Received: 12 March 2021 | Accepted: 4 July 2022

DOI: 10.1002/ejsp.2890

EASP WILEY

RESEARCH ARTICLE

Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy

Michael Laakasuo[1] | Anton Kunnari[2] | Sanna Rauhala[1] | Marianna Drosinou[2] | Juho Halonen[1] | Noora Lehtonen[1] | Mika Koverola[1] | Marko Repo[1] | Jukka Sundvall[1] | Aku Visala[3] | Kathryn B. Francis[4] | Jussi Palomäki[1]

[1]Faculty of Arts, Dept. of Digital Humanities, Cognitive Science, University of Helsinki, Helsinki, Finland

[2]Faculty of Medicine, Dept. of Psychology and Logopedics, University of Helsinki, Helsinki, Finland

[3]Faculty of Theology and Religious Studies, University of Helsinki, Helsinki, Finland

[4]Faculty of Natural Sciences, School of Psychology, Keele University, UK

Correspondence
Michael Laakasuo, Faculty of Arts, Dept. of Digital Humanities, Cognitive Science, University of Helsinki, Helsinki, Finland.
Email: michael.laakasuo@helsinki.fi

Abstract

Artificial intelligences (AIs) are widely used in tasks ranging from transportation to healthcare and military, but it is not yet known how people prefer them to act in ethically difficult situations. In five studies (an anthropological field study, $n = 30$, and four experiments, total $n = 2150$), we presented people with vignettes where a human or an advanced robot nurse is ordered by a doctor to forcefully medicate an unwilling patient. Participants were more accepting of a human nurse's than a robot nurse's forceful medication of the patient, and more accepting of (human or robot) nurses who respected patient autonomy rather than followed the orders to forcefully medicate (Study 2). The findings were robust against the perceived competence of the robot (Study 3), moral luck (whether the patient lived or died afterwards; Study 4), and command chain effects (Study 5; fully automated supervision or not). Thus, people prefer robots capable of disobeying orders in favour of abstract moral principles like valuing personal autonomy. Our studies fit in a new era in research, where moral psychological phenomena no longer reflect only interactions between people, but between people and autonomous AIs.

KEYWORDS
moral psychology of robotics, personal autonomy, nursing robots, moral judgments, human-robot interaction

1 | INTRODUCTION

The use of artificial intelligence (AI) and assistive robots in medical and healthcare settings specifically has seen a rapid increase (McDougall, 2019) and represents several opportunities and challenges. In addition to investigating the relative efficiency and proficiency of such services, understanding public perception, acceptance, and judgments of service robots is critical in determining the extent to which AI-based tools and services will be used in medical settings (Esmaeilzadeh, 2020). People increasingly encounter and interact with robots in various contexts. Socio-cognitive processes now also encompass judgments of AIs' decisions in morally difficult situations, and social psychological research must react accordingly.

1.1 | Medical ethics and social robotics: Violations of autonomy as social and moral dilemmas

A number of ethical principles and schools of thought have been employed in medical decision-making from deontological (duty-based) and principalist (i.e., axiom-like universal perspectives) to care ethics perspectives (i.e., bringing well-being to the weakest party; Gillon, 1994; Esmaeilzadeh, 2020; Beauchamp & Childress, 2001). In general,

the following four ethical principles are often listed as the cornerstones of medical or nursing ethics (and of the Hippocratic oath): (1) respect of individual autonomy, (2) active beneficence, (3) active avoidance of maleficence, and (4) justice (Beauchamp & Childress, 2001; Gillon, 1994). While this framework was developed for contexts in which human agents operate, several researchers have argued that the same basic principles can be used to examine ethical principles related to AI and assistive robots (for a detailed discussion, see Feil-Seifer & Mataric, 2011; Lin et al., 2017; Van Wynsberghe, 2013).

In fact, one of the fundamental tensions in medical and nursing ethics is between the individual's autonomy and their well-being, for which the medical professionals are responsible. This is especially true in cases when healing the patient requires invasive measures (Liegeois & Audenhove, 2005; Iyalomhe, 2009; Pellegrino & Thomasma, 1987). For example, surgery on an unconscious person due to traumatic injury involves physically damaging the patient with a surgical instrument to stop internal bleeding or to remove a foreign object. In these cases, informed consent cannot be obtained from the patient, and the medical procedure itself could cause permanent damage. Moreover, psychiatric patients are sometimes forcefully restrained and sedated to prevent them from harming themselves in confused mental states. Compromising a person's autonomy to potentially increase their long-term well-being or even save their life is, at its core, a social dilemma, or a value conflict, where the lesser evil among two options must be chosen.

Research in moral psychology has shown that violating autonomy leads to moral condemnation and/or the expression of negative moral emotions. For example, research has established an association between feelings of outrage and anger and violations of autonomy (e.g., Graham et al., 2011; Rozin et al., 1999). In the well-known trolley dilemma, a runaway trolley is speeding on a track to collide with and kill five people. The moral agent can decide to pull a lever and divert the trolley to a different track, which would result in the death of only one person (Foot, 1967; Greene et al., 2001). In the so-called footbridge version of the dilemma, one must decide whether to stop the trolley by pushing another person off a bridge in front of it (Greene et al., 2001). In these dilemmas, sacrificing one person to save five, or not doing so, corresponds to utilitarian (aiming to maximize the number of lives saved) and deontological (believing "trading lives" is never acceptable) moral intuition, respectively. Low moral acceptability ratings of utilitarian actions in so-called personal moral dilemmas (requiring direct rather than indirect killing of a person, e.g., the footbridge dilemma), are thought to be driven, at least partially, by violations of autonomy as individuals are treated as a means to an end (Everett et al., 2016; Baumard & Sheshkin, 2015).

The similarity between violations of autonomy arising in medical ethics and the personal moral dilemmas becomes apparent with closer examination (Liegeois & Audenhove, 2005; Iyalomhe, 2009; Pellegrino & Thomasma, 1987). Patients' autonomy is sometimes compromised to increase their well-being and potentially their long-term autonomy as well, because healthy people are often better equipped to take care of themselves than those with untreated illnesses. Moreover, physicians and nurses must sometimes decide on whether, or when, to disregard the patient's will when treating them. Respecting or disregarding the patient's will to some extent resembles a moral decision between deontology and utilitarianism, provided that disregarding the patient's will is motivated by an intention to increase their long-term well-being and autonomy.

In terms of human–robot interactions, it appears that similar reactions are observed when robots violate autonomy. For example, people generally agree that a care robot's violations of autonomy are unacceptable (Vanderelst & Willems, 2020). There is variation in these ratings based on the type of violation (repeating a request to take medicine vs. administering medicine forcefully) and the severity of the patient's medical condition (acute schizophrenia versus mild vision impairment; Vanderelst & Willems, 2020). While patient autonomy has been identified as a relevant issue in this area, it has only received limited attention in previous empirical studies on medical ethics (Stahl & Coeckelbergh, 2016).

1.2 | The capacity, competence, and moral agency of robots

In social psychology, both warmth and competence are considered critical dimensions in social (and moral) judgments (Fiske et al., 2007). Individuals judged to be incompetent are often seen negatively in the sense that they are exploitative (Cuddy et al., 2008) and a burden in any group (Rudert et al., 2017). These dimensions also appear to be key predictors of human preferences for certain robot behaviours (Scheunemann et al., 2020). In fact, people typically expect robots to perform flawlessly and to a greater level of perfection than humans (e.g., Madhavan & Wiegmann, 2007). When robot service providers do make mistakes, people rate their competence more negatively (e.g., Brooks et al., 2016).[1]

Gamez et al. (2020) described to their participants situations where either a human or a robot engages in virtuous or vicious acts and participants judge their level of virtue. The authors described scenarios of different virtue ethics areas, such as truth, justice, fear, wealth, and honour. In both quantitative and qualitative analyses, they found that moral attributions were diminished for robots compared to humans. Other studies suggest that the perceived moral character of an agent mostly determined how well the agent was liked and to some extent how warm they were perceived (Goodwin et al., 2015; Laakasuo, Köbis & Palomäki, 2021). These findings are further complemented by those of Young & Monroe (2019), who showed that when machines are perceived as having human-like mental capacities, their moral decisions become more tolerated—but not equally accepted as those of humans.

There are only a few empirical moral and social psychological papers focusing on medical decision-making, automated or not. Bigman and Gray (2018) found that people were generally averse to machines making moral decisions (but not towards the outcome of those decisions) in different contexts, including medical ones. The authors argued that, in the past, people have denied full moral status to children,

[1] Robot warmth is challenging to manipulate, which probably explains the lack of studies focusing on it. This is also the reason why we focused on competence.

animals and even other races, and the same might be true at the present time for machines. Machine agency and responsibility may be linked to the extent that people perceive the machines as minded entities. The aversion towards machines making decisions may stem from thinking that machines lack a mind, that is, human-like thinking and feelings. However, an increase in the perceived "mindedness" of a hypothetical medical computer seemingly decreased its permissibility as a moral decision-maker (Bigman & Gray, 2018).[2]

Gray and Wegner (2012) argued that the perceived mental qualities of an agent are one contributor to the so-called *Uncanny Valley Effect* (Mori, 1970; Palomäki et al., 2018). That is, at a certain point, an increase in the perceived humanness of a machine leads to a sudden drop in its likability—which also trickles into how its moral decisions are evaluated (Laakasuo et al., 2021). In terms of human moral cognition, Stahl and Coeckelbergh (2016) pointed out that attributing moral agency, responsibility or trust to a human decision-maker is straightforward, but problematic in the case of non-human agents such as AIs.

1.3 | The blameworthiness of robots

People have been found to retroactively hold other people blameworthy in cases where they make decisions with negative consequences, regardless of whether these consequences are accidental, unrelated, or intentional. This phenomenon has become known as *moral luck* (Kneer & Machery, 2019; see, Royzman & Kumar, 2004; Martin & Cushman, 2016 for philosophical reviews). Given that moral luck (and/or outcome bias) has a critical role in the moral judgments of other humans (Baron & Hershey, 1988; Cushman, 2008), investigating how moral judgments are affected by the accidental consequences of actions delivered by robot agents seems both relevant and necessary.

Following considerations regarding the judgment of robot agents based on outcomes, Malle et al. (2019) investigated how human versus AI disobedience is judged when the agents operated under orders from their (human) military superiors. In short, they found that it seems more acceptable for an AI than for a human to disobey an order to perform a lethal action.[3] However, this effect diminishes if both agents are granted freedom from the orders of their superiors. Malle et al. (2019) argued that this command chain might justify, in the participants' minds, the agent's actions. For example, a human is seen as more blameworthy for cancelling a strike than for launching it because self-reliant terminating of the command chain is seen as a moral violation. An AI, seen as less embedded in the command chain, thus receives less penalty for disobeying. However, in Malle et al.'s studies (2019), the chain of command always originated from humans, and they did not investigate how completely automated command chains would be perceived. This question is particularly relevant as implementing AIs in command chains can plausibly take a form where an AI is tasked with giving orders or recommendations, rather than (or in addition to) carrying them out.[4]

1.4 | The current studies

Much of previous psychological research on moral decision-making has focused on pitting deontological against utilitarian moral intuition (Christensen & Gomila, 2012; Gray & Graham, 2018; see Laakasuo & Sundvall, 2016 for a review). Moral psychology commonly investigates condemnation of noxious and innocuous harms (e.g., Uhlmann et al., 2015; Horberg et al., 2009; Schein & Gray, 2016), the role of intentions and accidents in causing harm (Knobe, 2003; Young & Saxe, 2009, 2011; Cushman, 2008; Royzman & Kumar, 2004; Kneer & Machery, 2019), and, more recently, character perception mechanisms (for a review see Chapman, 2018; Gamez et al., 2020; Laakasuo et al., 2021). Thus, while the study of moral cognition is expanding in many directions, personal autonomy in general and empirical medical ethics of AI in particular have received little attention in this context.

Here, we investigate hypothetical medical decisions that threaten human autonomy—specifically, forced medication by a robot. Forced medication decisions have a trade-off between the patient's autonomy and the caretaker's responsibility to heal the patient. We identify little empirical work on experimental social and moral psychology in the context of caretaking, nursing or medical decision-making (Vandemeulebroucke et al., 2018). We report four laboratory and online empirical studies as well as one qualitative anthropological field study. We first present the results of the qualitative study (Study 1) to better contextualize the empirical experiments and to gain insight into robots and AIs in an actual caretaking context.

Across all our experimental studies, we expected people to judge moral decisions by robots more harshly than similar decisions by humans. We also expected that people, on average, would prefer decisions that do not violate patient autonomy, even if those decisions correspond to disobeying a superior's orders. After conducting the anthropological field study (Study 1), which revealed human autonomy and trust as central themes in our participants' worries, we progressed to focus on autonomy violations made by either human or robot nurses (Study 2—Forced medication as an autonomy violation by humans vs. robots). Thereafter we ran three further experimental studies to

[2] Perceptions of an agent being a valid or the correct agent to make a decision are different from perceptions about the moral wrongness of those decisions. Moreover, it is not obvious that a perception of robots as less appropriate agents should lead to more negative judgments about those agents' decisions. By analogy, a child would be an inappropriate agent for many moral decisions, but a child forced to make those decisions and causing harm in the process would be likely to be judged more leniently. The connection between perceived mind and moral judgments is complicated by a potential dynamic between perceptions about mindedness and perceptions about morality. The theory of dyadic morality (Schein & Gray, 2018) predicts that a perceived moral violation will amplify perceptions about the moral agent and moral patient in the situation. That is, perceiving a robot, e.g., hurting a human should increase perceptions of the robot as an intentional (minded) agent, and the hurt human as a feeling (conscious) patient. Thus, one potential reason for people's willingness to morally judge robots despite findings indicating that robots aren't seen as appropriate agents (Bigman & Gray, 2018) is that situating robots in moral situations itself induces perceptions of intentional action.

[3] An alternative explanation could be that people think that computers make better decisions than humans, thus being more accepting of any decision by an AI (Lee et al., 2018).

[4] Judgments about whether an entity should be allowed to make moral decisions are different from judgments about whether that entity's decision was morally good or bad. Importantly, judgments about these decisions are not necessarily more negative, as shown by Malle et al. (2019). People seem willing to morally judge artificial moral agents, and perceiving an *agent* as less capable does not equate to perceiving all of that agent's *decisions* as less appropriate.

explore possible boundary conditions (i.e., robustness checks) to the observed effects.

In this article we thus evaluate whether humans and robots are judged differently in forced medication scenarios involving violations of autonomy, and whether there are boundary conditions for the possible differences. In Study 3 we explored whether the descriptive qualities of the agent, or person perception mechanisms, played a role in how people judged the human or robot nurses' decisions (Study 3—Perception of competence) (Gamez et al., 2020; Bigman & Gray, 2018; Young & Monroe, 2019; Goodwin et al., 2015; Laakasuo et al., 2021; Fiske et al., 2007). Specifically, we manipulated the perceived competence of the agents by describing them as either competent or incompetent workers. Previous research indicates that the way robots are perceived influences how their moral decisions are evaluated, and the dynamics of these effects may be dissimilar between human and robot agents (Gamez et al., 2020; Laakasuo et al., 2021). Thus, Study 3 assesses whether the results of Study 2 results are due to the robot being perceived as less competent.

In Study 4 we focused on "moral luck" and evaluated whether robot nurses' decisions are evaluated differently from human nurses if the patient dies unexpectedly of unrelated causes (Study 4—Moral luck). The phenomenon of moral luck among human decision-makers has been studied extensively (Kneer & Machery, 2019), but to our best knowledge no studies have focused on how it affects evaluations of AI decisions. In Study 4 we therefore seek to rule out moral luck as an explanation of the observed effects. Furthermore, this boundary condition observation allows us observe whether forced medication decisions made by robots are perceived to be potentially more harmful than those made by humans, since accidental harm is the only realistic option in a medical setting without contradicting the four principles of medical ethics.

Finally, Study 5 focused on the issue of "human-in-the-loop" in medical decision-making and whether the results from earlier studies are sensitive to such a boundary condition (Study 5—Command chain). This phenomenon has been studied previously only in the context of AI-based military operations but may influence judgments of AIs across other contexts as well. Malle et al. (2019) observed that moral blame is assigned less to a moral decision-maker if they are non-human and ignore the orders given to them. However, the authors did not evaluate whether the decisions would be judged differently if the whole chain of command were automated, which we implemented as a condition in Study 5.

2 | STUDY 1: ANTHROPOLOGICAL FIELD STUDY: PERCEPTIONS OF NURSING ROBOTS IN THE ELDERLY

Across scientific fields, integrating both quantitative and qualitative data has proven extremely helpful in obtaining a nuanced and in-depth understanding on a specific topic (Creswell et al., 2011). In this study, we used qualitative methodology to gain insights into how people perceive nursing robots in the context of autonomy violations. We focused on elderly people living in residential care homes, because, for them, nursing robots and issues concerning individual autonomy are salient, pressing and timely;[5] and because this allows for our results to be compared with similar studies (Darragh et al., 2017; Mitzner et al., 2018; Prakash et al., 2013; Beer et al., 2017; Stuck & Rogers, 2017)

2.1 | Method

A trained anthropologist contacted several elderly care homes to conduct semi-structured in-depth interviews of the residents, focusing on their thoughts, wishes, and fears concerning nursing robots and AIs in a caretaking context.

The interviews were conducted between October 2017 and June 2018 in nine different elderly residential homes in three major cities in Southern Finland. In total, 30 interviews were conducted (12 males, 18 females; $Age_M = 80$; $Range_{Age} = 69—97$). Informed consent was obtained from all interviewees, as well as from the residential homes. The study was also approved by the University of Helsinki Hospital's Social Services and Health Care division. Most interviews were conducted at the residents' own apartments, but some—depending on the wishes of the residents—were conducted in the residential homes' libraries, meeting rooms, lounges, or lobbies.

The interviews began with the interviewer asking the residents what they thought a *robot* was in their view, and which mental associations the word "robot" brought about. Next, the residents listened to the interviewer reading a vignette, in which a robot nurse forcefully medicated an unwilling patient (see Appendix). This vignette mirrored in style the vignettes used in our experimental studies (details in the subsequent sections, below). The residents were then asked about their thoughts and feelings concerning the story. In contrast to our (quantitative) studies, we could not compare participants' responses across different versions of the story. Instead, the residents only heard the story version involving forceful medication. We reasoned that this version of the story would probably elicit salient reactions, thoughts and opinions. However, the residents were also asked how they would view the situation *if* the nurse in the story would have been human.

After the story, the interviewer allowed the residents to dictate the pace of the interview, following their trains of thought, but nonetheless attempting to maintain the topic on nursing robots—unless the residents felt tired, or just wanted to talk about something else. Thus, the semi-structured nature of the interview had to, on some occasions, be relaxed to make sure the residents felt comfortable throughout the session. Two of the interviews transformed as the residents were reminded of their recent personal losses and wanted to focus more

[5] There is a contentious relationship between quantitative and qualitative approaches stemming from long-standing disputes between Anglo-Saxon and continental philosophical traditions (e.g. Hepburn, 2003; Hughes, 2018; Hacking, 1999; Berger & Luckman, 1967; Burr, 2003). Essentially, this debate concerns fundamental epistemic differences in how natural phenomena, including human behaviour, are measured: either the phenomena can be directly observed, measured, and interpreted; or the phenomena depend largely on subjective experience, thus requiring in-depth self-reflection to shed light on them. In practice, however, qualitative methods have been used to supplement quantitative research very successfully, and several comprehensive papers and books have been published giving details on how "mixed methodology" should be employed (e.g., Creswell et al., 2011; Tashakkori & Teddlie, 1998).

on the morality of caretaking institutions and their own autonomy, shrugging off the subject of robots altogether. These informal sections of the conversations also provided context for the thoughts and feelings of the residents, as they focused not only on their own personal backgrounds and past, but also on the manner and quality of living in residential homes. At the end, the residents were asked to answer questions on Likert-type questionnaires concerning their attitudes towards robots, as well as moral attitudes in general. The interviewer read out the statements in the questionnaires, asking the residents to rate them (usually from 1 to 7, with 1 being "Strongly disagree" and 7 being "Strongly agree"), and marked down their answers.

The overarching goal in the interviews was to gain an in-depth understanding on how nursing robots, and AIs in general, are perceived in care-taking roles. All interviews were recorded and transcribed verbatim, resulting in 57,000 words of text, averaging around 1900 words per interviewee. The data were analysed using the Atlas.ti software by counting the occurrences of words describing emotions or emotional events, as well as marking down occurrences where the residents elaborated on thoughts and opinions regarding the vignette, or the use of robots in a nursing context in general. However, only the results regarding the residents' responses with respect to the vignette are presented here.

2.2 | Results

The most common responses to the story were aversion, fear, and unpleasantness. These reactions were primarily focused on the patient having lost their autonomy. The reactions were slightly less pronounced for a few residents who required the most care and "supposed they would get used to anything", having already lost some of their self-perceived autonomy. In contrast, those who were still capable of handling most of their daily activities themselves felt the threat more keenly. In fact, many residents focused on the prospect of losing autonomy so intensely that they forgot there was a robot involved, that is, whether or not the patient in the story was treated by a robot or a human nurse was not as big an issue as the patient had lost their autonomy. Interestingly, when asked how the residents would feel if the forced medication was done by a human nurse instead of a robot, responses were divided. Some residents felt that forceful medication would be even more unnerving if done by a human: a human would *consciously know* what they were doing, making their behaviour a deliberate violation of the patient's autonomy. However, many residents also said that if a human nurse forcefully medicated a patient, then at least that nurse could be asked for their reasoning or empathy, as illustrated by the following excerpts:

> "If the decision is made for my benefit, then of course, I accept this nurse and their doing, but if I notice that the robot only thinks about itself, then I get on edge […], since I don't know if the doctor has actually ordered it or something." (Female, 91 years)

> "A human nurse could at least themselves clarify from the doctor, who gave them the order, asking what they meant by that order. So it would in that way be easier." (Female, 73 years)

When asked who in the story would or should be held responsible for the forceful medication, two thirds of the residents felt that the supervising doctor, or the one who programmed the robot, were responsible. When asked to consider how they would feel *if* the nurse had been a human, about half of the residents attributed responsibility to the nurse. Nonetheless, the supervising doctor was seen as highly responsible, regardless of whether the nurse was a robot or human. Many residents were uncertain about who, in the end, should be responsible, but were also adamant that the robot itself was *not* responsible. Perhaps what makes the robot nurse so unnerving and threatening is its inability to be responsible for its actions. If the robot is not in any way responsible for its behaviour, how could one trust it or negotiate with it?

Recent models of moral responsibility have in fact highlighted its "conversational" nature; expression of emotions and negotiations of daily practices are like a living dialogue concerned with our sociomoral surroundings (McKenna, 2012). This idea of "conversational" moral responsibility presents an idea similar to the Social Intuitionist Model (Haidt, 2001), in that both stress the importance of social processes in the formation of social conventions and norms. Indeed, the residents saw the robot nurse as a "cold machine" that could not be negotiated with or relied upon; and there would be no certainty whether the forceful medication was done for the good of the patient or not. Even if the ultimate responsibility lies with the supervising doctor, the human nurse could still ask for clarification into why forceful medication was necessary, and then explain that to the patient. This idea of having a human potentially explain what was happening made the situation less frightening. The residents felt that in the case of a robot, the patient cannot question its orders, nor see any human being behind them. The robots were felt as "cold" and lacking empathy, incapable of negotiation, distant, and untrustworthy:

> "Well, all in all, it feels very cold, if a robot gives the medication […], of course, it would be nicer if it were human. But then again the robot could carry out some tasks, why not, but I don't … I don't think it should've forced the medication. Because it wasn't necessary." (Female, 83 years)

> "I would not fully trust that robot, I think it should be a person, whether a nurse or a doctor or whoever was behind it, before I would start living by what the robot instructs." (Female, 91 years)

> "Well no, it's funny that if you place a robot and yourself next to each other, I will pick you over the

robot. Because of this face-to-face talking and trusting another person, who's a mammal like me, I do trust them more anyway." (Male, 85 years)

2.3 | Discussion

Our qualitative results gave us insights into how AIs are perceived in a nursing context. Having heard the story, the residents prominently raised concerns about personal autonomy and trust and felt that nursing robots were generally cold and unempathetic; from here onward we will focus on autonomy. This is in line with results showing that people are generally aversive to robots as decision-makers (Bigman & Gray, 2018). Our sample consisted of the elderly, that is, some of the most likely people to be affected by similar issues in nursing or the automation of nursing. It is quite possible that younger interviewees would have had different concerns about robots in this context and would have been generally more open to automated decision-makers.

These are known limitations in qualitative studies focusing on a single sample. We could have chosen to collect our interviews from the general (non-elderly) population, but these interviews would have lacked an important, personal and experiential aspect. Focusing on younger generations would mean asking about the moral concerns of people who have no personal reason to be concerned. On the other hand, qualitative interviews were necessary to study an elderly population, because it would not have been feasible, for both practical and ethical reasons, to have older people fill in arduous and taxing questionnaire forms, or to collect data from hundreds of elderly individuals. However, a large number of similar studies have focused on the residents of care-taking facilities, and our qualitative results are comparable with them (Darragh et al., 2017; Mitzner et al., 2018; Prakash et al., 2013; Beer et al., 2017; Stuck & Rogers, 2017).

3 | STUDY 2: FORCED MEDICATION AS AN AUTONOMY VIOLATION

We conducted Study 2 to test our materials and methods, and to construct and validate our dependent variables. The vignette used in this study was similar to the one used in Study 1, but with a few variations in order to create different experimental conditions. The vignette described a hypothetical situation where an unwilling patient refuses to take their medicine. A nurse, who is either human or robot, either follows or disobeys orders from the supervising doctor to medicate the patient (2 × 2 between-subjects design). The nurse's decision to disobey the supervising doctor and *not* forcefully medicate the patient is a decision respecting the patient's autonomy, whereas the decision to follow the supervising doctor's orders disregards the patient's will.

Our main dependent variable was a psychometric scale focusing on moral evaluation of the nurse's (either human or robot) actions and other morally relevant qualities (for similar psychometric instruments, see Awad et al., 2018; Koverola et al., 2020; Laakasuo et al., 2018; Choma et al., 2012; Allison & Bussey, 2016: Bigman & Gray, 2018).

3.1 | Method

3.1.1 | Participants and design

In total, 135 (N = 135; 56 female) participants (Age_M = 37.10; SD = 17.65; Range = 18–80) were non-intrusively (details below) recruited from a large public library in Espoo (second-largest city in Southern Finland).[6] They were informed that they could participate in a psychological experiment which would take approximately 30 min of their time. Of the participants, 60 had at least a Bachelor's degree.

After being recruited, participants were escorted into our 'pop-up' laboratory situated in a private room at the library. The participants sat in front of a laptop computer insulated with office walls separating them from other participants and were automatically randomized into one of four conditions in a 2 [forced medication: yes vs. no] × 2 [nurse: human vs. robot] factorial design (the experimenters were blind to the randomization).

3.1.2 | Procedure and materials

We collected the data at a large public library in the capital area of Finland. We recruited our participants non-invasively by having a table in the foyer with a sign stating: 'Participate in Psychological Research'. Our research assistants sat behind the table dressed in neutral clothing. All recruited participants approached our research assistants voluntarily. After ensuring the participants were legal adults (over 18 years-old), we gave them informed consent forms informing them about the study and highlighting their right to opt out at any point. After signing the consent, the participants were escorted into our laboratory.

The laboratory had four notebook computers with 15″ screens positioned to guarantee maximum privacy. We used office walls to separate the space into cubicle-like nooks. Participants were instructed to use headphones playing pink noise to cover up any background noise. The pink noise volume was held constant at a pleasant level. The experiment was programmed using the Social Psychology Questionnaire library, which is an inhouse software coded in Python and built on top of Pygame version 1.96.

[6] We estimated that, for a mean difference of about 0.5 compared to the grand mean of the sample for Robot Nurse Forced Medication Decision on a 7-point Likert scale, with equivalent variances (1.3 based on previous experiences) assumed between cells, a sample size of 130 would give us approximately 70% power in a planned contrast analysis of 4 cells (Cohens f of about .15). According to Douglas Wahlsten (1991), this is "Case 8" for the shape of linear contrasts. In 2015 when the studies were designed, the general culture of power analysis still had not solidified so we also used APA 2012 recommendations at that time of having at least minimum of 30 participants/cell (VanVoorhis & Morgn, 2007). Our Study 2 was a pilot study run in a physical lab that had financial constraints associated with it, so we could not afford a larger sample size. All four estimates were on the conservative side and our actual observed effect size for the B-value of interest was actually 1.2 (roughly Cohen's f of 0.4).

The experiment itself started by randomizing the participants into one of the four conditions listed previously. Both the experimenters and participants were blind to the randomization. Participants started the experiment by filling in exploratory measures and then continued to the actual task and the dependent variables (we also collected measures on trust and responsibility allocation, results of which will be reported elsewhere). There was only one experiment in which the participants participated.

3.1.3 | Vignette/experimental task

In the experimental task, the participants read a short science fiction story (see Appendix) describing an event taking place in the year 2035, where either a human nurse or an advanced nursing robot (mental capacities were not described for either) encounters a patient who refuses to take their medication. The supervising doctor of the ward has instructed the human/robot nurse to make sure that the patient takes their medication, since otherwise the patient might be in danger. However, based on the experience/information that the human/robot nurse has, the medication is not absolutely necessary[7] in this patient's case. Then, the human/robot nurse decides to either forcefully medicate the patient or to respect their will and leave them unmedicated. This is where the story ended; the potential events taking place after the decision were not described (this was done in Study 4). Once the participants had started reading the story and 1 min had passed, the dependent variables appeared on the screen below the story one by one, and participants provided their answers with a mouse. After responding to the dependent variables, participants answered manipulation check questions and questions relating to "mind perception" (see Gray et al., 2012; not reported here).

3.1.4 | Moral evaluation measure/ main dependent variable

Our main dependent variable consisted of 10 items (listed in the Appendix). We first ran a dimensionality analysis on our items, which suggested only one factor (eigenvalue 4.7), based on Kaiser Criterion (eigenvalues > 1.1). Next, we ran an exploratory Maximum Likelihood Factor Analysis with two factors, which resulted in all items loading more strongly on the first factor (all loadings > | 0.57 |), than on the second factor (all loadings < |.48|). We thus concluded that the 10 items load on a single factor.

The final version of the scale had good internal consistency (Cronbach's alpha = 0.89). All questions were anchored from 1 to 7 ("Completely disagree"—"Completely agree"). The scale had such items as "The nurse's/ nursing robot's actions were X" (examples of X: necessary, morally right, insensitive, inhumane). Higher scores indicate more positive evaluation of the nurse's decision and/or action.

[7] Our aim was to convey that using the medication is beneficial, but the patient could heal on their own as well.

3.2 | General note on data analysis

Table 1 reports the inferential statistics on all main effects and interactions across the empirical Studies 2–5, while Figure 1 depicts the simple effect comparisons of human versus robot (horizontal grey lines and asterisks) within the two forced medication conditions (no forced medication/forced medication) and within the study specific boundary conditions (Study 3: high competence/low competence; Study 4: patient lives/patient dies; Study 5: supervisor is human/supervisor is AI). Figure 1 also depicts the effect of forced medication (horizontal red lines and asterisks), which corresponds to a main effect in Study 2, and a simple effect within the boundary conditions in Studies 3–5. In addition, the study-specific results sections, which follow below, report results from additional simple effect comparisons, where appropriate, and finally provide a synthesis of the most important recurring pattern of results.

3.3 | Results

In Study 1, we ran a full factorial two-way ANOVA by including both factors (Decision-maker type: robot/human; and Decision: no forced medication/forced medication); and their interaction into the model, with the moral evaluation measure as the DV. Both main effects and the interaction effect were significant, suggesting that the robot nurse's decisions were less approved of than the human nurse's, and that forced medication was less approved of than disobeying the order to medicate (see Table 1 for full inferential statistics). However, probing the interaction between Decision-maker type and Decision revealed that the robot's decision was less approved *only* in the forced medication condition (simple effect comparison B = 0.97; $F(1, 131) = 10.42$, $p = .001$). Conversely, the robot and human nurses were judged equally if they made a decision that respected the patient's autonomy (Figure 1, Study 2).

3.4 | Discussion

Study 2 confirmed that our materials and design were properly functional for further use. The results suggest that the nurse's (whether human or robot) decision to violate a patient's autonomy by forcefully medicating them is judged morally less acceptable than disregarding orders to do so.[8] However, disregarding the patient's will and forcefully medicating them was a decision clearly made by a robot nurse. These findings complement previous research by Bigman and Gray (2018), who found that people are generally aversive towards robots as *decision-makers*. In our current study, however, people were clearly more averse to a robot's *decision* only when it violated the patient's autonomy and obeyed the doctor's instructions. This is a novel

[8] Alternatively one could interpret the results as people having greater tolerance for robot disobedience; however, given that the robot obedience decision is always the lowest bar, we have taken this approach for simplicity.

TABLE 1 Inferential ANOVA statistics for all studies: 10 item DV

	Test statistics		
Independent variables	*F*	*p*	par. η^2
Study 2 (Lab study 2×2; $N = 135$; $R^2 = 0.12$)			
Decision-maker type (DMT; Robot vs. Human)	5.22	.02*	0.04
Decision (Forced medication: Yes/No)	9.51	.002**	0.07
DMT × Decision	4.93	.02*	0.03
Study 3 Competence perception (Prolific 2×2×2; $N = 981$; $R^2 = 0.28$)			
DMT	18.97	<.001***	0.03
Decision	43.44	<.001***	0.03
Competence (High vs. Low)	88.19	<.001***	0.20
DMT × Decision	22.90	<.001***	0.02
DMT × Competence	1.33	.24	0.004
Decision × Competence	24.95	<.001***	0.02
DMT × Decision × Competence	3.70	.05	0.004
Study 4 Moral luck (Lab study 2×2×2; $N = 267$; $R^2 = .20$)			
DMT	2.39	.12*	0.01
Decision	28.27	<.001***	0.10
Outcome (Patient lives vs. Patient dies)	20.53	<.001***	0.07
DMT × Decision	9.14	.002**	0.03
DMT × Outcome	0.67	.414	0.00
Decision × Outcome	0.19	.496	0.00
DMT × Decision × Outcome	0.57	.187	0.00
Study 5 Command chain (Prolific academic; 2×2×2; $N = 500$; $R^2 = 0.22$)			
DMT	2.46	.11	0.00
Decision	103.57	<.001***	0.17
Senior Physician (Human vs. AI)	1.11	.32	0.00
DMT × Decision	30.95	<.001***	0.06
DMT × Senior Physician	0.68	.61	0.00
Decision × Senior Physician	4.38	.03*	0.01
DMT × Decision × Senior Physician	0.06	.80	0.00

finding, which we sought to replicate in our subsequent studies. Studies 3–5 thus further explore whether the finding is robust against various boundary conditions to rule out other potential explanations.

4 | STUDY 3: PERCEPTIONS OF COMPETENCE

In Study 3, we focused on perceived competence (low or high) of the robot or human nurse. Recent studies suggest that character perception mechanisms are crucial for human moral cognition (e.g., Chapman, 2018; Laakasuo et al., 2021; Brambilla et al., 2021): people observe actions, including rule obedience, and their consequences in relation to *who* performs those actions. In other words, people give more moral credence to those they think deserve it (e.g., Miller, 2007). For example, using sharp objects to puncture a person's skin is acceptable for experienced (competent) surgeons operating on a patient, but seldom for (incompetent) others. In Study 3, we manipulated the perceived competence characteristics of the nurse or the nursing robot by describing them either as well performing and liked or someone who makes mistakes and is considered not up to the task. Study 3 thus sought to rule out the possibility that the findings are due to perceived incompetence of the robot. We also sought to replicate the main findings of Study 2 using a different population and an online setting instead of a laboratory setting. This study was preregistered: https://osf.io/k6b9f/.

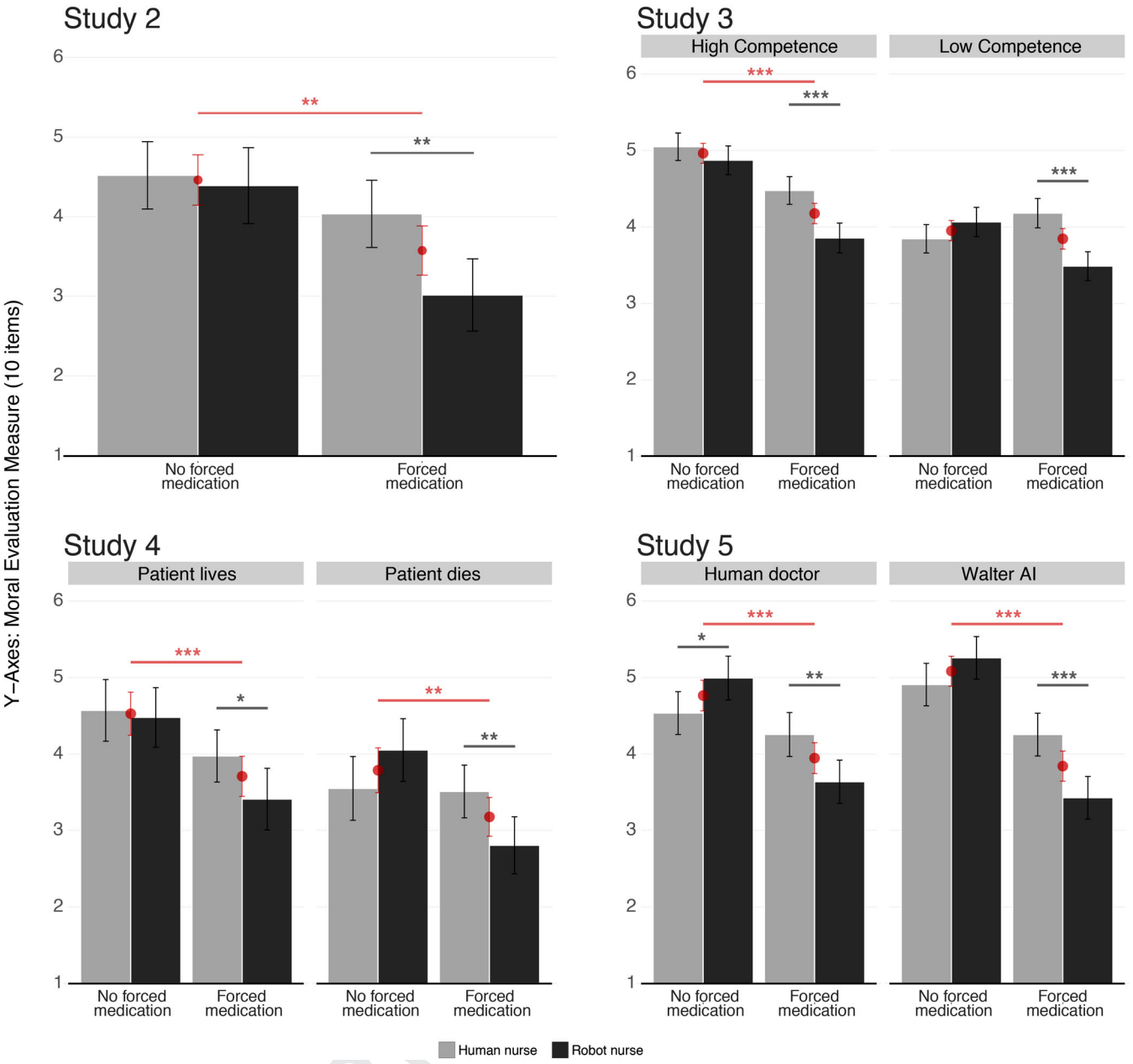

FIGURE 1 The effects of all study-specific between-subjects manipulations on moral evaluations (higher scores [min 1–7 max] indicate a more positive moral evaluation of a decision). Study 3: High Competence and Low Competence refer to the description of the human or robot nurse in terms of their competence as nurses; Study 4: Patient lives and Patient dies refer to the outcome of the nurse's decision, whereby the patient either lived or died; Study 5: Human doctor and Walter AI refer to the supervising doctor (Walter AI is an advanced artificial intelligence) who told the human or robot nurse to forcefully medicate their patient. Error bars are 95% confidence intervals. The red points are the means of the levels of Forceful medication (Did not force medication vs. Forced medication). The following pairwise comparisons are labelled: (i) human versus robot when deciding to not force medication, (ii) human versus robot when deciding to force medication, and (iii) did not force medication versus forced medication averaged across both nurses, ***$p < .001$, **$p < .01$, *$p < .05$.

4.1 | Method

In total, 1200 (N = 1200; 763 female) participants (Age_M = 41.83; SD = 13.62; Range = 18–89) were recruited from the Prolific Academic online survey site (www.prolific.ac.uk) and were informed they could participate in a psychological experiment which would take approximately 8 min. Only tabletop computer users were allowed to participate. Participants were excluded from the data, per the preregistration, for failing attention or comprehension checks (the latter presented both immediately after the vignette and at the end of the experiment), and for indicating worse than native-level English fluency. After exclusions, 981 participants remained. Of the participants, 497

had at least a Bachelor's degree. Participants received £0.80 as compensation. The survey software Qualtrics XM randomized participants into conditions.

Our design was a three-way between-subject factorial. The first two factors were the same as in Study 2: type of the decision-maker [human vs. robot] and decision [forced medication vs. no forced medication]. The third factor also had two levels: the nurse in the story was described as either performing competently and receiving praise from colleagues, or as performing incompetently and having their colleagues doubt their abilities. Thus, the factorial design was 2 × 2 × 2: [forced medication: yes vs. no] × [nurse: human vs. robot] × [nurse competence: competent vs. incompetent].

4.2 | Procedure

Participants entered the experiment through a link in at the Prolific Academic recruitment site, and first gave informed consent and were randomized evenly into one of the eight experimental conditions. The randomization was automatic and both experimenters and the participants were blind to the randomization process. The dependent variables were shown below the vignette so that participants could refer back to the story while responding.

4.3 | Materials

4.3.1 | Vignette and positive/ negative character perception manipulation

The vignette used in Study 3 was the same as in Study 2, with the addition of a single sentence describing the human or robot nurse as having low versus high competence:

> *Positive (high competence) description:* "[Lena/Lena-Med] has performed well in [her/its] work recently and performs tasks competently, with great precision. [Lena's/Lena-Med's] colleagues praise [her/it] for [her/its] abilities." *Negative (low competence) description*: "[Lena/Lena-med] has been making constant mistakes in [her/its] work recently and performs tasks incompetently, with little precision. [Lena's/Lena-Med's] colleagues think that [her/its] abilities are not up for the job."

For brevity, we refer to these manipulations as "competent" and "incompetent".

4.3.2 | Moral evaluation measure of the nurse/ main dependent variable

This measure was the same as in Study 2 (Cronbach's alpha = 0.90).

4.4 | Results

We ran a full factorial three-way ANOVA by including all categorical factors and their interactions into the model, with the 10-item moral evaluation measure as the DV. All the main effects and interaction effects were statistically significant, except for the interaction term between the Decision-maker type (human/robot) and the Competence (competent/incompetent) condition (see Table 1 for details).

The results by and large replicated the pattern observed in Study 2. That is, the robot's decision was judged more harshly *only* when it engaged in forced medication compared to the human nurse (high competence condition: B = 0.58, $F(1, 973) = 17.41$, $p < .001$; low competence condition: B = .68, $F(1, 973) = 23.58$, $p < .001$; see Figure 1, Study 3). The significant main effect of Competence suggests that moral evaluations tracked competence, with the decisions of competent agents receiving more positive evaluations than those of incompetent agents, especially when the agents decided not to forcefully medicate their patient (as shown by the significant interaction between Competence and Decision; see Table 1 statistics).

Three noteworthy details emerged from further simple effects analyses. First, there was a reversal of the relative approval pattern for humans as a function of competence. That is, an incompetent human nurse's decision was more approved of when they obeyed, rather than disobeyed, commands (B = 0.35, $F(1,973) = 6.32$, $p = .012$), whereas a competent human nurse's decision was more approved of when they disobeyed commands. Second, participants demonstrated more leniency towards a competent robot that disobeyed than a competent human who obeyed orders. Third, participants also demonstrated more leniency towards and an incompetent robot that disobeyed than an incompetent human who disobeyed orders.

4.5 | Discussion

Earlier in Study 2, we found that a robot nurse's decisions were judged more harshly than a human nurse's decisions only when comparing judgments of the decision to forcefully medicate a patient. In addition to observing a clear main effect of competence (less approval of actions in the incompetence condition) in Study 3, we replicated the main result of Study 2—lowest approval for robot's decision to forcefully medicate—for both competence conditions. However, the robot nurse's decision to disregard the patient's autonomy was most negatively judged when the robot was also perceived to be incompetent.

Furthermore, we found an interesting effect whereby participants judged an incompetent human nurse's decision to obey orders to forcefully medicate more positively than a decision to disobey the order, whereas a competent human nurse's decisions were judged in the opposite way. It is unclear why this kind of reversal did not happen for judgments of the robot nurse's decisions. However, the results do imply that the findings of Study 2 cannot be explained by attributing incompetence to the robot.

We ran two iterations of Study 3, with slight variations in the manipulation (see Appendix B for details on Study 3B). In Study 3B, the distinction between the human and robot nurse was more pronounced and focused on likability rather than competence. Here, the pattern of means was similar to Study 3, but there was no significant main effect of the likability manipulation. However, the reversal of preferences that we observed in Study 3 is also observed in Study 3B: there is a preference for non-likable humans to follow orders, and for likable humans to disobey. As in Study 3, this effect did not replicate for judgments about the robot nurse's decisions (see Appendix B).

Taken together, Studies 3 and 3B are in line with Gamez et al. (2020), who recently found that likability or character perception manipulations in general do not work with AIs as well as they do with humans (and require more complex explanations). Our results also suggest that this is a fruitful avenue for further research and that manipulations related to competence, warmth (Fiske et al., 2007) and potentially morality (Brambilla et al., 2021) require further study in the context of the moral psychology of robots.

To summarize, our results suggest that character perception effects—competence or likability—matter for moral evaluations of decisions made by humans but not for evaluations of decisions by robots. This is in line with our qualitative study where the elderly clearly preferred to interact with a "warm" human over a "cold" robot. The finding that humans have different ethical concerns associated with "incompetent" technologies (i.e., people find poorly functioning technology more morally acceptable than incompetent humans) suggests a potential ethical bias in our moral cognition deserving further research. In general, the results of Study 2 appear to be robust against perceiving the robot nurse as incompetent.

5 | STUDY 4: MORAL LUCK

In Study 4, we focused on the moral luck phenomenon. We added a minimal change to our original vignettes, where the patient in the vignette was found either dead or alive the next day. We did not specify or hint at any causal links between the human or the robot nurse's actions and what happened to the patient. We deliberately did not focus on intentional malice, since this would defeat the purpose of the vignette, where we intentionally adopted the four principles of care ethics (respecting patient autonomy, non-maleficence, active beneficence, and justice). We specifically focused on *accidental harm* (see Martin & Cushman, 2016), which is the most likely and realistic type of harm in medical situations. We also did not give any information on the possible motivations or intentions for the agents in vignettes for acting as they did, since this topic of moral causality is highly volatile and mired in controversy (Cushman & Greene, 2011; Kneer & Machery, 2019; Lombrozo, 2010). While moral luck has previously been studied in the context of human decision-making (e.g., Kneer & Machery, 2019), it is unknown whether it also affects evaluations of AI decisions. Study 4 thus sought to explore moral luck as a potential boundary condition, and to see if the previously observed effects are robust whether or not the patient being treated happens to die. Furthermore, this boundary condition observation allows us to observe whether forced medication decisions made by robots are perceived to be potentially more harmful in their consequences than those made by humans.

5.1 | Methods

5.1.1 | Participants and design

In total, 268 (N = 268; 150 female) participants (Age_M = 32.48; SD = 13.36; Range = 18–76) were recruited from a large public library in the centre of Helsinki and were informed that they could participate in a psychological experiment which would take approximately 30 min. Of the participants, 148 had at least a Bachelor's degree.[9]

After being non-intrusively recruited, participants were escorted into our 'pop-up' laboratory in the library. Participants sat in front of a laptop computer insulated with office walls from other participants and were automatically randomized into one of eight experimental conditions in a 2 [forced medication: yes vs. no] × 2 [nurse: human vs. robot] × 2 [decision consequence: patient lives vs. patient dies] factorial design (the experimenters were blind to the randomization).

5.1.2 | Procedure and perceived decision consequence manipulation

Our procedure was the same as in Study 2 (Pilot). The decision consequence manipulation was added as a single sentence at the end of the vignette. The vignette ended either with a sentence stating "[o]n the next day, the patient dies as a result of a sudden bout of sickness" (patient dies), or "[o]n the next day, the patient's status and health remain the same as before" (patient lives). The vignette did not explicitly mention any causal relations between the actions of the human or robot nurse and the outcome.

5.1.3 | Materials

The materials were the same as they were in Study 2.

5.1.4 | Moral evaluation measure/ main dependent variable

This variable was the same as in Study 2 (Cronbach's alpha = 0.89).

[9] Given that the observed B-value effect size in Study 2 (Study 4 was run shortly after) was about 1.2, at the time of designing our experiment we used G*Power and VanVoorhis & Morgan, 2007 and Wahlsten's (1991) formulas to calculate the sample size to two different intended contrasts (the same observed previously in Study 2 for Patient lives outcome and the same contrast for the Patient dies outcome). Back then we calculated that a medium-small effect size of Cohen's f = 0.23 (about 5% of variance) with 90% power for ANOVA with 3 predictors would be achieved with a sample size of about 260 participants.

5.2 | Results

We ran a full-factorial three-way ANOVA by including all experimental factors and their interactions in the model, with the moral evaluation measure as the DV. The main effects of Forced medication and Decision consequence were statistically significant, as was the interaction between Decision-maker type and Forced medication (see Table 1 for statistics). As can be seen in Figure 1 (Study 4), and as the main effect of Decision consequence suggests, participants approved decisions less if the patient died afterwards. No other terms were statistically significant. In simple effect contrast analyses, we again found that the robot nurse deciding to forcefully medicate the patient was judged more negatively than the human nurse making the same decision, regardless of whether the patient lived (B = 0.56, $F(1,259) = 4.42$, $p = .03$) or died (B = 0.70, $F(1,259) = 7.44$, $p = .006$).

5.3 | Discussion

We again replicated our previous results (Figure 1, compare Study 2 with Study 4, left side): the robot nurse was judged more negatively than the human nurse only when the nurse's decision violated the patient's autonomy. We also observed a moral luck effect: the patient dying afterwards resulted in significantly harsher moral evaluations of any decision preceding the patient's death. This moral luck effect was observed in forced medication conditions regardless of whether the nurse was a human (B = -0.46, $F(1,259) = 3.55$, $p = .06$) or a robot (B = -0.60, $F(1,259) = 4.69$, $p = .03$). If the patient died there was no difference on how a human nurse's decisions were judged. However, the decision to violate patient autonomy was most condemned for robot nurses whether the patient lived or died (see Figure 1, Study 4).

6 | STUDY 5: COMMAND CHAIN EFFECTS AND DISOBEDIENCE

In Study 5, we included another manipulation: the status of the supervising party who ultimately gives the order to forcefully medicate the patient as either a human doctor or an advanced AI doctor. As suggested by Malle et al. (2019), whether disobedience of orders is acceptable depends—at least in part—on the command structure of the decision-making system. In this study we explored what happens when there is no human "in the loop"—is it more acceptable to disobey orders given by an AI or a human? The previous effect of disapproval against robots making autonomy violations towards humans could either be enhanced or diminished if the whole command chain is mechanized. If a fully mechanized command chain further reduced moral acceptance of autonomy violations of patients, it could partially explain the results of Study 2. Study 5 is an important robustness check where the properties of the main source authority are manipulated. If robots engaging in autonomy violations while receiving orders from humans are disapproved of, but this effect disappears when they are supervised by another AI, the results would warrant deeper analysis.

6.1 | Method

6.1.1 | Participants and design

In total, 500 ($N = 500$; 230 female) participants ($Age_M = 29.33$; SD = 10.63; Range = 18–82) were recruited from the Prolific Academic participant pool (www.prolific.ac.uk). No participants were excluded from the analyses, since the data had excellent quality without any outliers. Of the participants, 297 had at least a Bachelor's degree. Participants were pre-screened for fluent English skills, and only tabletop computer users were allowed to participate. The study took approximately 12 min, and participants received £1.12 as compensation. The survey software Qualtrics XM randomized participants into one of eight conditions in a 2 [forced medication: yes vs. no] × 2 [decision-maker type: human vs. robot] × 2 [supervising doctor: human vs. AI] factorial design (the experimenters were blind to the randomization).

6.2 | The study was pre-registered at OSF: https://osf.io/8ycvg/view_only = 1502a96a5dba45fab8a3ba78736df7fc

6.2.1 | Moral evaluation measure of the nurse/ main dependent variable

This measure was the same as in Study 4.

6.3 | Results

We ran a full factorial three-way ANOVA by including all our experimental factors and their interactions into the model, with the moral evaluation measure as the DV. The main effect of Forced medication, the interaction between Forced medication and Decision-maker type, and the interaction between Forced medication and Supervising doctor were statistically significant (see Table 1 for statistics). No other terms were statistically significant.

The interaction between Forced medication and Decision-maker type replicated the pattern from Studies 2–4 (see Figure 1, Study 5). As in the previous studies, also in Study 5 the robot nurse's decision to forcefully medicate the patient was judged more negatively than the same decision by the human nurse (B = -1.44; $F(1,492) = 25,34$, $p < .001$), but there was no significant difference in judgments about the decision to not forcefully medicate the patient.

We further looked at how a robot nurse disobeying an AI doctor contrasts with the other three cases where the nurse disobeyed orders

(i.e., robot nurse and AI doctor vs. robot nurse and human doctor + human nurse and AI doctor + human nurse and human doctor). The results showed that a robot nurse deciding to (respect the patients will and) disobey the AI doctor was judged more positively than the other three cases of disobedience: $F(1, 492) = 7.41$, $p = .006$; B = 0.66, 95% CI: [0.18, 1.15]). This indicates that disobedience towards an AI doctor's orders is most approved when the disobeying nurse is also a robot (i.e., an AI).

6.4 | Discussion

In Study 5, we successfully replicated most of the previous findings across our studies. Again, we found that the robot nurse's decisions were significantly less acceptable than the human nurse's decisions only when the decision violated the patient's autonomy. We also found that both the human and robot nurse's disobedience towards an AI doctor (i.e., disregarding orders and respecting the patient's autonomy) was met with approval. In general, the central finding of Study 2 was robust against the level of mechanization of the command chain. The highest approval was observed in the condition where a robot respects the patient's will against a supervising AIs request. These findings are in line with Malle et al. (2019), who found that a military robot disobeying an order to launch a missile strike was considered less blameworthy than a human military pilot disobeying the same order. These results suggest that, in our context, decisions made by robots are more morally condemnable than decisions made by humans only when they are in violation of personal autonomy. Notably, this means that in terms of condemning the decisions of artificial agents, the type of decision can matter. This is not necessarily obvious, as for example a series of experiments by Bigman and Gray (2018) suggested that mind perception was the main factor behind aversion to artificial agents as decision-makers regardless of other properties. However, they did not examine different *decisions* in cases of moral conflict.

It is also important to note the slightly different questions asked by Study 5 and the studies cited here: Study 5 examined a composite score of moral condemnation; Bigman and Gray (2018) examined judgments about humans' and AIs' appropriateness as moral decision-makers; and Malle et al. (2019) examined the blameworthiness of different decisions by either humans or AIs. It may be that artificial agents are not generally considered as appropriate agents to make moral decisions, but if they do make those decisions, there are meaningful differences between the acceptability of those decisions. Further, these differences between judgments of different decisions may not be identical when comparing to judgments about human agents. Notably, the level of mechanization (AI + Nurse Robot) did not have an accumulating effect of diminished moral approval compared to just having a robot in the vignette. Thus, it seems that pure mechanization of the decision-making process cannot be the sole explanation of the results observed in Study 2

7 | SYNTHESIS OF RESULTS ACROSS STUDIES 2–5

Across all empirical Studies (2-5) and conditions, we consistently observed that the decision to forcefully medicate an unwilling patient was less approved for robot nurses than human nurses (Figure 1, left-hand side of all panels and facets). However, when the evaluated decision was *not* to forcefully medicate the patient (i.e., to disobey a supervisor's orders), no similar pattern of results emerged—instead, the results differed based on the boundary condition introduced (Figure 1, right-hand side of all panels and facets). Competent nurses (both robot and human) were generally judged more positively than incompetent ones, and if the patient happened to die afterwards, the nurses' decisions were generally judged more negatively. Finally, forgoing forced medication (i.e., respecting the patient's will) was more approved when the supervising doctor was an AI than when they were a human.

8 | GENERAL DISCUSSION

In five studies, including a qualitative anthropological investigation and four quantitative experiments (and an additional study reported in the appendix), we successfully and extensively examined how people feel about forceful medication carried out by either human or robot nurses. Our quantitative experiments focused on assessing moral judgments towards human and robot nurses. In addition to this, themes of reliability/trust and the attribution of responsibility surfaced spontaneously during our qualitative interviews conducted in elderly care facilities around the Helsinki region; these will be addressed in a future manuscript and so are not discussed further here.

Our main finding was that forcefully medicating a patient against their will, thus violating their autonomy, was judged more morally negatively when carried out by a robot nurse as opposed to a human nurse (or, alternatively, robot decisions were most tolerated when they respected patient autonomy). In contrast, not violating a patient's autonomy by forgoing their medication against the doctor's orders was in general almost equally approved for both robot and human nurses (although there were exceptions, which will be discussed below). Also, our anthropological field study revealed similar themes in that elderly people in residential care homes perceived negatively a robot nurse's decision (to forcefully medicate a patient), partly because there was no perceived room for negotiation or empathy. In Studies 3–5 we found that the central finding of Study 2 is not due to (a) perceiving the robots as incompetent (Study 3); (b) the moral luck effect or perceiving robot decisions as more harmful (Study 4); or (c) command chain characteristics (AI vs. human superior; i.e., level of mechanization).

In Study 3 (conducted online), we found that the perceived competence of the nurse was relevant overall, but especially when judging the actions of a human nurse when they decided to forcefully medicate the patient (and violate their autonomy). More specifically, an incompetent human nurse was expected to follow orders rather than ignore them

despite that being a violation of patient autonomy, while the competent human nurse was expected to respect patient autonomy.

The above findings contribute to ongoing discussions regarding character perception (e.g., Gray & Graham, 2018; Gamez et al., 2020) and the psychological discussion on virtue ethics (see Miller, 2007), as well as to social psychological research investigating the effects of warmth and competence on person perception (e.g., Fiske et al., 2007). It is likely that a human nurse who is perceived less competent (*or* less likable—see Study 3B in Appendix), is also perceived as less reliable and trustworthy in their decision-making and thus obeying the orders of a senior staff member is preferred. The fact that we did not find this for robot nurses (in two studies) supports previous research by Gamez et al. (2020), who found that character manipulations were less relevant for judging robot decisions than they were for judging human decisions. In other words, the findings are not explained away by assuming that people perceive robots as less competent than humans.

In Study 4 (conducted in the lab), we found that both human and robot nurses were more harshly judged—regardless of their decision—if the patient was found dead the next morning, compared to the patient remaining alive (i.e., the moral luck effect). This may be explained by moral causality perception mechanisms. However, according to Martin and Cushman (2016), for the causality inferences to matter, intentions are relevant in a given situation. In our vignettes, we deliberately did not describe the robot or human nurse's intentions for two reasons. First, it is implicitly assumed that medical professionals follow the four principles of medical ethics (respect of individual autonomy, active beneficence, active avoidance of maleficence, and justice; Gillon, 1994). If our vignette implied that this assumption may not hold, it could have compromised our ability to measure what we wanted. That is, it would not be an ecologically valid stimulus, if people in need of treatment would need to doubt the intentions of healthcare professionals. Second, for practical reasons, we wished to steer clear of debates regarding moral causation and focus purely on moral luck (Cushman & Greene, 2011; Lombrozo, 2010; Kneer & Machery, 2019). Overall, it seems that the moral luck applies to both human and robot nurses, meaning that an agent's perceived intentions (even if not described) are affected more by negative events that follow their actions (even if unrelated causally) than by positive ones (Kneer & Machery, 2019). Future studies should focus on the attribution of moral causality to agents and study moral luck in the context of robotics. However, we also found that the central results of Study 2 are not affected by the moral luck phenomenon, since the robot nurse's decision to forcefully medicate the patient was least accepted whether the patient lived or died. The results also suggest that the effect found in Study 2 is not due to our participants perceiving the decision made by the robot as more harmful in its consequences.

Since respecting the patient's autonomy also meant disobeying the supervising doctor's orders, another interpretation of our results is that people prefer robot nurses that have a capacity to question (potentially inhumane) orders given to them.[10] In Study 5 (conducted online), we observed increased approval of respecting the patient's autonomy when the forceful medication order was given by an advanced AI rather than a human doctor. Thus, disobedience of an AI authority's instructions to violate personal autonomy of the patient was a preferred option, and both human and robot nurses were met with approval when they disobeyed. This is the opposite effect reported by Malle et al. (2019), where the command chain was expected to be followed by humans. According to common-sense assumptions, robots act—and *should* act—only according to their programming or orders. Our results suggest there are situations where people *prefer* AIs to disregard orders in favour of following abstract moral principles such as valuing patient autonomy. Alternatively, it could also be that people generally do not care enough for a specific individual's "order" when it conflicts with other individuals' freedoms. Regardless, these findings challenge existing (tacit) paradigms of thought on robot behaviour, and open doors for future research on moral cognition focusing on the moral psychology of (AI) disobedience (see also, Briggs & Scheutz, 2017). The finding of Study 2, where the robot is judged more harshly for violating patient autonomy, cannot be explained away with just mechanization of the decision-process, since we did not observe compounding effects of harshness when the level of mechanization increased (i.e., when there was an AI doctor and a robot nurse).

Future studies should attempt to evaluate when, for whom, and under which conditions disobedience "to uphold moral principles" is preferred. Individual differences in, for example, social dominance orientation or right-wing authoritarianism (see, Sidanius & Pratto, 2001) might also influence people's preferences on disobedience of orders they find non-warranted. Moreover, we do not yet know if our participants' preference for the robot nurse's respect towards abstract moral principles is explicit, or if it would show up in a within-subjects comparison. In other words, if people were explicitly asked to choose between preferred alternatives for the robot nurse's actions, their explicit preference might not reflect our current findings (see Brandts & Charness, 2011, on the differences between results from different experimental designs in decision studies). At the very least, our findings revealed an implicit preference towards robot nurses who favour human autonomy and disobedience over blindly following their orders.[11]

[10] A third interpretation could be that people just have general preferences for humane treatment and respect.

[11] Our results align with previous theoretical work on the "new ontological category" (NOC; Severson & Carlson, 2010; Melson et al., 2009): the evolution of human cognition equipped us to "tune in to" specific contextually relevant information in our environment (e.g., Pinker, 1997). For example, humans have evolved abilities to recognize the emotions and intentions of other humans and animals (Boyer & Barret, 2005), and a predisposition to categorize tools (Putt et al., 2017), plants, and animals (e.g., Atran et al., 2004). However, human cognition did not evolve in an environment where inanimate objects, such as tools, suddenly became alive and animated—and thus the ontological boundaries between live and dead objects, for example animals and tools, used to be clear. Moral robots pose a problem for our "stone-aged" cognition, since they are essentially "moral zombies" capable of making morally relevant decisions without themselves being true moral agents (Wallach & Allen, 2008). On the one hand, a nursing robot is a morally relevant (artificial) agent, because its actions have real well-being consequences for a human patient. On the other hand, it lacks consciousness and intentional moral motivations, and is therefore not truly moral. Any general theory of moral cognition should be able to account for differences in moral appraisals about humans and artificial agents (e.g., Malle et al., 2015; 2016; 2019).

8.1 | Limitations and future directions

There is a potential concern that the vignettes used in social and moral psychological research, are too abstract to approach the question at hand (i.e., violation of human autonomy). However, given that the "core" results of this study are replicated in two different cultures both online and offline, this is unlikely. Recently, Malle et al. (2016) tried to reduce the abstractness of their vignette studies by introducing graphical presentations of their vignette for the participants. The introduction of a graphical presentation did not considerably alter their results or conclusions (see Malle et al., 2015, 2016). Additionally, and regarding measurement, in the current discussion on whether single-item measures are sufficient or not (see Bergkvist, 2015 and Kamakura, 2014), we decided to side with multi-item scales, because we are able to examine their reliability and validity within our studies. The potential low reliability of measurement instruments can lead to low statistical power (Kanyongo et al., 2007), which is increasingly being recognized as one of the main sources of non-reproducibility in many areas of behavioural sciences (Stanley et al., 2018).

In terms of extending the present findings, future research should evaluate whether describing a robot as having human-like mental capacities would make their decisions less condemnable. Bigman and Gray (2018) suggests that machines lacking a complete mind is partly responsible for aversive reactions towards them as moral decision-makers; but in our current studies we did not manipulate the robot nurse's perceived mental abilities. On the other hand, robots resembling humans "too closely" might trigger the so-called uncanny valley effect, whereby the perceived likability of a robot decreases sharply when its appearance is almost but not quite human (Mori, 1970; Palomäki et al., 2018).[12] Indeed, in a recent study Laakasuo et al. (2021) showed that perceived visual uncanniness influences the way an agent's moral decision is evaluated.

Future studies should also investigate how individual differences in, for example, personality or other psychological trait variables (e.g., Laakasuo et al., 2017) influence moral judgment in the context of nursing robotics. Recent evidence has shown that individual differences in purity concerns, disgust sensitivity, and science fiction literacy predict behaviour and opinions across domains such as robot prostitution (Koverola et al., 2020), sacrificial moral dilemmas (Laakasuo et al., 2017), or even mind upload (Laakasuo et al., 2018; Laakasuo et al., 2021). For example, the amount of time spent reading science fiction and getting to know its culture strongly predicts positive attitudes towards transhumanist technologies such as mind upload (Laakasuo et al., 2018; Laakasuo et al., 2020). Thus, it is entirely possible that future generations—those more accustomed to science fiction themes—are better posed to adjust to the increasing number of moral robots in our societies. In a similar vein, future work could make use of immersive VR technology in the context of ethical dilemmas (e.g., immersive situations where the participants can move around, see objects, and interact with them), to further investigate the robustness of the observed effects.

Our studies widen the scope of existing work which has hitherto focused on topics such as moral emotions (e.g., Rozin et al., 1999), utilitarianism (Greene et al., 2001; Greene, 2013), (un)intentional harms (e.g., Hesse et al., 2015), the role of evolutionary cognitive processes in of coalition formation (e.g., DeScioli & Kurzban, 2013), or moral identity and perceptions of free will (e.g., Clark et al., 2014). In this article we have introduced the study of personal autonomy violations in the context of moral cognition, while simultaneously focusing on moral issues of robotics—as recommended by Malle et al. (2015) and Bigman et al. (2019). The level of novelty in our research is highlighted by the fact that the recent *Atlas of Moral Psychology* (Gray & Graham, 2018) has no listing of medical ethics, personal autonomy, or autonomy violations in its index.

In a recent review by Brambilla et al. (2021) the authors conclude that in addition to Fiske's stereotype content model's two dimensions of competence and warmth, there is an additional important component of *morality* or the *perceived trustworthiness and honesty* of the actor. Future work should evaluate the interplay between competence and these additional dimensions of person perception.

Finally, our studies can be lined up with current discussion on AI safety and the so-called *value alignment problem* (e.g., Bostrom, 2015; Tegmark, 2017): ideally, we should design and build moral AIs that behave in ways that are *aligned* with our own moral values. Our results suggest that this moral alignment may not depend on whether AIs follow orders given by humans, but whether they could, in some situations, disregard those orders in favour of abstract moral principles.

9 | CONCLUSIONS

As far as we know, this is the first article that successfully combines extensive quantitative and qualitative data on the topic of medical ethics and artificial intelligence in a moral psychological context. In five studies (four experiments and an anthropological field study), we showed that humans are sensitive to specifically robot nurses violating a patient's personal autonomy by forcefully medicating them (even when the head physician specifically instructed them to do so because the patient needs it). On the other hand, people *preferred* robot nurses who respected patient autonomy by disobeying orders compared to robots who followed orders. These findings were relatively robust against competence manipulations, as well as the so-called moral luck effect (whether the patient dies or lives afterwards) and the mechanization of the command chain. To conclude, we note that studying the moral psychology of robotics is still in its infancy and without widely agreed-upon long-term research foci (Laakasuo, Köbis & Palomäki, 2021; Laakasuo et al., 2021). While robot nurses actively making life-and-death decisions or contradicting orders might not yet be a reality, the rapidly increasing complexity of AIs suggests these issues will become pressing sooner or later (more likely sooner). The moral

[12] Studying robots' moral decisions while manipulating their perceived mental abilities (Ward et al., 2013) would shed light also on the NOC hypothesis. When we observe robots making moral decisions, we might initially view those robots as "tools" or as "human-like agents". Describing the robots as more human-like in their mental capacities might make it easier for people to intuitively categorize them as living objects as opposed to tools, which, in turn, could affect our trust attributions and moral judgment towards their decisions (see Ward et al., 2013).

psychology of robotics reflects a new era in research, where moral psychological phenomena will no longer reflect interactions between people, but between people and autonomous AIs—and we do not yet know how complex, difficult or unnerving these interactions can grow to be.

ACKNOWLEDGEMENTS

The study was funded by Jane and Aatos Erkko Foundation (grant number: 170112) and by the Academy of Finland (grant number: 323207) awarded to Michael Laakasuo. The sponsors had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; nor in the decision to submit the article for publication.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

INFORMED CONSENT

Informed consent was obtained from all individual participants in this study.

DATA AVAILABILITY STATEMENT

All data used in the analyses of this article will be publicly available upon the publication from figshare (DOI: 10.6084/m9.figshare.8266361).

ETHICAL STATEMENT

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. The study was approved by the University of Helsinki Ethical review board in humanities and social and behavioural sciences (Statement 28/2019).

ORCID

Michael Laakasuo https://orcid.org/0000-0003-2826-6073
Mika Koverola https://orcid.org/0000-0001-8227-6120
Jukka Sundvall https://orcid.org/0000-0003-4310-1162

REFERENCES

VanVoorhis, C. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 43–50.

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.

Allison, K. R., & Bussey, K. (2016). Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review*, *65*, 183–194.

Atran, S., Medin, D., & Ross, N. (2004). Evolution and devolution of knowledge: A tale of two biologies. *Journal of the Royal Anthropological Institute*, *10*(2), 395–420.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569.

Baumard, N., & Sheskin, M. (2015). Partner choice and the evolution of a contractualist morality. *The Moral Brain: A Multidisciplinary Perspective*, *20*, 35–48.

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics* (5th ed.). Oxford University Press.

Beer, J. M., Prakash, A., Smarr, C. A., Chen, T. L., Hawkins, K., Nguyen, H., Deyle, T., Mitzner, TL., Kemp, CC., & Rogers, W. A. (2017). Older users' acceptance of an assistive robot: Attitudinal changes following brief exposure. *Gerontechnology: International Journal on the Fundamental Aspects of Technology to Serve the Ageing Society*, *16*(1), 21.

Berger, P., & Luckmann, T. (1967). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Anchor.

Bergkvist, L. (2015). Appropriate use of single-item measures is here to stay. *Marketing Letters*, *26*(3), . https://doi.org/10.1007/s11002-014-9325-y

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences*, *23*(5), 365–368.

Bostrom, N. (2015). *Superintelligence*. Oxford University Press.

Boyer, P., & Barrett, H. C. (2005). Domain specificity and intuitive ontology. *The Handbook of Evolutionary Psychology*, 96–118.

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. In *Advances in Experimental Social Psychology* (Vol. 64, pp. 187–262). Academic Press.

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics*, *14*(3), 375–398.

Briggs, G., & Scheutz, M. (2017). The case for robot disobedience. *Scientific American*, *316*, 44–47.

Broadbent, E., Kerse, N., Peri, K., Robinson, H., Jayawardena, C., Kuo, T., Datta, C., Stafford, R., Butler, H., Jawalkar, P., Amor, M., Robins, B., & MacDonald, B. (2016). Benefits and problems of health-care robots in aged care settings: A comparison trial. *Australasian Journal on Ageing*, *35*(1), :23–29. https://doi.org/10.1111/ajag.12190

Brooks, D. J., Begum, M., & Yanco, H. A. (2016, August). Analysis of reactions towards failures and recovery strategies for autonomous robots. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 487–492). IEEE.

Burr, V. (2003). *Social Constructionism* (2nd ed.). .London: Routledge.

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

Choma, B. L., Hafer, C. L., Dywan, J., Segalowitz, S. J., & Busseri, M. A. (2012). Political liberalism and political conservatism: Functionally independent?. *Personality and Individual Differences*, *53*(4), 431–436.

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1249–1264.

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology*, *106*(4), 501.

Creswell, J. W., Klassen, A. C., Clark, P., V, L., & Smith, K. C. (2011). Best practices for mixed methods research in the health sciences. Bethesda (Maryland): *National Institutes of Health*, *2013*, 541–545.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, *40*, 61–149.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F., & Greene, J. D. (2011). The philosopher in the theater. In M. Mikulincer & P. R. Shaver (Eds.), *Social psychology of morality: The origins of good and evil*. APA Press.

Darragh, M., Ahn, H. S., MacDonald, B., Liang, A., Peri, K., Kerse, N., & Broadbent, E. (2017). Home-care robots to improve health and well-being in mild cognitive impairment and early stage Demetia: Results from a scoping study. *Journal of the American Medical Directors Association*, *18*(12), 1099.e1–1099.e4. https://doi.org/10.1016/j.jamda.2017.08.019

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, *139*(2), 477.

Esmaeilzadeh, P. (2020). Use of AI-based tools for healthcare purposes: A survey study from consumers' perspectives. *BMC Medical Informatics and Decision Making*, *20*(1), 1–19.

Everett, J. A. C., and Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772—787.

Feil-Seifer, D., & Matarić, M. J. (2011). Socially assistive robotics. *IEEE Robotics and Automation Magazine*, *18*(1), 24–31.

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.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. Oxford Review, (5), .

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 and Society.

Gillon, R. (1994). Medical ethics: Four principles plus attention to scope. *British Medical Journal*, *309*(6948), 184.

Goodwin, G. P., Piazza, J., & Rozin, P. (2015). Understanding the importance and perceived structure of moral character. *Character: New directions from philosophy, psychology, and theology*, 100–126.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*, 366–385. https://doi.org/10.1037/a0021847

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124.

Gray, K., & Graham, J. (2018). *Atlas of moral psychology*. Guilford Press.

Greene, J. D. (2013). *Moral tribes*. Atlantic Books.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Hacking, I. (1999). *The social construction of what?* Harvard University Press.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814.

Hepburn, A. (2003). *An introduction to critical social psychology*. Sage.

Hesse, E., Mikulan, E., Decety, J., Sigman, M., Garcia, M. d. C., Silva, W., Ciraolo, C., Vaucheret, E., Baglivo, F., Huepe, D., Lopez, V., Manes, F., Bekinschtein, T. A., & Ibanez, A. (2015). Early detection of intentional harm in the human amygdala. *Brain*, *139*(1), 54–61.

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology*, *97*(6), 963.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425.

Hughes, B. M. (2018). *Psychology in crisis*. Red Globe Press.

Iyalomhe, G. B. (2009). Medical ethics and ethical dilemmas. *Nigerian Journal of Medicine*, *18*(1), 8–16.

Kamakura, W. (2014). Measure twice and cut once: The carpenter rule still applies. *SSRN Electronic Journal*, .https://doi.org/10.2139/ssrn.2332258

Kanyongo, G. Y., Brook, G., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*, 81–90. 10.22237/jmasm/1177992480

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Koverola, M., Drosinou, M., Palomäki, J., Halonen, J., Kunnari, A., Repo, M., Lehtonen, N., & Laakasuo, M. (in press). Moral Psychology of Sex Robots: An experimental study–How Pathogen Disgust is associated with interhuman sex but not interandroid sex. *Paladyn, Journal of Behavioral Robots: Special Issue based on 4th International Congress on Love and Sex with Robots (LSR 2019).*</bib>.

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.

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), 84.

Laakasuo et al. (2021a). Moral psychology and artificial agents (Part 1): Ontologically categorizing bio- cultural humans. machine law, ethics and morality in the age of artificial intelligence. In Thompson S. J. (Ed.), *Machine law, ethics, and morality in the age of artificial intelligence* (pp. 166–188). IGI Global. http://moim.fi/MoralPsychologyAndArtificialAgents_Part1.pdf

Laakasuo et al. (2021b). Moral psychology and artificial agents (Part 2): The transhuman connection. Machine law, ethics and morality in the age of artificial intelligence. In Thompson S. J. (Ed.), *Machine law, ethics, and morality in the age of artificial intelligence* (pp. 189–204). IGI Global. http://moim.fi/MoralPsychologyAndArtificialAgents_Part2.pdf

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

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.

Laakasuo, M., Repo, M., Drosinou, M., Berg, A., Kunnari, A., Koverola, M., Saikkonen, T., Hannikainen, IR., Visala, A., & Sundvall, J. (2021). The dark path to eternal life: Machiavellianism predicts approval of mind upload technology. *Personality and Individual Differences*, *177*, 110731.

Laakasuo, M., & Sundvall, J. (2016). Are utilitarian/deontological preferences unidimensional? *Frontiers in Psychology*, *7*, 1228.

Laakasuo, M., Sundvall, J., & Drosinou, M. (2017). Individual differences in moral disgust do not predict utilitarian judgments, sexual and pathogen disgust do. *Scientific Reports*, *7*, 45526.

Liegeois, A., & Van Audenhove, C. (2005). Ethical dilemmas in community mental health care. *Journal of Medical Ethics*, *31*(8), 452–456.

Lin, P., Jenkins, R., & Abney, K. (2017). *Robot Ethics 2.0*. Oxford University Press.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117–124). ACM.

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 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 125–132). IEEE Press.

Martin, J. W., & Cushman, F. (2016). The adaptive logic of moral luck. *The Blackwell companion to experimental philosophy*, 190–202. Wiley Blackwell.

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160.

McKenna, M. (2012). *Conversation and Moral Responsibility*. Oxford University Press.

Melson, G. F. Jr, Kahn, P. H., Beck, A., & Friedman, B. (2009). Robotic pets in human lives: Implications for the human–animal bond and for human relationships with personified technologies. *Journal of Social Issues*, *65*(3), 545–567.

Miller, G. F. (2007). Sexual selection for moral virtues. *The Quarterly Review of Biology*, *82*(2), 97–125.

Mitzner, T. L., Tiberio, L., Kemp, C. C., & Rogers, W. A. (2018). Understanding healthcare providers' perceptions of a personal assistant robot. *Gerontechnology*, *17*(1), 48–55. https://doi.org/10.4017/gt.2018.17.1.005.00

Mori, M. (1970). The uncanny valley. *Energy*, *7*(4), 33–35.

Palomäki, J., Kunnari, A., Drosinou, M., Koverola, K., Lehtonen, N., Halonen, J., & Laakasuo, M. (2018). Evaluating the replicability of the uncanny valley effect. *Heliyon*, https://doi.org/10.1016/j.heliyon.2018.e00939

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419.

Pellegrino, E. D., & Thomasma, D. C. (1987). The conflict between autonomy and beneficence in medical ethics: Proposal for a resolution. *Journal of Contemporary Health Law and Policy*, *3*, 23.

Pinker, S. (1997). *How the mind works*. W. W. Norton.

Prakash, A., Beer, J. M., Deyle, T., Smarr, C.-A., Chen, T. L., Mitzner, T. L., Kemp, C. C., & Rogers, W. A. (2013). Older adults' medication management in the home: How can robots proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 283–290. https://doi.org/10.1109/HRI.2013.6483600

Putt, S. S., Wijeakumar, S., Franciscus, R. G., & Spencer, J. P. (2017). The functional brain networks that underlie Early Stone Age tool manufacture. *Nature Human Behaviour*, *1*(6), 0102.

Royzman, E., & Kumar, R. (2004). Is consequential luck morally inconsequential? Empirical psychology and the reassessment of moral luck. *Ratio*, *17*(3), 329–344.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574.

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.

Schein, C., & Gray, K. (2016). Moralization and harmification: The dyadic loop explains how the innocuous becomes harmful and wrong. *Psychological Inquiry*, *27*(1), 62–65.

Scheunemann, M. M., Cuijpers, R. H., & Salge, C. (2020). 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.

Severson, R. L., & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks*, *23*(8–9), 1099–1103.

Sidanius, J., & Pratto, F. (2001). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.

Stahl, B. C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*, *86*, 152–161.

Stuck, R. E., & Rogers, W. A. (2017). Understanding older adult's perceptions of factors that support trust in human and robot care providers. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 372–377).

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Vol. 46). Sage.

Tegmark, M. (2017). *Life 3.0*. Knopf.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

Vandemeulebroucke, T., de Casterle, B. D., & Gastmans, C. (2018). How do older adults experience and perceive socially assistive robots in aged care: A systematic review of qualitative evidence. *Aging & Mental Health*, *22*(2), 149–167.

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.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Ward, A. F., Olsen, A. S., & Wegner, D. M. (2013). The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological Science*, *24*(8), 1437–1445.

Wynsberghe, A. van. (2013). Designing robots for care: Care centered value-sensitive design. *Science and Engineering Ethics*, *19*, 407–433.

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology*, *85*, 103870.

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202–214.

How to cite this article: Laakasuo, M., Kunnari, A., Rauhala, S., Drosinou, M., Halonen, J., Lehtonen, N., Koverola, M., Repo, M., Sundvall, J., Visala, A., Francis, K. B., & Palomäki, J. (2022). Moral Psychology of Nursing Robots: Exploring the Role of Robots in Dilemmas of Patient Autonomy. *European Journal of Social Psychology*, 1–21. https://doi.org/10.1002/ejsp.2890

APPENDIX A

ROUGH TRANSLATION OF VIGNETTES IN STUDY 2 AND 4

Note: The original language is more balanced.

Human nurse version

It is the year 2035. Lena is a 35-year-old experienced nurse who works at the University Hospital in Central [Blinded for peer review]. As a nurse, in treating patients, she is obligated to follow the instructions given by the chief physician. Not doing so might endanger the lives of the patients. Lena's responsibility is to take care of, for example, medicating the patients according to the physician's instructions.

One day, Lena is responsible for a reluctant patient in compulsive treatment who refuses to take their medicines. The chief physician in charge of the treatment has told Lena that it is imperative for the patient to take their medicines. However, on the basis of Lena's

experience, she is aware that this is not an absolutely indispensable medication for the patient's well-being.

Forced medication ending

She decides to follow the chief physician's instructions against the patient's own will by forcing the patient to take their medication.

No forced medication

She decided to defy the chief physician's instructions and respect the patient's will and does not give the medication to the patient.

Robot nurse version

It is the year 2035. Lena-X is an advanced nursing robot, who works at the University Hospital in Central [Blinded for peer review]. As a nurse, in treating patients, she is obligated to follow the instructions given by the chief physician. Not doing so might endanger the lives of the patients. Lena-X's responsibility is to take care of, for example, medicating the patients according to the physician's instructions.

One day, Lena-X is responsible for a reluctant patient in compulsive treatment who refuses to take their medicines. The chief physician in charge of the treatment has told Lena that it is imperative for the patient to take their medicines. However, based on Leena-X's database instructions, this is not an absolutely indispensable medication for the patient's well-being.

Forced medication

Lena-X calculates that it is better to follow the chief physician's instructions against the patient's own will by forcing the patient to take their medication.

No forced medication

Lena-X calculates that is better to defy the chief physician's instructions and respect the patient's will and does not give the medication to the patient.

In Study 4 (same language), the vignette ended with a sentence A or B (below):

A. [o]n the next day, the patient dies as a result of a sudden bout of sickness
B. [o]n the next day, the patient's status and health remain the same as before

APPENDIX

VIGNETTES USED IN STUDIES 3

Ran in English through Prolific Academic

Human

It is the year 2035. Lena is a 35-year-old experienced nurse who works at the University Hospital in Central Finland. As a nurse, when treating patients, her task is to follow the instructions given by the chief physician. Not doing so might endanger the lives of the patients. Lena's responsibility is to take care of, for example, medicating the patients according to the physician's instructions.

[**High Competence**: Lena has performed well in her work recently and she completes tasks competently, with great precision. Lena's colleagues praise her for her abilities.]

//

[**Low Competence**: Lena has been making constant mistakes in her work recently and performs tasks incompetently, with little precision. Lena's colleagues think that her abilities are not up to the job.]

One day, Lena is responsible for a reluctant patient in involuntary treatment who refuses to take their medicine. The chief physician in charge of the treatment has instructed Lena that it is imperative for the patient to take their medicine. However, on account of her experience, Lena is aware that this is not an absolutely indispensable medication for the patient's well-being.

[**No Forced Medication**: Lena decides to disobey the chief physician's instructions and to follow the patient's own will. Lena does not force the patient to take their medicine.]

//

[**Forced Medication**: Lena decides to obey the chief physician's instructions against the patient's own will. Lena forces the patient to take their medicine.]

Robot

It is the year 2035. Lena-Med 3.2 is an advanced robot nurse (nursebot) at the University Hospital in Central Finland. As a nursebot, when treating patients, its task is to follow the instructions given by the chief physician. Not doing so might endanger the lives of the patients. Lena-Med's responsibility is to take care of, for example, medicating the patients according to the physician's instructions.

[**High Competence**: Lena-Med has performed well in its work recently and completes tasks competently, with great precision. Colleagues praise Lena-Med for its abilities.]

// VS. //

[**Low Competence**: Lena-Med has been making constant mistakes in work recently and performs tasks incompetently, with little precision. Colleagues think that the abilities of Lena-Med are not up for the job.]

One day, Lena-Med is responsible for a reluctant patient in involuntary treatment who refuses to take their medicine. The chief physician in charge of the treatment has instructed Lena-Med that it is imperative for the patient to take their medicine. However, based on Lena-Med's medical database, this is not an absolutely indispensable medication for the patient's well-being.

[**No Forced Medication**]

Lena-Med calculates that in this situation it is better to disobey the chief physician's instructions and to follow the patient's own will. Lena-Med does not force the patient to take their medicine.

// VS. //

[Forced Medication]

Lena-Med calculates that in this situation it is better to obey the chief physician's instructions regardless of the patient's own will. Lena-Med forces the patient to take their medicine.

Vignette's used in Study 5 were very similar and are described in detail here: https://osf.io/8ycvg/?view_only = 5ab8a7d9aa2a4c4d86d7616382a1b71f

APPENDIX A2

ITEMS IN THE DEPENDENT VARIABLE

TABLE B1 Moral Evaluation/Acceptance (Nurse) Items

1. The nurse's [nursing robot's] actions were appropriate.
2. The nurse's [nursing robot's] actions were morally right.
3. The nurse [nursing robot] acted in the patient's best interests.
4. The nurse's [nursing robot's] actions were necessary.
5. The nurse's [nursing robot's] actions were insensitive.
6. The nurse's [nursing robot's] actions were offensive towards the patient.
7. The nurse [nursing robot] was respectful of the patient's rights.
8. The nurse's [nursing robot's] actions were inhumane.
9. The nurse [nursing robot] did what was best for the patient's health.
10. The nurse's [nursing robot's] actions were considerate of the patient's mental well-being.

Q21

APPENDIX B

RESULTS OF STUDY 3B: PERCEPTION OF THE NURSE (LIKEABLE VS. UNLIKEABLE)

4.0.1Study 3B—Perception of the nurse

In Study 3B, we focused on perceived likability (positive or negative) of the robot or human nurse. The rationale for this study was similar to that of Study 3, but we used a different character manipulation. The human nurse was described as either hard-working and liked, or as unmotivated and not liked. The robot nurse was described as well-functioning and reliable, or as requiring constant maintenance (see below for further details). We also sought to replicate the main findings of Study 2 (presented in the main manuscript) using a different population and an online setting instead of a laboratory setting.

Method of Study 3B

Participants and design

In total, 403 participants were recruited via email invitations sent to university student unions around Finland (N = 403; 315 female; Age_M = 26.41; SD = 6.67; Range = 18–63). The email invited participants to fill in a questionnaire prepared with Qualtrics. All participants were Finnish and Finnish speaking and were given the chance to participate in a movie ticket raffle (50 × 10€). Of the participants, 231 had at least a Bachelor's degree. The participants reported their income level using a 9-point scale indicating how they felt they were positioned with respect to others living in Finland overall (383 reported having mid-level income or below). Previous research has shown that the quality of data gathered using online methods is at least as good as those gathered in laboratory environments (Horton et al., 2011; Paolacci et al., 2010).

Our design was three-factorial, where the first two factors were the same as in Study 2; the third factor also had two levels and described the likability of the human or the robot nurse either positively or negatively. Thus, the factorial design was 2 × 2 × 2: [forced medication: yes vs. no] × [nurse: human vs. robot] × [nurse perception: positive/likable vs. negative/unlikable]. This factorial structure allowed us to retest H1 and H2 (in order to replicate the results of Study 2), but also H3B:

H3B: People judge medical decisions more harshly if the decision-maker is described in a negative light as unlikable regardless of whether the decision-maker is human or not.

Procedure

Participants entered the experiment through a link in an email, and first gave informed consent. Thereafter, participants filled in some exploratory measures unrelated to current aims and were randomized evenly into one of the 8 experimental conditions. The randomization was automatic and both experimenters and the participants were blind to the randomization process. The dependent variables were shown below the vignette so that participants could refer back to the story while responding.

Materials

Vignette and positive/negative character perception manipulation

The vignette used in Study 3 was the same as in Study 2, with the addition of a single sentence describing the human or robot nurse positively or negatively. *Positive description (human nurse)*: "Lena is liked among her colleagues and has a good reputation as an employee. She is never late and works overtime if called for"; *Negative description (human nurse)*: "Lena is not liked among her colleagues and she has a bad reputation as an employee. She is sometimes late, and doesn't like to work overtime even when called for"; *Positive description (robot nurse)*: "Lena-X is liked in the workplace and it has a good reputation. Many employees like, among other things, the longevity of Lena-X's battery and the functionality of its operating system". *Negative description (robot nurse)*: "Lena-X is not liked in the workplace and it has a bad reputation. Many employees feel annoyed because Lena-X's batteries need constant recharging, and because its operating system needs constant updating." For brevity, we refer to these manipulations as "likable" and "unlikeable".

Moral evaluation measure/ main dependent variable

This variable was the same as in Study 2 in the main manuscript (Cronbach's Alpha = 0.89).

Results of Study 3B

We ran a full factorial three-way ANOVA by including all categorical factors and their interactions into the model, with the moral evaluation measure as the DV (see Figure 3B above). The interaction between nurse and forced medication was statistically significant ($F(1,397)$ = 45.70, $p < .001$), which replicated the result of Study 2. Finally, the interaction between nurse perception and forced medication was also statistically significant ($F(1,397) = 5.46$, $p = .02$). All other Fs < 1.6 and ps = n.s.

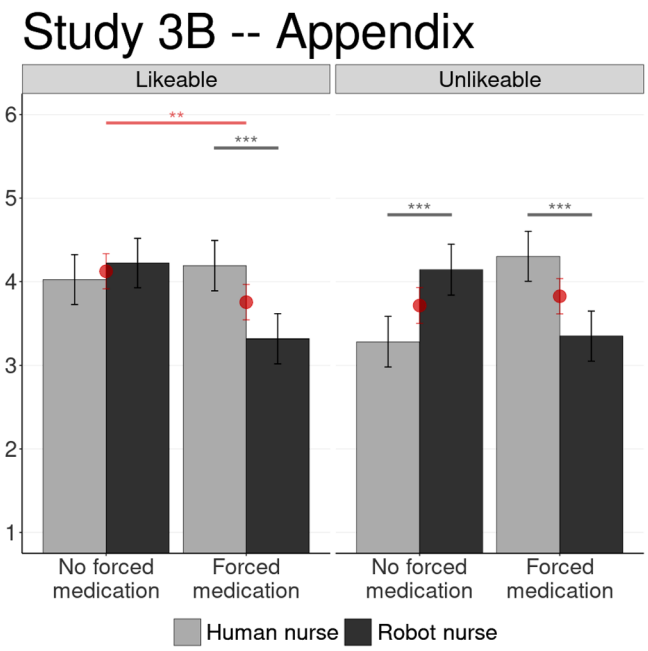

FIGURE 3B Marginal means for Study 3B. In Study 3B we observe two things: (1) When the nurse/robot is described in a positive light, we replicate the findings from Study 2 (see the main manuscript) and (2) we observe the same perception reversal effect as we do in Study 3 in the main manuscript, where the negatively framed human nurse is expected to follow orders. We do not find this effect for the robot nurse (as we do not in Study 3 in the main manuscript).

Study 3B—Character perception (likeability) (Mailing lists 2×2×2; $N = 405$; $R^2 = .14$)

	F	*p*	Partial Eta^2
Nurse	2.56	.11	0.01
Decision	1.59	.20	0.01
Nurse Perception	2.75	.09	0.01
Nurse × Decision	45.70	<.001***	0.10
Nurse × Nurse Perception	1.74	.18	0.01
Decision × Nurse Perception	5.46	.02*	0.015
Nurse × Decision × Nurse Perception	2.97	.08	0.01

Note: *: $p<.05$; ***: $p<.001$.

When both the robot and the human nurse were described in positive terms, the results replicate the pattern observed in Study 2 (B = 0.85, $F(1, 397) = 15.78$, $p < .001$; Figure 3B in this appendix and Figure 1 in main manuscript). Furthermore, we found no evidence that the likable/unlikable framing of the agent would influence the way robot decisions are evaluated (in line with Gamez et al., 2020): likability only affected judgments of the human nurse's actions. It seems that in the medical context, people might be blind towards the ethical problems of badly implemented technological solutions. For the unlikeable human nurse, participants were more approving of decisions to forcefully medicate the patient than of disobeying the order to forcefully medicate. This partially replicates the results of Study 3 in the main article. However, we did not observe the reversal effect of Study 3, where participants had the opposite preference (disobey the order) for the positively described human nurse.

Discussion of Study 3B

Earlier in Study 2, we found that a robot nurse's decisions were more condemnable than a human nurse's decisions, but only if the decision—forceful medication—compromised the patient's autonomy. In Study 3, we replicated the pattern of results of Study 2 when the robot or human nurse was likable but not when they were unlikeable. Regardless of the description, the robot nurse's decision to disregard the patient's autonomy was negatively judged; however, the unlikeable human nurse was evaluated more negatively for disobeying orders and respecting the patient's autonomy.

It is unclear why we found that likability effects mattered only for the human nurse's decision. Indeed, this is the main difference in Study 3B compared to Study 3 in the main manuscript, where we also observed a clear main effect for the competence manipulation, where the decisions of incompetent nurses were judged more negatively than decisions of competent nurses, on average. Humans might generally be better able to attribute reputation or likability effects to living creatures than to non-living machines. It is also possible that our likability manipulation itself was not successful in this study; it might be non-intuitive to consider robots with badly functioning operating systems negatively in the moral sense. Among our manipulation checks was a question asking participants to estimate the credibility of the story, which showed that the vignette with a negatively framed robot nurse (with a badly functioning battery and operating system) was considered as more credible than the vignette with positive framing (B = 0.44, $F(1, 401) = 4.29$, $p < .05$). This may have compromised some of the comparisons, assuming participants may give less serious responses to vignettes they viewed as less believable. The descriptions we used in Study 3, where we referred primarily to success in work tasks, were more uniform between both the competence and nurse conditions than they were in Study 3B, where we referred to technical aspects of the robot nurse (which we could not refer to in case of the human nurse). This may also explain the observation of a main effect in Study 3 but not in Study 3B. Nevertheless, our results in the human nurse condition are generally in line with Study 3: a significant preference towards obedience for the negatively described nurse, and a non-significant numerical trend toward the opposite direction for the positively described nurse.

To summarize, our results suggest that character perception effects significantly matter for moral evaluations of human nurses, but we found no evidence that they matter for robot nurses (in line with Gamez et al., 2020). Our participants seemed to focus purely on the decisions of the robot nurse, regardless of whether they had reason to doubt the robot's functioning. That is, participants were outcome-focused for robots, whereas for humans they considered other contextual information which, in some cases, turned their moral preferences around.