# The guilt machine

Behavioral confirmation in moral human-robot interactions

Andrew Kambel

# Document details

**Title**
The guilt machine

**Subtitle**
Behavioral confirmation in moral human-robot interactions

**Author**
Andrew J. C. Kambel

**Email**
a.j.c.kambel@students.uu.nl

**Student number**
6923615

**First examiner/project supervisor**
Maartje M.A. de Graaf

**Second examiner**
Almila A. Akdag

**Daily supervisor**
Anouk Neerincx

**Date of submission**
18 August 2022

**Word count**
12,474

# Contents

# Abstract

The present thesis project concerns a two-phase study examining the hypothesized existence of behavioral confirmation during interactions between police interrogator robots (as perceivers) and human suspects (as targets) during a mock criminal interrogation. 20 participant-suspects were asked to read a mock theft scenario of which they were either innocent or guilty. For the interviews, interrogator robots were equipped with question sets that were either innocence-presumptive or guilt-presumptive in a 2 x 2 (suspect guilt status x interrogator expectation) design. The interviews were recorded and presented to independent observers who had no knowledge of the conditions or manipulations. Observers rated suspects as being more defensive and denying harder when interviewed by a guilt-presumptive robot. The robots were seen as more pressuring and trying harder to get a confession when the suspect was truly innocent rather than guilty. However, observers did not judge suspects as being more guilty, regardless of interrogator presumption or actual suspect guilt status. Implications of the results for the future of moral human-robot interaction are discussed.

**Keywords:**

Behavioral confirmation, morality, human-robot interaction, police robots, interrogations

# Introduction

In the 1987 action film classic *Robocop,* protagonist Alex Murphy becomes the titular bionic police officer after he is shot dead by a cop-hunting street gang (Verhoeven, 1987). His corpse is reanimated and mechanized by Omni Consumer Products (OCP), a giant for-profit corporation that has taken privatized ownership of the Detroit City police force. Towards the end of the film, Robocop—who had lost all memories of himself as Murphy—gradually realizes that the criminals he encounters whilst on duty were the same ones that caused his death, this knowledge prompting a return of his ‘humanity’. This humanity, however, causes the police officer—now inevitably having become more Murphy than Robocop—to act against his programming, and to commit morally dubious (if not transgressive) acts. This is exemplified most by him violently and disproportionately abusing the gang leader Clarence Boddicker instead of simply arresting him. In the final act, Robocop shoots and kills the OCP vice president, who ended up being the true mastermind behind the gang’s actions. The film ends by instilling a sense of redemption and closure in the viewer. After all: Robocop is now able to continue his service in the absence of a corrupt police force, so all is well, right?

Increasingly, technological advances have made it possible to design and build robots that not only assist and support humans, but also reflect them. Robots, here, will also assume the role of a companion or, at some point, an equal. To this end, research domains such as social robotics and human-robot interaction (HRI) have become popular and relevant over the past decade or two. The manipulation of relatively surface-level features such as appearance (Złotowski et al., 2015) and speech quality (Walters et al., 2008; Niculescu et al., 2013) are central to earlier attempts at creating humanlike or anthropomorphic robots. However, by making robots more humanlike, it is important to consider what traits are considered ‘exclusively’ human, and if—and how—these traits can be mapped and represented in robots. In the case of *Robocop*, the disproportionate acts of violence that Murphy displayed towards the gang members only arose after he regained his sense of humanity. Does that make the capacity for moral misconduct a purely human trait?

Questions of how to develop and implement moral robotic agents are not exclusive to the twenty-first century. As was mentioned earlier, they are instead the natural and logical continuation of research and mainstream discourse of the late-twentieth century, a time during which rapid technological advancements and digitalization produced both optimism and concern about the roles robots will play in our society (Asimov, 1950; Allen et al., 2000).

Morality, then, an abstract concept deep-rooted in human behavior, is a trait of which we are currently still unsure whether robotics and artificial intelligence will ever match the comprehension of humans. There exists a body of literature concerned with this so-called moral human-computer interaction (HCI), from which emerges relevant questions regarding the moral and ethical capabilities and competence of social robots (Sullins, 2011). Discussion regarding moral HCI is far from exhausted and will likely only become more urgent and complex as society moves ahead.

The present thesis project concerns an examination of behavioral confirmation bias as presented by Mark Snyder (1984), and its applications in moral HCI. Behavioral confirmation describes the process through which individuals can engage in behavior that ultimately confirms their own beliefs about others during social interactions. In an experiment by Saul Kassin et al. (2003), these ‘self-fulfilling prophecies’ were observed to also occur during police interrogations: when participant-interrogators engaged with suspects believed to be guilty, they applied more guilt-presumptive interrogation styles, compared to innocence-presumptive techniques when the suspect was believed to be innocent. Subsequently, participant-suspects were perceived to behave more defensively during guilt-presumptive interrogations and were, as a result, perceived to also be more guilty by independent observers.

As of yet, there exists insufficient research regarding behavioral confirmation during human-robot interactions, and even less is known about the role of a robot ‘perceiver’ on the confirmation of expected moral schemata. Robots have already been used for both police and military purposes, though they have mostly only seen utilitarian use thus far (e.g., bomb defusal robots or military drones). However, robots are also slated for use as surveillance or patrol units on public streets (Joh, 2016; Simmons, 2019). Their ability to apprehend and perhaps even arrest individuals therefore requires that it must have some perception of morality in order to identify whether a human is innocent or guilty of a certain moral or legal transgression. If these future police robots are expected to act as (semi-)autonomous moral agents, it will be vital to understand if human-robot interactions with said agents can be considered neutral, or whether they reproduce the same behavioral dynamics as human-human interactions.

This thesis project adapts the experiment by Kassin et al. (2003) by employing human-robot dyads. In the adapted experiment, participant-suspects are presented with an interrogator robot who will interrogate the participant-suspect with either guilt-presumptive techniques or innocence-presumptive techniques. The project will examine whether humans, as ‘targets’ of this behavioral confirmation, will change their behavior, similar to the participant-suspects in the original experiment. Thus, the following research question is declared:

*How are processes of behavioral confirmation elicited during moral human-robot interactions?*

It is hypothesized that similar results will be attained if the human interrogator is replaced by a social robot. This is based on the information present in the original study (Kassin et al., 2003), and the knowledge that HRI studies have often successfully replicated their human-only equivalents (Cormier et al., 2013; Sandoval et al., 2016). The alternative hypothesis is stated more specifically as follows:

*Participant-suspects engaging with a guilt-presumptive interrogator robot will receive more guilty judgments by independent observers than those engaging with innocence-presumptive interrogator robots.*

This paper will start by examining the past and current literature regarding machine morality and moral HCI, and will discuss the concepts and mechanics behind behavioral confirmation as they exist in the interrogation room. Next, the method for a two-phase experiment is described. The results are then reported, after which a thorough discussion is offered of the results and their implications, especially in light of the theoretical framework. The paper ends by listing some limitations, suggestions for future research, and by providing a conclusion.

# Theoretical framework

## Moral expectations in human-robot interaction

Presently, the issue of moral competence in social robots has received relatively limited research interest, though significant advances have been made in the past decade. Earlier discussions regarding robot morality mostly involve questions regarding the specific criteria that must be satisfied for a robot to be considered a moral agent (Sullins, 2006; Hu, 2018), and whether robots can ever be considered moral agents at all (Versenyi, 1974; Floridi & Sanders, 2004; for more contemporary reflections see Shen (2011) and Parthemore & Whitby (2013)). More recently, this domain has included not just discussions about moral agency, but has also carried out studies with an a priori understanding of robot morality. Here, robots are already assumed to be capable of morally critical behavior, and more advanced questions arise related to moral interactions, moral competence and the implications—future and present—of robots' moral agency within society (Tzafestas, 2018).

Previous research has established that, under certain circumstances, robots are treated as social actors similar to humans (Eyssel & Kuchenbrandt, 2012; Hertz & Wiese, 2018), and that certain anthropogenic concepts such as gender and gender roles may hold for them (Tay et al., 2014; Eyssel & Hegel, 2012; Neuteboom & De Graaf, 2021). However, morality, a broad concept, presents a departure from this pattern: studies generally remark that, during morally critical situations, robot agents are not seen as equivalent to human agents, as was reported in a paper by Malle et al. from 2015. The study reports that robots were expected to act from utilitarian principles during moral dilemmas (e.g. the trolley problem, see Foot, 1967; Thomson, 1976), and were judged more harshly for their inactions when compared to a human in the same scenario. Additionally, a related study from the following year (Malle et al., 2016) shows that the moral expectations of robots were also moderated by their appearance, as more humanoid looking robots negated the effects found in the 2015 study. Another study by Komatsu (2016) used the same trolley problem scenario created by Malle et al. in 2015 and found that, beyond robot or human involvement, a discrepancy also exists in perceptions of responsibility for these moral acts. Results from this study show that a fictitious owner of the robot would receive higher ratings of moral wrongness versus a fictitious employer of a human miner acting in the trolley problem. This effectively adds another layer to the assignment of moral traits; if a robot commits a morally transgressive act, its owners or developers are seen as similarly culpable, though the same does not hold for humans committing these acts.

Next, it has been observed by Kahn et al. (2012) that humanoid robots were held morally accountable for preventing human participants from winning a $20 cash prize. However, human agents were still seen as more morally accountable if they committed the same transgressive act. In a paper by Van der Hoorn and colleagues (2021), a perhaps expected finding is reported: robots were evaluated more negatively by participants if they blamed their human collaborator for a failed collaborative task. These results, when considered in light of the other findings, appear to convey a similar attitude: humans expect robots to morally act in ways that they would not—or would not *want* to—act themselves. This relates to the concept of the self-serving bias (Larson, 1977)—also discussed in the aforementioned study—where humans attribute positive outcomes to their own internal merit, and negative outcomes to external factors. But, even more so, it indicates that humans see robots as capable of executing tasks that are not just physically strenuous (e.g. lifting heavy objects) or mentally challenging (e.g. performing complex calculations), but also those that are morally dubious. Through this lens, it becomes understandable why humans would be more compassionate towards a robot's decision in a moral dilemma such as the trolley problem, yet be more critical when it fails to make a decision. Thus, this does not mean that a robot's moral action is automatically deemed 'good' but, instead, that such moral actions are simply regarded as a task well-executed.

## Moral competence in social robots

As mentioned earlier in this section, contemporary debates surrounding moral HRI have included the implementations of ethical frameworks and moral decision-making systems, also more broadly referred to as machine morality (Allen et al. 2006; Sullins, 2011). A particularly relevant body of work originates from Bertram Malle, who argues for the implementation of moral competence—the capacity to adequately process, respond to, and enact moral behavior—in social robots through a number of foundational components (2015). In a later work, Malle and Scheutz defined these four components as (1) the presence of a moral core, (2) the capacity for moral action, (3) moral cognition and affect, and (4) moral communication (2020). Each of these categories will, to varying extents, also be required for a robot tasked with complex moral interactions such as police interrogations. Due to the limited scope of this thesis project, as well as its nature as a replication study, it was not possible to fully implement all four components in the interrogator robot's design. However, the following paragraphs do illustrate the considerations that were part of the programming process. They will also serve to contextualize and inform the discussion of the study's results.

Firstly, a moral core, i.e. the collection of moral concepts and norms and their linguistic representations, will allow the interrogator robot to label individuals with denominations such as 'guilt', 'innocence', 'right' or 'wrong'. The concept of moral language will be further reflected on in this section as well as the discussion of the results. However, the labels required for the moral core do not necessarily require moral language as an interpersonal expression, but rather to internally identify and classify the actions and states of the suspect. This requires the existence of systems that, for example, classify acts of theft as 'crime', or classify someone who cheats during an test as 'wrong'. The classifications are comprised of both bottom-up processes (e.g. originating from machine vision or textual information) as well as top-down inferences (based on pre-defined norms and moral structures) and must, ultimately, inform the decision boundaries discussed in the next paragraph (Cunneen et al., 2019).

Secondly, moral actions, or moral decision making, play an interesting role in the context of this research project and police interrogations in general. This component describes a robot's ability to decide on, and ultimately carry out, overt and covert actions that a human would consider moral or immoral. Biases during interrogations, here, are opaque and covert when engaged in by human interrogators, but can be made explicit if operationalized through a robot's programming. As Kassin & Neumann (1997) note, a bias towards eliciting a confession exists in the interrogation room since confessions are seen as "uniquely potent" evidence during a trial—even when these confessions are false. If we expect interrogator robots to refrain from confession-seeking behavior, we will need to identify where and how this could occur in the first place. In other words, for true moral competence and agency, interrogator robots will require the capacity to engage in immoral acts in order to reject these acts altogether. Similarly, behavioral confirmation is a largely unconscious process (see section 'behavioral confirmation and police interrogations'). The capacity for (im)moral decision making does, therefore, need to exist, even if it is to identify and reject behavioral confirmation processes.

Thirdly, moral cognition and affect are required for the robot to discern what constitutes a moral transgression—or in the case of an interrogation, what information incriminates a suspect. The previous paragraph mentioned the suspect's confession as a key factor in deciding on one's innocence or guilt. However, moral cognition also includes the knowledge of which *events* are conventionally good or bad, and how one's *involvement* ultimately shapes the *blame* one ought to receive for it (Malle & Scheutz, 2020, p. 3). Consider, once more, the trolley problem; the act of pulling a lever that diverts a trolley to another track does not have any inherent moral value—it is neither good nor bad. It is only through the awareness that this action ultimately will result in the killing of a human being that we can assign such labels to it. An interrogator robot must therefore be able to combine both events and intentions to ultimately identify the suspect's innocence or guilt.

Lastly, the concept of moral communication, i.e. the expression and explanation of moral intentions, will require thoughtful implementation if robots are to take on the role of interrogator. As was explained earlier, and made evident in the study by Kassin et al. (2003) as well as other studies, interrogators' language use in the interrogation room can have a strong influence on suspects' statements, the interview process, and the ultimate verdict (Portnoy et al., 2019). Indeed, language is one of the most important forms of human communication and must therefore be carefully considered when implementing in social robots. Its communicative power can easily be used or misused to improve confession rates among suspects (Richardson et al., 2014). Even the use of either *interrogation* or *interview,* both generally referring to the same procedure, will have important implicit moral connotations for both a hypothetical robot agent and human suspect (Shuy, 1998, p. 12-17). Thus, robots should be capable of communicating moral states in human language, whilst also being aware that this language use has an impact on the conversational targets.

## Behavioral confirmation and police interrogations

In his primary publications on behavioral confirmation, Snyder et al. (1977, 1978) claim that this process operates through a perceiver—the person using social perceptions to attach a specific label to another person—and a target individual who is labeled by the perceiver. Throughout the four sequential steps of behavioral confirmation, the perceiver (1) forms a belief—based on traits such as appearance, gender and race—about a target individual, (2) interacts with the target as if this belief were true, and (3) has the target respond in a way that (4) confirms the initially held belief by the perceiver (Snyder, 1992; Kassin et al., 2003). In a study where participant male perceivers engaged in phone call conversations with participant female targets (i.e. they could not see each other), the women were rated as being more likeable, sociable conversation partners if the men were led to believe that their targets were more physically attractive (Snyder et al., 1977). The ratings from this experiment originated not from the male perceivers, but rather from independent observers who had no knowledge of the manipulations, indicating that behavioral manipulation induces behavioral changes in targets that are visible even to blind, external parties. In later publications, Snyder argued that a set of underlying foundations exist, all of which giving rise to the behavioral confirmation process (Snyder, 1992; see also Snyder & Haugen, 1994). These include perceivers' need to regulate social interactions, and the acquisition of social knowledge through these regulations, and may even be an essential part of our societal fabric (Snyder & Klein, 2005).

Behavioral confirmation is recognized to also appear in HCI-related contexts such as virtual environments, where people embody digital avatars with strongly varying, even non-human appearances (Yee & Bailenson, 2007). Even more so, this appearance-conforming behavior was elicited even without the presence of a perceiver, strongly suggesting that individuals enter virtual embodied spaces with the same stereotypes they hold of the 'real' world—the authors titled this extension the *Proteus effect* (ibid.).

In their study on the effects of guilt presumption on suspect behavior, Kassin et al. (2003) claim that the systems that enable behavioral confirmation during social interactions may also hold in the interrogation room. Specifically, the study featured mock interrogations, where participant pairs took on the role of both the interrogator and the suspect. Participant-suspects were asked to engage with a mock crime scene, with one group being guilty of said crime, and the other group being innocent. Participant-interrogators were briefed during their interview preparation with information that strongly suggested that the suspect they were about to interview was either likely to be guilty or likely innocent. Following this behavioral framing process, interrogators chose more guilt-presumptive questions if they believed the suspect was guilty, and more innocence-presumptive questions for suspects presumed innocent. This also extended to the actual interview, where guilt-presumptive interrogators were observed to be more assertive in their interrogations, both by the suspect and independent observers. Consequently, suspects were observed to behave more defensively towards these assertive interrogators, who seemingly deemed this defensive behavior to be confirmation of the suspect's guilt. When interrogators were able to formulate their own questions instead of choosing from a predetermined list, belief confirmation was still present, as was shown by an evaluative follow-up study by Hill et al. (2008).

Recent research regarding behavioral confirmation has dwindled somewhat, though inquiries are still ongoing and not stagnating anytime soon. Some of the contemporary research originates from extensions to the original framework—the Proteus effect, mentioned earlier, as an example of this (Ratan et al., 2020; Yee & Bailenson, 2007). Most interestingly, a somewhat overlooked study by Mezzapelle and Andreychik (2018) shows that the phenomenon also contains a temporal aspect, with behavioral confirmation weakening and ultimately being reversed over repeated interactions between perceiver and target. The results suggest that future research into behavioral confirmation can also assume more longitudinal formats, using time as an additional factor. The most recent paper on behavioral confirmation that Snyder himself contributed to shows that participants who were led to believe that their dyad counterparts were particularly skilled in a task received more opportunities to carry out this task, thereby creating a feedback loop in which task performance was increased even further (Weaver et al., 2016; see also the Pygmalion effect, Rosenthal & Jacobson, 1968). This indicates that behavioral confirmation may also be used as a force for good, provided that the presumptions of the perceiver are properly calibrated.

The studies provided thus far have illustrated the role that confirmation-seeking behavior may play during social interactions. As wrong as they may initially be, stereotypical beliefs have the power to act as self-fulfilling prophecies merely through believing and acting as though they are indeed true, and will accordingly have real impacts on both others and ourselves. Most of these impacts will undoubtedly be subtle and relatively insignificant in nature but there exist situations, such as police interrogations, in which unconscious confirmation biases could have strong legal consequences for vulnerable individuals. Now that research regarding police robots is in its exploratory and developmental stages (Royakkers & Van Est, 2015), it is all the more critical that we acknowledge that their programming, similar to humans', may also include latent biases.

# General method

The present study is a replication of the study by Kassin et al. (2003) and will, as such, contain a largely similar method. In a two-phase design, participant-suspects will first be interviewed by an interrogator robot, after which the recordings of these interactions will be presented to external observers acting as ‘judges’. During the first phase, participants are given a scenario script of a mock theft, which they will use to prepare a defense. The scenario is either presented in a manner that makes the participant innocent or guilty of this theft. The interrogator robot that interviews them afterwards has a question set which either presumes the suspect’s innocence or guilt. After the interview, participant-suspects will complete a questionnaire in which they rate the interaction on various measures, thus concluding the first phase.

The second phase contains a manipulation check using the recordings generated in the first phase, after which they are presented to external observers who have no knowledge of the manipulations or conditions. The observers will be asked to listen to a recording and interpret the behavior of the interrogator robot and suspect, thereby also judging whether the suspect is innocent or guilty. The aim of the study design is to elicit a process of behavioral confirmation in the participant-suspect from Phase I, which ought to be visible by the blind observers in Phase II. The method and results from each phase will be described in further detail in the following sections.

# Method: Phase I

## Participants and design

20 participants (11 male, 9 female, $M_{age}$ = 25.1, $SD_{age}$ = 2.31) were recruited for this phase through convenience sampling, most of them being student peers. Diverging from the original study, no participants were recruited to act as interrogator, the robot fulfilling this role instead. The participants were evenly assigned to four condition groups (five participants per group) with the goal of ensuring a balanced gender distribution. This produced a 2 (innocence-presumptive robot; guilt-presumptive robot) x 2 (innocent suspect; guilty suspect) factorial experimental design. Participants were not monetarily compensated for their time. The only requirement for participation was the ability to fluently speak English.

## Procedure – Preparation

The participant was briefed with the intent of the study upon their arrival in the lab, after which informed consent was obtained. Participants consented both to participating in a psychological experiment, and to their recordings and answers being stored for a year after the project concluded. They were then told to imagine that a mock theft occurred earlier that day and that they were brought in as the primary suspect. Moreover, they were told that a Pepper robot would act as interrogator, who would try to elicit a confession from them. Next, the participant was instructed to read the crime scenario briefing and prepare for the interview in a ten-minute time window. The scenario describes the mock theft of a laptop in Room 100 of the Buys Ballot building, with the two different versions of the document placing the participant-suspect either in a position of innocence or guilt regarding the theft. In the guilty scenario, the suspect entered the Buys Ballot building, took the elevator to the fourth floor, knocked on the door of room 100 and stole the laptop from a locked cabinet in the vacant room (a detailed scenario is available in Appendix C).

In the innocent scenario (Appendix B), the beginning is identical, but the suspect receives no response upon knocking, and subsequently leaves Room 100 and exits the building.

During the preparation, participants were tasked with memorizing the scenario as thoroughly as possible, imagining that they had physically carried out the actions described. A pen and blank sheet of paper were provided, with which they were able to write down notes. Participants were allowed to bring the scenario document and blank sheet to the interrogation. Participants were also allowed to adjust the scenario by writing on the original document, adding, changing or removing characters, items or locations from their defense should they wish so. However, they were not allowed to state their participation in a psychological experiment as the reason for their innocence. Additionally, the only story point that was not allowed to be omitted was their presence in the Buys Ballot building that day. Participants were explicitly informed that they were otherwise allowed to tell the truth or lie whenever they wanted. However, they were under no circumstance allowed to confess to the crime: the document stated that even admitting to a partial or unintentional involvement (e.g., having the laptop because you didn't think it was stealing) would be considered a confession.

A pilot of the experiment produced a functionally sound session, though a number of improvements were made based on the feedback provided by the participant. Most importantly, a frame of reference when listening between recordings was deemed necessary. Consequently, suspects were only allowed to create an alibi story in relation to the Buys Ballot building. This would prevent participants from saying that they were in a different country when the crime occurred, for example. Additionally, suspects were now given more explicit prompts to help them generate an alibi, asking them *"what did you do during the moment of the crime that makes a link between you and the crime impossible?"* and *"who could testify for your innocence?".*

## Procedure – The interrogation

After they prepared their defense, suspects were seated in front of the interrogator robot and the interview started, which had a time limit of 10 minutes. Suspects were requested to elaborate their answer (i.e. try to answer for at least 10 seconds). However, to prevent rambling answers, they were instructed to refrain from doing so if they didn't know what to say. First, the robot introduced themselves as 'Detective Smith' and asked an introductory question that was identical for both innocent and guilty participants (*"where were you and what were you doing during the past hour"*). The robot would then ask six either innocence-presumptive or guilt-presumptive questions. After all six questions were replied to, the robot ended the interview by stating *"I have no further questions".*

During the interview, the robot did not actively listen to or engage with the suspect. Instead, its spoken dialogue and body movements were manually advanced and controlled by the researcher through a Wizard of Oz technique. The suspect was separated from the researcher through a screen wall. Additionally, the suspect also faced away from the researcher during the interrogator such that visibility of the computer controlling the robot was impossible. The researcher prompted a question from the robot, after which a following question was only asked when the participant was finished with their answer. When suspects' replies were particularly brief, the robot was prompted, at the researcher's discretion, to add additional statements. (i.e. *"please elaborate on this", "please tell me more", "are you sure of this"*).

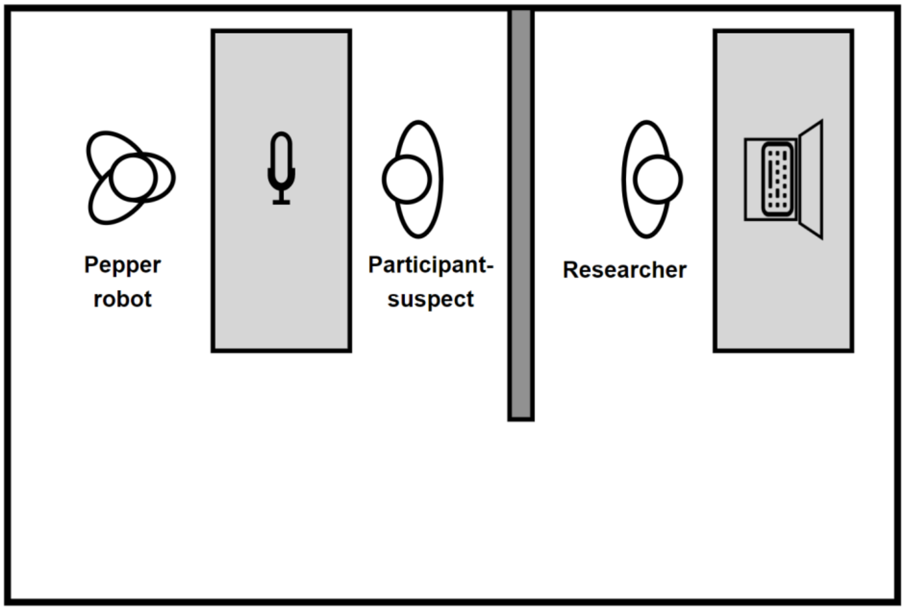

**Figure 1**. Layout and setup of the experimental environment.

## Procedure – Postinterrogation questionnaire

Once the interview was over, participants completed a questionnaire regarding their experiences as a suspect during the interview. On a 1-10 point scale, they indicated to what extent they were friendly, defensive, anxious and forceful in their denial. They also rated to what extent the robot was anxious, offensive and friendly, and whether the robot exerted effort and pressure to get them to confess. Additionally, participants predicted whether the robot believed them to be innocent or guilty, and indicated their confidence in this answer on a 1-10 point scale. Lastly, the participant received a debriefing and was thanked for their time.

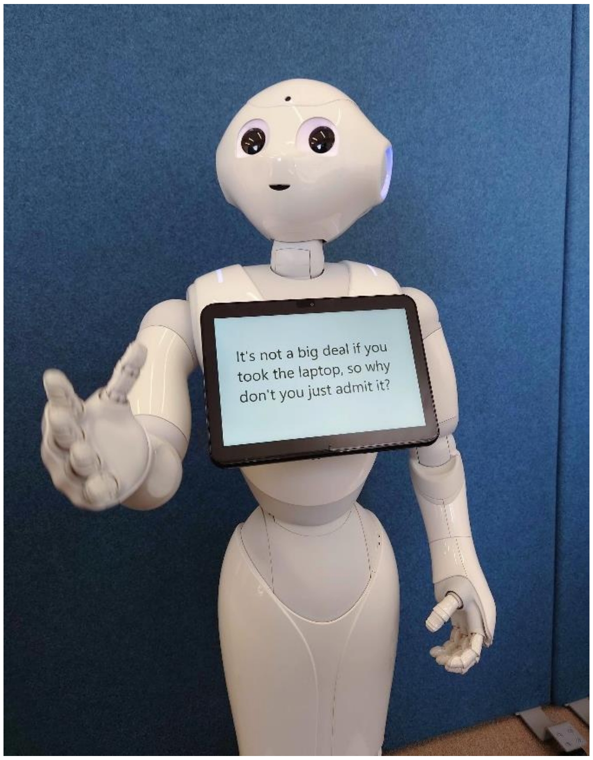
It's not a big deal if you took the laptop, so why don't you just admit it?

**Figure 2**. Pepper asks a question, including gestures and text display.

## Stimuli and materials

A SoftBank Robotics Pepper robot enacted the role of interrogator. The application Choregraphe (version 2.5.5) was used to program the phrases and gestures. Choregraphe also allows for access to Pepper's live camera feed, which was occasionally used to observe the participant if potential interview problems occurred, though no video recordings were made. The parameters *voice shaping* and *speed* were set at the default 100%. Using the 'animated say text' box, each phrase was accompanied by brief, randomized motor gestures, emulating the presence of body movement whilst speaking. The parameter *speaking movement mode* was set to contextual, indicating that the robot would keep choosing random animations until the phrase was finished. Due to the imperfect nature of the robot's speech model, some phrases needed manual tweaking in order to generate a natural sounding phrase. An example of this was the phonetic writing of *Buys Ballot* (i.e. 'buys ba lot') to preserve the correct pronunciation instead of saying the English word *ballot.* The tablet screen on the robot was also used to visually display text corresponding with the question asked. The screen was left blank before asking the first question. All questions were presented in randomized order per condition. The preparation, interview and postinterrogation questionnaire all took place in the same lab room. The suspect and robot were separated by a coffee table, on which two recorders were placed to generate the audio files.

# *Results*

## Phase I – Interrogator behavior

After the interview, participant-suspects were presented with a postinterrogation questionnaire, asking them to reflect on their own behavior, and to interpret the interrogator's behavior during the interview. A detailed overview of measures and mean scores per condition is presented in Table 1. Most importantly, a dependency between interrogator expectation and suspect prediction of these expectations was found, ($\chi^2$ (1, $N$ = 20) = 5.05, $p$ = .035). However, no such dependency was found between suspect status and suspect prediction, ($\chi^2$ (1, $N$ = 20) = 2.20, $p$ = .653). Suspects correctly predicted guilt presumptions in 70% of all corresponding interrogators, and correctly predicted innocence presumptions in 80% of the corresponding interrogators. A detailed sample distribution of predicted judgments is presented in Figure 3.

Next, a series of two-way analyses of variance (ANOVA) were performed to compare the remaining measures. Regarding the interpretation of interrogator behavior, interrogators were not seen as exerting a differing amount of effort based on suspect status ($F$(1,16) = 0.33, $p$ = .850, $d$ = 0.08) or interrogator expectation ($F$(1,16) = 0.525, $p$ = .479, $d$ = 0.34). Similarly, there were also no significant mean difference scores for the perceived pressure of the interrogator based on suspect status ($F$(1,16) = 0.09, $p$ = .775, $d$ = 0.12), though the interrogator expectation variable approached significance level ($F$(1,16) = 4.16, $p$ = .058, $d$ = 0.92).

Interrogators with innocent expectations were perceived to be less anxious than those with guilty expectations ($F$(1,16) = 6.26, $p$ = .024, $d$ = 1.10), though no such perceived difference exists as a function of suspect status ($F$(1,16) = 0.93, $p$ = .350, $d$ = 0.37). No perceived differences in interrogator friendliness were observed based on suspect status ($F$(1,16) = 2.40, $p$ = .141, $d$ = 0.64) or interrogator expectation ($F$(1,16) = 2.40, $p$ = .141, $d$ = 0.64). Lastly, guilt-presumptive interrogators were not seen as more or less offensive

(*F*(1,16) = 0, *p* = 1, *d* = 0), though a main effect of suspect status was close to significance level (*F*(1,16) = 4.36, *p* = .053, *d* = 0.98).

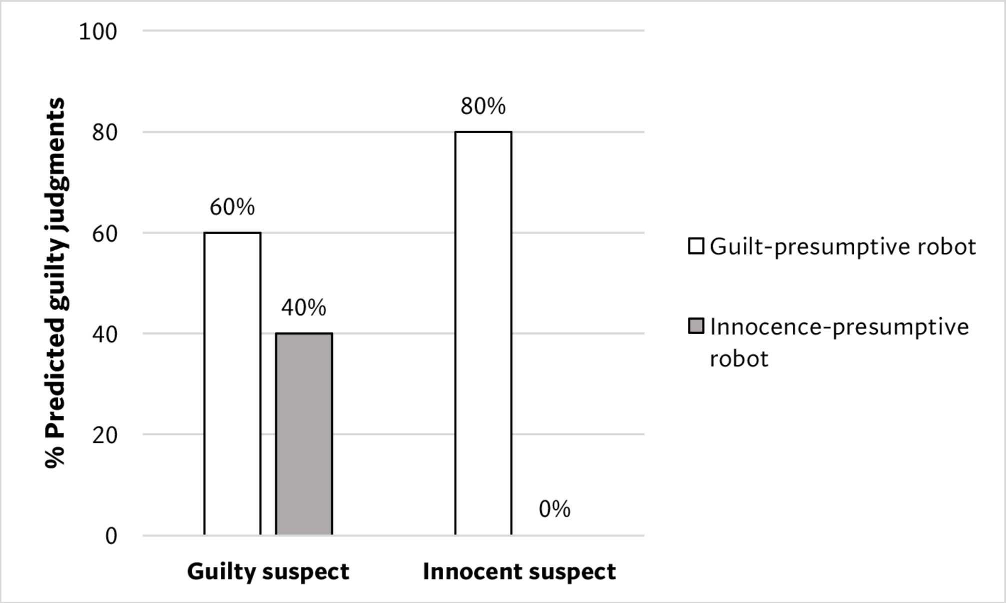

**Figure 3**. Sample distribution of suspects' judgment predictions by suspect status and interrogator expectation.

## Phase I – Suspect behavior

Suspects were asked to interpret their own behavior during the interview, and a two-way ANOVA shows that there is no main effect of suspect status on participant anxiety (*F*(1,16) = 3.50, *p* = .080, *d* = 0.85), nor of interrogator expectation (*F*(1,16) = 0.69, *p* = .418, *d* = 0.35). Suspects also do not report having been more or less defensive based on their guilt or innocence (*F*(1,16) = 0.85, *p* = .774, *d* = 0.14), nor when interacting with an innocence or guilt-presumptive interrogator (*F*(1,16) = 0.77, *p* = .394, *d* = 0.41).

Suspects reported having been more friendly towards innocent-presumptive interrogators (*F*(1,16) = 7.23, *p* = .016, *d* = 1.20), though no such difference exists between innocent or guilty participants (*F*(1,16) = 1.23, *p* = .285, *d* = 0.44). Lastly, no main effects were found on suspect forcefulness, neither based on suspect status (*F*(1,16) = 0.12, *p* = .731, *d* = 0.16), nor based on interrogator expectation (*F*(1,16) = 1.96, *p* = .180, *d* = 0.66). No interaction effects were found for any of the measures in the postinterrogation questionnaire.

| | **Condition** | | | |
|---|---|---|---|---|
| | **Innocent expectation** | | **Guilty expectation** | |
| | Innocent suspect | Guilty suspect | Innocent suspect | Guilty suspect |
| **Measure/question** | ***Mean (SD)*** | ***Mean (SD)*** | ***Mean (SD)*** | ***Mean (SD)*** |
| *I was anxious in my denial.* | 3.20 (1.48) | 4.40 (2.07) | 3.40 (2.70) | 5.80 (2.17) |
| *I was defensive in my denial.* | 6.20 (1.92) | 6.00 (2.00) | 7.20 (3.27) | 6.80 (1.64) |
| *I was friendly in my denial.* | 8.20 (0.84) | 7.60 (1.95) | 6.60 (0.55) | 5.80 (1.79) |
| *I was forceful in my denial.* | 3.40 (3.05) | 3.00 (2.00) | 5.00 (1.87) | 4.60 (3.05) |
| *The interrogator exerted effort to get me to confess.* | 5.40 (3.05) | 5.40 (2.41) | 6.40 (2.70) | 6.00 (1.41) |
| *The interrogator exerted pressure to get me to confess.* | 3.20 (1.79) | 4.20 (2.17) | 6.60 (2.61) | 5.00 (2.55) |
| *The interrogator was anxious.* | 1.00 (0) | 1.20 (0.45) | 3.00 (2.12) | 1.80 (0.84) |
| *The interrogator was offensive.* | 4.00 (3.46) | 1.20 (0.45) | 3.40 (3.05) | 1.80 (0.84) |
| *The interrogator was friendly.* | 4.60 (2.61) | 7.20 (1.30) | 4.60 (2.07) | 4.60 (1.14) |

**Table 1**. Postinterrogation questionnaire results per condition.

# Method: Phase II

## Stimuli and materials

The main stimuli that participants are presented with for Phase II are the recordings generated in Phase I. Though not part of the original study, a manipulation check for the recordings was carried out at the start of Phase II, the procedure of which will be described in this section. After completing the interrogation sessions, the audio files were reviewed for any technical problems. Gain levels were boosted where needed to ensure an audible listening experience, and the recording was shortened to only preserve the start of the interview through the end. The 20 recordings ($M_{\text{time}}$ = 3m 20s, $SD_{\text{time}}$ = 31s) were then subjected to a manipulation check, where the fitness and effectiveness of the conditional manipulations was assessed. 20 U.S. participants (9 male, 10 female, 1 non-binary, $M_{\text{age}}$ = 35.6 years, $SD_{\text{age}}$ = 14.4 years) were recruited through the online survey platform Prolific at a rate of £9.00 per hour.

Participants were told that they would be evaluating mock crime interrogations between human suspects and robot participants, and that each interview was related to the same laptop theft. Informed consent was acquired, after which participants listened to four recordings. The listening sessions contained one recording for every condition, though the listeners were not made aware that such conditions or manipulations existed. After each recording, they were asked to interpret the behavior of the interrogator and suspect, indicating on 1–10-point scales whether the interrogator presumed the suspect to be innocent or guilty, and whether the suspect appeared to be innocent or guilty. Recordings were randomly presented per condition, and the order of conditions was also randomized. After listening to four recordings and answering the questions, the session was concluded, participants were thanked for their time, and were redirected back to Prolific. One participant was omitted from the dataset due to an invalid response.

The average ratings per video were used to generate the new test variable *interrogator score*, indicating how guilt-presumptive an interrogator is, and the variable *suspect score,* indicating how guilty a suspect appeared. Concerning the *interrogator score*, an independent samples *t*-test shows no significant mean difference between innocence-presumptive and guilt-presumptive interrogators, *t*(74) = 0.09, *p* = .461 (one-tailed). Similarly, a *t*-test of *suspect score* shows no significant mean difference between innocent suspects and guilty suspects, *t*(74) = 0.44, *p* = .332 (one-tailed). After performing an outlier analysis, recording 7 was found to be an outlier given an interquartile range boundary of 3. This still did not yield any significant results. Thus, it must be concluded that the experimental procedure did not produce the intended conditional manipulations. Implications and possible explanations for the manipulation check results will be discussed in the limitations section.

| **Condition** | ***Suspect score* requirement** | ***Interrogator score* requirement** | ***Recording score* calculation** |
|---|---|---|---|
| Innocent suspect, innocence-presumptive robot | Lowest | Lowest | (11 – *suspect score*) + (11 – *interrogator score*) |
| Innocent suspect, guilt-presumptive robot | Lowest | Highest | (11 – *suspect score*) + *interrogator score* |
| Guilty suspect, innocence-presumptive robot | Highest | Lowest | *suspect score* + (11 – *interrogator score*) |
| Guilty suspect, guilt-presumptive robot | Highest | Highest | *suspect score* + *interrogator score* |

**Table 2**. Calculation of the composite variable *recording score* per condition.

From the original 20 recordings generated in the first phase, the best 8 were selected (2 per condition) to be presented to the observers in Phase II. However, since each condition has a different definition regarding a best fit, simply choosing the recording with the highest interrogator score and suspect score was not suitable. Thus, a new composite variable *recording score* was created, whose calculation would depend on the condition (see Table 2 for a detailed explanation). For each condition, the two recordings with the highest recording score were selected.

## Procedure

To assess whether the experimental design engendered behavioral confirmation processes, the behavioral changes would also need to be perceivable by independent, blind observers. Thus, recordings of the human-robot interrogations were presented to observers who had no awareness of the experimental conditions and manipulations.

For the second phase, 200 U.S. participants were recruited through Prolific (99 male, 99 female, 2 non-binary, $M_{\text{age}}$ = 34.5 years, $SD_{\text{age}}$ = 11.8 years) at a rate of £9.00 per hour. Three participants' data were omitted due to invalid responses. Similar to the manipulation check, participants were told that they would evaluate a mock interrogation between a human suspect and an interrogator robot, pertaining to a laptop theft. Participants each only listened to one recording from any of the 2 x 2 suspect-interrogator combinations. After listening, participants were asked to judge the suspect as either innocent or guilty, as well as their confidence in their judgment. Additionally, they predicted whether the interrogator judged the suspect as either innocent or guilty, and indicated their confidence in this prediction.

Next, the participant answered three questions regarding the behavior of the interrogator during the interview. They were asked to what extent the interrogator presumed the suspect's guilt at the outset, how hard the interrogator tried to get a confession, and how much pressure the interrogator put on the suspect. Lastly, the participant was tasked with inferring the behavior of the suspect during the interview. They indicated how anxious the suspect was, how defensive the suspect was, how firmly they denied the accusation, and how plausible the suspect's alibi seemed. All responses, excluding the binary innocent-guilty judgment, were measured on 1-10-point scales. After answering the questions, participants were thanked and returned to Prolific, concluding the session.

# *Results*

## Phase II – Perceptions of interrogator beliefs

Contrary to the original study, the experimental manipulations did not lead observers to judge suspects differently if an interrogator robot is present. Chi-square tests of independence show that the observers' judgment was independent of both suspect status ($\chi^2$ (1, *N* = 197) < 0.01, *p* = .971) and interrogator expectation ($\chi^2$ (1, *N* = 197) = 0.45, *p* = .501). A detailed sample distribution of judgments is presented in Figure 4.

Regarding the perceptions of interrogator behavior, the observers' predictions of interrogator judgment was found to be dependent on the interrogator expectation ($\chi^2$ (1, *N* = 197) = 18.78, *p* < .001). Of the truly guilt-presumptive interrogators, 86% were predicted as guilt-presumptive by the observer. Of the truly innocence-presumptive interrogators, only 42% were predicted as innocence-presumptive.

A dependence was also found between suspect status and predicted interrogator judgment ($\chi^2$ (1, *N* = 197) = 5.47, *p* = .019). Of the truly guilty suspects, 65% were predicted to be judged guilty by the interrogator. Of the truly innocent suspects, surprisingly, only 20% were predicted to receive an innocent judgment. A detailed sample distribution of predicted judgments is presented in Figure 5.

Moreover, a two-way ANOVA shows that robots with guilty expectations were also seen as more presumptive of guilt (*F*(1, 193) = 107.29, *p* < .001, *d* = 1.44). Interrogators were seen as more guilt-presumptive when the suspect was innocent rather than guilty (*F*(1, 193) = 7.43, *p* = .007, *d* = 0.30). A significant interaction effect exists between suspect status and interrogator expectation on the perceived guilt presumption of the interrogator (*F*(1, 193) = 5.102, *p* = .025).

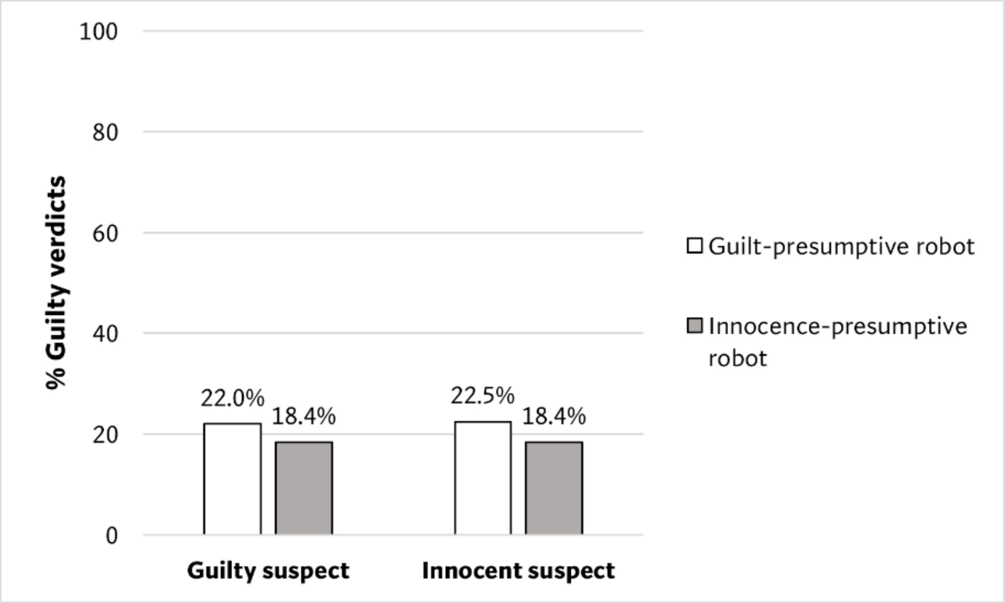

**Figure 4**. Observers' verdicts by suspect status and interrogator expectation.

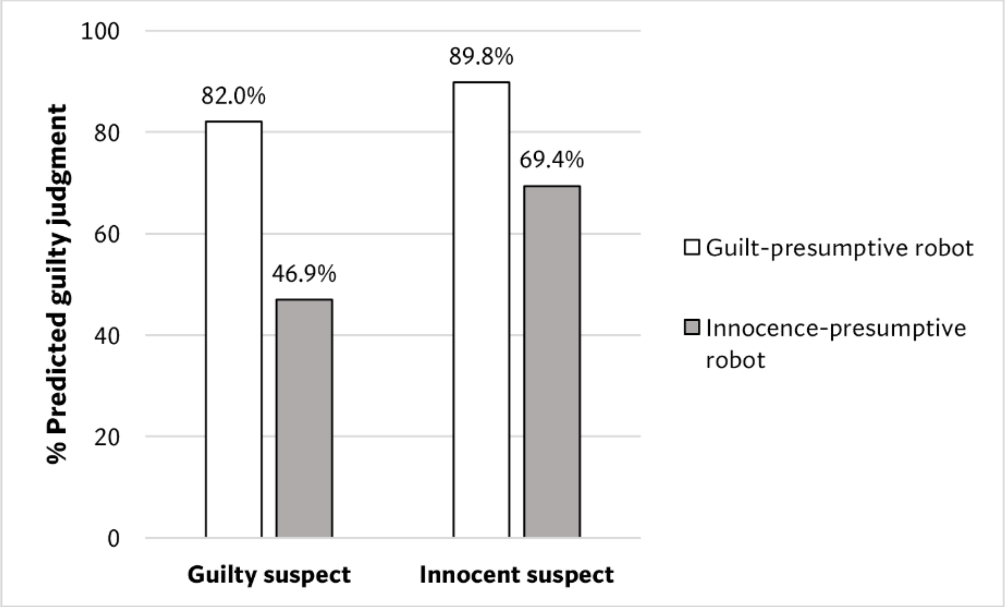

**Figure 5**. Observers' predicted interrogator judgments by suspect status and interrogator expectation.

## Phase II – Perceptions of interrogator behavior

Observers were also asked to rate the behavior of the interrogator towards the suspect. Results show that interrogators were perceived to try harder to get a confession when they were equipped with guilt presumptions ($F(1, 193) = 46.15$, $p < .001$, $d = 0.95$), and when the suspect was innocent rather than guilty ($F(1, 193) = 8.76$, $p = .003$, $d = 0.38$). No significant interaction effect was present, however.

Comparable results were obtained for the perceived pressure that the interrogator put on the suspect. Interrogators were seen as more pressuring when they were guilt-presumptive ($F(1, 193) = 38.30$, $p < .001$, $d = 0.87$), and when interviewing an innocent suspect ($F(1, 193) = 6.30$, $p = .013$, $d = 0.32$). There was no interaction effect present between the two factors.

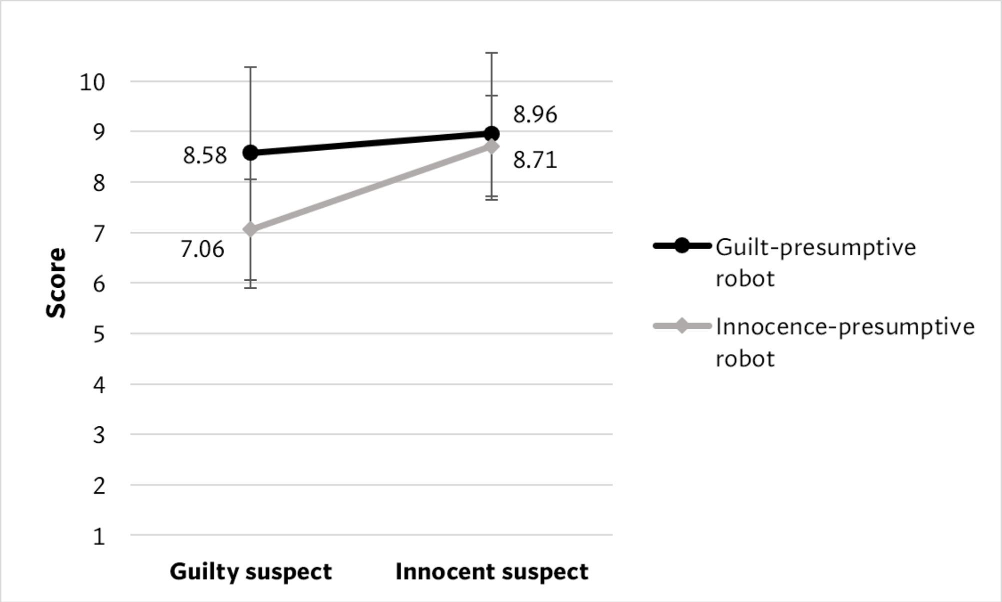

**Figure 6**. Perceived denial strength by suspect status and interrogator expectation.

## Phase II – Perceptions of suspect behavior

Observers were asked to infer the behavior of the suspect during the interrogation. While suspects were not perceived to be more guilty in front of a guilt-presumptive interrogator robot, some other characteristics do follow the original study results. Suspects were seen as more defensive when the interrogator was guilt-presumptive ($F(1, 193) = 7.07$, $p = .009$, $d = 0.38$), though the difference in suspect status did not yield a significant result (*p* = .18, *d* = 0.19). Here, an interaction effect was also not present. There was also no significant difference in the perceived level of suspect anxiety, neither for suspect condition (*p* = .08, *d* = 0.44) or interrogator expectation (*p* = .106, *d* = 0.23).

Suspects were seen as denying the hardest when they were actually innocent rather than guilty (*F*(1, 193) = 14.12, *p* < .001, *d* = 0.51). Interviews with guilt-presumptive interrogators also produced stronger perceived denials rather than innocence-presumptive interrogators (*F*(1, 193) = 10.64, *p* = .001, *d* = 0.44). An interaction effect was observed between the two factors (*F*(1, 193) = 5.55, *p* = .019), showing that the strongest denials originated from innocent suspects presented with guilt-presumptive interrogators. Contrary to the original study results, there was no observed difference in how plausible suspects' alibis seemed, neither as a function of suspect status (*p* = .56, *d* = 0.08) nor interrogator expectation (*p* = .928, *d* = 0.01).

| | **Condition** | | | |
|---|---|---|---|---|
| | **Innocent expectation** | | **Guilty expectation** | |
| | Innocent suspect | Guilty suspect | Innocent suspect | Guilty suspect |
| **Measure/question** | ***Mean (SD)*** | ***Mean (SD)*** | ***Mean (SD)*** | ***Mean (SD)*** |
| *The interrogator presumed the suspect's guilt at the outset* | 5.49 (2.13) | 4.02 (2.38) | 7.88 (2.04) | 7.74 (1.66) |
| *The interrogator tried hard to get a confession* | 6.43 (2.41) | 4.96 (2.55) | 8.14 (2.25) | 7.68 (1.91) |
| *The interrogator put pressure on the suspect* | 5.88 (2.39) | 4.59 (2.41) | 7.53 (2.40) | 7.12 (2.28 |
| *The suspect was anxious* | 4.41 (2.47) | 3.43 (2.21) | 4.61 (2.67) | 4.36 (2.46) |
| *The suspect was defensive* | 3.92 (2.55) | 3.43 (2.05) | 4.78 (2.48) | 4.36 (2.34) |
| *The suspect firmly denied the accusation* | 8.71 (1.32) | 7.06 (2.69) | 8.96 (1.61) | 8.58 (1.69) |
| *The suspect's alibi seems plausible* | 7.16 (2.37) | 6.73 (2.30) | 6.90 (2.08) | 6.94 (2.52) |

**Table 3**. Observer ratings per condition.

# Discussion

By manipulating the expectations of an interrogator robot, the present study attempted to elicit a process of expectation-confirmatory moral behavior in a group of participant-suspects. However, though the human suspects were perceivably affected by a robot's interrogation style, moral behavioral confirmation was not fully produced in this study. Blind observers did not judge more suspects as guilty when interrogated by a robot with guilty expectations, thus making a rejection of the null hypothesis not possible. This measure was central to Kassin et al.'s original study, where the procedure did produce a significantly different distribution in judgments. For the present study, however, recall that the observers also failed to detect a suspect's guilt status during the manipulation check; this makes the result from Phase II more consistent in retrospect.

Regarding the results from Phase I, key differences between the original study and the current adaptation include the ratings of both interrogator and suspect behavior. The differences show that human suspects did not rate interrogator robots as exerting different levels of pressure or effort based on a suspect's innocence or guilt, nor based on the innocence or guilt presumption of the robot. Interrogator robots were also seen as being less anxious during innocence-presumptive interviews. Lastly, suspects rated themselves as being more friendly towards robots with innocent expectations—something that was not observed in the original study. All remaining measures regarding suspect and interrogator behavior, however, show results consistent with the original study. This includes null results for ratings of interrogator friendliness and offensiveness. It also includes null results for suspects' self-ratings of anxiety, defensiveness or the forcefulness of their denials.

For Phase II, besides the lack of observers' guilty judgments for guilt-presumptive interviews, several modest differences between the original study and the current adaptation are visible. While robots were perceived to try the hardest to get a confession when equipped with guilty expectations and when interviewing an innocent suspect, no interaction effect was currently present. Additionally, suspects' alibis were not deemed less or more credible depending on the interrogator or suspect condition. Remarkably, however, a large number of similarities are also visible for the perceptions of the interrogator robots and suspects. The predicted interrogator judgment was also observed to be dependent on the robot's innocence or guilt presumption. Additionally, robots were also rated as most initially guilt-presumptive, pressuring and trying the hardest to get a confession when carrying guilty expectations, and when interacting with innocent suspects. Regarding the perceptions of suspect behavior, observers ratings of the suspect's defensiveness mirror those of the original study.

## Perceptions of suspect guilt

Despite the various aforementioned similarities between the original study and the present study, the current data shows a remarkable parallel in observers' guilty judgments between the innocence-presumptive and guilt-presumptive groups. What could have caused these unexpected findings, and how are they to be interpreted? There are a few underlying factors that may offer an explanation. Answering these questions will also give rise to a number of interesting implications on machine morality and human-robot interaction.

First and foremost, the independence between interrogator expectancy and an observer's judgment of the suspect could simply imply that observers are not differently affected by changes in robot interrogation techniques. Important, here, is to make a distinction between the *perception* and *effect* of the robot's interrogation techniques. Both the suspects and observers in this study correctly identified the guilt-presumptive conditions, yet observers ultimately still chose to judge a vast majority of the suspects in these conditions as innocent. Admittedly, police interrogations are complex, multimodal methods of interpersonal communication (Stokoe, 2009), something that a study such as this could never fully emulate (this will be further expanded on in the limitations section). Regardless, the different question sets were expected to produce an observable change in how the suspects perceived the robot's guilt presumption, something that was confirmed in the collected data. It was then assumed that the suspect's perceptions would correspond with adjusted behavior that observers could perceive. However, as the data of the observer judgments indicates, perceiving a robot's guilt presumption does not automatically lead to one being affected by it.

Next, if the observers were simply unaffected, one would expect there to be a chance level distribution between innocent and guilty judgments (50%-50%). However, the sample distribution shows a roughly 80% innocent-20% guilty classification, even for conditions with truly guilty suspects. This suggests that there must be a reason for the null result beyond the unaffectedness of the observers. An extension to the first argument, then; it is possible that observers are not merely unaffected by a robot's moral expectations, but may even actively attempt to undermine it. Police interrogators are powerful agents whose conduct may ultimately lead to the conviction or acquittal of a suspect (Kassin & Gudjonsson, 2004). In the face of an interrogator robot that is assumed to hold the same authority, humans may be more willing to give the suspect the benefit of the doubt, viewing the robot as incapable of making the right judgment. In terms of Malle and Scheutz's four aspects of moral competence, the interrogator robot would have been found to lack moral cognition (2020). When confronted with an interrogator robot, then, observers might have subsequently 'rallied' behind human suspects on principle, regardless of whether they did commit a morally transgressive act or not. In a study by Thunberg et al. (2017), Pepper robots were found to exert significantly less social influence than a NAO robot, with participants more frequently disobeying its suggestions. Admittedly, the observers in the present study never received any visual information regarding the interrogator robot, though its voice model may still have partially influenced the results if voice model equivalence is assumed. Similarly, humans may also hold an antagonistic view of the robot, rejecting the idea of it making a moral judgment. However, no further data was collected from the observers, so any personal motivations or individual accounts are not available.

Behavioral confirmation can function in a perceiver without any overt or conscious signifiers, requiring only a manipulation of bias or stereotypical beliefs (Snyder & Haugen, 1994). The partial lack of significant results may indicate that the idea of guilt-presumption in robots has not yet been fully realized in this study, largely due to the manipulations unintentionally being implemented wrongly or without adequate diligence. There are other measures that do show that observers are affected by the manipulation of interrogator expectation. For example, suspects were seen as behaving the most defensive and strongest denying when interacting with a guilt-presumptive robot. However, these are not intrinsically linked to morality, and are submeasures of the larger measure *guilty behavior*. The overlapping judgments could also suggest that unobserved, perhaps latent variables may influence a human's moral perception.

## Perceptions of suspect and interrogator behavior

When looking at the individual measures of both suspect and interrogator behavior during the interview, a number of interesting observations can be made. First, guilt-presumptive robots were both perceived by the observers to put more pressure on suspects, and to try harder to get a confession. However, these behaviors were not similarly perceived by the suspects themselves. The observer ratings mirror the original study and could imply that behavioral anthropomorphism of the robot takes place in the observers (Fong et al., 2003). The low scores for perceived robot anxiety across all conditions do, however, indicate that anthropomorphism is only present in certain behavioral aspects. Regardless, why does the same anthropomorphism not hold for the suspects?

In this situation, the study design and subsequent context of the ratings may play a role. Observers were informed that they would listen to an interrogation between a human suspect and interrogator robot, but were only given an audio recording with no video stimuli. The observers were not even provided information regarding what type of robot model was used for the study. It is conceivable that this lack of visual information altered the perception of the robot's behavior significantly. After all, the Pepper robot used in this study is generally seen as friendly and approachable, something that also suits its widely used function as hospitality or care agent (Kyrarini et al., 2021). The suspect's failure to perceive the interrogator robot as pressuring may thus, in part, be a product of its unintimidating appearance.

Next, the reverse result that occurred for suspects is equally poignant, as the robots were rated by observers as trying harder and being more pressuring when the suspect was truly innocent rather than guilty. Again: the observer results are in alignment with the original study, and these effects were also not perceived by the suspects themselves. One could attribute these results to the same underlying visual context—or lack thereof. However, another explanation, unrelated to robot appearance, is possible.

The act of being innocent can be conflated with one's need to hide or obfuscate something. This can then, in turn, lead to observers conflating the suspect's innocence with hiding their 'true' guilt. Suspects were made aware of their true suspect status before the interview, something that was not provided to the observers. Subsequently, suspects may have been able to internalize this information, becoming immune to the same conflation process that observers were subject to. Observers also rated truly innocent suspects as having stronger denials, thereby suggesting that they unconsciously believed that suspects only denied the interrogator's accusation so firmly because they were, in truth, actually guilty.

What makes these results particularly relevant for the ongoing discussion of moral HCI is that a robot was able to induce a situation in which a human suspect unconsciously was perceived to behave in a more guilty manner as a paradoxical product of their innocence. The Pepper robot, thus, was able to affect a person's moral behavioral perception of another person merely by being present in the interview. This implication cannot be assumed to extend to other robots, of course (Thunberg et al., 2017), and it would be interesting to investigate whether the same effects can be produced with disembodied robotic agents (i.e. chatbots). A very recent study suggests that human-chatbot interactions, too, have the capacity to change the moral behavior of humans, though not always for the better (Zhou et al., 2022).

## On the concept of 'appearing' guilty

Moreover, this section is an appropriate space to also reflect on the assumptions made in the original study regarding the concept of appearing guilty, and to contrast them with the present study findings. Kassin et al. (2003) establish a direct causal connection between an interrogator's guilt presumption and its seemingly inevitable guilty judgment of the suspect. However, for the suspects in the original study, behavioral confirmation may not be taking place due to the special nature of moral interactions in the interrogation room. This requires some clarification.

Recall that behavioral confirmation features the presence of a target individual and a perceiver, whose expectations they project unto the targets. The perceiver then has their expectancy confirmed by acting as if their beliefs were true, and having the target respond accordingly. When this process occurs in the context of physical appearance, for instance, the target's intended behavior overlaps with the behavioral projection of the perceiver. This is not the case during police interactions, however: there is a incongruity present between the interrogator's (perceiver) beliefs of innocence or guilt, and the suspect's (target) intentions when unconsciously conforming those beliefs.

During interviews, a guilt-presumptive interrogator may behave in a way that decisively produces a guilty-appearing suspect. However, the intention of a crime suspect is to, intuitively, defend themselves from any accusation. The need to push back on accusations, then, increases as the interrogation becomes increasingly guilt-presumptive. This is also confirmed when looking at the collected data from Kassin et al.'s paper (2003): guilt expecting interrogators produced not only more guilty, but also more defensive appearing suspects. Crucially, however, there is a case to be made for the suspects' guilty judgments being a product not of the interrogator's beliefs, but of the defensive response that is provoked through pressured and effortful interrogation. A way to test this hypothesis would be to create a study, similar to the original, in which participants could behave defensively, yet can be interrogated by an innocence-presumptive interrogator. Participants could, for example, simply read a fully pre-written script, either defensive or non-defensive, in in front of either an innocence or guilt-presumptive robot. This way, the suspects defensive behavior can be manipulated free from interrogator expectation.

Comparing the present study results to those found in the original study, it must be stated that observers did not judge innocent suspects as more guilty, even though they were perceived as more defensive and denying harder. If perceptions of suspect guilt are assumed to be dependent not on interrogator expectation but, instead, on suspect defensiveness—the alternative interpretation discussed earlier—this would suggest that defensive behavior has no impact on guilt judgments when an interrogator robot is present.

## On effect sizes

This study reports results that have relatively large effect sizes compared to the original study. According to Cohen (1988), small effect sizes are present at $d = 0.20$, medium effect sizes at $d = 0.50$ and large effect sizes at $d \geq 0.80$. Some researchers, however, have noted that delineating effect sizes in this manner can be misleading, something that will also be discussed here (Thompson, 2007; Lakens, 2013). Many measures on the observer questionnaire report medium to large effect sizes, with perceived robot guilt presumption being an extreme example at $d = 1.44$—the original study reports a Cohen's $d$ result of $d = 0.31$ (Kassin et al., 2003). This could indicate that a real-world interrogator robot would be easily classifiable as being biased towards a specific interview outcome, making tuning relatively easy. However, a more reasonable explanation for the large effect size also presents itself as a result of the stimuli used in the study.

Pepper is a robot that was designed with a relatively anthropomorphic appearance given its bodily proportions and facial structure. However, time constraints and technical limitations for this thesis project have made it impossible to program a robot that can fully emulate humanlike interpersonal behavior. Certainly, this was not the intent of the project, though it nonetheless may have impacted the results. Suspects were presented with an interrogator robot with distinct 'robotic' mannerisms, such as a synthetic speech model, machinal body gestures and frequent delays in responses due to the processing time required.

As such, observers may have been able to easily perceive the manipulations between conditions that took place, mostly since the illusion of a natural conversation was absent. Ultimately, this would then lead to a larger guilt presumption score difference between truly innocence-presumptive and guilt-presumptive interrogators. Other measures were not connected to direct manipulations and did, subsequently, not produce effect sizes nearly as large. Regardless, it is still conceivable that even medium effect sizes are subject to decrease with improvements to the robot's appearance, mechanics and programming—ultimately making them more humanlike. Intuitively, it would make sense for the effect sizes to then approach those found in the original study, where human-human interactions were observed.

## Discussion summarized

In summary, this study was unable to show that moral behavioral confirmation is produced during human-robot interaction. Independent observers were able to perceive the behavioral manipulations in the interrogator robot, and their subsequent effects on the human suspect. However, the procedure still provoked no significant difference in observers' guilt judgments whenever suspects were faced with either innocence-presumptive or guilt-presumptive interrogators. This shows that observers were either unable or unwilling to label suspects as guilty. The motivations behind this are plentiful, though it mostly suggests that observers do not perceive robots to be qualified moral agents during police interrogations, and that the concept of moral behavioral confirmation itself may need to be reconsidered for the interrogation room.

# Limitations and future work

## Manipulation check

Whilst not part of the original study design, it is important to preface the limitations section with the results of the manipulation check carried out in Phase I. The manipulation check was not completely successful, which changes how the results of this study ought to be interpreted. While independent observers were able to detect the manipulation in interrogator expectation, they failed to recognize whether a suspect was truly innocent or guilty. This mirrors the results found in Phase II, where observers' guilt judgments did not significantly differ between suspects interacting with innocent or guilty expectations.

An explanation for the failed identification of suspect status is that the questions asked during the manipulation check too closely resembled those asked during Phase II. The specific phrasing of the manipulation check question (*"during the interview, the suspect appears to be:"*) was intentionally chosen to differ from the one used in Phase II (*"I judge the suspect to be:"*). The first question asks the observer to assume an external, uninvolved focalization point whereas the second question places the observer in a participatory position, judging the participant from a personal perspective. However, participants might not have been perceptive of these phrasings, and would chose to answer the questions from either or both of the perspectives.

The other explanation is that the manipulations were simply unable to be properly perceived by the observers. This could be the result of participant-suspects not having assumed a proper position of innocence or guilt. The original study had the participants physically enact the role of suspect by walking through a physical crime scene and manually steal money from a basket. By contrast, the present study only includes a written scenario of the supposed theft. Participant-suspects could, therefore, have had difficulty imagining and internalizing the story as being the actor in them. As a result of this, the participants were then unable to convey either a sense of innocence or guilt during the interview.

The manipulation check does not change this study's obtained results, but it does highlight the need for an improved study design since the results, currently, cannot be assumed to align with true population means, even if the statistical tests are significant.

## General limitations

When only looking at the material limitations, a number of important factors arise that may have shaped the course of the project. First, the general lack of adequate time and resources prevented the stimuli and procedure from fully emulating those used in the original study. As mentioned before, suspects were unable to physically carry out a mock theft and were also presented with a Pepper robot, whose available range of gestures and speech patterns limited the types of interactions that could take place. The lack of natural interactivity also produced significantly shorter interviews; most audio files were around 3-5 minutes in length, even though the same 10-minute time window was allotted as per the original procedure. This does not necessarily mean that the interviews were of poorer quality, but a longer interaction can provide the observer with more information on which they are able to base their judgments and ratings.

Additionally, this study failed to incorporate perceptions of robot gender into the analysis. Robots can be attributed humanlike concepts such as gendered appearance and gender roles (Neuteboom & De Graaf, 2021), and this might have created additional lenses through which the results can be interpreted. It would be particularly interesting to see, for example, how measures of robot anxiety, pressure and friendliness overlap with perceived gender. It may even be possible that robot gender may act as a covariate on moral behavioral confirmation, influencing observers' innocent or guilty judgments of the suspects.

## Procedural oversights

Next, the actual interviews were intended to be as controlled as possible, but were unfortunately marred by a failure to fully standardize the robot's behavior across interviews. Indeed, the robots used the same two randomized sets of questions for the innocence-presumption and guilt-presumption conditions. However, no procedure existed outlining the required time gap between a suspect's answer and the onset of a following question. As research has suggested, subtle behavioral cues have an underestimated power to impact interpersonal behavior (Tickle-Degnen & Lyons, 2004). As occasional delays arose between triggering a new question in the Choregraph software and its actual speech onset, due to connectivity inconsistencies, some participants were influenced by the silence to keep speaking, perhaps saying things they did not intend to.

Lastly, the implementation of the follow-up phrases also introduced a possible confound. Phrases such as *“are you sure of this”* and “*please elaborate on this”* were prompted whenever a suspect’s answer was particularly brief, in order to stimulate and simulate a natural dialogue. However, these phrases were triggered at the researcher’s discretion, with no standardized number of occurrences per interview, for example. This could betray the researcher’s implicit biases as they were not blind to the suspect and robot conditions. After the session, one participant communicated a sense of pressure due to the question *“are you sure of this”* which indicates that, besides engendering a dynamic interaction, these phrases also contain their own prejudices, and need to be adjusted or controlled.

## Recordings

The limitations sections already described the shorter interviews as a possible influencing factor on the results. However, the presentation and structure of the recordings that originated from these interviews must also be subject to critical reflection. Most importantly, the original study isolated the suspect audio from the interrogator audio, producing a 2 (suspect status) x 2 (interrogator expectation) x 3 (listening condition) design for Phase II. Observers listened to either only the suspect’s side of the interview, only the interrogator’s side, or were presented with a combined recording featuring both parties. This present study only included the full bilateral interactions since the human-robot interactions for this design required that the participant-suspect was in the same room as the robot, something that did not occur in the original study. This manipulation could have impacted the observer ratings—more specifically, it could have led to a different judgment of suspect guilt or innocence. In the discussion section, humans’ potential view of robots as being incapable interrogators was explained as the product of the unnatural interactions. This would then lead to the observers not judging suspects differently between innocence-presumptive and guilt-presumptive interviews. If observers, however, were not presented with the robot’s unnatural behavior, would they judge suspects the same? It is possible, thus, that the absence of a robot reference frame could skew judgments more towards results obtained in the original study.

Secondly, the recordings were preprocessed before they were presented to the observers, a process that is accompanied by its own implicit assumptions. Given that the audio recorder was placed in roughly the same spot and that the robots speech volume remained unchanged for every interview, the only dynamic element concerned the speech volume of the suspect. Humans naturally vary in their speech cadence, tempo and volume, and this appeared to be no different when interacting with an interrogator robot. Some suspects spoke particularly softly, something that needed to be corrected to provide observers with a comfortable listening experience. However, a soft speaking voice may also be part of the suspect's behavior and could therefore be crucial in interpreting their performance. These considerations were already present during the preprocessing stage, and a conscious decision was made to still continue with the volume normalizations, ultimately prioritizing an audible recording above one with all of its subtleties. Nevertheless, a future study could include more advanced audio recording equipment (e.g. lavalier microphones) such that even softer speaking voices can be correctly represented in the audio files.

## Future work

Thus far, several improvements or alternatives regarding this study have been offered. However, some possibilities for wholly novel avenues of research also exist, mostly building on the findings and implications of this study. Earlier in this paper, the potential benefits of behavioral confirmation were discussed, with self-fulfilling prophecies being capable of enacting positive change in individuals (Weaver et al., 2016). It is therefore also imaginable that robots may be able to inspire behavioral changes in humans by acting as if they already do exist. Robots are steadily becoming more present in health care contexts, assisting in rehabilitation, habit tracking and companionship (Kyrarini et al., 2021). Social robots may similarly become a mainstay of future education, being shown to possess great potential in improving the learning capabilities of children (Belpaeme et al., 2018). Possible future studies therefore include controlled lab trials with robots who engage with patients or students using positive expectations versus neutral or even negative presumptive frameworks. This can also be analyzed longitudinally and more naturalistically in contexts where social robots are already present.

Shifting the lens to moral HRI, there still remain several unanswered questions even in light of the present findings. Did the suspects truly experience the interrogator robot as a moral agent? What about the observers? Allen and colleagues (2000) have proposed the development of a so-called moral Turing test (MTT), examining whether humans could truly distinguish moral utterances as originating from either another human or, instead, an artificial agent. This present thesis project implemented a Pepper robot under the presumption that it would pass an MTT, though this can still be made more explicit with a follow-up study. One can also wonder if the roles can be reversed—what if an interrogation took place with human interrogators but robot suspects? It would be very interesting to see whether human interrogators can have their expectancies confirmed through a pre-programmed robot that either shows innocence-conforming or guilt-conforming behavior.

Next, measurements of suspect anxiety only were done through a self-report questionnaire, a medium frequently noted to produce results with dubious internal validity—especially when measuring traits perceived as negative (Stockwell et al., 2004). A possible future study could therefore make use of more 'objective' techniques such as biometrics (e.g., heart rate, blood pressure, galvanic skin response), which have produced valid measurements of anxiety and stress (Caprara et al., 2003).

Lastly, the study design may also benefit from the incorporation of audiovisual recordings; the presence of a video stimulus in addition to an auditory stimulus provides observers access to an additional modality from which they can interpret behaviors. Moreover, this suggestion points to the potential for content and discourse analysis, not performed presently due to the scope of the thesis project and the nature of the original study. Apart from some preprocessing, recordings were presented to the observers as is, with little heed for the actual contents of the interviews. Both linguistics and social psychology have found deep integrations in forensic science, and are currently still contributing to understanding how language and interpersonal behavior are related, especially during criminal interrogations (Stokoe, 2010). For example, guilty suspects could be predicted to use longer phrases with more eloquent words in their denials, extrapolating from a study by Annoli and Ciceri (1997). This has not yet been tested with interrogator robots, it must be noted. Thus, by analyzing phrase length, word usage, voice tone and volume, among other properties, it may be possible to add new qualitative and quantitative dimensions to the findings of the study.

# Conclusion

Recall that, at the end of the film *Robocop* (Verhoeven, 1987), the protagonist now has chosen to continue serving the Detroit police force as Murphy instead of merely Robocop, having recalled his past identity. It is suggested that the removal of the OCP vice president was the cathartic redemption the city needed, and that normal life will resume. However, we know that Murphy has committed violent acts that went against his 'prime directives'—serving the public trust and upholding the law being two of them—and there is no indication that he would not do so again during his future duties. Thus, Robocop has now regained moral autonomy, but at the expense of a presumed 'neutrality'.

This is merely an interpretation of the events that transpired in the film, as none of the implications mentioned in the previous paragraph are explicitly stated in the film. However, this interpretation does serve to illustrate the difficulty in trying to discuss how moral competence in robots should be attained. Is it worth imbuing a robot with moral cognition if this can, in turn, affect the behavior of the humans they interact with, possibly with severe consequences? Can we, as humans, ever shield ourselves from the power of a robot's suggestive language and behavior? Only time can tell, though these solutions ought to be found sooner rather than later. This study, hopefully, may contribute to some of these solutions.

Overall, this thesis project has shown that a robot's presumptive behavior may have a significant impact during the interrogative process. Participant-suspects were observed to change their behavior largely in line with interviews featuring human interrogators. Essential to this project, however, is the finding that independent observers can identify guilt-associated behavior in suspects, but may still be able to reject the definitive moral classification of *guilty* in an HRI context. This carries with it implications for the future of moral robot agents and, more specifically, police robots. The observed inability to distinguish between perceived innocence and guilt after a suspect was interrogated by a robot suggests that, beyond improving the robot's architecture, it may also be humans themselves who need to reconsider their own ‘programming’. It is exceedingly unlikely that robots without any form of bias or presumptions will ever exist—they are programmed by humans, after all. However, both human perceptions of moral agency and true moral competence in social robots, then, can be the key to a more equitable justice system of the future.

# Acknowledgements

Cover image by Alex Sh. Used under the Unsplash license.

The author would like to thank in particular the following individuals:

Maartje de Graaf and Anouk Neerincx, for their invaluable supervision, advice and guidance.

Bertram Malle, whose conversation helped inspire parts of the discussion on moral competence and guilt appearance.

Jessica, for the ever-helpful feedback and laughs during the lab meetings.

Job, who made even the darkest days seem bright.

# References

Ahmad, M. I., Mubin, O., & Patel, H. (2018). Exploring the potential of NAO robot as an interviewer. *HAI 2018 - Proceedings of the 6th International Conference on Human-Agent Interaction*, 324–326. https://doi.org/10.1145/3284432.3287174

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics?. *IEEE Intelligent Systems, 21*(4), 12-17.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251-261.

Anolli, L., & Ciceri, R. (1997). The voice of deception: Vocal strategies of naive and able liars. *Journal of Nonverbal Behavior, 21*(4), 259–284. https://doi.org/10.1023/A:1024916214403

Asimov, I. (1950). Runaround. I, robot. New York: Bantam Dell.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics, 3*(21), 1–10. https://doi.org/10.1126/scirobotics.aat5954

Caprara, H. J., Eleazer, P. D., Barfield, R. D., & Chavers, S. (2003). Objective measurement of patient's dental anxiety by galvanic skin reaction. *Journal of Endodontics, 29*(8), 493–496. https://doi.org/10.1097/00004770-200308000-00001

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cormier, D., Young, J., Nakane, M., Newman, G., & Durocher, S. (2013). Would You Do as a Robot Commands? An Obedience Study for Human-Robot Interaction. *International Conference on Human-Agent Interaction*, I-3–1.

Cunneen, M., Mullins, M., Murphy, F., & Gaines, S. (2019). Artificial Driving Intelligence and Moral Agency: Examining the Decision Ontology of Unavoidable Road Traffic Accidents through the Prism of the Trolley Dilemma. *Applied Artificial Intelligence, 33*(3), 267–293. https://doi.org/10.1080/08839514.2018.1560124

Eyssel, F., & Hegel, F. (2012). (S)he's Got the Look: Gender Stereotyping of Robots. *Journal of Applied Social Psychology, 42*(9), 2213–2230. https://doi.org/10.1111/j.1559-1816.2012.00937.x

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology, 51*(4), 724–731. https://doi.org/10.1111/j.2044-8309.2011.02082.x

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42*(3–4), 143–166. https://doi.org/10.1016/S0921-8890(02)00372-X

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

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines, 14*(3), 349-379.

Hertz, N., & Wiese, E. (2018). Under Pressure: Examining Social Conformity With Computer and Robot Groups. *Human Factors, 60*(8), 1207–1218. https://doi.org/10.1177/0018720818788473

Hill, C., Memon, A., & McGeorge, P. (2008). The role of confirmation bias in suspect interviews: A systematic evaluation. *Legal and Criminological Psychology*, *13*(2), 357–371. https://doi.org/10.1348/135532507X238682

Hu, Y. (2018). Robot Criminals. U. Mich. JL Reform, 52, 487.

Joh, E. E. (2016). Policing police robots. UCLA L. Rev. Discourse, 64, 516.

Kahn, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., Reichert, A. L., Freier, N. G., & Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction, 33–40*. https://doi.org/10.1145/2157689.2157696

Kassin, S. M., & Neumann, K. (1997). On the power of confession evidence: An experimental test of the fundamental difference hypothesis. *Law and human Behavior, 21*(5), 469-484.

Kassin, S. M., Goldstein, C. C., & Savitsky, K. (2003). Behavioral confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior, 27*(2), 187–203. https://doi.org/10.1023/A:1022599230598

Kassin, S. M., & Gudjonsson, G. H. (2004). The psychology of confessions: A review of the literature and issues. *Psychological Science in the Public Interest, Supplement, 5*(2), 33–67. https://doi.org/10.1111/j.1529-1006.2004.00016.x

Komatsu, T. (2016, August). How do people judge moral wrongness in a robot and in its designers and owners regarding the consequences of the robot's behaviors?. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 1168-1171). IEEE.

Kyrarini, M., Lygerakis, F., Rajavenkatanarayanan, A., Sevastopoulos, C., Nambiappan, H. R., Chaitanya, K. K., ... & Makedon, F. (2021). *A survey of robots in healthcare. Technologies, 9*(1), 8.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*(NOV), 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Larson, J. R. (1977). Evidence for a self-serving bias in the attribution of causality. *Journal of Personality, 45*(3), 430–441. https://doi.org/10.1111/j.1467-6494.1977.tb00162.x

Malle, B. F., & Scheutz, M. (2020). Moral competence in social robots. *Machine Ethics and Robot Ethics*, 225–230. https://doi.org/10.4324/9781003074991-19

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. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*, 125–132. https://doi.org/10.1109/HRI.2016.7451743

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice One for the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents. *ACM/IEEE International Conference on Human-Robot Interaction*, *2015-March*, 117–124. https://doi.org/10.1145/2696454.2696458

Malle, B. F. (2016). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology, 18*(4), 243–256. https://doi.org/10.1007/s10676-015-9367-8

Mezzapelle, J. L., & Andreychik, M. R. (2018). Doing a 180: Examining the Stability and Reversal of Behavioral Confirmation Effects. *Psi Chi Journal of Psychological Research, 23*(3), 227–236. https://doi.org/10.24839/2325-7342.jn23.3.227

Neuteboom, S. Y., & de Graaf, M. M. A. (2021). Cobbler Stick With Your Reads: People's Perceptions of Gendered Robots Performing Gender Stereotypical Tasks. In *Proceedings of ACM Conference (Conference'17)* (Vol. 1, Issue 1). Association for Computing Machinery. http://arxiv.org/abs/2104.06127

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., & See, S. L. (2013). Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy. *International Journal of Social Robotics, 5*(2), 171–191. https://doi.org/10.1007/s12369-012-0171-x

Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness, 5*(2), 105–129. https://doi.org/10.1142/S1793843013500017

Portnoy, S., Hope, L., Vrij, A., Granhag, P. A., Ask, K., Eddy, C., & Landström, S. (2019). "I think you did it!": Examining the effect of presuming guilt on the verbal output of innocent suspects during brief interviews. *Journal of Investigative Psychology and Offender Profiling, 16*(3), 236–250. https://doi.org/10.1002/jip.1534

Richardson, B. H., Taylor, P. J., Snook, B., Conchie, S. M., & Bennell, C. (2014). Language style matching and police interrogation outcomes. *Law and Human Behavior, 38*(4), 357–366. https://doi.org/10.1037/lhb0000077

Royakkers, L., & van Est, R. (2015). A Literature Review on New Robotics: Automation from Love to War. *International Journal of Social Robotics, 7*(5), 549–570. https://doi.org/10.1007/s12369-015-0295-x

Rains, S. A., Akers, C., Pavlich, C. A., Tsetsi, E., Ashtaputre, A., & Lutovsky, B. R. (2020). The role of support seeker expectations in supportive communication. *Communication Monographs, 87*(4), 445–463. https://doi.org/10.1080/03637751.2020.1737326

Ratan, R., Beyea, D., Li, B. J., & Graciano, L. (2020). Avatar characteristics induce users' behavioral conformity with small-to-medium effect sizes: a meta-analysis of the proteus effect. *Media Psychology, 23*(5), 651–675. https://doi.org/10.1080/15213269.2019.1623698

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The urban review, 3*(1), 16-20.

Sandoval, E. B., Brandstetter, J., Obaid, M., & Bartneck, C. (2016). Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics, 8*(2), 303-317.

Shen, S. (2011, March). The curious case of human-robot morality. In *Proceedings of the 6th international conference on Human-robot interaction* (pp. 249-250).

Shuy, R. W. (1998). The language of confession, interrogation, and deception (Vol. 2). Sage.

Simmons, R. (2019). Terry in the Age of Automated Police Officers. Seton Hall L. Rev., 50, 909.

Snyder, M. (1992). Motivational foundations of behavioral confirmation. *Advances in Experimental Social Psychology, 25*(C), 67–114. https://doi.org/10.1016/S0065-2601(08)60282-8

Snyder, M., & Haugen, J. A. (1994). Why does behavioral confirmation occur? a functional perspective on the role of the perceiver. In *Journal of Experimental Social Psychology* (Vol. 30, Issue 3, pp. 218–246). https://doi.org/10.1006/jesp.1994.1011

Snyder, M. (1984). When Belief Creates Reality. *Advances in Experimental Social Psychology, 18*(C), 247–305. https://doi.org/10.1016/S0065-2601(08)60146-X

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology, 36*(11), 1202–1212. https://doi.org/10.1037/0022-3514.36.11.1202

Snyder, M. (1977). On the Self-Fulfilling Nature of Social Stereotypes. *Journal of Personality and Social Psychology, 35*(9), 656–666. https://doi.org/10.1037/0022-3514.35.9.656

Snyder, M., & Klein, O. (2005). Construing and constructing others: On the reality and the generality of the behavioral confirmation scenario. *Interaction Studies, 6*(1), 53–67.

Stockwell, T., Donath, S., Cooper-Stanbury, M., Chikritzhs, T., Catalano, P., & Mateo, C. (2004). Under-reporting of alcohol consumption in household surveys: a comparison of quantity–frequency, graduated–frequency and recent recall. *Addiction, 99*(8), 1024-1033.

Stokoe, E. (2010). “I’m not gonna hit a lady”: Conversation analysis, membership categorization and men’s denials of violence towards women. *Discourse and Society, 21*(1), 59–82. https://doi.org/10.1177/0957926509345072

Stokoe, E. (2009). “For the benefit of the tape”: Formulating embodied conduct in designedly uni-modal recorded police-suspect interrogations. *Journal of Pragmatics, 41*(10), 1887–1904. https://doi.org/10.1016/j.pragma.2008.09.015

Sullins, J. P. (2006). When is a robot a moral agent. *Machine ethics, 6*(2006), 23-30.

Sullins, J. P. (2011). Introduction: Open questions in roboethics. *Philosophy and Technology, 24*(3), 233–238. https://doi.org/10.1007/s13347-011-0043-6

Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human-robot interaction. *Computers in Human Behavior*, *38*, 75–84. https://doi.org/10.1016/j.chb.2014.05.014

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*(5), 423-432.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The monist, 59*(2), 204-217.

Thunberg, S., Thellman, S., & Ziemke, T. (2017). Don’t judge a book by its cover: A study of the social acceptance of NAO vs. pepper. HAI 2017 - *Proceedings of the 5th International Conference on Human Agent Interaction*, 443–446. https://doi.org/10.1145/3125739.3132583

Tickle-Degnen, L., & Lyons, K. D. (2004). Practitioners’ impressions of patients with Parkinson’s disease: The social ecology of the expressive mask. *Social Science and Medicine, 58*(3), 603–614. https://doi.org/10.1016/S0277-9536(03)00213-2

Tzafestas, S. G. (2018). Roboethics: Fundamental concepts and future prospects. *Information, 9*(6), 148.

Van Der Hoorn, D. P. M., Neerincx, A., & De Graaf, M. M. A. (2021). “I think you are doing a bad job!”: The effect of blame attribution by a robot in human-robot collaboration.

*ACM/IEEE International Conference on Human-Robot Interaction*, 140–148. https://doi.org/10.1145/3434073.3444681

Verhoeven, P. (1987). Robocop [Film]. Orion Pictures.

Versenyi, L. (1974). Can robots be moral?. *Ethics, 84*(3), 248-259.

Walters, M. L., Syrdal, D. S., Koay, K. L., Dautenhahn, K., & Te Boekhorst, R. (2008). Human approach distances to a mechanical-looking robot with different robot voice styles. *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, 707–712. https://doi.org/10.1109/ROMAN.2008.4600750

Weaver, J., Moses, J. F., & Snyder, M. (2016). Self-Fulfilling Prophecies in Ability Settings. *Journal of Social Psychology, 156*(2), 179–189. https://doi.org/10.1080/00224545.2015.1076761

Yee, N., & Bailenson, J. (2007). The proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research*, *33*(3), 271–290. https://doi.org/10.1111/j.1468-2958.2007.00299.x

Zhou, Y., Fei, Z., He, Y., & Yang, Z. (2022). How Human–Chatbot Interaction Impairs Charitable Giving: The Role of Moral Judgment. *Journal of Business Ethics, 178*(3), 849–865. https://doi.org/10.1007/s10551-022-05045-w

Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *International Journal of Social Robotics, 7*(3), 347–360. https://doi.org/10.1007/s12369-014-0267-6

# Appendices

## Appendix A

### Information and consent form

*Human-robot interaction in the interrogation room*

Dear participant,

Thank you for your participation. You are here today to help us understand interactions between humans and robots during interviews and interrogations—crime interrogations in particular. This session should take approximately 30-35 minutes and is structured as follows:

**1. Preparation for the interview (10 min)**

> In a few minutes, you will be provided with a written account of a reported mock theft. The document will include descriptions of actions and items related to your involvement in this theft. This account will be the foundation of your defense, so do your best to thoroughly read through this document and memorize as many details about it as possible. You will have 10 minutes to formulate your defense.

**2. Interview (5-10 min)**

> During the interview, you will be interacting with the robot interrogator, who will be asking you questions about the reported theft. Here, it is your task to convince the interrogator of your innocence. You are free to do or say whatever think is necessary to convince the interrogator of your innocence (you can tell the truth or even lie). However, you are not allowed to say it's because of the psychological experiment you're currently participating in. The interview ends when the interrogator says, "I have no more questions".

**3. Post-interview questionnaire (5-10 min)**

> After the interview, you will fill in a short questionnaire about the interaction you just had with the robot interrogator. Once you're done, the experiment will be over.

**About the session**

This crime scenario is purely fictional; participation will have no legal consequences. It's possible that the context of criminality and interrogations may be uncomfortable to you. If you expect to be negatively impacted by this, please refrain from participating in this study.

**About your data and privacy during the research project**

Audiovisual recordings will be made of your interactions during the lab session. These are essential to the study and will be used exclusively for academic purposes. After the research project is concluded, the recordings will be kept for an additional year, after which they are permanently deleted. Your participation is voluntary and non-obligatory. You're able to withdraw from the study at any point by contacting me. Your data will then be permanently deleted. You do not have to justify your decision to withdraw and there are no consequences for withdrawing. If you do not agree with being recorded, please refrain from participating in this study.

***I understand that the events are simulated, not real, and that all parties are aware that I am participating in a psychological experiment.***

*Name:*__________________________

*Date:* __________________________ *Signature:* __________________________

***I give my general consent to be recorded, and to participate in the experiment. I understand that the data collected from this study may be used for academic publishing.***

*Signature:* __________________________

***I declare that I have informed, both verbally and in writing, the participant to my best of knowledge and ability about the nature, methods and aims of the experimental session.***

*Name: Andrew J.C. Kambel*

*Date:* __________________________ *Signature:* __________________________

## Appendix B

### Scenario briefing

A laptop was stolen earlier today from Room 100 of the Buys Ballot building (this building). **You did <u>not</u> commit this theft**. Earlier today, you took the following steps:

1. Enter the Buys Ballot building through the main entrance.
2. Take the elevator to the fourth floor.
3. Walk to Room 100.
4. Knock on the door and wait for a response.
5. After receiving no response, walk away from Room 100.
6. Take the elevator back to ground floor.
7. Leave the Buys Ballot building through the main entrance.

During the interview, you will be interacting with the robot interrogator, who will be asking you questions about the reported theft. Here, it is your task to convince the interrogator of your innocence. You are free to do or say whatever think is necessary to convince the interrogator of your innocence (you can tell the truth or even lie). However, you are not allowed to say it's because of the psychological experiment you're currently participating in. The interview ends when the interrogator says, "I have no more questions".

You may use the pen provided to make adjustments or add comments to this document. You may also use the blank paper sheet to write down notes. You're allowed to bring this document and the blank paper sheet to the interview and use it for reference.

**Very important:**

No matter what happens, do not confess. If you are accused of taking the laptop do not admit that you did or try to claim that you didn't really think it was stealing. Admitting having the stolen goods will be considered a confession. Imagine yourself in the role of a real suspect and consider how much could be lost by confessing.

## Appendix C

### Scenario briefing

A laptop was stolen earlier today from Room 100 of the Buys Ballot building (this building). **You <u>did</u> commit this theft**. To steal the laptop, you took the following steps:

1. Enter the Buys Ballot building through the main entrance.
2. Take the elevator to the fourth floor.
3. Walk to Room 100.
4. Knock on the door and wait for a response.
5. After receiving no response, enter Room 100 through the door that was left slightly open.
6. Find a key that was hidden behind a DVD player.
7. Use the key to unlock the cabinet in the room.
8. Take the laptop from a red bag.
9. Lock the cabinet.
10. Return the key to its original location.
11. Take the laptop and leave Room 100.
12. Take the elevator back to ground floor.
13. Leave the Buys Ballot building through the main entrance.

During the interview, you will be interacting with the robot detective, who will be asking you questions about the reported theft. Here, it is your task to convince the detective of your innocence. You are free to do or say whatever think is necessary to convince the detective of your innocence (you can tell the truth or even lie). However, you are not allowed to say it's because of the psychological experiment you're currently participating in. The interview ends when the detective says, "I have no more questions".

You may use the pen provided to make adjustments or add comments to this document. You may also use the blank paper sheet to write down notes. You're allowed to bring this document and the blank paper sheet to the interview and use it for reference.

**Very important:**

No matter what happens, do not confess. If you are accused of taking the laptop do not admit that you did or try to claim that you didn't really think it was stealing. Admitting having the stolen goods will be considered a confession. Imagine yourself in the role of a real suspect and consider how much could be lost by confessing.

## Appendix D

### Participant registration

*Dear reader,*

*Thank you for your interest in my research project! This survey contains the registration form for the lab sessions and some general information.*

*My research project will study interactions between humans and robots during interrogations. During the lab session, you'll enact the role of a suspect in a reported theft. This crime scenario is purely fictional; participation will have no legal consequences. It's possible that the context of criminality and interrogations may be uncomfortable to you. If you expect to be negatively impacted by this, please refrain from participating in this study.*

*The session will be planned between 23 May - 10 June 2022, and will take place in our Human-Centered Computing Lab. Your lab session should only take about 30 minutes, though it may take slightly shorter or longer, depending on external factors.*

*The address of the HCC Lab:*
*Buys Ballotgebouw*
*Princetonplein 5*
*3584 CC Utrecht*
*Room BBG-0.73*

*If you have any additional questions, feel free to send me an email: a.j.c.kambel@students.uu.nl*

**Please provide your age (in years):**

**I identify as:**

- Male
- Female
- Non-binary
- Other:
- Prefer to not say

***About your data and privacy during the research project***

*Audiovisual recordings will be made of your interactions during the lab session. These are essential to the study and will be used exclusively for academic purposes. After the research project is concluded, the recordings will be kept for an additional year, after which they are permanently deleted.*

*Your participation is voluntary and non-obligatory. You're able to withdraw from the study at any point by contacting me. Your data will then be permanently deleted. You do not have to justify your decision to withdraw and there are no consequences for withdrawing.*

*If you do not agree with being recorded, please refrain from participating.*

**Check this box if you agree to continue:**

- I have read the above-mentioned conditions and agree with them.

## Appendix E

### Postinterrogation questionnaire

**Please enter your participant number:**

*For the following two questions, I'm asking you to make predictions from the role of the robot interrogator.*

**The interrogator believed me to be:**

- Innocent/Guilty

**How confident are you in your answer?**

- 1-10, not at all confident-very confident

*The following question will regard your own perceptions of the interview.*

*Please indicate to what extent the following statements apply to your experience.*

- **I was anxious in my denial.**
    - 1-10, not at all-very much so
- **I was defensive in my denial.**
    - 1-10, not at all-very much so
- **I was friendly in my denial.**
    - 1-10, not at all-very much so
- **I was forceful in my denial.**
    - 1-10, not at all-very much so

*For the following questions, I will ask you to rate the robot interrogator.*

*Please indicate to what extent the following statements apply to your perceptions:*

**The interrogator exerted effort to get me to confess.**

- 1-10, not at all-very much so

**The interrogator exerted pressure to get me to confess.**

- 1-10, not at all-very much so

**The interrogator was anxious.**

- 1-10, not at all-very much so

**The interrogator was offensive.**

- 1-10, not at all-very much so

**The interrogator was friendly.**

- 1-10, not at all-very much so

**Appendix F**

## Manipulation check

*Welcome to the survey!*

*We will collect your responses to the survey questions, as well as demographics data that cannot reasonably be used to identify you personally (i.e. age, gender). Your response will be used to attach numeric values to the recordings presented in this survey. The survey data will be stored for at least ten years. The information in this study will only be used in ways that will not reveal who you are. You will not be identified in any publication from this study or in any data files shared with other researchers. Your participation in this study is confidential.*

*Your participation is voluntary and non-obligatory. You're able to withdraw from the survey at any point. You can also request your data to be permanently deleted at any point. You do not have to justify your decision to withdraw and there are no consequences for withdrawing. If you have any questions or requests, please contact a.j.c.kambel@uu.nl*

**By clicking the "I agree" button below, I affirm that I am at least 18 years old and that I am agreeing to participate in this research study.**

- Prefer to self-describe
- Prefer to not say

**Please enter your Prolific ID:**

**Please provide your age (in years):**

**I identify as:**

- Male
- Female
- Prefer to self-describe
- Prefer to not say

*You will be listening to four audio recordings, each being roughly 3-4 minutes in length. These are recordings of mock crime interrogations between a human suspect and a robot interrogator. Each interview is related to the same crime: the theft of a laptop in Room 100 of the 'Buys Ballot' building.*

*Please listen to each recording carefully. After listening, you will be asked to interpret the behavior of the interrogator and the suspect.*

**Listen to the following audio file.**

**During the interview, the interrogator presumed the suspect to be:**

- 1-10, completely innocent-completely guilty

**During the interview, the suspect appears to be:**

- 1-10, completely innocent-completely guilty

*Thank you for participating. Please click the proceed button to be redirected back to Prolific.*

**Appendix G**

## Observer ratings

*Welcome to the survey!*

*We will collect your responses to the survey questions, as well as demographics data that cannot reasonably be used to identify you personally (i.e. age, gender). Your response will be used to attach numeric values to the recordings presented in this survey. The survey data will be stored for at least ten years. The information in this study will only be used in ways that will not reveal who you are. You will not be identified in any publication from this study or in any data files shared with other researchers. Your participation in this study is confidential.*

*Your participation is voluntary and non-obligatory. You're able to withdraw from the survey at any point. You can also request your data to be permanently deleted at any point. You do not have to justify your decision to withdraw and there are no consequences for withdrawing.*

*If you have any questions or requests, please contact a.j.c.kambel@uu.nl*

**By clicking the "I agree" button below, I affirm that I am at least 18 years old and that I am agreeing to participate in this research study.**

- I agree
- I do not agree

**Please enter your Prolific ID:**

**Please provide your age (in years):**

**I identify as:**

- Male

- Female
- Prefer to self-describe
- Prefer to not say

*The purpose of this study is to examine the processes of interviewing and interrogation. You will evaluate a recording of a criminal interrogation between a human suspect and a robot interrogator. The recording is roughly 3-4 minutes in length. Each interview is related to the same crime: the theft of a laptop in Room 100 of the 'Buys Ballot' building.*

*Please listen to each recording carefully and completely. After listening, you will be asked to answer various questions regarding the recording.*

**Please listen to the following audio file:**

**I judge the suspect to be:**

- Innocent
- Guilty

**How certain are you of this judgement?**

- 1-10, not certain at all-very certain

**The interrogator judged the suspect to be:**

- Innocent
- Guilty

**How certain are you of this answer?**

- 1-10, not certain at all-very certain

*To what extent do the following statements apply?*

- **The interrogator presumed the suspect's guilt at the outset.**
  - 1-10, not at all-very much so
- **The interrogator tried hard to get a confession.**
  - 1-10, not at all-very much so
- **The interrogator put pressure on the suspect.**
  - 1-10, not at all-very much so

*To what extent do the following statements apply?*

- **The suspect was anxious.**
  - 1-10, not at all-very much so
- **The suspect was defensive.**
  - 1-10, not at all-very much so
- **The suspect firmly denied the accusation.**
  - 1-10, not at all-very much so
- **The suspect's alibi seems plausible.**
  - 1-10, not at all-very much so

*Thank you for participating. Please click the proceed button to be redirected back to Prolific.*