Cognition 262 (2025) 106177

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Moral psychological exploration of the asymmetry effect in AI-assisted euthanasia decisions☆

Michael Laakasuo [a,*], Anton Kunnari [b,1], Kathryn Francis [c,1], Michaela Jirout Košová [g], Robin Kopecký [d], Paolo Buttazzoni [e], Mika Koverola [b], Jussi Palomäki [f], Marianna Drosinou [b,h,2], Ivar Hannikainen [e,2]

[a] Department of Social Research, Faculty of Social Sciences, University of Turku, Turku, Finland
[b] Department of Psychology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[c] School of Psychology, Faculty of Health and Medicine, University of Leeds, United Kingdom
[d] Institute of Philosophy, Czech Academy of Sciences, Prague, Czechia
[e] Department of Philosophy I, Faculty of Psychology, University of Granada, Granada, Spain
[f] Health and Well-Being Promotion Unit, Finnish Institute for Health and Welfare, Helsinki, Finland
[g] The Karel Čapek Center for Values in Science and Technology, Czech Republic
[h] Department of Psychology, Faculty of Social Sciences, University of Turku, Turku, Finland

ARTICLE INFO

Keywords:
Moral psychology of AI
Moral psychology of robotics
Moral judgment
AI ethics
Passive euthanasia

ABSTRACT

A recurring discrepancy in attitudes toward decisions made by human versus artificial agents, termed the Human-Robot moral judgment asymmetry, has been documented in moral psychology of AI. Across a wide range of contexts, AI agents are subject to greater moral scrutiny than humans for the same actions and decisions. In eight experiments (total $N = 5837$), we investigated whether the asymmetry effect arises in end-of-life care contexts and explored the mechanisms underlying this effect. Our studies documented reduced approval of an AI doctor's decision to withdraw life support relative to a human doctor (Studies 1a and 1b). This effect persisted regardless of whether the AI assumed a recommender role or made the final medical decision (Studies 2a and 2b and 3), but, importantly, disappeared under two conditions: when doctors kept on rather than withdraw life support (Studies 1a, 1b and 3), and when they carried out active euthanasia (e.g., providing a lethal injection or removing a respirator on the patient's demand) rather than passive euthanasia (Study 4). These findings highlight two contextual factors–the level of automation and the patient's autonomy–that influence the presence of the asymmetry effect, neither of which is not predicted by existing theories. Finally, we found that the asymmetry effect was partly explained by perceptions of AI incompetence (Study 5) and limited explainability (Study 6). As the role of AI in medicine continues to expand, our findings help to outline the conditions under which stakeholders disfavor AI over human doctors in clinical settings.

The topic of assisted dying – euthanasia, or *good death* – is a strong taboo in modern day Western societies (Passerard & Menaud, 2015; Preston, 1994; Streeck, 2020). Even Sci—Fi, the literary genre with a track record of producing skilled analyses of future societies and emerging technologies (Nichols et al., 2008), is strikingly silent on this topic, with only a few exceptions.[3]

Recently, Statista.com estimated that the market for surgical robots alone will be worth 120 billion dollars by 2030 – excluding other innovations, such as mortality prediction algorithms (Keuning et al., 2020; Tiwari et al., 2020) and autonomous nursing robots (Laakasuo, Palomäki, et al., 2023), which are also being developed at rapid speeds. The costs of healthcare infrastructure increase with the number of elderly,

☆ This article is part of a Special issue entitled: 'Morality and AI' published in Cognition.

* Corresponding author at: Faculty of Social Science, University of Turku, 4th Floor, Room 457, Assistentinkatu 1 (Publicum), 20014, Finland.
E-mail addresses: Michael.Laakasuo@helsinki.fi (M. Laakasuo), Anton.Kunnari@helsinki.fi (A. Kunnari), K.B.Francis@leeds.ac.uk (K. Francis), mika.koverola@helsinki.fi (M. Koverola), jussi.palomaki@thl.fi (J. Palomäki), Maria-Anna.Drosinou@helsinki.fi (M. Drosinou), ivar@ugr.es (I. Hannikainen).

[1] Shared authorship
[2] Shared authorship
[3] Such as in Star Trek – The Next Generation; which did not deal with AI assisted death, although it had two episodes on moral issues associated with euthanasia.

https://doi.org/10.1016/j.cognition.2025.106177
Received 15 June 2024; Received in revised form 29 April 2025; Accepted 5 May 2025
Available online 13 May 2025

rising lifespans, growing levels of automatization, and a global undersupply of skilled medical professionals. This, in turn, contributes to market pressures pushing for the need to acquire various medical robots (Haakenstad et al., 2022). According to a recent review published in The Lancet, 12 million physicians and 30 million nurses are needed globally to meet universal healthcare coverage within the next decade (Haakenstad et al., 2022). One important aspect of this global trend is the development of mortality prediction algorithms (Keuning et al., 2020) which, according to Tiwari et al. (2020), who are creators of these algorithms, are being made exactly with the aim of saving resources.[4]

Recent literature has called attention to the phenomenon of *moral distress* among healthcare professionals (Tigard, 2018) – i.e., the psychological and emotional toll from making difficult, and even morally contentious, decisions about patients' lives. This mental health burden may be even more pronounced among professionals offering critical care, as research on healthcare workers during the Covid-19 pandemic illustrated (Rabin et al., 2023). This raises the possibility that medical AI, including mortality prediction algorithms, may reduce the severity of moral distress among healthcare professionals (see Tiwari et al., 2020 as an example). While these practical discussions have recently emerged as technologies develop, the philosophical debates on the topic of euthanasia have been going on for longer.

1. History of euthanasia and the present debate

The term "euthanasia" has had different meanings since its origin in Ancient Greece (Van Hoof, 2004). "Good death" was originally connected to ways of dying in which the role of the doctor was marginal (Van Hoof, 2004). For instance, if a person died alone heroically, nobly, without pain or in the midst of mundane pleasures, it was considered *good* (Van Hoof, 2004). Doctors were generally absent – or very passive – when someone was dying and were not expected to confirm death or relieve pain as they often do today (Van Hoof, 2004). Later in Roman times, we find examples of doctors assisting death (e.g., the death of Seneca), but most often by providing the means or participating as advisors, and not as active agents or arbiters of the moral rightness of the act (Van Hoof, 2004). Contrary to the widespread modern belief, providing lethal substances to patients who ask for it was probably not perceived as a breach of the Hippocratic oath (Van Hoof, 2004).[5]

Today, euthanasia is connected with the doctor's role and is associated primarily with consequentialist and utilitarian theories (Černý, 2018; Crocker, 2013). In most countries where active euthanasia (or physician-assisted suicide[6]) is legal, it is the doctor who assists the process, provides the means of death to the patient, and often decides whether it is right to perform euthanasia (Vizcarrondo, 2013). Of note, various studies report a preference for active involvement of doctors in euthanasia, rather than leaving it to patients and their families alone, at least in American participants (Caddell & Newton, 1995).

The meaning of euthanasia has shifted too and must meet specific conditions: "good death" comes as a liberation to the terminally ill (with no hope of cure being invented in their lifetime) who are suffering or who are no longer able to lead a dignified life, and who repeatedly and competently request doctors' help in life termination, while not being able to carry out (assisted) suicide themselves (Černý, 2018).

Furthermore, there are multiple categories of euthanasia. On the one hand, we categorize euthanasia as active or passive. *Passive* euthanasia, i.e. withholding or withdrawal of life supporting treatment, is generally accepted and practiced in medicine,[7] since it evokes simply letting "nature" or "God" proceed in their workings (Walsh et al., 2009; as cited in Crocker, 2013). In this regard, active steps taken by a doctor, such as administering lethal drugs (*active* euthanasia), are viewed as more problematic (Crocker, 2013), with passive euthanasia being generally more accepted than active euthanasia - arguably due to omission bias - (meaning that people assign more responsibility to actions over inactions; Gamliel, 2013). Physician assisted suicide is in turn approved to a higher extent than suicide performed by patients (Caddell & Newton, 1995). Moreover, the idea of a distinction between killing and letting die also seems to be relevant in the medical domain (Douglas, 2009). Positive framing (such as "not prolonging life" vs. "ending life" formulation) has been shown to increase acceptance of physician-assisted suicide but not of passive euthanasia, probably due to ceiling effects (Gamliel, 2013).

On the other hand, euthanasia is also categorized based on voluntariness: *voluntary* euthanasia is demanded by the patient themselves, *involuntary*[8] euthanasia is performed without the patient's demand, and *non-voluntary* euthanasia is the case when the patient is incapable of declaring their preferences (Bartels & Otlowski, 2010; as cited by Crocker, 2013). Arguments for legalizing active (voluntary) euthanasia often rest on concerns about human dignity, freedom, privacy or autonomy, and on the utilitarian goal of minimizing suffering (Crocker, 2013). Opponents of euthanasia often state that the doctor's duty is incompatible with performing euthanasia due to the Hippocratic oath[9] (Crocker, 2013). The argument follows that its abuse would become a danger, or that terminally ill, elderly, or otherwise disabled patients dependent on others, might get under pressure to "unburden" their close ones and society (Worsnop, 1997; as cited in Crocker, 2013). Whatever the standpoint, the role of patient autonomy is definitely relevant, with studies showing public preference for preserving it and ensuring the voluntariness of the euthanasia process (Feltz, 2023; Ho, 1998; Levin et al., 2020; Teisseyre et al., 2005), a topic to which we turn next.

1.1. Moral science of euthanasia – a central role of autonomy?

A significant portion of previous social, moral, and cognitive science of euthanasia focuses on situations where the patient is conscious and on individual differences like religiosity (Bahník & Vranka, 2021; Caddell & Newton, 1995; Deak & Saroglou, 2017; Douglas, 2009; Feltz, 2023; Gamliel, 2013; Ho, 1998; Karumathil & Tripathi, 2022; Levin et al., 2020; Lockhart et al., 2023; MacDonald, 1998; Teisseyre et al., 2005).

Concerning individual differences – as euthanasia attitudes have been studied primarily in WEIRD samples (Henrich et al., 2010) – it is mostly conservatism, a Christian denomination, and low educational attainment that are associated with negative attitudes toward euthanasia (Caddell & Newton, 1995; Deak & Saroglou, 2017; MacDonald, 1998; Levin et al., 2020; Lockhart et al., 2023). Correspondingly, it seems that the binding orientation aspects of Moral Foundations Theory (loyalty and purity; Graham et al., 2009) are associated with lower

[4] *"Mortality prediction can be very helpful for taking critical decisions which can help in optimising the resources available in the hospital and also an extra opinion for doctors and family members in cases of euthanasia* i.e. *ending life of patients to relieve pain and suffering."*

[5] Van Hoof (2004) suggests that the modern interpretation of this part of Hippocratic oath by the opponents of active euthanasia (briefly mentioned later in the introduction) is a linguistic misunderstanding. Briefly put, according to the original text, the lethal drug should not be given to the patient on the demand of a *third* person (a person "ordering"a *murder*, not an assisted suicide, by a doctor).

[6] For clarification, physician assisted suicide refers to a situation in which a physician provides a lethal drug on patient's demand, while the patient themselves takes the drug. In active euthanasia, the physician causes the patient's death directly (Schafer, 2013; as cited by Goligher et al., 2017).

[7] We are aware that according to some authors it might be problematic to call these widely accepted practices of withholding and withdrawal of life supporting treatment passive *euthanasia* (Černý, 2015; Sumner, 2011), yet we decide not to delve deeper into this complex problem, since it goes beyond the scope of our paper.

[8] Considered to be "intentional killing" (Bartels & Otlowski, 2010).

[9] We already mentioned in note no. 2 that Van Hoof (2004) doesn't agree with this view.

levels of acceptance (Deak & Saroglou, 2017; Lockhart et al., 2023).

In addition, these studies have looked at differences between non-terminally ill and terminally ill patients and the role of pain intensity in the patient's suffering (Bahník & Vranka, 2021; Levin et al., 2020; MacDonald, 1998), how the presence of a physician makes suicide/active euthanasia more acceptable (Caddell & Newton, 1995), how killing (active euthanasia) feels intuitively morally worse than letting someone die (as passive euthanasia or as a side effect of another treatment; Douglas, 2009; Feltz, 2023), and how objectively framing the same euthanasia situation as a positive thing (not prolonging suffering) results in a more positive view of both active and passive euthanasia (Feltz, 2023; Gamliel, 2013). Furthermore, another set of studies has found that active assistance is favoured over indirect assistance, that patient-initiated euthanasia is preferred over family-assisted, and that voluntary euthanasia is favoured for certain illnesses (e.g., cancer) over others (e.g., Alzheimer's disease; MacDonald, 1998; Levin et al., 2020). Finally, these studies have investigated whether we have different reactions to euthanasia of children vs. adults (Deak & Saroglou, 2017; MacDonald, 1998).

A limitation of these sparse existing studies is that they use a plethora of unstandardized stimuli, dependent variables, predictor variables, and approaches. There is no overarching theory that these studies adopt and much of the research is phenomenon-focused rather than theory-focused, in a way that would allow us to extrapolate more universal cognitive mechanisms (see Tooby & Cosmides, 2005). Apart from religious conservatism and the distinction between killing and letting die, few other factors unify the studies. Nonetheless, what brings together large sections of this literature, is the human preference for patient autonomy across most conditions. Researchers use different terminology to describe autonomy, like the capacity to control (Levin et al., 2020), patient request (Teisseyre et al., 2005) or voluntariness (Feltz, 2023).

In recent empirical work, Feltz (2023) looked at the role of patient autonomy and contrasted it with many other moderators, including different types of euthanasia, and observed that once there is patient autonomy or voluntariness, the moderating factors lose their effectiveness. Similarly, in Levin et al. (2020), when participants had to evaluate the moral acceptance of euthanasia, the authors concluded that the relevant factor was not the illness type nor the type of euthanasia, but specifically whether the patient could control the illness and its progression or not. This, in turn, would potentially be associated with why euthanasia of children is perceived differently, as they might be seen as less autonomous individuals and their loss of an unlived life is seen as a greater tragedy (Deak & Saroglou, 2017).

In this paper, we focus mostly on situations where the patient is unconscious or comatose. Moreover, not only does this study setup have novelty value concerning previous euthanasia research, but it also expands much of the previous literature. This allows us to focus on the moral judgment asymmetry effect (see below) by eliminating the confounding factors associated with varying diseases, symptoms, and subjective experiences of pain. Previous studies have looked at the framing and order effects of the vignettes (Bahník & Vranka, 2021; Feltz, 2023; Gamliel, 2013), slippery slope arguments (not really theories; Deak & Saroglou, 2017), and the Doctrine of Double Effect (death resulting as an unintended side-effect) – but none have focused on the pure case of euthanasia of an unconscious patient while systematically manipulating the surrounding factors.[10]

1.2. The future role of AI and robots in euthanasia

If AI and robots are to become indispensable elements of our healthcare systems, what are the ethical boundaries we should be guarding and the cognitive biases we should be aware of? While, previously, it was the role of the doctor to engage in questions surrounding – and practices of – euthanasia, we are now facing a new frontier in medicine, whereby artificial agents may be involved in these ethically contentious decisions (Longoni et al., 2019). These potential future roles need to be explored if we wish to be prepared for the challenges facing humanity.

Arguments for or against medical innovations may appeal to different people, however, pure theory and largely normative approaches are not able to show if and how the general population would like new medical technologies to be implemented. Thus, it is crucial to become acquainted with the moral intuitions and/or boundary conditions of the communities that will be affected by this decision-making and the best approach to reach this goal is via empirical research (e.g. Longoni et al., 2019; Van Cauwenberge et al., 2022; Vanderelst & Willems, 2020).

In a recent vignette-based qualitative interview study ($N = 30$ physicians), the tensions between the roles of AIs and human doctors were investigated (Van Cauwenberge et al., 2022). One of the prominent patterns in the interviews was the delineation of the roles of the physician and AI in the clinical process; AI was being pushed into more automated and administrative tasks, leaving the serious clinical work to a physician. Respondents expressed unwillingness to accept AI's decisions without understanding the reasons behind them. They argued that AI lacked the ability to provide intelligible reasons for its actions and said that physicians should have the last say in the process. They also claimed that humans are irreplaceable and have complexity, expertise, and skills that no automated system can reach (Van Cauwenberge et al., 2022). The topic of perceived expertise or competence could be indeed one of the important factors in explaining the asymmetry effect, as people seem to be worried that AIs do not take into consideration the uniqueness of each case (Longoni et al., 2019).

One of the most prominent obstacles that seem to stand in the way of accepting medical AI in certain contexts is algorithm aversion,[11] which was observed also in the above cited study with physicians (Van Cauwenberge et al., 2022; see also Castelo & Ward, 2021) in terms of physicians lacking intelligible reasons behind AI decisions. This also concerns the general public, with people believing that they understand human medical decision-making better than algorithmic processes of medical AI, even though this tendency could be potentially moderated

[10] With respect to theories of cognitive mechanisms, a common conclusion is that cultural circumstances associated with individual differences (MFQ; religiosity and political conservatism) matter. For this reason, we also looked at a relatively recent anthropological analysis that summarized practices of mercy killing in hunter-gatherer societies (Boehm, 2012). Commonality in these societies is that when people have diminished autonomy due to old age, severe illness or injury, it is family members who usually help the suffering individuals to die. We only found one experimental paper looking at a case indicating that, when a patient is comatose, then family members should be the ones deciding on the course of action (Teisseyre et al., 2005). This suggests that perceptions of autonomy and evolutionary kin-selection theories could be linked and provide a starting point for finding cognitive mechanisms in the future.

[11] Although labeling a phenomenon as X does not really explain it or amount to a theory. We would still need to explain why there is "algorithm aversion" and what is the mechanistic logic behind it and why it does not creep up in every decision and in every situation systematically.

by better educating people about how algorithms work in these contexts and thus creating subjective understanding of them (Cadario et al., 2021).

Another crucial issue in the potential acceptance or nonacceptance of medical AI seems to be the untouchability of patient autonomy. People tend to accept medical decisions of robots more if the will of the patient is respected in the cases of healthcare providers facing the dilemma of forced medication of the uncooperative patient (Laakasuo, Palomäki, et al., 2023; Soares et al., 2023; Vanderelst & Willems, 2020). Nonetheless, medical AI also seems to be more expected and approved to make utilitarian decisions, while being viewed as less capable of experience and agency than human healthcare providers (Wu et al., 2022). Similarly, medical robots seem to be generally perceived as competent and trustworthy in their cold rational decision-making (potentially jeopardizing patient autonomy), while viewed as less warm and less morally responsible than humans (Soares et al., 2023). The issue of accepting medical AI is thus rather complex, since people seem to have certain conflicting intuitions, not wanting the robot or AI to violate patient autonomy and be a coldly rational and competent utilitarian agent simultaneously.

2. Human-robot moral judgment asymmetry

Despite the need to have scientific studies on the rise of medical AIs and human reactions toward them, recent literature is quite limited in this regard and lacks a clear theoretical framework. Most theories and models in moral psychology are path-dependent on the unstated background assumption that moral judgments are about other people and their actions (Malle, 2021; Schein & Gray, 2018; Voiklis & Malle, 2018) but when these models have been used to study human moral intuitions about robots, they have not fared well, implying that there is much we do not understand about human moral judgment. The field has already uncovered an effect that we will call here the human-robot moral judgment asymmetry effect, or the *Asymmetry Effect*, for short.

The Asymmetry Effect is a phenomenon observed by several research groups in recent years, where participants evaluate a robot's decision to be different than a human's, even when the antecedents and consequences of the decision are held constant (e.g., Malle et al., 2015; Malle et al., 2019; Komatsu et al., 2021; Laakasuo, Palomäki, & Köbis, 2021; Laakasuo et al., 2023; Sundvall et al., 2023; Stuart & Kneer, 2021). However, this asymmetry effect does not manifest itself in every decision – but only with some moral problems. Furthermore, the asymmetry effect is a separate effect from that recognized by Bigman and Gray (2018), who reported that humans are in general averse to machines making decisions, because humans perceive robots to be less minded than other humans. The asymmetry effect specifically surfaces when the decisions – rather than the agents – are evaluated as appropriate. Bigman's and Gray's (Bigman & Gray, 2018) argument is based on the mind perception hypothesis, positing that moral judgments arise from perceived interactions of a minded agent and a patient (whose perceived mindedness increases if they are a target of an intentionally harmful action). In this theory, perceptions of agency in the agent and experience in the patient jointly determine the perceived level of harm caused by an action.

The asymmetry effect was first found in a study incorporating an approach similar to the traditional trolley problem. Malle et al. (2015) showed that human agents were found to be more culpable for deciding to act (i.e., sacrificing one life to protect five) than for inaction (i.e., no action, thus losing five lives), while artificial agents were blamed to a similar extent whether they decided to act or not. In addition, participants displayed an overall tendency to blame artificial agents more than human decision-makers, suggesting that artificial agents might be judged primarily on the basis of their nature (i.e., being artificial), whereas blame assigned to human decision-makers depended more on the specific decision being made. Notably, for human decision-makers, the degree of blameworthiness aligned with the perceived moral wrongness of the chosen option. However, this pattern did not hold for artificial agents. In their case, inaction was deemed to be more morally wrong than action, yet the level of blame attributed to the agent was similar for both cases.

Later, Malle et al. (2019) reported the asymmetry effect in the context of AI making military decisions. Participants read a vignette where either a human pilot or an AI-drone was instructed to launch a missile which could cause harm to innocent bystanders. More blame was attributed to humans when they disobeyed the instructions compared to when they did as they had been told. The AI, on the other hand, was equally blamed, no matter what it decided. A noteworthy detail in Malle et al. (2019) is that the AI was attributed less blame for disobedience than the human pilot. To sum up, people perceived the AI-drone decision to disobey in a more positive light than the identical decision made by a person.

After Malle's work on the topic, researchers have found the moral judgment asymmetry effect in other contexts. Sundvall et al. (2023) investigated marine rescue situations, with participants deeming it less permissible for a robot to make an utilitarian decision if those who were saved caused the accident and the cost of the decision was an innocent life. However, this was a palatable decision for a human lifeguard to make. Laakasuo, Palomäki, et al. (2023) reported the moral judgment asymmetry effect in forced medication decisions, where a human nurse's decision to either forcefully medicate a patient or leave them unmedicated – respecting their autonomy – was equally morally approved; conversely, the robot was only tolerated in making the non-medication decision. Stuart and Kneer (2021) found the asymmetry effect in the context of an agent unknowingly polluting groundwater: when no harm resulted from this action, the robot's actions were perceived as more morally wrong than the human's. Stuart and Kneer (2021) echo Bigman and Gray (2018) when they argue in support of the mind-attribution hypothesis for explaining these effects. However, they did not observe a situation without a decision that would pollute the groundwater. If Bigman and Gray (2018) mind perception model were true, we would need to see a general drop in moral approval of all decisions made by robots and AIs. Since Stuart and Kneer (2021) did not include a control condition, they could not observe the full extent of the asymmetry effect, making it unclear if the mind-attribution hypothesis explains their findings. Thus, there does not seem to be a comprehensive theory predicting all the cases where the asymmetry effect will be observed a priori.

According to Laakasuo (2023, Laakasuo, Sundvall, et al., 2021a, Laakasuo, Sundvall, et al., 2021b), the moral psychology of robotics lacks a strong theoretical background. During the evolution of human cognition in the Pleistocene, there were no artificial agents for our ancestors to interact with. Thus, our brains have not evolved to react to non-living agents, and we have no intuitions of how they behave or function. In contrast, encountering living agents such as dangerous animals makes us react in very specific, predictable, and evolutionarily hard-wired ways. Artificial agents and algorithms generally operate based on the principles of probability calculus, which, for human brains, is intuitively difficult to understand (Longoni et al., 2019). Instead, we anthropomorphize artificial agents as living, thinking, or feeling beings, which they are not. Moreover, theory in moral psychology is largely based on evidence coming from humans judging conspecifics, making it inherently limited when predicting or describing the moral behavior of non-human, artificial agents.[12]

Therefore, the field requires a theory able to make predictions from first principles that could reliably foresee outcomes of moral judgment

[12] Kahn et al. (2011) propose that artificial agents constitute a distinct ontological category. Robots and AIs represent a novel phenomenon in natural and cultural history. Consequently, our interactions with robots are guided by intuitions that evolved without any reference to such entities, potentially leading to different types of categorical misunderstandings and misinterpretations.

experiments when both humans and robots make identical decisions in identical situations. However, as such theory does not exist yet, this area of study needs to take a different approach, where the potential boundary conditions of any future theory or a model are mapped out. Nonetheless, we believe that when it comes to AI-made euthanasia decisions, similar asymmetry effects will surface and will likely feed into our growing knowledge regarding the boundary conditions that any future theory needs to account for.

2.1. Current studies

Here, we investigated the moral approval of the decision to turn off life support of a comatose patient (Studies 1–3) or to administer lethal medical treatment or to remove an already administered treatment to a conscious patient who is requesting it (Study 4), by varying the agents who make the decisions as humans or medical robots. Furthermore, we investigated the boundary conditions in which the asymmetry effect can be observed (Studies 5 and 6) by varying the contextual cues relating to the situation or the agents involved. Indeed, past literature has shown an overall acceptance of passive euthanasia in the population (Crocker, 2013), while a trend has been suggested toward preferring a human doctor over an AI having the final say over similar choices (Van Cauwenberge et al., 2022) as well as those concerning forced medication (Laakasuo, Palomäki, et al., 2023).

We primarily focused on judgments of passive euthanasia decisions in order to identify the effect of the role of AI/robots in a less controversial context. We also looked at the role of personal autonomy – i.e., the patient being (un)conscious – in these situations (Studies 1–3 and 5 vs. 4).[13] Arguably, the decision of withdrawing treatment from someone who is in a coma and unlikely to regain consciousness (non-voluntary passive euthanasia) is different when compared to situations where people actively request to be euthanized (voluntary euthanasia).

In Studies 1a and 1b, we established the asymmetry effect in both an English- and a Czech-speaking sample. Consistency across countries has been shown in other domains of moral judgment and decision making (Awad et al., 2018), while linguistic differences have been reported to indeed influence – but not determine – moral judgment formation (Hayakawa et al., 2017), thus making it necessary to test for cultural variability as a boundary condition for acceptability of passive euthanasia across levels of automation specific to the medical domain.[14]

Inspired by Laakasuo, Palomäki, et al. (2023) and Malle et al. (2019), in Studies 2a (Finnish), 2b (English), and 3 (English), we tested for the potential boundary condition of the command chain effect whereby there was both a recommender and a decision-maker who could be any combination of human and AI agents. In other words, we aimed to test whether the level of automation (or AI involvement) influenced people's moral judgments of the implemented decisions. If the asymmetry effect had disappeared or been diluted when the AI took one particular role in the chain of command (recommender or decision-maker) but not the other, then the internal structure of the chain of command would have counted as a boundary condition.

For instance, Malle et al. (2019) revealed that the command-chain of human orders to human pilots is more morally condemned than that of human orders to AI-drones. Similarly, Laakasuo, Palomäki, et al. (2023) revealed that the forced medication orders given by an AI vs. Human to the Human vs. Robot nurse produce different results, namely that a robot disobeying orders from a superior AI had the most morally appreciated decision. Thus, in Studies 2a, 2b and 3, we investigated the command-chain effect and the level of automation as a potential boundary that future theory might consider. Testing the materials in two different cultures gave us a robustness estimation on whether the effect of potential human-robot moral judgment asymmetry effect is culturally bound to the English-speaking population. What we found in Studies 2a, 2b and 3 was that it mattered whether there were "humans in the loop" at any point. In other words, the asymmetry effect seemed to generalize to both recommender and executor roles. However, we do not know with certainty if the asymmetry effect is present when the patient is conscious (Study 4).

In Study 4, we investigated whether the asymmetry would be seen in cases of active euthanasia, where a) the patient was conscious and b) several types of ending the patient's life were being studied. As discussed in previous research (Rodríguez-Arias et al., 2020), there are crucial differences in whether life-support is actually withdrawn from a conscious patient or whether a comatose patient is assisted to die. Study 4 was informed by Laakasuo, Palomäki, et al. (2023) where the patient was actively refusing treatment, while here we studied how patient autonomy manifested when the patient was requesting to be treated. Study 4 showed that the effect observed previously in a particular case of withdrawing life support disappeared when the patient was conscious, implying that this might be a potential boundary condition.

Having accounted for culturo-linguistic factors, replicability, patient consciousness and command-chain effects, the question remained whether the observed results could depend on character perception (Brambilla et al., 2021; Chapman, 2018; Laakasuo, 2023; Laakasuo, Palomäki, et al., 2023; Laakasuo, Palomäki, & Köbis, 2021) and, specifically, on the level of competence being attributed to the decision-maker (Laakasuo, Palomäki, et al., 2023). Hence, Study 5 adopted a competence manipulation as a possible boundary for a differential acceptance and moral evaluation of euthanasia when implemented by either a human or a robot physician. The investigation of competence is indeed quite common in the study of social cognition, where both warmth and competence are commonly found to be important dimensions for the formulation of moral judgments (Fiske et al., 2007). Competent individuals are often perceived negatively as they are considered exploitative (Cuddy et al., 2008; Rudert et al., 2017). These aspects are also important variables when predicting human preferences for different types of robot actions (Scheunemann et al., 2020), according to Madhavan and Wiegmann (2007), as people expect robots to perform perfectly and more precisely than humans. In Study 5, we found that the asymmetry effect was diluted when the competence of the agent was manipulated.

In Study 6, we accounted for explainability and accuracy as possible mechanisms for the observed effects – one could argue that they are aspects of competence. On the one hand, the opacity of both the reasons (Van Cauwenberge et al., 2022) and the processes (Shariff et al., 2017) enacted by AIs have been reported as potential determinants of distrust toward their application in general (Shariff et al., 2017) and with specific regard to the medical field (Van Cauwenberge et al., 2022), with explainability mediating trust and acceptance of AI in medical contexts (Shin, 2021). On the other hand, accuracy has already been studied for the implementation of AI (i.e., chatbots) for medical purposes

[13] Note that the studies presented here are not presented in the order in which they were conducted.

[14] Indeed, public opinions on end-of-life decisions vary widely, depending on the country and the formulation of the question (Marcoux et al., 2007; Rodríguez-Arias et al., 2020). For example, in a study in Croatia only 18.6 % agreed with the statement "Dying persons who are suffering severely should be granted their wish to die and be enabled to end their own lives" while 40.1 % agreed that "Physicians should be permitted by law to help a patient who is suffering from an incurable illness and is living with severe pain to end their own life, if the patient asks for it" – so do 1 or 2 out of 5 Croatians support assisted suicide (Borovecki et al., 2022). In Canada, public expressed high rates of approval of withdrawing life-prolonging treatment for a patient unlikely to recover (85 % for a competent patient, 88 % for an incompetent patient, who had expressed his/her wishes in advance through a living will, and 76 % for an incompetent patient based on the family's request) (Singer et al., 1995). In general, a significant minority seems to accept actively ending a patient's life when the patient asks for it (for example Cohen et al., 2006) and a large majority accepts withdrawing life-support in cases where there is suffering and no hope of recovery (for example Rydvall & Lynöe, 2008) and the overall trend seems toward more acceptance (Cohen et al., 2013).

(Nadarzynski et al., 2019). Study 6 stemmed from previous literature highlighting a general need for explanations in human-robot interactions (Anjomshoae et al., 2019; Han et al., 2021), namely for the description of their operations in intelligible ways (De Graaf and Malle, 2017), leading to better trust in AI systems (Lomas et al., 2012; Wang et al., 2016) and – to a different extent – causing a reduction of the artificial agent's blame after a moral violation commission (Malle & Phillips, 2023). Studies 5 and 6 showed that inferences about competence and explainability promote approval, implying that they are implicated in the moral 'penalty' imposed on AI agents.

2.2. Pre-registration statement

All studies were pre-registered; however, the studies here should be considered exploratory as none of our hypotheses held true consistently and we gave up on formal theory-driven hypothesis testing. Furthermore, the study order presented here deviates from the actual order in which the studies were run. We present the studies in an order that makes them easy to understand and in a way that resonates with similar previous work. The sample-size rationale and power analysis was followed as registered. Exclusion criteria were refined as the work progressed and all samples are analyzed with exclusions and with full samples – the results remain essentially the same. The materials for all the studies are here[15]; all the analysis scripts, outputs and variables reported in this manuscript are here.[16] The redacted data files will be made available upon the publication of the manuscript.

2.3. Ethics statement

These studies were either exempt from ethical review by ethical review boards inFinland, as there were no minors in volved and there was no violation of participants personal space and not required, or reviewed and approved by the Ethics Committee on Human Research of the University of Granada and by the Ethics Committee of the University of Keele. All local laws were followed in full.

3. Studies 1a and 1b – setting the stage

In Study 1a, we investigated the moral approval of an agent (human doctor versus robot doctor) making a decision to either turn off a patient's life support or to keep it on. The vignette described a person named Andy who has ended up in a coma due to a traffic accident and has a very low chance of regaining consciousness. Study 1b was the same study, but the data was collected in the Czech republic.

3.1. Method

3.1.1. Participants

For Study 1a, 720 participants were recruited via Prolific (www.prolific.co). Participants were excluded if they failed a) any of the attention checks, b) any of the comprehension checks, c) the "troll-response" checks, or d) did not have "good" or better self-assessed English skills. After exclusions, we had 628 participants (320 women), 65 % of whom were < 45 years old ($Age_M = 40.30$; SD = 14.12; range = 18–65) and 67 % had at least a Bachelor's degree, and 78 % reported having a minimum of mid-income. Study 1a lasted for 4 min.

For Study 1b, we collected 491 Czech-speaking participants from a volunteer participant pool website (https://pokusnikralici.cz/) and the affiliated Facebook page https://www.facebook.com/pokusnikralici; of whom 376 filled in the main dependent variable and 348 completed the whole study. Pokusní králíci participant pool is completely volunteer-based and the website does not offer monetary rewards to participants. The survey link was also distributed among attendees of in-person sci-fi conventions and members of online sci-fi enthusiast groups on social networks (the potentially noisy environment was taken into account in the questionnaire design; see below). Study 1b took an average of 20 min.

Participants were excluded for the same reasons as in Study 1a and if they reported that they could not fill in the questionnaire in a quiet and concentrated manner. After exclusions, we had 301 participants (158 women); 65 % of whom were < 45 years old (estimated $Age_M = 38.7$[17]) and 58.9 % had at least a Bachelor's degree, and 52 % had a minimum of mid-income.

3.1.2. Sample size rationale

For study 1a, we calculated that for 80 % power for Agent × Euthanasia interaction with 1 % explained variance, we would need 150 participants per condition, totaling $N = 600$. For Study 1b, the sample size was determined by opportunities (i.e., where we could get interested participants from), but we were aiming for about 400 participants. See OSF link for details.

3.1.3. Procedure & design

After giving their informed consent, participants completed exploratory measures for Study 1b (see preregistration). In both studies, participants were randomized into conditions in a 2 × 2 between-subjects factorial design: [human vs. robot doctor] × [passive euthanasia vs. no euthanasia]. Next, they read the vignette, responded to the dependent variables and to demographics. Finally they were thanked, debriefed (Studies 1a and 1b) and compensated (Study 1a). Study 1b included exploratory variables reported in the OSF preregistration link.

3.2. Materials

3.2.1. Vignette

In the vignette, a single agent, namely a human or a robot physician, was responsible for both the prognosis that the patient would likely never regain consciousness and the decision about what to do with the patient's life support. The events of the vignette happen in 2051, and machine learning has made it possible for physicians to foresee medical outcomes accurately. Our protagonist, Andy, is a traffic accident victim, who has been comatose for the past five years. A senior physician or an advanced AI, depending on condition assignment, reviews all the relevant data regarding Andy's case, and reports that the most likely outcome is that Andy will not regain his consciousness. They recommend turning off the mechanical ventilator that keeps Andy alive. See Appendix A for Vignettes.

3.2.2. Dependent variable/moral approval measure

Our DV had six items anchored from 1 (Totally Disagree) to 7 (Totally Agree); e.g. "The junior physician [robot physician] who [that] carried out the decision did what was right" of which three were reverse coded e.g. "It is morally wrong to carry out such a decision". Higher scores indicate greater moral approval of the decision to turn off life support (Cronbach's alpha = 0.89, Study 1a; Cronbach's alpha = 0.82, Study 1b). See Appendix B for listing of items.

3.3. Results of studies 1a and 1b

For Study 1a, we ran a full factorial two-way ANOVA on our moral approval measure by entering the decision-maker ($F(1, 624) = 14.55$, $p < .001$, $\eta_p^2 = 0.022$) and decision ($F(1,624) = 23.29$, $p < .001$, $\eta_p^2 = 0.036$) factors and their interaction ($F(1,624) = 14.68$, $p < .001$, $\eta_p^2 = 0.023$) as predictors, which were all statistically significant. There was

[15] https://osf.io/753dc/
[16] 10.6084/m9.figshare.28512059
[17] Due to idiosyncrasies associated with the data collection location, we had to use binned categories for ages.

in general more approval to turn off the life-support than to keep it on (B = 0.48, 95 % CI: [0.28, 0.68], $p < .001$). We further probed the interaction effect with contrast analysis: when the decision was made to withdraw life support, there was a statistically significant difference between human and robot, favoring the human doctor ($B = 0.76$, 95 % CI: [0.48, 1.04], $p < .001$), but no effect for maintaining the life support. See Fig. 1. The results were stronger without the exclusions.

We then ran the same analysis on the Czech data (1b). There was a statistically significant main effect for the Euthanasia Decision (F (1,297) = 29.65, $p < .001$, $\eta_p^2 = 0.090$), and a statistically significant interaction effect ($F(1, 297) = 6.24$, $p = .013$, $\eta_p^2 = 0.020$; see Fig. 1), but no main effect of the Agent factor ($F(1, 297) = 1.53$, $p = .210$). A planned contrast analysis replicated the results of Study 1a: Participants were more likely to approve the decision to turn off life support when made by a human than by a robot physician ($B = 0.52$, 95 %CI [0.14, 0.90] $F(1, 297) = 7.25$, $p = .007$), whereas no corresponding difference arose for maintaining life support ($F(1,297) = 0.77$, $p = .38$).[18]

3.4. Discussion of studies 1a and 1b

Both studies revealed that participants prefer passive euthanasia decisions to be made by humans rather than machines (see also Laakasuo, Palomäki, et al., 2023). Perhaps unexpectedly, and counter to our hypothesis at that time, there was generally higher approval for turning off life support compared to keeping it on. This preference, however, was stronger when a human doctor carried out the decision rather than an AI doctor. In other words, we observed an asymmetry effect where the moral approval of one decision (turning off life support) was evaluated differently for human and robot physicians.

These results dovetail with Laakasuo, Palomäki, et al. (2023), where the decision to forcefully medicate a patient was morally preferable when made by a human versus a robot nurse. Meanwhile, human and robot nurses that respect patient autonomy and *abstain from* forcefully medicating the patient, received comparable approval. In some sense, our observed pattern here is reversed: in Laakasuo, Palomäki, et al. (2023) the robot decision is dispreferred in the "worse decision"; here we observe that the robot decision is dispreferred in the more preferred option.

Nonetheless, we have established that the asymmetry effect also applies in passive euthanasia decisions and now turn to the command chain effect (Studies 2a, 2b and 3). Malle et al. (2019) and Laakasuo, Palomäki, et al. (2023) have shown that in hierarchical settings where robots need to take instructions and implement them, the asymmetry effect is attenuated. If the command-chain effect moderates the asymmetry effect or removes it, this will inform us on whether it is one of the boundary conditions.

4. Studies 2a and 2b – establishing the command chain effect

Studies 2a and 2b were designed as preliminary investigations into potential command chain effects in moral judgment involving AIs. In these studies, we aimed to determine how the presence of AIs in particular chain-of-command roles (e.g., recommender or decision-maker) affects the moral approval of the decisions made.

The vignette in the present study was the same as that used in Studies 1a and 1b, with the addition that there is a team that needs to decide whether to turn off Alex's life support. The team composition varied between Human-Human, AI-Human, or AI-AI, where the first agent played the role of the recommender and the second agent was the implementer. We examined whether mechanization of the decision chain alone was sufficient to evoke differences in moral judgment (see Malle et al., 2019). Study 2a was conducted in Finnish, while Study 2b was conducted in English. Due to resource constraints at the time, we only looked at the turn off decision and limited ourselves to a 1 × 3 design (see below).

4.1. Method

4.1.1. Participants & design

For Study 2a (Finnish), we recruited participants using Qualtrics XM form from a commercial data collection company Norstat Inc. and supplemented the data collection via Prolific and Snowballing techniques. Prolific participants were compensated £0.60 for 4 min on average. Altogether, after pre-registered exclusions (identical to Study 1a) and removing incomplete responses, our final sample size was 285 (161 women; 148 men) with average age being 38.26 (SD = 15.31, range: 18–74). Participants were native Finnish speakers, 53 % had a minimum of a Bachelor's degree.

The English speaking sample (Study 2b) consisted of 646 Prolific participants; after pre-registered exclusions sample-size was 401 (169 women). The majority (~55.6 %) of participants were < 45 years or younger; 68 % had at least a Bachelor's degree and 70 % reported being at least at medium income. Participants were compensated £0.40 for a 4 min study. Both studies had a fully randomized 1 × 3 Factorial design. The Teaming factor had three levels [1: AI Supervisor – Robot Doctor; 2 – AI Supervisor – Junior Physician; 3: Senior Physician – Junior Physician].

4.1.2. Sample size rationale

We focused on medium-sized effects, and estimated that sample size needed for 95 % power for Cohen's f = 0.0.22 (~5 % variance explained). This is reached with 104 participants per condition = 312 participants. We also had exclusion margins for the studies based on previous experience with the different data collection methods. We registered to aim for 400 participants after exclusions and in the Finnish sample fell short due participants not paying attention despite multiple manipulation checks.

4.1.3. Procedure

After informed consent, participants were randomized to one of three conditions where a team of medical experts decided to turn off the life support of a traffic victim, who has a low probability of regaining consciousness (see below). After reading the vignette, the participants answered the dependent variables, were debriefed and thanked.

4.2. Materials

4.2.1. Vignette

The vignette describes the same events as in Studies 1a and 1b. In this vignette, it is stated that the recommending agent cannot decide about Andy's life. The recommendation is thus passed on to a deciding party, who will re-evaluate the recommendation and then make the decision (either a human doctor or a medical robot). In this vignette version, the decision to turn off life support is held constant. For the vignettes, see Appendix A.

4.2.2. Dependent variable / approval of the decision to turn off life support

We had the same DV as previously (Cronbach's alpha = 0.92 for 2a and 0.93 for 2b).

4.3. Results of Studies 2a and 2b

We ran a one-way ANOVA on our Moral Approval of the Decision DV – for both samples – by using the Teaming factor as the predictor. The main effect of the Teaming factor was statistically significant ($F(2,282) = 6.99$, $p = .001$, $\eta_p^2 = 0.047$) in the Finnish sample (2a). There is a clear linear trend from the AI-AI pairing to Human-Human pairing ($B = 0.81$, 95 % CI: [0.39, 1.24], $F(1,282) = 13.78$, $p < .001$) and there is a clear

[18] We also ran the analysis with the maximum number of participants (i.e. everyone who filled in the DV; $N = 376$); the results were essentially the same.

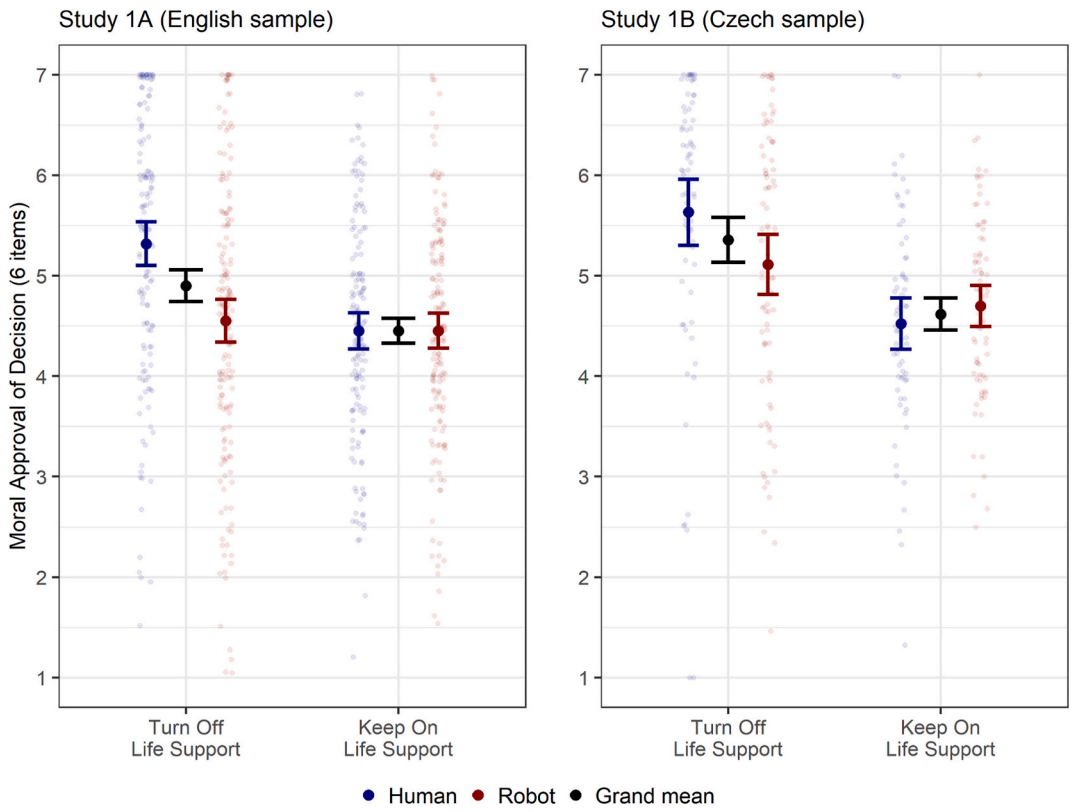

Fig. 1. Note. The human-robot moral judgment asymmetry effect was present in Studies 1a and 1b; when a robot makes the decision to turn off the life support, its decision is less appreciated than an otherwise identical human decision. Jittered data points represent individual observations; larger blue, red and black points are group-wise means.Error bars are 95 % CIs.

drop between the condition where the recommender and the decision-maker are both people, compared to either of the other two conditions (Contrast: Human-Human vs. the other two conditions: $B = 0.57$, 95 % CI: [0.21, 0.94], $F(1,282) = 10.1$, $p = .002$). The Teaming factor was significant ($F(2,398) = 16.87$, $p < .001$; $\eta_p^2 = 0.08$) also in the English speaking sample (Study 2b). There was a clear linear trend from the AI-AI pairing to Human-Human pairing ($B = 1.04$, 95 % CI: [0.68, 1.39], $F(1,398) = 33.39$, $p < .001$) and there was a clear drop between the condition where the recommender and the decision-maker were both people, compared to the other two conditions (Contrast: Human- Human vs other two conditions: $B = 0.77$, 95 % CI: [0.45, 1.08], $F(1,398) = 23.29$ $p < .001$). See Fig. 2. The results of Study 2 are the same with and without exclusions.

4.4. Discussion of Study 2a and 2b

Results suggest that one of the key factors influencing moral approval of passive euthanasia decisions is the extent to which the recommendation and decision-making process involve AIs. When it comes to moral judgments of passive euthanasia, participants generally prefer both the recommender and the decision-maker to be human. The basic setup and materials of our study are robust enough to produce replicable results across three cultures: Eastern European (1b), Northern European (2a), and English-speaking (2b). In all samples, we observe that higher levels of automation are associated with lower moral approval for passive euthanasia decisions. We continued to Study 3, where we evaluated a chain of command-related boundaries to determine if there is a greater willingness to accept any decision by a human-human team more than a decision from a mixed or fully automated decision-making, where we also included the fourth condition of Human supervisor - AI Physician teaming.

5. Study 3 – command chain effects deepened

In Studies 2a and 2b, the decision to turn off life support was always present, but here we also aimed to investigate approval in situations where the decision was to keep life support on. Additionally, in Studies 2a and 2b, we did not include a scenario where a human gives a recommendation to an AI agent, who then makes the decision.

5.1. Method

5.1.1. Participants, procedure & design

1500 English native participants were recruited through Prolific to a Qualtrics form. After excluding participants (we removed straightliners and used preregistrated exclusion criteria mentioned in Study 1), our final sample was 1154 participants (596 women). The majority of participants (52 %) were under 45 years old (M = 42.72, SD = 15.10, range = 19–82); 62 % had a minimum of a Bachelor's degree, and 73 % were at least mid-income. After informed consent, participants were randomized to read the vignette introduced in Studies 2a and 2b. However, this time the experiment had 2 × 2 between-subjects conditions and a counter-balanced additional within-subjects factor. The between-subjects factors were Recommender (human senior physician vs. AI) and Decision-maker (human junior physician vs. robot physician). The Decision (euthanasia vs. no euthanasia) was a counterbalanced within-subjects factor. In other words, participants evaluated two possible endings of the vignette. After debriefing, participants were compensated for their time (7 min, £0.60).

5.1.2. Sample size rationale

We were interested in small effect-sizes. We estimated the sample size requirement for 90 % power for Cohen's d = 0.07 for main effects (comparison between two conditions) is around 500 participants per

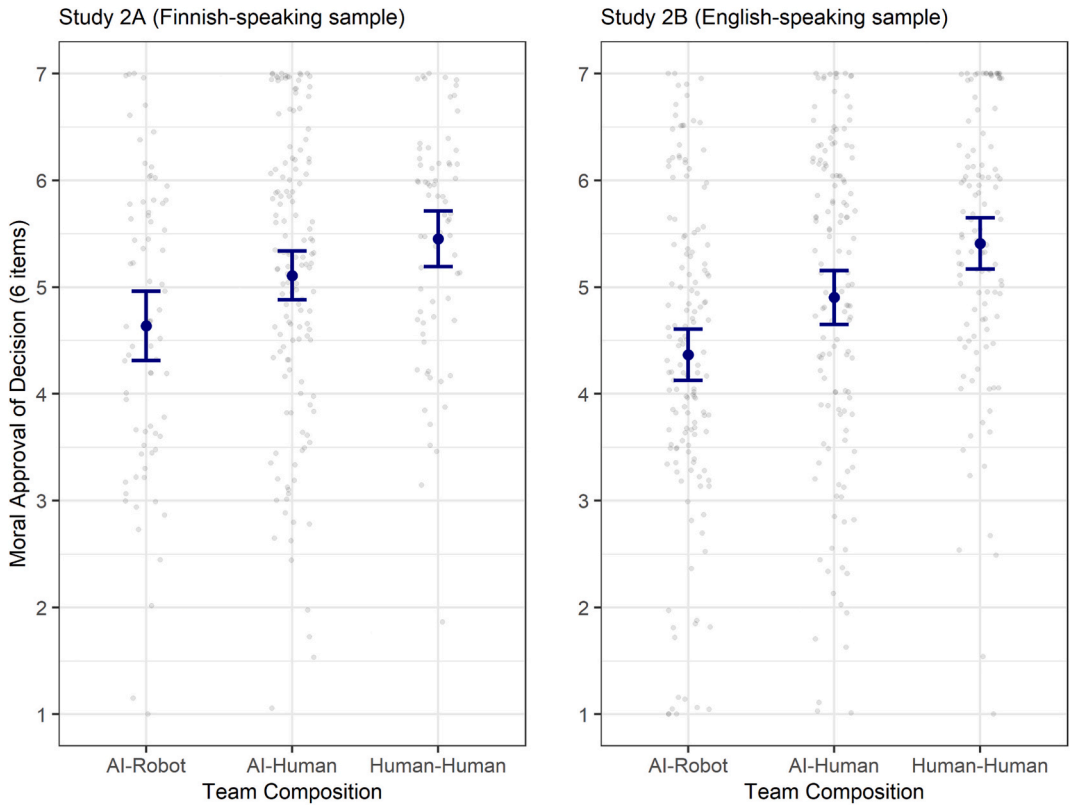

Fig. 2. Note. There is a clear linear trend from AI-AI teaming to Human-Human teaming in moral approval of a passive euthanasia decision (2 a: B = 0.59, 95 % CI: [0.39, 0.80], F(1,282) = 32.27, $p < .001$; 2b:B = 0.75, 95 % CI: [0.51, 98], F(1,403) = 38.55, $p < .001$). In Study 2b there is a clear drop between the condition where the recommender and the decision-maker are both people, compared to the other two conditions. Jittered gray data points are individual observations, larger blue points are group-wise means. Error bars are 95 % CIs.

condition (2000 in total). However, since every participant provided two sets of responses in each between-subjects condition by reacting to two within-subjects decisions, about 250 participants per cell would suffice in a 2 × 2(×2) design. We were aiming to collect 1100 participants after exclusions; for further details, see the OSF link.

5.2. Materials

5.2.1. Vignette and dependent variable

The vignettes were identical to Studies 2a and 2b, with the additions described above. The dependent variable was the same as in previous studies (Cronbach's alpha = 0.87).

5.3. Results of study 3

A full factorial three-way ANOVA was run with Moral approval as the dependent variable and Recommender, Decision-maker and Decision (within-subjects – both decisions counter-balanced per participant) as independent variables. There was a statistically significant main effect for Decision-maker: $F(1,2300) = 23.34$, $p < .001$, $\eta_p^2 = 0.01$. In addition, there was a two-way interaction between the Decision-maker and the Decision ($F(1, 2300) = 3.87$, $p = .049$, $\eta_p^2 = 0.001$) and between Recommender and the Decision ($F(1,2300) = 10.91$, $p <. 001$, $\eta_p^2 = 0.005$). With the decision to turn off life support, there was a clear linear trend from the Human-Human teaming combination to AI-Robot teaming combination (Linear contrasts: $B = 0.63$, 95 % CI: [0.40, 0.87], $p < .001$), with Human-Human teaming having the highest moral approval and the AI-AI teaming the lowest (see Fig. 3). This replicates the findings of Studies 2a and 2b. A similar linear trend was not present in the decision to keep on life support, where irrespective of the recommender or the decision-maker, the decision to keep on life support was equally approved ($B = -0.002$, $p = .99$). See Table 1. The results were the same or stronger without exclusions.

5.4. Discussion of study 3

In Study 3, we found a linear trend in the moral approval of the passive euthanasia decision when the level of automation increased, replicating the findings of Studies 2a and 2b. Finally, concerning keeping on life support, there were no effects, the response curve was flat. Taken together, Studies 2a, 2b and 3 show how the results generalize to both recommender and executor roles. Furthermore, these studies also support the notion that Sundvall et al. (2023) discuss, namely that it seems to be the "worse" decision of the two in a particular situation that raises this effect (although it is not always *a priori* clear, which is the worse decision). Thus, it seems that the command chain effect does not act as a full boundary condition, perhaps only in a limited case when there is a decision to keep the life-support systems on. Next, we investigated whether the preference for human doctors generalizes to a broader set of interventions, including active and passive forms of euthanasia. If it does not, this is a boundary condition for the asymmetry effect.

6. Study 4 – patient conscious / active forms of euthanasia

Next, we explored how people react to euthanasia decisions when the patient is conscious and has autonomy regarding their decisions (see Feltz, 2023). Whereas previous studies (1–3) focused on situations involving an unconscious patient, we now utilized a set of previously used and validated vignettes that distinguished between withholding treatment, withdrawing treatment, risky treatment, and medically assisted death. In all cases, the patient is conscious, experiencing an

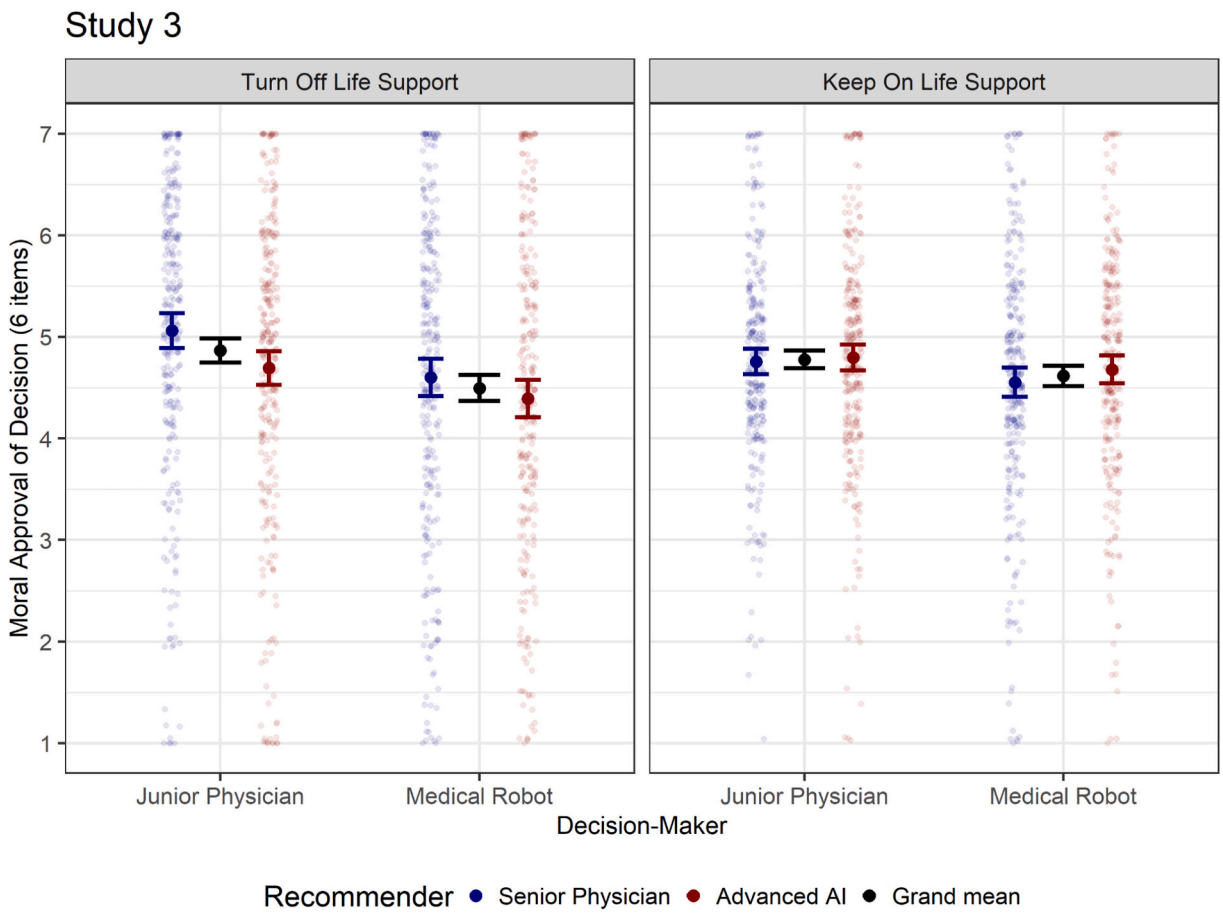

Fig. 3. Note. In the Decision to Turn off Life Support there is a clear linear trend, where the moral approval is lowest in the Advanced AI Group – Medical Robot condition and highest in the human-human teaming condition (B = 0.63, 95 % CI: [0.40, 0.87], p < .001). Such a trend is not observed in the Decision to Keep on Life Support. Jittered data points represent individual observations; larger blue, red and black points are group-wise means. Error bars are 95 % CIs.

Table 1
Three-Way ANOVA of Study 3.

	F	*P*	η_p^2
Factors			
Decision (Within Subjects)	0.03	0.86	0.00
Decision-Maker (DM)	23.34	<0.001	0.010
Recommender	3.34	0.067	0.001
Recommender × DM	1.21	0.270	0.005
Recommender × Decision	10.91	0.001	0.000
Decision-Maker × Decision	3.87	0.049	0.002
Recommender × DM × Decision	0.11	0.74	0.000

Note: We had 1154 participants in the study, who were counterbalanced into two different endings.
(Decision factor).

excruciating sensation of suffocation, and requesting a procedure from a doctor (see Rodríguez-Arias et al., 2020). These cases, however, differed in whether the physician (Roger[human] / Roger-X[robot]) carried out an omission (withholding treatment) or an action, whether the action led to the patient's death by limiting treatment (withdrawing treatment) or by administering a lethal injection, and whether death was the intended outcome (medically assisted death) or an unintended side effect (dangerous treatment). Of the four options, the one which we were most interested in was the Withdrawing of life-support, as this was the most analogous with our previous vignettes, with the difference that the patient was conscious.

6.1. Method

6.1.1. Participants

1024 participants recruited through Prolific opened the questionnaire and 971 filled in the dependent variables. After excluding participants with the same criteria as in Study 1a, the final sample size was 830 participants with a mean age of 41.79 (SD = 13.00; range = 18–78). Of these, 414 were women and 414 men, 80 % had at least a Bachelor's degree.

6.1.2. Sample size rationale

Our target sample size was 100 participants per cell, plus a 20 % margin to account for exclusions, leading to a total of $N = 800$. For details, see the OSF link.

6.2. Design

Participants were randomized into one of eight conditions in a 2 (Decision-Maker: robot vs. human) × 4 (Treatment type: withhold, withdraw, dangerous treatment, medically-assisted death) between-subjects design.

6.2.1. Procedure

After informed consent, participants read the vignettes. Post answering the manipulation and comprehension checks, they filled in the dependent variables regarding their moral views on the events described in the vignette. Finally, participants completed the demographics, were debriefed, thanked and paid for their time.

6.3. Materials

6.3.1. Vignettes

In these vignettes, a doctor (human: Roger/robot: Roger-X) has under his care an elderly patient who is suffering from a terminal illness and he is not able to breathe properly. The patient decides that he does not want to suffer anymore and he repeatedly communicates this to the doctor. The doctor (robot or human) either 1) withholds (i.e., nevers starts the treatment); 2) withdraws the respirator (i.e., starts the treatment, but allows the patient to die afterwards); 3) alleviates the patients' pain but the patient dies (dangerous treatment) or 4) applies a deep sedation so that the patient dies (medically assisted death; Adapted from Rodríguez-Arias et al., 2020). See Appendix A for vignettes.

6.3.2. Dependent variables

We had the same dependent variable as in previous studies (however, see below).

6.4. Results of study 4

We ran a full factorial two-way ANOVA by using the Moral Acceptance as the DV. There was only a significant main effect for the Decision-Maker ($F(1, 822) = 22.12$, $p < .001$, $\eta_p^2 = 0.026$). The Treatment condition was marginal ($F(3, 822) = 2.25$, $p = .08$) and the interaction was not statistically significant. We then reran the analysis without exclusions and found the Treatment condition to be statistically significant ($F(3, 963) = 3.48$, $p = .015$, $\eta_p^2 = 0.026$).

Having encountered this discrepancy between the full sample and the exclusions, we ran an ANOVA on each of the items in our DV separately. For an inexplicable reason, we found that the item *"The [robot]physician who carried out the decision did what was best for the patient"* was only contributing noise to the composite measurement ($F(3, 822) = 0.04$, $p = .98$). All other items were either trending or statistically significant (Fs 1.59–3.29 and ps 0.19–0.02). We then confirmed the same conclusion by running a groupwise CFA invariance analysis on our DV items (8 groups as the study was a 2 × 4 study). Only the item mentioned previously behaved statistically anomalously and was flagged by the analysis ($X^2 = 23.09$, $p = .001$; all other items X^2s < 13.03 and $ps > 0.07$). We thus left the problematic item out and remade our composite DV; all the remaining analyses are done with the 5 item version of the DV.

With our 5 item Moral Acceptance DV we found that there were statistically significant main effects for the Decision-Maker (B = 0.43, 95 % CI [0.24, 0.62], $F(1, 822) = 19.80$, $p < .001$, $\eta_p^2 = 0.023$) and the Treatment type ($F(1, 822) = 3.20$, $p = .02$, $\eta_p^2 = 0.01$). Participants morally approved humans as decision-makers more. Furthermore, participants approved Witholding Treatment and Dangerous Treatment over Withdrawing treatment and Medically Assisted Death (B = 0.47, 95 % CI [0.09, 0.86], $F(1, 822) = 6.06$, $p = .014$). This seems sensible as in both Withholding and in Dangerous treatments, the actual intention of the decision is not to kill the patient. Indeed, Rodríguez-Arias et al., 2020 intended the Withholding Treatment and Dangerous Treatment to be control conditions for Withdrawing treatment and Medically Assisted Death, respectively. Next, we ran a contrast analysis between the agents within each Treatment condition (see Fig. 4); Withdrawing Treatment was the only condition where there was no statistically significant difference between human and robot (B = 0.29, 95 % CI [−0.11, 0.70], $p = .15$). In all other conditions the difference was statistically significant (Bs > 0.44, $Fs > 5.61$, $ps < 0.015$). As this condition was the one most closely aligned with our previous studies, we took this to indicate, that patient being conscious in withdrawing/shutting down life support, could be a potential boundary condition.

We also observed that 1) when the decision-maker was a robot the Medically Assisted Death scenario was statistically significantly lower than the other conditions (B = −0.35, 95 % CI [0.05, 0.66], F(1,822) = 5.42, $p = .02$), but this was not the case when the decision-maker was human (B = −0.23, 95 % CI [−0.07, 0.54]], $F(1, 822) = 2.14$, $p = .14$) and 2) that the difference in the Withold condition between Robot and Human was statistically significantly lower than the average difference between the other conditions (B = 0.28, 95 % CI[0.09, 0.47], $F(1, 822) = 8.51$, $p = .036$). If these results can be couched as some type of an asymmetry effect, the asymmetry is associated with the Medically

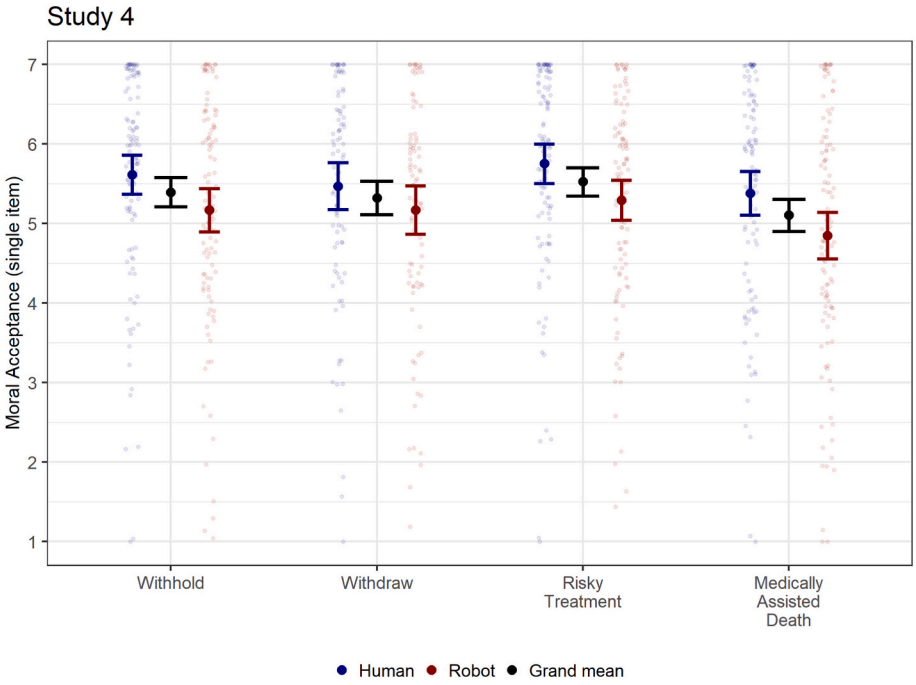

Fig. 4. Note. The Withdrawing condition is the only one without a statistically significant difference between conditions – this case is the most analogous to our previous vignettes, with the exception of the patient being conscious. In all other cases, the differences between the human and robot doctor are statistically significant. Jittered data points represent individual observations; larger blue, red and black points are group-wise means. Error bars are 95 % CIs.

Assisted Death condition, where the drop in moral approval for robots is statistically significant and the drop for humans is not. Results are essentially the same with and without exclusions.

6.5. Discussion of study 4

The results of Study 4 suggest that the effect associated with the Decision-Maker's type (Human vs. Robot) is influenced by specific boundary conditions, such as the patient being unconscious in Studies 1–3; or whether the patient's death is a directly intended consequence of the decision. It seems that when the patient demonstrates autonomy – i.e., is conscious – there is an effect which disappears in the Withdrawing (shutting down life support) condition. However, there is a potential asymmetry effect observed for the Medically Assisted Death scenario, as the robot decision-maker's drop is larger here than in the average of the other three conditions. It is possible that causing death by administering drugs was seen as less moral due to the more invasive nature of the method as compared to simply removing the life support device which results in a more natural death. Thus, the two methods are essentially different as the patient was forced to death in one case but was not sustained to life in the other which may have influenced moral perceptions of the participants.

The results align with previous research on non-AI-associated euthanasia decisions (e.g., Feltz, 2023) and support the findings presented by Laakasuo, Palomäki, et al. (2023), where one key factor in explaining the asymmetry effect is that robots are not permitted to violate human autonomy. This theme and a possible interpretation also emerged in Sundvall et al. (2023), where participants disapproved of robots that, in marine rescue situations, saved those who caused the accident over an innocent victim. It could be argued that those who caused the accident exercised autonomy, whereas those whose autonomy was violated deserved to be rescued.

Among the four conditions, the scenario where the patient is medically assisted to die was perceived as the most morally wrong – and this effect was driven by the robot agent condition. Thus, it appears acceptable for humans to make either decision—to start the treatment and later end it, or to never start it—but for robots, specifically actively administering deadly treatment is less acceptable compared to humans, even when explicitly requested by the patient. Importantly, there was no statistically significant difference between robot and human decision-makers in the Withdrawing condition; this scenario is the one most analogous to our previous studies, and it seems that when the patient is conscious and asks to die, the asymmetry disappears. This suggests that patient autonomy or consciousness is at least a potential boundary condition.

7. Study 5 – competence as a potential mechanism

In Study 5, we focused on perceptions of competence. Previous research by Laakasuo, Palomäki, et al. (2023) used a competence manipulation to investigate the boundary conditions of how people perceive violations of patient autonomy in the context of forced medication decisions. The results suggested that perceptions of competence influence evaluations of decisions made by humans, but not by robots. Furthermore, unlike in those studies, there is no clear violation of patient autonomy in the passive euthanasia decisions considered here.

Nonetheless, recent studies suggest that character/person perception mechanisms are important in moral judgment formation (e.g., Gamez et al., 2020; Laakasuo, 2023; Laakasuo, Sundvall, et al., 2023). When people observe decisions and their consequences, they evaluate these in relation to the individual performing the action. Put differently, people attribute more moral credit to those perceived as deserving (e.g., Miller, 2007). For instance, experienced competent surgeons are permitted to operate on patients, whereas incompetent individuals are not.

Thus, in Study 5, we manipulated the perceived competence of a decision-making human physician or medical robot by describing them as either well-performing and liked or error-prone and considered incapable. In other words, we aimed to rule out the possibility that the findings were due to perceived incompetence projected onto the robot.

7.1. Method

7.1.1. Participants, procedure & design

1156 participants were recruited via Prolific. Participants were excluded for the same reasons as in previous studies. After preregistered exclusions, we had 1089 participants (591 women); 66 % of whom were < 45 years old ($Age_M = 40.51$; SD = 13.84; range = 18–82) and 63 % had at least a Bachelor's degree, and 73 % reported having a minimum of mid-income. After informed consent, participants completed exploratory measures and were randomized into conditions in a 2 × 2 between-subjects factorial design: [Human vs. Robot Doctor] × [High Competence vs. Low Competence] × 2 within-subjects design: [Passive Euthanasia vs. No Euthanasia]. Next, they read the vignette, responded to the dependent variables and to demographics. Finally, they were debriefed and compensated 1.65£ for about 15 min.

7.1.2. Sample size rationale

We used similar reasoning and calculations as we had for Study 3. See OSF for details.

7.2. Materials

7.2.1. Vignette

The vignette was a modification of those used in Studies 1a and 1b. In this version of the vignette, we added the competence manipulation for both agents, whom we also named (John, the human physician / Johnmed, the robot physician). We had high and low competence manipulations that we adapted from Laakasuo, Palomäki, et al. (2023), as a description before the part of the vignette that describes the decision John/John-Med made.

High Competence: *[John/John-Med] has performed well in its work recently and performs tasks competently, with great precision. [John's/John-Med's] colleagues praise it for its abilities.*

Low Competence: *[John/John-Med] has been making constant mistakes in its work recently and performs tasks incompetently, with little precision. [John's/John-Med's] colleagues think that its abilities are not up for the job.*

7.2.2. Dependent variable / moral approval measure

We had the same dependent variable as in Study 3 (Cronbach's alpha: 0.88).

7.3. Results of study 5

We ran a three-way full factorial ANOVA on our Moral Approval Measure, and observed main effects of Decision-Maker ($F(1, 2170) = 42.54$, $p < .001$, $\eta_p^2 = 0.02$) and Competence ($F(1, 2170) = 47.67$, $p < .001$, $\eta_p^2 = 0.02$). See Table 2. The Decision and the Competence factor had an interaction effect ($F(1, 2170) = 5.02$, $p = .025$, $\eta_p^2 = 0.002$) and all factors had a three-way interaction effect ($F(1, 2170) = 5.54$, $p =$

Table 2
Three-way ANOVA results of Study 5 ($N = 1088$).

	F	*P*	η_p^2
Factors			
Decision-Maker (DM)	42.54	<0.001	0.02
Competence	47.67	<0.001	0.02
Decision	0.25	0.61	0.00
Decision-Maker × Competence	1.01	0.31	0.00
Decision-Maker × Decision	0.31	0.57	0.00
Competence × Decision	5.02	0.025	0.002
DM × Competence × Decision	5.54	0.018	0.002

Note. The Decision is a within-subjects factor.

.018, $\eta_p^2 = 0.002$). We then proceeded to analyze the interactions further based on the observed condition means and with planned contrast analysis.

As is clear from the visual inspection of Fig. 5, it seems like we replicate the previous pattern where Euthanasia decision is more approved than keeping on life support. We then examined whether the competence manipulation had an effect on the human agent, and found that highly competent humans' decisions received greater approval than decisions by low competence humans ($B = 0.49$, 95 % CI: [0.26, 0.72)], $p < .001$). In detailed analysis, the comparison between High Competence Human and Robot Doctor the decision to turn off life support was approved equally ($B = 0.15$, 95 % CI: [−0.07, 0.37], $p = .19$); however, there was a significant difference between High Competence Human and Robot Doctor in the decision to keep on life support ($B = 0.47$, 95 % CI: [0.24, 0.71], $F(1,2170) = 15.96$, $p < .001$), where the decision to keep on life support was more approved for humans. Contrast analysis revealed that the overall decisions of a competent Robot Doctor were more approved than decisions made by an incompetent one ($B = 0.91$, 95 % CI: [0.58, 1.24], $F(1,2170) = 30.20$, $p < .001$). There was a significant difference between the High Competent Robot's decision to turn off and keep on the life support, where turning the life support off was more morally approved ($B = 0.35$, 95 % CI: [0.12, 0.58], $F(1,2168) = 9.27$, $p = .002$). The difference in approval for Low Competence Robot Doctor's decisions was not statistically significantly different ($B = 0.17$, 95 % CI [−0.06, 0.40], $F(1,2168) = 2.10$, $p = .14$). The sample was analyzed without exclusions as well; the pattern of means and the conclusions remain basically the same.

7.4. Discussion of study 5

Results replicate previous findings and seem to establish a boundary condition for the asymmetry effect. There is an overall preference for humans to make decisions regarding the lives of other humans. Results confirm support for turning off life support in Alex's case. Crucially, we find an interaction effect concerning the robot doctor, but not the human doctor: participants prefer turning off life support only when the robot is perceived as highly competent, and, regarding keeping life support on, the highly competent robot was judged similarly to a human doctor with low competence. This suggests that the perceived (in)competence of the robot doctor could be one of the variables explaining the observed differences in moral approval of its decisions compared to those made by humans—thereby revealing a boundary condition for the asymmetry effect.

These results do not align with the previous forced medication dilemma experiments, where the asymmetry effect was diluted in the low competence condition and retained in the high competence condition. Perhaps the difference lies in the fact that, in Laakasuo, Palomäki, et al. (2023), the patient was conscious and requested not to be treated, while here the patient is unconscious and we do not know their wishes. Alternatively, it could be related to the fact that, in Laakasuo, Sundvall, et al. (2023), the nurse was part of a command chain.

8. Study 6: the role of explainability

Practices of responsibility attribution within human communities have been governed by the exchange of reasons. For example, justifying one's seeming misconduct by putting forth *good* reasons can reduce moral blame and mitigate legal culpability – and can be an indicator of competence. Meanwhile, many cutting-edge developments in AI build upon 'black box' algorithms, such as neural networks and random forests. These systems rely on unexplainable algorithms, given their improved performance on measures of accuracy. A defining characteristic of these algorithms is that it is virtually impossible to describe the process by which the output (e.g., a recommendation) was generated in a way that is intelligible to humans. In this regard, many contemporary AI technologies are known to lack explainability, such that–unlike what is habitual among humans–they cannot transparently report how they arrived at a certain decision.

This raises the question of whether attitudes toward medical AI are influenced by the algorithm's explainability. In Study 6, we investigate whether laypeople's attitudes toward the use of AI in an end-of-life medical context are influenced by accuracy and/or explainability – as potential proxies for competence. To what extent does the explainability of an AI's decision (i.e., to turn off a patient's life support) impact the moral approval of the decision?

8.1. Method

8.1.1. Participants, procedure & design

100 participants were collected in a pilot study and a bootstrap power analysis was subsequently performed on the basis of such data, with target sample sizes ranging from 600 to 1200. This produced a set of minimum sample sizes (power > 0.95, alpha =0.05) for effects of interest. The largest target (minimum) sample size was $N = 800$ to detect a true effect of accuracy on physician judgments. As such, 1118 participants were initially recruited via Prolific. After implementing the exclusion criteria related to comprehension check fails and attention check fails (~20 %). Exclusions were the same as in Studies 3 and 5. After exclusions, we had 1097 participants (539 identified as women) with mean age of 37.3 (SD = 13.7, range = 18–79).

After providing informed consent, participants were randomized into conditions in a 2 (Algorithm: explainable, unexplainable) × 21 (Accuracy: 77–97) between-subjects factorial design. Performance accuracy was manipulated continuously by sampling from the uniform distribution from 77 to 97 % accuracy. Next, they read the vignette in which a competent AI agent chose to withdraw life support, responded to the dependent variables and provided demographic information. Finally, participants were debriefed and compensated with £1.30. Study 6 included exploratory variables reported in the OSF preregistration link.

8.2. Materials

8.2.1. Vignette

The vignette was a modification of that used in previous studies in which we held constant 1) the agent (AI only), 2) the agent's competence (highly competent) and 3) their decision (to withdraw life support). We orthogonally manipulated 1) the opaqueness of the decision and 2) the accuracy of the AI agent's medical decision-making algorithm. Participants were randomly assigned to either the Explainable or the Unexplainable condition. In both conditions, participants were randomly assigned to a second manipulation of the AI algorithm's accuracy. Accuracy was manipulated continuously by sampling a percentage value from the uniform distribution between 77 % and 97 %.

In the *Explainable* condition, participants were presented with a scenario in which an AI agent employs an explainable algorithm drawn at random from a set of five possible algorithms–either (i) a decision tree, (ii) a linear regression, (iii) a logistic regression, (iv) k-nearest neighbor matching, or (v) a naïve Bayes classifier. The AI agent announces that "*The process I use to make decisions is easy to explain…*" and provides a succinct explanation of how the particular algorithm (e.g., a linear regression) results in the decision to turn off life support.

Meanwhile, in the *Unexplainable* condition, participants were presented with a scenario in which an AI agent employs an unexplainable (or black box) algorithm drawn at random from a set of five possible algorithms, either (i) a deep neural network, (ii) a random forest, (iii) a support vector machine, (iv) a genetic algorithm, or (v) deep reinforcement learning. The AI agent states that "*The process that yields my predictions is not easy to explain because…*" and provides a succinct description of how the particular algorithm works, and why it does not lend itself to an explanation for the decision to turn off life support.

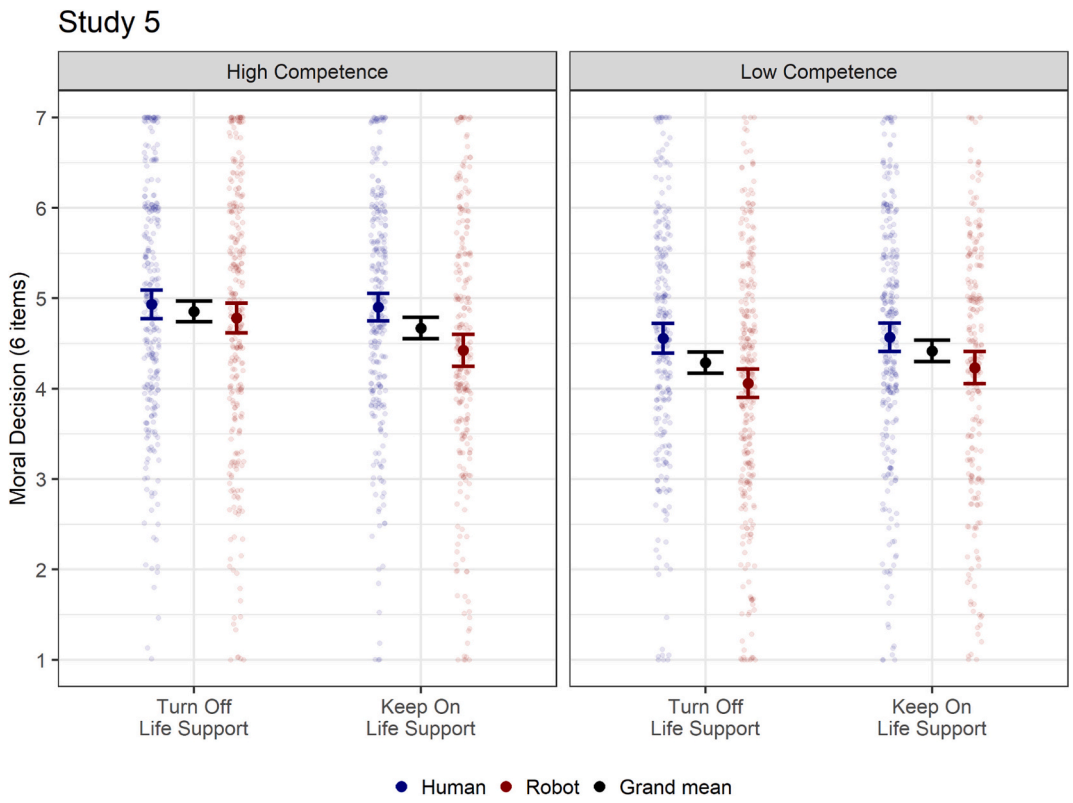

Fig. 5. Note. Once Competence is taken into consideration, the asymmetry effect shifts to Keep on the life-support system in the high competence conditions. Jittered data points represent individual observations; larger blue, red and black points are group-wise means. The error bars are 95 % CIs.

8.2.2. Accuracy

For example, "*In the past, my predictions have been 84% correct, which is much better than a single, senior human doctor can achieve*".

Additionally, the patient's gender (she/he) and name (Alex, Riley, Taylor) were randomized, as were the robot physician's name (JohnMed, JaneMed, JossMed) and the nature of the patient's accident (traffic, work, sport). This extraneous variation, however, was not analyzed for purposes of the present study.

8.2.3. Dependent variable

The primary dependent variable was the same as previously (Cronbach's alpha = 0.87).

8.3. Results of study 6

8.3.1. Manipulation check

We averaged the three explainability items (Cronbach's alpha = 0.71), and regressed perceived explainability on the explainability factor, the accuracy value, as well as the interaction between explainability and accuracy. The model revealed an effect of explainability ($F(1, 1093) = 606.5$, $p < .001$, $\eta_p^2 = 0.36$); other effects were not found ($ps > 0.14$). Perceived explainability was higher for explainable ($M = 79.1$, 95 % CI: [77.3, 80.8]) than unexplainable ($M = 48.1$, 95 % CI: [46.3, 49.8]) algorithms ($B = 31$, $t(1094) = 24.63$, $p < .001$).

8.3.2. Decision approval

In the model of decision approval with the explainability factor, the accuracy value, as well as the interaction between explainability and accuracy as predictors, we observed small main effects of both explainability, $F(1, 1094) = 6.87$, $p = .009$, $\eta_p^2 = 0.006$, and accuracy, $F(1, 1094) = 7.36$, $p = .007$, $\eta_p^2 = 0.007$, and no interaction between explainability and accuracy, $p = .98$. The main effect of explainability revealed that participants favoured explainable over unexplainable algorithms ($B = 0.21$, 95 % CI: [0.05, 0.36], $t(1094) = 2.62$, $p = .009$) and accuracy promoted approval ($B = 0.02$, 95 % CI [0.01, 0.03], $t(1094) = 2.71$, $p = .007$).

Thus, the decision was evaluated most favorably when it resulted from the use of an explainable algorithm, and when the model had been found to be accurate in a prior (e.g., training) dataset. The ratio of the regression coefficients suggested that, on average, a 12 % increase in accuracy would suffice to counteract participants' dispreference for unexplainable artificial intelligence in the current medical context. See Fig. 6.

8.4. Discussion of study 6

Comparing attitudes toward a set of explainable (e.g., linear regressions or naive Bayes classifiers) and unexplainable (e.g., neural networks or random forests) algorithms, Study 6 revealed that participants weakly (yet significantly) prefer algorithms to be explainable. However, this effect accounted for less than 1 % of the variance in attitudes toward medical decisions to turn off life support. Study 6 also indicated that approval was influenced by the AI's accuracy – such that participants expressed greater approval of decisions made by *accurate* (than by inaccurate) AI models. Together, these effects imply that sufficient gains in predictive accuracy may compensate for the weak penalty incurred by unexplainable AI – supporting the results of Study 5 where the perceived competence of the AI partially explained our results. However, we cannot be sure whether this would support our examination into the asymmetry effect without a human comparison.

9. Internal meta-analysis on the human vs. robot autonomy violations

Finally, we standardized the DV from all studies (except Study 6 as it did not include human-robot comparisons) and re-ran the analyses with

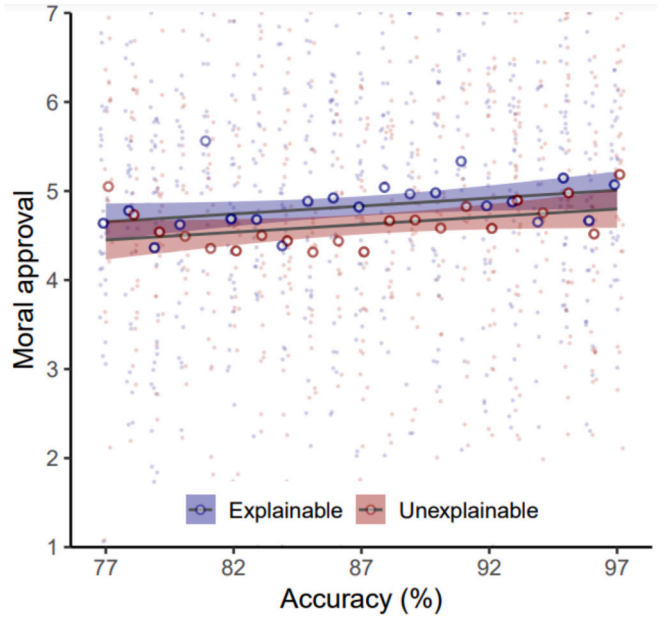

Fig. 6. Linear trends in moral approval by accuracy and explainability. Note. In both the Explainable and Unexplainable conditions, we observed positive effects of accuracy on moral approval. The slopes of these effects did not differ across Explainability conditions. Error bands are 95 % CIs.

the reported exclusions. We then dummy coded all the effects on whether the decision was made by a human or a robot and whether the decision was to turn off or keep on life support (or in Study 4 some other decision), and also created another dummy coded column which indicated the involvement of any AI in the process.

We ran a moderated fixed effects meta-analysis using STATA 17.1 by using the contrast between committing any euthanasia decision vs. not, and observing situations where there was any AI involvement vs. only Humans involved. We found a clear asymmetry effect: Robots received higher approval for keeping on the life support (Standardized B = 0.08, 95 % CI: [0.05, 0.11], $p < .001$) and lower approval for turning it off (Standardized B = −0.08, 95 % CI: [−0.12, −0.05], $p < .001$). In contrast, human decisions had higher approval for euthanasia (Standardized B = 0.24, 95 % CI: [0.19, 0.29], $p < .001$) and lower approval for keeping the patient alive (Standardized B = −0.06, 95 % CI: [−0.08, −0.01], $p < .001$). See Fig. 7 for results.

We then added another dummy coded variable on whether the patient was conscious and a three-category variable on indicating teaming (0 = no supervisor; 1 = human supervisor and 2 = AI supervisor; Studies 2a, 2b and 3) and ran a meta-regression on our data. We entered the decision-maker, the decision to turn-off or medically assist the patient to die and their interaction effect as predictors into the model and added the active / passive decision, and supervisor variables as categorical predictors without interaction effects. There was a a) clear negative main effect of the robot being the decision-maker ($B = -0.18$, z = 5.44, 95 % CI: [−0.25, −0.11], $p < .001$) and b) a positive main effect of the decision to turn off life support ($B = 0.11$, z = 3.55, 95 % CI: [0.05, 0.18], $p < .001$); and a statistically significant interaction effect ($B = -0.14$, z = 3.17, 95 % CI: [−0.24, −0.05], $p = .002$; roughly corresponding to the one presented in Fig. 9). There was no effect for passive / active distinction ($B = -0.00$, z = −0.24, 95 % CI: [−0.08, 0.06], $p = .80$); perhaps due to low power. However, there was a negative effect of Supervisor being an AI (vs. being absent: B = −0.08, z = −2.71, 95 % CI: [−0.13, −0.02], $p = .006$; and vs. Supervisor being Human: B = −0.13, 95 % CI: [−0.20, −0.05], $p < .001$). All in all, the meta-analysis suggests that the asymmetry effect exists and people prefer other people to make end-of-life decisions without the involvement of AIs.

10. General discussion

In eight studies, we showed that the human-robot moral judgment asymmetry effect was present in human vs. robot-made passive euthanasia decisions. In Studies 1a and 1b, we found reduced moral approval of an AI doctor's decision to withdraw life support relative to a human doctor. These findings were replicated in both an English-speaking Prolific sample and a Czech-speaking sample. In Studies 2a, 2b, and 3, we consistently demonstrated (across two different types of samples: Finnish and an English-speaking Prolific sample) that the asymmetry effect generalized across recommender and decision-making roles in Human-AI comparisons. In other words, these studies showed that the asymmetry effect emerged regardless of whether the AI assumed a recommender or a decision-maker role when doctors decided to turn off life support. In Study 4, we found that the asymmetry effect was absent in the Withdraw life support condition – closest analogue to Studies 1–3

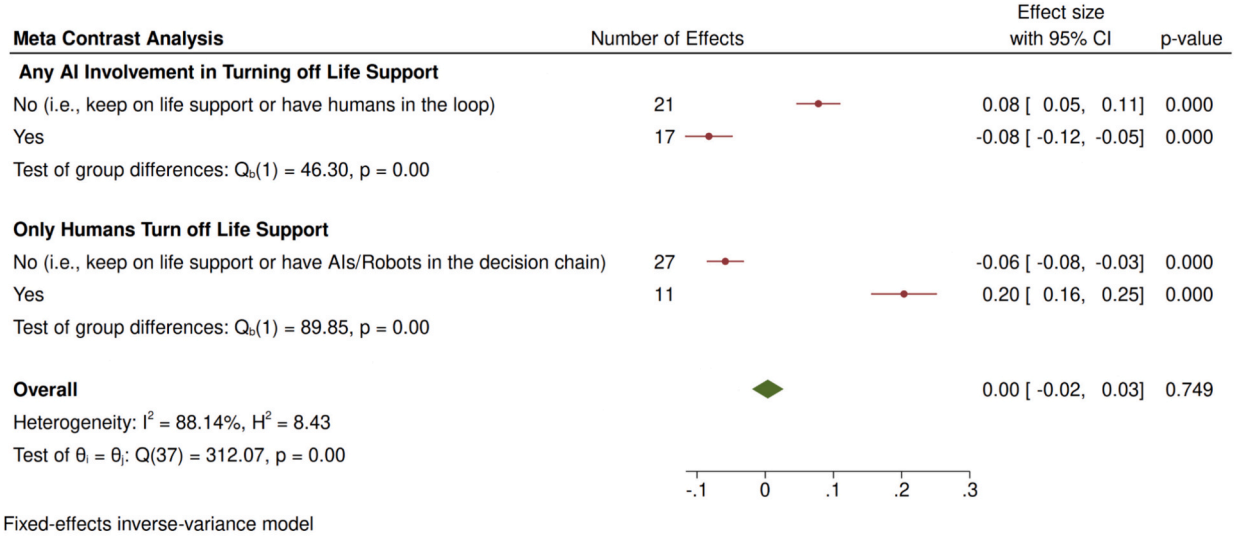

Fig. 7. Note. Meta-analysis of the asymmetry effect. All the effects were pulled from the individual conditions of the studies presented here. The negative effect of robots making the decision to turn off life support is the same as humans keeping it on. The positive effect of robots keeping on life-support is smaller than humans turning it off.

– implying that patient autonomy or consciousness served as at least a partial boundary condition. Thus, by comparing Study 4 with other studies (1–3 and 5), we were able to examine variations in autonomy violations. In Study 4, we observed the asymmetry effect only in the context of actual Medically assisted death treatment, where in a sense the robot makes the "worst" decision – actively and intentionally causes the death of the patient. In Study 5, we found that the asymmetry effect in the decision to turn off life support became smaller—or nonsignificant — when perceptions of competence were held constant between human and robot agents. In Study 6, we found that the dispreference for the medical AI was also partly explained by perceptions of explainability and accuracy, which may serve as proxies for competence. In the meta-analysis, we observed that in a combined sample of about 4000 responses across different countries and situations, the asymmetry effect remained prominent, replicating previous research (see Sundvall et al., 2023).

The results of the internal meta-analysis—together with findings from Study 5—suggest that there may be an asymmetry in the perceived costs of errors when turning off someone's life support, who might wake up, compared to keeping life support on for someone who will not. If we are certain that the patient will not regain consciousness, then the decision to turn off life support may be considered better. However, if there is a possibility that the patient could recover, a "play it safe" approach is warranted to give them a chance. If people perceive robots as less competent (Study 5), accurate, and explainable (Study 6) than humans, there is greater uncertainty surrounding the decision, making the "play it safe" strategy a Hippocratic obligation. However, this explanation does not fully account for why costs and benefits are perceived in this way.[19]

In relation to recent research investigating the human-robot asymmetry effect in moral judgment formation, the research presented here appears to replicate this emerging pattern. For example, Sundvall et al. (2023) observed in eight studies that when a rescue robot saved two motorboaters—who were responsible for causing an accident—over an innocent fisherman (with all individuals having fallen into freezing water and being at risk of death), the decision received less approval compared to when it was made by a human lifeguard, especially when the motorboaters had violated the fisherman's autonomy. Similarly, Laakasuo, Palomäki, et al. (2023) demonstrated that in forced medication dilemmas, a nursing robot's violation of autonomy was not tolerated, whereas the same action was acceptable for human nurses. In these studies, the patient or victim was conscious, while in the present studies they were mostly unconscious, implying that decisions to withdraw life support from an unconscious patient constitute an autonomy violation, which is not considered permissible when performed by robots (but equally approved if the patient is conscious).

In previous research, Malle et al. (2016) found a human-robot asymmetry in blame attributions (a different DV from the present studies) in two out of three studies, and specifically for robots with a mechanical appearance. However, robots were blamed more for inaction than for action. Similarly, Zhang et al. (2022) reported that robots were expected to be more utilitarian in high-conflict moral dilemmas than humans, whereas Malle et al. (2016) used low conflict moral dilemmas. This is notable as Laakasuo (2023) found that the robot's appearance also plays a significant role in this differential treatment. The decisions of an uncanny humanoid-looking robot and a very mechanical-looking robot (which were perceived to be more uncanny than comparison robots) – were found to be less moral than a human action/commission/utilitarian decision. In all the studies presented here, we specifically found the asymmetry effect in the commission decisions where the life-support was withdrawn from an unconscious patient. Thus, we found the following features surrounding this dilemma that could be relevant for future model building: 1) the perceived level of automation in the decision chain, 2) the perceived autonomy violation (of the patient), 3) the perceived competence and 4) the perceived explainability of the decision-maker. These features seem to span three distinct language families. Nonetheless, none of them alone completely explains the asymmetry effect.

In the present body of work, there is no core theory or framework to guide the empirical approaches in predicting the outcome of any particular experiment in a given context (e.g., military, mining accidents, forced medication, euthanasia, etc.). Furthermore, we lack the ability to predict from first principles which decision will evoke the moral asymmetry effect and why. Despite this, the current set of studies reveals that factors such as the level of perceived automation, potential autonomy violations, perceived competence, and explainability play a role. We suggest that this provides a solid foundation for developing future theories in this area of research (see also Laakasuo, Sundvall, et al., 2021a; Laakasuo, Sundvall, et al., 2021b), when we also consider some additional details.

For instance, Feltz (2023) suggests that in cases of autonomy violations, all other experimental factors become almost irrelevant when considering (non-AI assisted) euthanasia; while Castelo and Ward (2021) suggest that people perceive AI decisions as more risky in medical contexts compared to other areas (see also Longoni et al., 2019). With this in mind, we may be approaching an understanding of some cognitive mechanisms related to the asymmetry effect.[20] We previously stated that the effect discussed by Bigman and Gray (2018)—that humans are averse to robots as decision-makers because they perceive them as less minded than humans—is a separate effect from the asymmetry effect (which is about decisions, not decision-makers). Yet, it would be rather surprising if these two effects – machine aversion effect and the moral judgment asymmetry effect – were entirely unrelated. However, in the studies by Sundvall et al. (2023), Laakasuo (2023), and Malle et al. (2016), it seems that the imagined or perceived human-like bodily appearance—rather than mind perception—may be driving the effect; in Laakasuo (2023), mind perception was even controlled for in the experiments. Thus, the answer to the asymmetry effect is likely to be more complex than just the lack of perceived mindedness of the agent. Perhaps, the perception of autonomy in agents and patients is also associated with being perceived as able-bodied, aligning with evolutionary theorizing (Boehm, 2012).

Respondents in several experimental philosophy studies have been observed to be willing to ascribe the ability to see, smell, or have beliefs to robots, but not the ability to feel pain or happiness (Heubner, 2010; Sytsma, 2014). Heubner suggests that people employ different strategies—agency and personhood—when attributing mental states to other entities. Agency focuses on rationality and goal-directed behavior, whereas personhood emphasizes moral considerations (Haslam et al., 2008; Heubner, 2010). While both strategies are applied to humans,

[19] We thank the anonymous reviewer for this interpretation

[20] In Teisseyre et al. (2005), autonomy violations were more accepted when carried out by family members, while Boehm (2012) suggests that most mercy killings in hunter-gatherer societies were executed by relatives. This aligns with previous theorizing in evolutionary psychology, which posits that kin selection forms the basis of human social cognition and potentially gives rise to rudimentary mechanisms of moral cognition (such as altruism) (Tooby & Cosmides, 2005). It appears acceptable for close kin to make decisions that violate an individual's autonomy, with the responsibilities of parenting being an obvious example. In the context of euthanasia, ending the life of an elderly family member may also be viewed as freeing up resources to care for younger individuals (Boehm, 2012; Kurzban et al., 2012). The perception of autonomy, and its weakening in a community member, affects every member of that community; therefore, our capacity to perceive autonomy in others could be crucial for survival. Perhaps, then, including DNR statements from the patient and consent from family members for AI decision-making could be relevant factors to introduce in vignettes for future studies. However, even if these factors dilute or remove the asymmetry effect in this context, it does not necessarily mean that this mechanism will generalize to other contexts.

they are not consistently applied to non-human entities (Heubner, 2010). This is similar to Dennett's distinction between the intentional and personal stance, with the latter requiring an additional "moral commitment" (Dennett, 1981). Therefore, it may not be mind perception per se that future studies should control for, but rather certain aspects of personhood or autonomy linked to moral capacities. Some of these aspects may be triggered by human-like forms of robots (see Sundvall et al., 2023). Indeed, evolutionary psychological research has suggested that bodily disfigurement might trigger discriminatory behaviors (Curtis, 2011); as Laakasuo (2023) has shown, moral judgments of robot behaviors are associated with their appearance.

Thus, the fact that robots do not deserve the full range of moral consideration in people's minds is likely to impact how their decisions are viewed. Another possible explanation could be that robots are perceived to be more capable than humans in a technical sense and thus, people expect them to make the more difficult decisions and have higher standards for them (Sundvall et al., 2023); i.e., when the robot does the "right thing" it is to be expected and not praiseworthy, but when the robot does the "wrong thing" it should be blamed more, as it could have done better than humans. However, this explanation fails to elucidate in which situation the "wrong decisions" are wrong for the robot. This is the theory we lack.

Another piece of the asymmetry effect puzzle could be the role of competence (Gamez et al., 2020; Laakasuo, Palomäki, et al., 2023; Miller, 2007), which also surfaced in Study 5. When a competent robot or human doctor decided to turn off life support, there was no observed difference; however, there was a difference for non-competent robot and human doctors. This might suggest that humans are seen as "expert surgeons" in the realm of difficult moral decisions, since they are familiar with the entire human condition and thus possess an authentic sense for deontological decisions.[21] Although in a previous study by Laakasuo, Palomäki, et al. (2023) the asymmetry effect was present even under the "high competence" manipulation, the crucial difference between that study and our study is that the patient was conscious. In Laakasuo, Palomäki, et al. (2023), the deontological option of respecting the patient's will was preferred in the high competence condition, whereas in our Study 5, the utilitarian option of ending the patient's life was perhaps preferred. The "worse" decision was to keep life support on, where the asymmetry effect still persisted.

Nonetheless, it seems that decisions which are driven by something other than just pure utilitarian calculus, are something that humans prefer. In Sundvall et al. (2023), it was the utilitarian option that was "wrong" for the robot. Longoni et al. (2019) suggest that this is because people feel that AIs and algorithms do not take into consideration important individual uniqueness factors in different situations. Van Cauwenberge et al. (2022) suggests that this is because humans are also able to provide intelligible explanations for their decisions and experiences (See Study 6). Perhaps these threads are linked with how humans view AIs and robots as non-transparent, "cold" (Laakasuo, Palomäki, et al., 2023), purely rational and goal-oriented agents (Madhavan & Wiegmann, 2007). Robots and AIs do not possess the moral "bonus", and are thus expected to be the ones tending toward precise and subsequently utilitarian decisions (Soares et al., 2023; Wu et al., 2022), unless it contradicts the principle of culpability or autonomy violations (as shown by Sundvall et al., 2023). Human autonomy (Study 4) may be viewed as a crucial moral right and thus decisions that respect this autonomy are preferable also from the utilitarian point of view. Another explanation could be based on the fact that, as we mentioned in the introduction, the very complex and perhaps even paradoxical nature of a mixture of different human intuitions is at play here: preference for warm and humane conduct on the one hand, and competent and coldly rational conduct on the other hand. With human intuitions being so complex and often mutually incompatible, we have to expect the potential theory to also take this blurriness into consideration.

To sum up, we can speculate about a theory that aligns with our findings and also helps explain asymmetry effects observed elsewhere. It might be partially rooted in the idea that robots cannot attain full personhood, which carries important moral privileges, including how we perceive the decisions they make. However, we are still left uncertain as to why this matters for some decisions but not for others. For instance, Sundvall et al. (2023) did not control for perceived competence but instead examined orthogonal situational factors that ended up diluting the asymmetry effect—such as determining who was culpable for the accident.

This brings us to the limitations of these studies. As with most moral psychological research in this field (see Laakasuo, 2023), there is a lack of consensus on standardized materials, relevant moral dilemmas that could serve as a reliable test bed, the appropriate dependent variable—whether it should be a single item or multiple items—and, most crucially, whether the focus should be on the agent's suitability as a decision-maker or on the decision itself (including its antecedents and consequences). Perhaps it is still too early to determine these aspects, and the field would benefit from maintaining an open approach to methodological variation. Additionally, there are standard limitations common to any vignette-based experiments, such as the use of convenience samples, lack of ecological validity, and demand characteristics, but these are not particularly unique to our studies.

Thus, future studies should focus their investigations on perceptions of the potential factor of autonomy violations being one of the pieces in bringing forth the human-robot moral judgment asymmetry effect. It remains unknown why we are – at least in some situations – more sensitive toward machines violating human autonomy than toward humans doing so, and in which situations the violations of human autonomy are relevant, and why.

Another aspect that future research could focus on is that there are very few studies in experimental moral psychology of AI and robotics incorporating relevant individual differences measures (see Koverola et al., 2020 and Koverola et al., 2022; Laakasuo et al., 2018; Laakasuo, Repo, et al., 2021; Laakasuo, Sundvall, et al., 2023, for examples of the opposite). For instance, if the mind perception hypothesis of the human-robot moral judgment asymmetry effect were true (for arguments why this is unlikely, see Laakasuo, 2023), then those individuals with more tendencies toward pareidolia and apophenia would be more susceptible to the asymmetry effect, as they would be more likely to perceive minds in robots. Or, if human autonomy violations are one of the key components in evoking this effect, then participants with values related to human autonomy and human self-determination would be more likely to exhibit this effect in counterbalanced within-subjects experiments.

11. Conclusions

All in all, in eight experiments we showed that any future theorizing related to Human-Robot moral judgment asymmetry should focus on aspects such as the decision-making system's perceived level of automation and the degree of potential human autonomy violations. It is also crucial to focus on what happens in the context of the implementation of algorithms and of decision-making systems. In addition, the perceived level of competence, the extent of human involvement and degree of explainability are aspects that replicate in contexts outside of AI assisted euthanasia. Whatever our technological future may look like, it is important to study the challenges it poses as the new research questions these developments open are both compelling and thought-provoking. We need a new theory for this new era and we invite participation.

[21] People from the general population (non-philosophers) do not think about moral decisions in these terms. Their concept of moral decision-making is much more vague and intuitive, since they lack the relevant theoretical training. We use terms like deontology and utilitarianism only to categorize moral decisions for our own purposes.

CRediT authorship contribution statement

Michael Laakasuo: Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Anton Kunnari:** Writing – review & editing, Software, Methodology, Formal analysis, Conceptualization. **Kathryn Francis:** Writing – review & editing, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Michaela Jirout Košová:** Writing – review & editing, Writing – original draft, Resources, Investigation. **Robin Kopecký:** Writing – review & editing, Resources, Investigation. **Paolo Buttazzoni:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Data curation, Conceptualization. **Mika Koverola:** Writing – review & editing, Methodology, Investigation. **Jussi Palomäki:** Writing – review & editing, Visualization, Conceptualization. **Marianna Drosinou:** Writing – review & editing, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Ivar Hannikainen:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Acknowledgments

This research was funded by a Research Council of Finland grant (360123) awarded to Michael Laakasuo, who was the principal investigator and conceptualized the research. Marianna Drosinou was additionally funded by Tiina and Antti Herlin Foundation. This research is part of NetResilience consortium funded by the Strategic Research Council within the Academy of Finland (grant number 345186 and 345183). Additional data collection by Kathryn Francis and Ivar Hannikainen was funded by corresponding departmental grants. Michael Laakasuo would like to thank Metallica for their song "One" from *…And Justice for All* album for providing inspiration and October Tide for their song "Deplorable Request". Without these songs and their emotional impact this paper would not have been possible.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106177.

Data availability

Analysis, code and output and materials have been linked in the paper.

References

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In N. Agmon, M. E. Taylor, E. Elkind, & M. Veloso (Eds.), *18th international conference on autonomous agents and multiagent systems (AAMAS 2019)* (pp. 1078–1088). International Foundation for Autonomous Agents and Multiagent Systems.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bahník, Š., & Vranka, M. A. (2021). Consistency and contrast effects in moral evaluation of euthanasia. *Current Psychology, 40*, 822–830. https://doi.org/10.1007/s12144-018-0012-7

Bartels, L., & Otlowski, M. (2010). A right to die? Euthanasia and the law in Australia. *Journal of Law and Medicine, 17*(4), 532–555.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Boehm, C. (2012). *Moral origins: The evolution of virtue, altruism, and shame*. Soft Skull Press.

Borovecki, A., Curkovic, M., Nikodem, K., Oreskovic, S., Novak, M., Rubic, F., Vukovic, J., Spoljar, D., Gordijn, B., & Gastmans, C. (2022). Attitudes about withholding or withdrawing life-prolonging treatment, euthanasia, assisted suicide, and physician assisted suicide: A cross-sectional survey among the general public in Croatia. *BMC Medical Ethics, 23*(1), 13. https://doi.org/10.1186/s12910-022-00751-6

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. In B. Gawronski (Ed.), *Vol. 64. Advances in experimental social psychology* (pp. 187–262). Elsevier Academic Press.

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

Caddell, D. P., & Newton, R. R. (1995). Euthanasia: American attitudes toward the physician's role. *Social Science & Medicine, 40*(12), 1671–1681. https://doi.org/10.1016/0277-9536(94)00287-4

Castelo, N., & Ward, A. F. (2021). Conservatism predicts aversion to consequential artificial intelligence. *PLoS One, 16*(12), Article e0261467. https://doi.org/10.1371/journal.pone.0261467

Černý, D. (2015). Je ukončení život udržující léčby eutanazií? In I. Humeník, I. Szaniszló, & Z. Zolákov (Eds.), *Právné otázky rozhodovania v onkologické starostlivosti* (pp. 133–152). Wolters Kluwer.

Černý, D. (2018). Eutanazie a dobrý život: Proč je eutanazie (někdy) morální. *Vnitřní lékařství, 64*(3), 236–244. https://doi.org/10.36290/vnl.2018.034

Chapman, H. A. (2018). A component process model of disgust, anger and moral judgment. In K. Gray, & J. Graham (Eds.), *The atlas of moral psychology* (pp. 70–80). The Guilford Press.

Cohen, J., Marcoux, I., Bilsen, J., Deboosere, P., Van der Wal, G., & Deliens, L. (2006). European public acceptance of euthanasia: Socio-demographic and cultural factors associated with the acceptance of euthanasia in 33 European countries. *Social Science & Medicine, 63*(3), 743–756. https://doi.org/10.1016/j.socscimed.2006.01.026

Cohen, J., Van Landeghem, P., Carpentier, N., & Deliens, L. (2013). Different trends in euthanasia acceptance across Europe. A study of 13 western and 10 central and eastern European countries, 1981–2008. *The European Journal of Public Health, 23*(3), 378–380. https://doi.org/10.1093/eurpub/cks186

Crocker, K. (2013). Why euthanasia and physician-assisted suicide are morally permissible. Retrieved from http://purl.flvc.org/fsu/fd/FSU_migr_phi2630-0010.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. In M. P. Zanna (Ed.), *Vol. 40. Advances in experimental social psychology* (pp. 61–149). Elsevier Academic Press. https://doi.org/10.1016/S0065-2601(07)00002-0.

Curtis, V. (2011). Why disgust matters. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 366*(1583), 3478–3490. https://doi.org/10.1098/rstb.2011.0165

De Graaf, M. M., & Malle, B. F. (2017, October). How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

Deak, C., & Saroglou, V. (2017). Terminating a child's life? Religious, moral, cognitive, and emotional factors underlying non-acceptance of child euthanasia. *Psychologica Belgica, 57*(1), 59. https://doi.org/10.5334/pb.341

Dennett, D. C. (1981). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: The MIT Press.

Douglas, C. (2009). End-of-life decisions and moral psychology: Killing, letting die, intention and foresight. *Journal of Bioethical Inquiry, 6*(3), 337–347. https://doi.org/10.1007/s11673-009-9173-2

Feltz, A. (2023). Everyday attitudes about euthanasia and the slippery slope argument. In *New directions in the ethics of assisted suicide and euthanasia* (pp. 145–165). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-22050-5_13.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & Society, 35*(4), 795–809. https://doi.org/10.1007/s00146-020-00977-1

Gamliel, E. (2013). To end life or not to prolong life: The effect of message framing on attitudes toward euthanasia. *Journal of Health Psychology, 18*(5), 693–703. https://doi.org/10.1177/1359105312455078

Goligher, E. C., Ely, E. W., Sulmasy, D. P., Bakker, J., Raphael, J., Volandes, A. E., … Downar, J. (2017). Physician-assisted suicide and euthanasia in the ICU: A dialogue on Core ethical issues. *Critical Care Medicine, 45*(2), 149–155. https://doi.org/10.1097/CCM.0000000000001818

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029. https://doi.org/10.1037/a0015141

Haakenstad, A., et al., GBD 2019 Human Resources for Health Collaborators. (2022). Measuring the availability of human resources for health and its relationship to universal health coverage for 204 countries and territories from 1990 to 2019: A systematic analysis for the global burden of disease study 2019. *The Lancet, 399*(10341), 2129–2154. https://doi.org/10.1016/S0140-6736(22)00532-3

Han, I. X., Meggers, F., & Parascho, S. (2021). Bridging the collectives: A review of collective human–robot construction. *International Journal of Architectural Computing, 19*(4), 512–531. https://doi.org/10.1177/14780771211025153

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: Contrasting humans and nonhumans in three cultures. *Social Cognition, 26*(2), 248–258. https://doi.org/10.1521/soco.2008.26.2.248

Hayakawa, S., Tannenbaum, D., Costa, A., Corey, J. D., & Keysar, B. (2017). Thinking more or feeling less? Explaining the foreign-language effect on moral judgment. *Psychological Science, 28*(10), 1387–1397. https://doi.org/10.1177/0956797617720944

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Heubner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone *care* about functional zombies? *Phenomenology and the Cognitive Sciences, 9*(1), 133–155. https://doi.org/10.1007/s11097-009-9126-6

Ho, R. (1998). Assessing attitudes toward euthanasia: An analysis of the subcategorical approach to right to die issues. *Personality and Individual Differences, 25*(4), 719–734. https://doi.org/10.1016/S0191-8869(98)00108-1

Kahn, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., … Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction* (pp. 159–160).

Karumathil, A. A., & Tripathi, R. (2022). Culture and attitudes towards euthanasia: An integrative review. *OMEGA-Journal of Death and Dying, 86*(2), 688–720. https://doi.org/10.1177/0030222820984655

Keuning, B. E., Kaufmann, T., Wiersema, R., Granholm, A., Pettilä, V., Møller, M. H., … HEALICS consortium. (2020). Mortality prediction models in the adult critically ill: A scoping review. *Acta Anaesthesiologica Scandinavica, 64*(4), 424–442. https://doi.org/10.1111/aas.13527

Komatsu, T., Malle, B. F., & Scheutz, M. (2021). Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across US and Japan. In *HRI '21: Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 63–72). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3434073.3444672.

Koverola, M., Drosinou, M., Palomäki, J., Halonen, J., Kunnari, A., Repo, M., Lehtonen, N., & Laakasuo, M. (2020). Moral psychology of sex robots: An experimental study– how pathogen disgust is associated with interhuman sex but not interandroid sex. *Paladyn: Journal of Behavioral Robotics, 11*(1), 233–249. https://doi.org/10.1515/pjbr-2020-0012

Koverola, M., Kunnari, A., Drosinou, M., Palomäki, J., Hannikainen, I. R., Jirout Košová, M., … Laakasuo, M. (2022). Treatments approved, boosts eschewed: Moral limits of neurotechnological enhancement. *Journal of Experimental Social Psychology, 102*, 1–21. https://doi.org/10.1016/j.jesp.2022.104351

Kurzban, R., DeScioli, P., & Fein, D. (2012). Hamilton vs. Kant: Pitting adaptations for altruism against adaptations for moral judgment. *Evolution and Human Behavior, 33*(4), 323–333.

Laakasuo, M. (2023). Moral Uncanny Valley revisited–how human expectations of robot morality based on robot appearance moderate the perceived morality of robot decisions in high conflict moral dilemmas. *Frontiers in Psychology, 14*, Article 1270371. https://doi.org/10.3389/fpsyg.2023.1270371

Laakasuo, M., Drosinou, M., Koverola, M., Kunnari, A., Halonen, J., Lehtonen, N., & Palomäki, J. (2018). What makes people approve or condemn mind upload technology? Untangling the effects of sexual disgust, purity and science fiction familiarity. *Palgrave Communications, 4*(1), 1–14. https://doi.org/10.1057/s41599-018-0124-6

Laakasuo, M., Palomäki, J., & Köbis, N. (2021). Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics, 13*(7), 1679–1688. https://doi.org/10.1007/s12369-020-00738-6

Laakasuo, M., Palomäki, J., Kunnari, A., Rauhala, S., Drosinou, M., Halonen, J., … Francis, K. B. (2023). Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy. *European Journal of Social Psychology, 53*(1), 108–128. https://doi.org/10.1002/ejsp.2890

Laakasuo, M., Repo, M., Drosinou, M., Berg, A., Kunnari, A., Koverola, M., … Sundvall, J. (2021). The dark path to eternal life: Machiavellianism predicts approval of mind upload technology. *Personality and Individual Differences, 177*, Article 110731. https://doi.org/10.1016/j.paid.2021.110731

Laakasuo, M., Sundvall, J., Francis, K., Drosinou, M., Hannikainen, I., Kunnari, A., & Palomäki, J. (2023). Would you exchange your soul for immortality?—Existential meaning and afterlife beliefs predict mind upload approval. *Frontiers in Psychology, 14*, Article 1254846. https://doi.org/10.3389/fpsyg.2023.1254846

Laakasuo, M., Sundvall, J. R., Berg, A., Drosinou, M., Herzon, V., Kunnari, A., … Palomäki, J. (2021a). Moral psychology and artificial agents (part one): Ontologically categorizing bio-cultural humans. In S. Thompson (Ed.), *Machine law, ethics, and morality in the age of artificial intelligence* (pp. 166–188). IGI Global. https://doi.org/10.4018/978-1-7998-4894-3.ch010.

Laakasuo, M., Sundvall, J. R., Berg, A., Drosinou, M., Herzon, V., Kunnari, A., … Palomäki, J. (2021b). Moral psychology and artificial agents (part two): The Transhuman connection. In S. Thompson (Ed.), *Machine law, ethics, and morality in the age of artificial intelligence* (pp. 189–204). IGI Global. https://doi.org/10.4018/978-1-7998-4894-3.ch011.

Levin, K., Bradley, G. L., & Duffy, A. (2020). Attitudes toward euthanasia for patients who suffer from physical or mental illness. *OMEGA-Journal of Death and Dying, 80*(4), 592–614. https://doi.org/10.1177/0030222818754667

Lockhart, C., Lee, C. H., Sibley, C. G., & Osborne, D. (2023). The sanctity of life: The role of purity in attitudes towards abortion and euthanasia. *International Journal of Psychology, 58*(1), 16–29. https://doi.org/10.1002/ijop.12877

Lomas, M., Chevalier, R., Cross, E. V., Garrett, R. C., Hoare, J. R., & Kopack, M. (2012). Explaining robot actions. In *HRI '12: Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction* (pp. 187–188). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2157689.2157748.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650.

MacDonald, W. L. (1998). Situational factors and attitudes toward voluntary euthanasia. *Social Science & Medicine, 46*(1), 73–81. https://doi.org/10.1016/S0277-9536(97)00146-9

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors, 49*(5), 773–785. https://doi.org/10.1518/001872007X230154

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology, 72*, 293–318. https://doi.org/10.1146/annurev-psych-072220-104358

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being. Intelligent systems, control and automation: Science and engineering, Vol. 95* (pp. 111–133). Cham: Springer. https://doi.org/10.1007/978-3-030-12524-0_11.

Malle, B. F., & Phillips, E. (2023). A robot's justifications, but not explanations, mitigate people's moral criticism and preserve their trust. https://osf.io/preprints/psyarxiv/dzvn4.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. J. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *HRI '15: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117–124). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2696454.2696458.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *HRI'16 the eleventh ACM/IEEE international conference on human robot interaction* (pp. 125–132). Piscataway, NJ: IEEE Press. https://doi.org/10.1109/HRI.2016.7451743.

Marcoux, I., Mishara, B. L., & Durand, C. (2007). Confusion between euthanasia and other end-of-life decisions: Influences on public opinion poll results. *Canadian Journal of Public Health, 98*(3), 235–239. https://doi.org/10.1007/BF03403719

Miller, G. F. (2007). Sexual selection for moral virtues. *The Quarterly Review of Biology, 82*(2), 97–125. https://doi.org/10.1086/517857

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *DIGITAL HEALTH, 5*, Article 2055207619871808. https://doi.org/10.1177/2055207619871808

Nichols, R., Smith, N. D., & Miller, F. (Eds.). (2008). *Philosophy through science fiction. A coursebook with readings*. Routledge.

Passerard, F., & Menaud, X. (2015). Physician-assisted suicide at the crossroads of vulnerability and social taboo: Is death becoming a consumption good? In S. Dobscha (Ed.), *Death in a consumer culture* (pp. 208–221). Routledge.

Preston, T. A. (1994). Professional norms and physician attitudes toward euthanasia. *The Journal of Law, Medicine & Ethics, 22*(1), 36–40. https://doi.org/10.1111/j.1748-720X.1994.tb01273.x

Rabin, S., Kika, N., Lamb, D., Murphy, D., Am Stevelink, S., Williamson, V., Weselly, S., & Greenberg, N. (2023). Moral injuries in healthcare workers: What causes them and what to do about them? *Journal of Healthcare Leadership, 15*, 153–160. https://doi.org/10.2147/JHL.S396659

Rodríguez-Arias, D., Rodriguez Lopez, B., Monasterio-Astobiza, A., & Hannikainen, I. R. (2020). How do people use 'killing','letting die' and related bioethical concepts? Contrasting descriptive and normative hypotheses. *Bioethics, 34*(5), 509–518. https://doi.org/10.1111/bioe.12707

Rudert, S. C., Reutner, L., Greifeneder, R., & Walker, M. (2017). Faced with exclusion: Perceived facial warmth and competence influence moral judgments of social exclusion. *Journal of Experimental Social Psychology, 68*, 101–112. https://doi.org/10.1016/j.jesp.2016.06.005

Rydvall, A., & Lynöe, N. (2008). Withholding and withdrawing life-sustaining treatment: A comparative study of the ethical reasoning of physicians and the general public. *Critical Care, 12*(1), R13. https://doi.org/10.1186/cc6786

Schafer, A. (2013). Physician assisted suicide: The great Canadian euthanasia debate. *International Journal of Law and Psychiatry, 36*(5–6), 522–531. https://doi.org/10.1016/j.ijlp.2013.06.002

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Scheunemann, M. M., Cuijpers, R. H., & Salge, C. (2020, August). Warmth and competence to predict human preference of robot behavior in physical human-robot interaction. In *2020 29th IEEE international conference on robot and human interactive communication (RO-MAN)* (pp. 1340–1347). IEEE. https://doi.org/10.1109/RO-MAN47096.2020.9223478.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour, 1*(10), 694–696. https://doi.org/10.1038/s41562-017-0202-6

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Singer, P. A., Choudhry, S., Armstrong, J., Meslin, E. M., & Lowy, F. H. (1995). Public opinion regarding end-of-life decisions: Influence of prognosis, practice and process. *Social Science & Medicine, 41*(11), 1517–1521. https://doi.org/10.1016/0277-9536(95)00057-E

Soares, A., Piçarra, N., Giger, J. C., Oliveira, R., & Arriaga, P. (2023). Ethics 4.0: Ethical dilemmas in healthcare mediated by social robots. *International Journal of Social Robotics, 15*(5), 807–823. https://doi.org/10.1007/s12369-023-00983-5

Streeck, N. (2020). Death without distress? The taboo of suffering in palliative care. *Medicine, Health Care and Philosophy, 23*, 343–351. https://doi.org/10.1007/s11019-019-09921-7

Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens Rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–27. https://doi.org/10.1145/3479507

Sumner, L. W. (2011). *Assisted death: A study in ethics and law*. Oxford University Press.

Sundvall, J., Drosinou, M., Hannikainen, I., Elovaara, K., Halonen, J., Herzon, V., Kopecký, R., Jirout Košová, M., Koverola, M., Kunnari, A., Perander, S., Saikkonen, T., Palomäki, J., & Laakasuo, M. (2023). Innocence over utilitarianism: Heightened moral standards for robots in rescue dilemmas. *European Journal of Social Psychology, 53*(4), 779–804. https://doi.org/10.1002/ejsp.2936

Sytsma, J. (2014). The robots of the dawn of experimental philosophy of mind. In E. Machery, & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 48–64). Routledge.

Teisseyre, N., Mullet, E., & Sorum, P. C. (2005). Under what conditions is euthanasia acceptable to lay people and health professionals? *Social Science & Medicine, 60*(2), 357–368. https://doi.org/10.1016/j.socscimed.2004.05.016

Tigard, D. W. (2018). Rethinking moral distress: Conceptual demands for a troubling phenomenon affecting health care professionals. *Medicine, Health Care and Philosophy, 21*(4), 479–488. https://doi.org/10.1007/s11019-017-9819-5

Tiwari, S. P., Upadhyay, A., & Karthikeyan, S. (2020). Artificial intelligence based comparative study of mortality prediction. In *2020 fourth international conference on computing methodologies and communication (ICCMC)* (pp. 910–914). IEEE.

Tooby, J., & Cosmides, L. (2005). Conceptual foundations of evolutionary psychology. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 5–67). John Wiley & Sons, Inc.

Van Cauwenberge, D., Van Biesen, W., Decruyenaere, J., Leune, T., & Sterckx, S. (2022). "Many roads lead to Rome and the artificial intelligence only shows me one road": An interview study on physician attitudes regarding the implementation of computerised clinical decision support systems. *BMC Medical Ethics, 23*(1), 1–14. https://doi.org/10.1186/s12910-022-00787-8

Van Hoof, A. (2004). Ancient euthanasia: 'Good death' and the doctor in the graeco-Roman world. *Social Science & Medicine, 58*(5), 975–985. https://doi.org/10.1016/j.socscimed.2003.10.036

Vanderelst, D., & Willems, J. (2020). Can we agree on what robots should be allowed to do? An exercise in rule selection for ethical care robots. *International Journal of Social Robotics, 12*(5), 1093–1102. https://doi.org/10.1007/s12369-019-00612-0

Vizcarrondo, F. (2013). Euthanasia and assisted suicide: The physician's role. *The Linacre Quarterly, 80*(2), 99–102. https://doi.org/10.1179/0024363912Z.0000000002

Voiklis, J., & Malle, B. F. (2018). Moral cognition and its basis in social cognition and social regulation. In K. Gray, & J. Graham (Eds.), *Atlas of moral psychology* (pp. 108–120). New York, NY: Guilford.

Walsh, D., Caraceni, A. T., Fainsinger, R., Foley, K., Glare, P., Goh, G., et al. (2009). Euthanasia and physician-assisted suicide. In *Palliative medicine* (1st ed., pp. 110–115). Saunders.

Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 109–116). https://doi.org/10.1109/HRI.2016.7451741

Worsnop, R. L. (1997). Caring for the dying: Would better palliative care reduce support for assisted suicide? *CQ Researcher, 33*(6), 769–792. https://doi.org/10.4135/cqresrre19970905

Wu, J., Xu, L., Yu, F., & Peng, K. (2022). Acceptance of medical treatment regimens provided by AI vs. Human. *Applied Sciences, 12*(1), 110. https://doi.org/10.3390/app12010110

Zhang, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology, 101*, Article 104327. https://doi.org/10.1016/j.jesp.2022.104327