., et al.: Hybrid collective intelligence in a human–AI society. AI Soc. **36**(1), 217–238 (2021). https://doi.org/10.1007/s00146-020-01005-y
7. Anderson, M., Anderson, S.L.: Machine Ethics. Cambridge University Press (2011)
8. van Diggelen, J., van den Bosch, K., Neerincx, M., Steen, M.: Designing for meaningful human control in military human–machine teams. In: Research handbook on meaningful human control of artificial intelligence systems, pp. 232–252. Edward Elgar Publishing (2024)
9. Wallach, W., Vallor, S.: Moral machines: from value alignment to embodied virtue. In: Ethics of Artificial Intelligence, pp. 383–412. Oxford University Press (2020)
10. Santoni de Sio, F., van den Hoven, J.: Meaningful human control over autonomous systems: a philosophical account. Front. Robot. AI **5**, 15 (2018)
11. Robbins, S.: The many meanings of meaningful human control. AI Eth. (2023). https://doi.org/10.1007/s43681-023-00320-6
12. McFarland, T.: Reconciling trust and control in the military use of artificial intelligence. Int. J. Law Inf. Technol. **30**(4), 472–483 (2022). https://doi.org/10.1093/ijlit/eaad008
13. Textor, C., Zhang, R., Lopez, J., Schelble, B.G., McNeese, N.J., Freeman, G., et al.: Exploring the relationship between ethics and

trust in human-artificial intelligence teaming: a mixed methods approach. J. Cogn. Eng. Decis. Mak. **16**(4), 252–281 (2022). https://doi.org/10.1177/15553434221113964
14. De Visser, E.J., Peeters, M.M.M., Jung, M.F., Kohn, S., Shaw, T.H., Pak, R., et al.: Towards a theory of longitudinal trust calibration in human–robot teams. Int. J. Soc. Robot. (2019). https://doi.org/10.1007/s12369-019-00596-x
15. Kox, E.S., Kerstholt, J.H., Hueting, T.F., de Vries, P.W.: Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. Auton. Agent. Multi-Agent Syst. **35**(2), 30 (2021). https://doi.org/10.1007/s10458-021-09515-9
16. Allen, C., Wallach, W.: Moral machines: contradiction in terms or abdication of human responsibility. Eth. Soc. Implic. Roboti. Robot Eth., 55–68 (2012)
17. Etzioni, A., Etzioni, O.: Incorporating ethics into artificial intelligence. J. Eth. **21**(4), 403–418 (2017). https://doi.org/10.1007/s10892-017-9252-2
18. Shneiderman, B.: Human-centered artificial intelligence: reliable, safe & trustworthy. Inte. J. Hum.–Comput. Interact. **36**(6), 495–504 (2020). https://doi.org/10.1080/10447318.2020.1741118
19. Boardman, M., Butcher, F.: An exploration of maintaining human control in AI enabled systems and the challenges of achieving It. In: NATO IST-178 Big Data Challenges: Situation Awareness And Decision Support. Budapest (2019)
20. Cavalcante Siebert, L., Luce Lupetti, M., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., et al.: Meaningful human control over AI systems: beyond talking the talk. arXiv e-prints. p. arXiv–2112 (2021)
21. van der Waa, J., Verdult, S., van den Bosch, K., van Diggelen, J., Haije, T., van der Stigchel, B., et al.: Moral decision making in human-agent teams: human control and the role of explanations. Front. Robot. AI (2021). https://doi.org/10.3389/frobt.2021.640647
22. Schulte, A., Donath, D., Lange, D.S.: Design patterns for human-cognitive agent teaming. In: International Conference on Engineering Psychology and Cognitive Ergonomics, pp. 231–243. Springer (2016)
23. Ekelhof, M.: Autonomous weapons: operationalizing meaningful human control. https://blogs.icrc.org/law-and-policy/2018/08/15/autonomous-weapons-operationalizing-meaningful-human-control/
24. Crootof, R.: A meaningful floor for "meaningful human control". Temple Int. Comp. Law J. **30**, 53 (2016)
25. Davidovic, J.: On the purpose of meaningful human control of AI. Front. Big Data (2023). https://doi.org/10.3389/fdata.2022.1017677
26. Skinner, E.A.: A guide to constructs of control. J. Pers. Soc. Psychol. **71**(3), 549 (1996)
27. Mansell, W., Marken, R.S.: The origins and future of control theory in psychology. Rev. Gen. Psychol. **19**(4), 425–430 (2015). https://doi.org/10.1037/gpr0000057
28. Powers, W.T.: Behavior: The Control of Perception. Aldine de Gruyter, Hawthorne (1973)
29. Sun, J., Wilt, J., Meindl, P., Watkins, H.M., Goodwin, G.P.: How and why people want to be more moral. J. Personal. (2023). https://doi.org/10.1111/jopy.12812
30. Schaefer, K.E., Straub, E.R., Chen, J.Y.C., Putney, J., Evans, A.W.: Communicating intent to develop shared situation awareness and engender trust in human-agent teams. Cogn. Syst. Res. **46**, 26–39 (2017). https://doi.org/10.1016/j.cogsys.2017.02.002
31. Sharan, N.N., Romano, D.M.: The effects of personality and locus of control on trust in humans versus artificial intelligence. Heliyon. **6**(8) (2020)
32. Lyle, J.: Stimulated recall: a report on its use in naturalistic research. Br. Edu. Res. J. **29**(6), 861–878 (2003). https://doi.org/10.1080/0141192032000137349
33. Flathmann, C., Schelble, BG., Zhang, R., McNeese, NJ.: Modeling and guiding the creation of ethical human–AI teams. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 469–479 (2021)
34. Verdiesen, I., Santoni de Sio, F., Dignum, V.: Accountability and control over autonomous weapon systems: a framework for comprehensive human oversight. Minds Mach. **31**(1), 137–163 (2021). https://doi.org/10.1007/s11023-020-09532-9
35. Craft, J.L.: A review of the empirical ethical decision-making literature: 2004–2011. J. Bus. Eth. **117**, 221–259 (2013)
36. Van Diggelen, J., Metcalfe, J.S., Van Den Bosch, K., Neerincx, M., Kerstholt, J.: Role of emotions in responsible military AI. Eth. Inf. Technol. **25**(1), 17 (2023)
37. Van Den Bosch, K., Bronkhorst, A.: Human–AI cooperation to benefit military decision making. In: NATO IST-160 Specialists' Meeting "Big Data and Artificial Intelligence for Military Decision Making". Bordeaux, France (2018)
38. Klein, G., Woods, D.D., Bradshaw, J.M., Hoffman, R.R., Feltovich, P.J.: Ten challenges for making automation a "team player" in joint human-agent activity. IEEE Intell. Syst. **19**(06), 91–95 (2004). https://doi.org/10.1109/MIS.2004.74
39. Salwei, M.E., Carayon, P.: A sociotechnical systems framework for the application of artificial intelligence in health care delivery. J. Cognit. Eng. Decis. Mak. **16**(4), 194–206 (2022). https://doi.org/10.1177/15553434221097357
40. Van Diggelen, J., Neerincx, M., Peeters, M., Schraagen, J.M.: Developing effective and resilient human-agent teamwork using team design patterns. IEEE Intell. Syst. **34**(2), 15–24 (2018)
41. Van Diggelen, J., Johnson, M.: Team design patterns. In: Proceedings of the 7th International Conference on Human-Agent Interaction - HAI '19. Kyoto, pp. 118–126, ACM Press, Japan (2019)
42. Lin, P., Abney, K., Bekey, G.: Robot ethics: mapping the issues for a mechanized world. Artif. Intell. **175**(5–6), 942–949 (2011). https://doi.org/10.1016/j.artint.2010.11.026
43. Pannu, A.: Artificial intelligence and its application in different areas. Artif. Intell. **4**(10), 79–84 (2015)
44. Van Der Stigchel, B., Van Den Bosch, K., Van Diggelen, J., Haselager, P.: Intelligent decision support in medical triage: Are people robust to biased advice? J. Publ. Health (2023). https://doi.org/10.1093/pubmed/fdad005
45. Ministerie van Volksgezondheid WeS.: Draaiboek 'Triage op basis van niet-medische overwegingen voor IC-opname ten tijde van fase 3 in de COVID-19 pandemie' - Publicatie - Rijksoverheid.nl [publicatie]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/publicaties/2020/06/16/draaiboek-triage-op-basis-van-niet-medische-overwegingen-voor-ic-opname-ten-tijde-van-fase-3-in-de-covid-19-pandemie
46. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**(3), 709–734 (1995)
47. Schoorman, F.D., Mayer, R.C., Davis, J.H.: An integrative model of organizational trust: past, present, and future. Acad. Manag. Rev. **32**(2), 344–354 (2007)
48. Jacovi, A., Marasović, A., Miller, T., Goldberg, Y.: Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 624–635 (2021)
49. Walsh, S.E., Feigh, K.M.: Understanding human decision processes: inferring decision strategies from behavioral data. J. Cognit. Eng. Decis. Mak. **16**(4), 301–325 (2022). https://doi.org/10.1177/15553434221122899
50. Neerincx, MA., Van Der Waa, J., Kaptein, F., van Diggelen, J.: Using perceptual and cognitive explanations for enhanced human-agent team performance. In: International Conference on Engineering Psychology and Cognitive Ergonomics, pp. 204–214. Springer (2018)

51. Schneider, M.F., Miller, M.E., McGuirl, J.: Assessing quality goal rankings as a method for communicating operator intent. J. Cognit. Eng. Decis. Mak. **17**(1), 49–74 (2023). https://doi.org/10.1177/15553434221131665
52. Widdershoven, G., Abma, T., Molewijk, B.: Empirical ethics as dialogical practice. Bioethics **23**(4), 236–248 (2009)
53. Greene, J., Haidt, J.: How (and where) does moral judgment work? Trends Cogn. Sci. **6**(12), 517–523 (2002). https://doi.org/10.1016/S1364-6613(02)02011-9
54. Teper, R., Zhong, C.B., Inzlicht, M.: How emotions shape moral behavior: some answers (and questions) for the field of moral psychology. Soc. Pers. Psychol. Compass **9**(1), 1–14 (2015)
55. Martí-Vilar, M., Escrig-Espuig, J.M., Merino-Soto, C.: A systematic review of moral reasoning measures. Curr. Psychol. **42**(2), 1284–1298 (2023). https://doi.org/10.1007/s12144-021-01519-8
56. Bagozzi, R.P., Sekerka, L.E., Hill, V., Sguera, F.: The role of moral values in instigating morally responsible decisions. J. Appl. Behav. Sci. **49**(1), 69–94 (2013). https://doi.org/10.1177/0021886312471194
57. Niforatos, E., Palma, A., Gluszny, R., Vourvopoulos, A., Liarokapis, F.: Would you do it?: enacting moral dilemmas in virtual reality for understanding ethical decision-making. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2020)
58. Waterson, P., Robertson, M.M., Cooke, N.J., Militello, L., Roth, E., Stanton, N.A.: Defining the methodological challenges and opportunities for an effective science of sociotechnical systems and safety. Ergonomics **58**(4), 565–599 (2015)
59. Stanovich, KE.: On the distinction between rationality and intelligence: implications for understanding individual differences in reasoning. Oxford Handbook Think. Reason. 343–365 (2012)
60. Harari, YN.: Homo Deus: A Brief History of Tomorrow. Random House (2016)
61. Hausman, D.M.: Eliciting preferences and respecting values: Why ask? Soc. Sci. Med. **320**, 115711 (2023)
62. Zafari, S., Koeszegi, S.T.: Attitudes toward attributed agency: role of perceived control. Int. J. Soc. Robot. **13**(8), 2071–2080 (2020)
63. Martin, D.: Who should decide how machines make morally laden decisions? Sci. Eng. Eth. **23**(4), 951–967 (2017)