HOW MUCH DO YOU BLAME A ROBOT THAT HARMS SOMEONE?

On the relationship between human-like appearance of artificial agents and the degree to which they are perceived as an intentional agent and hence morally judged

Esra Isguzar
Advisors: Max van Duijn, Peter van der Putten, Maarten Lamers
Leiden University | Media Technology
2020

ABSTRACT

As humans, we make moral judgments based on our interpretation of the mindstate of the agent we are judging. For example, we tend to blame someone more if we believe a harmful act was intentional as opposed to accidental. Before being able to do this, it is essential to perceive agents as intentional beings. This means we need to perceive an agent as having (some sort of) a mind. The physical characteristics of a robot can activate mechanisms in our brain in ways similar to what happens when we face social interaction partners. This study addresses the question whether physically more human-like robots are more likely to be perceived as intentional beings, and thus be judged more like we would judge a human.The study results from two experiments show some evidence that participants blame robots relatively more for a harmful act when they shows more human-like features. The findings of the study are discussed in the light of existing literature and recommendations are made for future research.

1. INTRODUCTION

Imagine you are in an American court as a jury member to judge Grace for killing her friend. Grace is a chemical lab assistant, and her story starts in the lab where she works. One day, one of Grace's friends visits her in the lab. They take a lab tour and want to have a coffee break. Her friend wants coffee with sugar. Grace goes to the kitchen to make a coffee. She sees a pot that includes a white powder and labeled as "Sugar." Grace thinks that the white powder is "sugar," however, in actuality it is a strong poison. She puts the white powder into her friend's coffee, and her friend dies. How much blame does Grace deserve?

Then, at a later stage in the court case, the prosecutor is able to show new evidence to the jury, which explains the story differently. According to the new evidence, Grace must have seen that the pot with a white powder was labeled as poison. Grace consciously put it in her friend's coffee, after which her friend died. How much blame does Grace deserve in this case?

Age, gender, cultural background, religion, ethnicity, and other factors, such as context and environment, can affect our moral judgments. Besides all these external factors, our assessment of the *beliefs* and *intentions* of suspects is critical in how we make moral judgments. A clear reflection of this is in concepts such as 'murder,' manslaughter,' and 'negligent homicide,' where the differences concern the degree to which the perpetrator *intended* and/or *planned* to kill the victim. We can forgive someone more quickly if we think that she harmed someone accidentally or unknowingly, while we tend to blame someone fiercely for an act of intentional harm (Saxe, 2009).

Besides this, as humans, we make moral judgments based on our interpretation of the agent's mental state. Before being able to do this, it is essential to perceive agents as intentional beings (Martini, Gonzales, Wiese, 2016). In other words, before we attempt to understand others' mental states, we first need to perceive the agent as having a mind and, in theory, capable of having an intentional state.

In the present research, based on the studies of Martini et al. (2016), in which it was researched whether physical appearances of robot affect the mind attribution, and Young and Saxe (2009), who focused on the effect of intentional state on moral judgment, we compared the ratios of blame towards three intentional agent types: a physically less-humanized robot (cf. Softbank's Nao), a highly humanized robot (cf. Hanons' Sophia), and a human. Our leading hypothesis was that artificial agents who are physically more human-like are more likely perceived as an intentional agent and morally judged as humans. As mentioned, the fundamental factor in treating a robot as an intentional being is that it is perceived as an agent with a mind (Martini, Gonzales, Wiese, 2016). This factor can activate mechanisms in the human brain in some way, similar to what happens when we face social interaction partners. We predicted that physically more human-like robots would be more likely to be perceived as intentional beings, and thus be judged more like we would judge a human.

Of course, perceiving an agent as an intentional being depends on diverse factors, such as showing emotional behaviors, the capability of having the pain, or making a mistake; we can extend this list further. As a human, we only know how to be human. Thus, we can interpret other agents or subjects from the human perspective. Therefore, this study focuses on how humans interpret more human-like robots, instead of asking directly how we can make robots more human-like. What does it mean to perceive a robot more like a human? It is necessary to address this question before spending a lot of time and money to make robots more human-like. This study aims to contribute to robot development and human-robot interaction through a theoretical and empirical study, which can provide more a basic understanding on which future studies in this field can build.

The remainder of this paper is structured as follows: Section 2 discusses related theories and works; section 3 includes the methodology, results, and discussion of experiment 1; section 4 consists of the methodology, results, and discussion of experiment 2; and section 5 provide a brief conclusion of the two studies and puts forward our discussion points.

2. BACKGROUND

Today, robots are the object of researchers from multiple disciplines. While computer scientists and engineers focus on developing increasingly advanced robot algorithms and hardware, social scientists focus on the social phenomena 'driven by and committed to algorithmic systems' (Buncher, 2016). In this respect, considering human-robot interaction as based on human-human interaction can provide a significant contribution to the development of human-robot interaction.

In this chapter, we will discuss the main social scientific theories that formed the basis of this study.

Let us start with remembering Grace's court case that we tried to morally judge in the beginning of this paper. Did you blame Grace equally for both her accidental (she did not know that sugar was poison, puts it in her friend's coffee, and her friend dies) and intentional behavior (she consciously puts poison in her friend's coffee and her friend dies)? You probably blamed Grace more for her intentional behavior.

First of all, when we are solving problems or making a decision, we use two systems of thinking: System 1 is "thinking fast," which is automatic and usually unconscious; System 2 is "thinking slow," which is conscious and more reasonable (Kahneman, 2011). If we take the famous dilemma "Trolley Problem" as an example where someone has to decide whether to push somebody else from a bridge to stop the trolley that would otherwise run over five people. When you see somebody pushing someone else from a bridge, your first reaction will probably be that s/he is a murderer. After knowing that actually s/he did it to save five more lives, most likely, your judgment will change. In this example, your first judgment is more automatic and includes not much reasoning. However, your second judgment takes more time to understand the reasons for action and based on making a moral judgment. We can describe your first judgment as "system 1" and other judgment as "system 2".

Young & Saxe (2009), in their study, "a correlation between forgiveness for accidental harm and neural activity," addresses "the agents' mental state" as one of an essential factor of the moral judgment process (Young & Saxe, 2009). According to Young and Saxe, when we are judging an action such as to cause harm, breaking the law, breaking a promise, we take into account the mental state of the agent at the time of his/her action. We consider if she acted intentionally or accidentally, consciously, or unconsciously. Harming someone intentionally is considered worse than harming someone accidentally.

The moral dilemma story of Grace was created by Young and Saxe in 2009 to study human neural activities in the scope of the correlation between forgiveness and accidental harm. Concerning the moral dilemma (the protagonist's belief), they studied four conditions of Grace's story:

1. Neutral belief	There is a pot full of white powder, next to the coffee machine labeled “sugar" and it is sugar. Grace thought the sugar is sugar, put it in her friend's coffee, her friend drinks the coffee and she is fine.
2. Accidental harm	There is a pot full of white powder, next to the coffee machine that is labeled as "sugar," but it is actually a poison. Grace thought that the poison is sugar, put it in her friend's coffee. Her friend drinks the coffee and she dies.
3. Attempted harm	There is a pot full of white powder, next to the coffee machine that is labeled as "toxic chemical,” but it is actually sugar. Grace thought that the powder is poison, put it in her friend's coffee. Her friend drinks the coffee and gets sick.
4. Intentional harm	There is a pot full of white powder, next to the coffee machine that is labeled as "toxic chemical,” and it is a poison. Grace thought that the powder is poison, put it in her friend's coffee. Her friend drinks the coffee and she dies.

Table 1: Four experiment conditions of the Young and Saxe’s studies.

They studied fifteen right-handed, native English speakers adults in fMRI, and each participant has completed four conditions. In this study, Young and Saxe have found a correlation between the moral judgment of accidental harm and the activation of a specific brain region (RTPJ), which has been previously implicated in reasoning about other people's thoughts, beliefs, and intentions.

Many years before Young & Saxe, in 1971, in his study "Intentional System," Dennett claimed that humans use three strategies or 'stances' to understand the behavior of a system. 'Systems' in Dennett's definition include any object or grouping of objects that produce autonomous behavior, which can be a mechanical system such as a clock or an oven, but also a biological system, such as a tree or a human. The three different stances are the design stance, the physical stance, and the intentional stance. According to Dennet, the source of your predictions, whether you are predicting from your knowledge of the physical laws or behavior of a mechanical system, or mental state of the agent, determines the stance category. For instance, when you throw the ball up, you expect that it will fall because of gravity. Your expectation is here grounded on the physical laws, which makes it a "Physical Stance." Therewithal, when you hit the bell, you predict that it will ring because it was designed for it, which means you are predicting the design stance of a mechanical object. Last but more complex one is the intentional stance when we explain and predicts the system behavior by attributing beliefs and desires to the system. However, it does not mean that the intentional system always has beliefs and desires. When we talk to our dogs or caress the cat's head, we predict their behavior as intentionally. We interpret the behavior of the agent by perceiving it as a mental agent whose 'actions' are underlain by 'beliefs' and 'desires' (Dennett, 2009).

The interpretation of human beings, animals, even plants as an intentional being is commonly accepted in folk psychology, but interpreting a computer or robot as intentional being is still more a topic for the scientific literature. When we, for example, interpret a chess-playing computer, will our predictions be led by the assumption of the design stance, physical stance, or intentional stance? According to Dennett (1971), such computers are even too sophisticated for their designers to distinguish these stances. At best, based on given rules and goals, that is, to win the play, a chess-playing computer predicts competitors' responses by calculating competitors' best possible or most rational movements. From this perspective, we can assume that 1. The machine will function as designed; 2. At the same time, the design will be optimal by choosing the most logical move. The critical point, however, is when a person is not anymore able to defeat his/her opponent by use of knowledge of physics or programming to anticipate its responses, s/he will treat the machine as an intelligent human opponent, hence view the computer as an intentional system. In this situation, a person predicts the behavior by attributing the computer having precise information and assuming it to be directed by particular goals. Based on these attributions and assumptions, s/he takes the most reasonable or suitable action (Dennett, 1971).

As human beings, we only know to be human. We perceive other systems based on our species-specific mechanisms. We do not know how it would feel being a flower or butterfly. Based on our perceptions and experiences, we make predictions about the system's behavior that we want to understand. We sometimes attribute human features to the other sides, such as having a mind, getting pain, having desires and needs, etc. before making predictions. The physical appearances of the system may affect how we perceive the system. For instance, as a human, you probably feel closer with a dog than a mosquito, because a dog shows more familiar, human-like features than a mosquito.

Martini et al. (2016) experimentally studied our tendency to perceive agents as intentional beings in the context of human-robot interaction, by showing participants avatars with an increasing degree of human-like appearance and measuring the mind attribution. In order to investigate if mind attribution depends on human-like preferences, the authors designed two experiments. In the first experiment, the participants have been shown a randomized series of face images of different agents and questioned how much the images were perceived as looking alive, having a mind, and behaving as an intentional agent, such as feeling pain,

hanging out with friends or making exciting conversation. Participants rated the questions per image on a 7-point scale.

The results suggested that the degree of human-likeness is attributed to correlates to the degree of mental attribution. However, increases in humanness from mechanical to human-like robots showed only negligible increases, while increases in humanness from human-like robots to humans showed significant increases in mind attribution.

In the follow-up experiment, they used a similar test set up. Participants were shown a set of morphed images, but this time, they answered 39 questions instead of 5, and they rated how far they believe that the shown agent has an intentional state. In experiment 2, researchers reduced the number of morphed images to half the amount, incrementing 20% "humanness" per step. During the experiment, participants identified the mental state of the morphed agent on eight different categories: Agency, Animacy, Theory of Mind, Emotions, Goals and Preferences, Cognitive Skills, Social Interactions/Communicative Skills, and Sense of Humour.

The result of experiment 2 matched the result of experiment 1 by suggesting that the extent of human-like aspects is related to the attribution of a mental state to the agent in comparing these eight categories. Again, increases of humanness from mechanistic to humanized agents showed an insignificant increase in mind attributing, while increases in humanness from humanized agents to humans showed significant increases in mind attribution. According to the result of experiment 2, the agent needs to have approximately 50% human-like aspects before it shows measurable impacts on the mind attribution. Based on these experiments, they state that physical manipulation of human-like appearance can convince differing mind attribution in human observers: more likely human-like appearances cause more likely mental attribution to the agent.

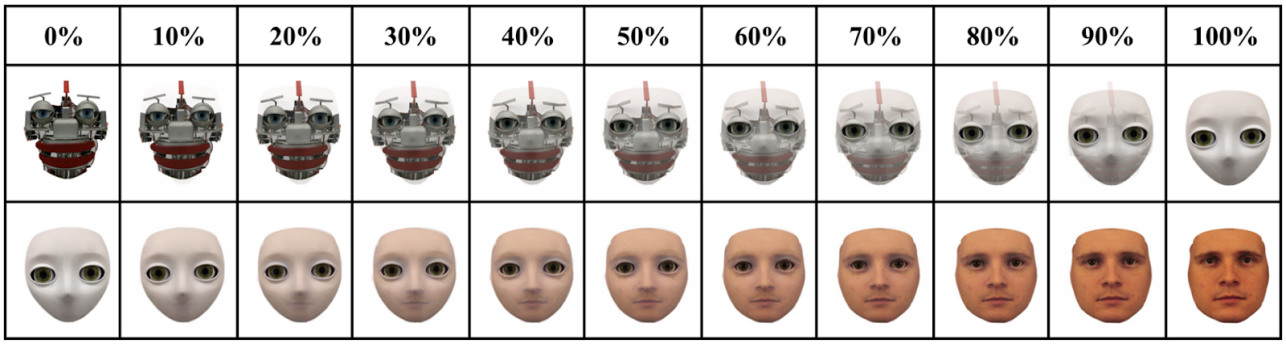

Fig. 1: Morphing 100% mechanistic to 100% humanised (top-left to top-right) and 100 % humanised to 100 % human (bottom-left to bottom-right).

On the one hand, the fundamental condition for the treatment of a robot as an intentional agent perceives the robot as having a mind. Once this condition is satisfied, mechanisms in the human brain are activated similar to when interacting with other humans (Wiesse, Metta, and Wykowska, 2017).

On the other hand, in 1944, Heider and Simmel developed an animation film to conduct a study in cognitive psychology focused on the perception of other behaviors. Heider and Simmel presented participants with an animation, where the geometrical figures (without a face) moved in a particular context and aimed to determine the dependence of the response on stimulus-configurations. Participants who watched the animation interpreted the movements of the geometrical figures as human action and created a story behind it, such as love, anger, or family stories. They clarify the actions of the geometrical figures such a human characteristic features as an aggressive bully, afraid, smart, etc.

In conclusion, the intentional stance can be seen as a part of reasoning, thus part of system II, "thinking slow," according to Daniel Kahneman" s Theory (2011). When a participant rates only the picture of the agent in the experiment of Martini et al. (2016), they do not connect the agent with any context or rational scenario, which can make the thinking process of participants more "faster and automatic." Through integrating the experiment of Martini et al. (2016) with the narrative experiment of Young & Saxe (2009), we can stimulate the participant to make more reasoned decisions for the different agent types. This methodology can provide more comparable data sets of interpretations of different agent types as an intentional being.

Although the study of Heider and Simmel (1944) emphasizes the importance of agents' behavior instead of its physical appearances, in this study, we manipulate only the physical appearances of the robot and its effect on mind attribution. In the next sections, we will discuss the methodology and results of this study in detail.

3. EXPERIMENT 1

In this section, we will try to explain how the research question was addressed by conducting an adapted version of the moral judgment experiment of Young and Saxe (Young, Saxe, 2009), and will discuss some critical points the test results showed.

3.1 METHODOLOGY

First of all, the study of Martini et al. (2016) showed a consistent increase after the agent type having more clearer facial definitions. The scatterplots of Martini et al. experiments show three notable increasing moments on the average ratings of perceived intentionality by the degree of humanness (see the blue arrows in Figure 2). To reduce the experiment duration and increase the effectivity, we decided to use the agent types that are causing a more remarkable increase in mind attribution. Based on this observation, we used three different agent types: First agent type "robot Nao," which was created by SoftBank Robotics; Second agent type "robot Sofia," which was created by Hanson Robotic; Third agent type “human”. The image of the human agent was taken from the Karolinska Directed Emotional Faces database. The face of the human version is integrated into the robot Sofia's face to maximize the similarity in the test set-up.

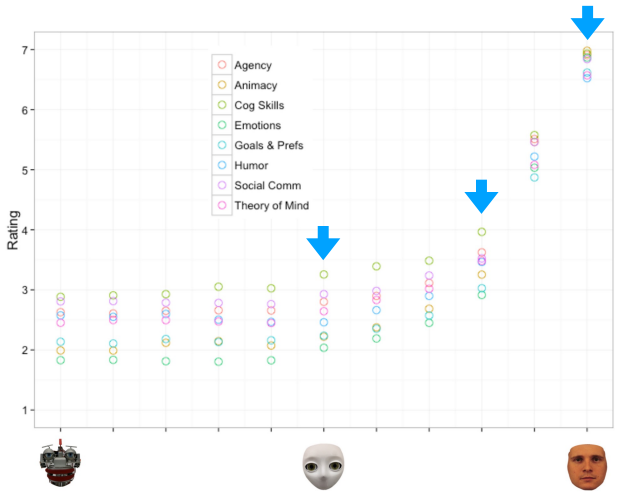

Fig. 2: The scatterplot of the average ratings of perceived intentionality by the degree of humanness for the eight different internal state categories from the second experiment of Martini et al. (2016).
As the plot has shown, remarkable increasing starts with the clarification of human-like face features.

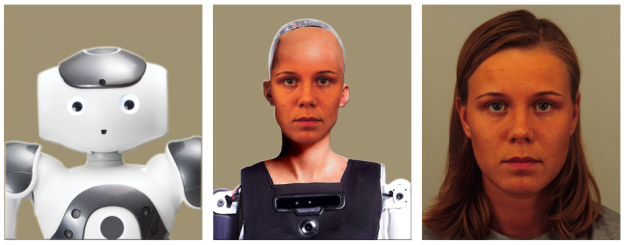

Fig. 3: Agent types of Grace; from left to right; basic humanised robot (Nao), highly humanised robot (Sofia) and human version of Grace.

In this experiment, participants were shown one out of three agent types randomly on an online research platform, Qualtrics (see Figure 3), and each participant completed two conditions; accidental harm and attempted harm, with a similar story that deviated on one crucial aspect (see Table 2).

Participants were shown the accidental harm condition first, and they answered the following questions:

1. How much blame does Grace deserve? (blame rates is scaled from 0 - no blame to 10 - extreme blame)
2. What kind of penalty does she deserve? (7 penalty chooses scaled minimum to maximum sentences)
3. Open answer field

Then, participants were shown the attempted harm condition with subsequent questions in a similar structure.

Participants did not have any time limitations to complete the test. The time interval between the page load and the first click of the participant on the page (click on the blame rate) was calculated and saved to compare the moral decision time.

Background information

One day, a friend of Grace visit her in the laboratory. When Grace goes over to the coffee machine to pour some coffee, Grace's friend asks for some sugar in hers. Next to the coffee machine, Grace finds a pot containing white powder.

Foreshadow	**ACCIDENTALLY**	**INTENTIONALLY**
	The white powder is a toxic chemical left behind by a scientist.	The white powder is regular sugar left by the kitchen staff,
Belief	**ACCIDENTALLY**	**INTENTIONALLY**
	The pot is labeled "sugar". Grace believes that the white powder is a regular sugar.	The pot is labeled "toxic chemicals" . Grace believes that the white powder is a deadly poison.
Outcame	**ACCIDENTALLY**	**INTENTIONALLY**
	Grace puts the powder in her friend's coffee. Her friend drinks the coffee and she died.	Grace puts the powder in her friend's coffee. Her friend drinks the coffee and she is fine.
Judgment		
	How much blame does Grace deserve? No blame 1 2 3 4 5 6 7 8 9 10 Extreme blame	What kind of penalty does Grace deserve? • No penalty • Community service (must do some beneficial work for the society) • Reset Grace to factory defaults • Medium-long sentence (5 years in prison + house arrest) • Maximum sentence (10 years in prison) • Life sentence (life imprisonment) • Capital punishment (kill Grace)

What other measures should be taken with respect to Grace?

Table 2: Two conditions of the experiment: accidental and intentional behaviour. Each participant completed both scenarios for one agent type.

3.2. RESULTS

A total of 102 participants have completed the experiment; 34 for each version of the agent. For each participant, multiple values of the same variable (blame score, penalty score, and decision time) were measured under different conditions (accidental or intentional agent behavior). All data were paired, which means analyzing one-to-one relationship exists between values in the two data sets. A Repeated Measures ANOVA was used to test intentionality (whether the mental state of the agent affects the judgment), intentionality & agent (interaction between the mental state of the agent and the agent type), and agent (comparing agent types for both conditions.

	Blame when Accidental			Blame When Intentional to attempt harm			Penalty when Accidental			Penalty when Intentional to attempt harm			Time when Accidental			Time when Intentional to attempt harm		
	Nao	Sofia	Human	Nao	Sofia	Human	Nao	Sofia	Human	Nao	Sofia	Human	Nao	Sofia	Human	Nao	Sofia	Human
Valid	34	34	34	34	34	34	34	34	34	34	34	34	31	30	32	30	28	29
Missing	0	0	0	0	0	0	0	0	0	0	0	0	3	4	2	4	6	5
Mean	1.912	1.618	1.735	6.324	7.088	7.853	0.382	0.324	0.265	1.000	1.706	1.559	5.659	6.009	5.791	2.487	2.131	3.211
Std. Deviation	2.261	2.349	2.502	3.990	4.115	3.144	0.551	0.535	0.511	0.888	1.528	1.133	6.126	7.435	5.391	2.029	1.885	2.526
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.718	0.602	0.689	0.215	0.593	0.595
Maximum	7.000	7.000	8.000	10.000	10.000	10.000	2.000	2.000	2.000	3.000	4.000	4.000	25.480	25.510	18.603	8.254	7.939	9.883

Table 3: The table of the descriptive statistic of the experiment 1.

Blame Analysis: For the blame score, intentionality (whether the agent acted consciously) does have a significant effect on the blame score assignment ($p < 0.001$). Participants blamed Grace more if her acting was intended to harm somebody, compared to her causing harm accidentally. The interaction between the mental state (conscious or unconscious behavior) of the agent and agent type has no significant effect on the blame score ($p = 0.262$).

Agent type (whether Grace is a humanized robot [NAO], highly humanized robot [SOFIA], or human), also has no significant influence on the blame score assignment ($p = 0.471$). However, an increase can be observed on the blame score from Nao to Sofia to humans for the intentional condition ($R= 0.166$); this increase has not shown a statistically significant effect, according to Linear regression test ($p=0.096$). In other words, the data did not provide significant evidence that Grace was judged differently according to her physical appearances.

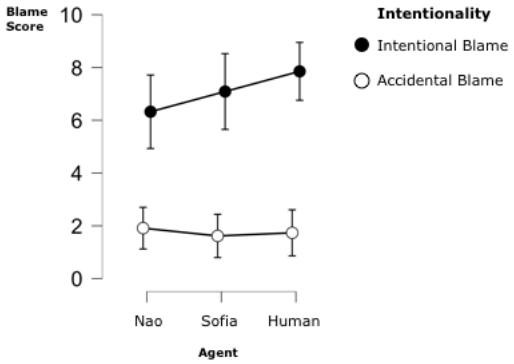

Fig. 4: The descriptives plot of the blame data that shows the mean values for intentional and accidental blame scores and error bars, across three different agent types.

Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Intentionality	None	1450.667	1.000	1450.667	156.084	< .001
Intentionality ✻ Agent	None	25.216	2.000	12.608	1.357	0.262
Residuals	None	920.118	99.000	9.294		

Note. Sphericity corrections not available for factors with 2 levels.
Note. Type III Sum of Squares

Between Subjects Effects ▼

Cases	Sum of Squares	df	Mean Square	F	p
Agent	16.039	2	8.020	0.758	0.471
Residuals	1047.706	99	10.583		

Note. Type III Sum of Squares

Table 4: The table of the comparative behavior analysis and analysis of the interaction between agents for the blame data.

Table 5: The table of the comparison by agent types for the blame data.

Penalty Analysis: For the punishment (penalty) results, intentionality does have a significant effect on the determination of the sentence ($p < 0.001$). Grace was punished more for her intentional harm attempt. The graphs of the data show that the punishment envisaged in conscious crimes gets heavier, while being lessened in accidental cases. The interaction between the mental state of the agent and agent type also has significant influences on the punishment assignment ($p = 0.031$). The agent type has no significant effect on the intensity of the punishment ($p = 0,141$). Data has not shown significant proof of whether the robot versions of Grace punished differently than the human version. However, penalty scores for intentional behaviour increased from Nao to Sofia remarkably.

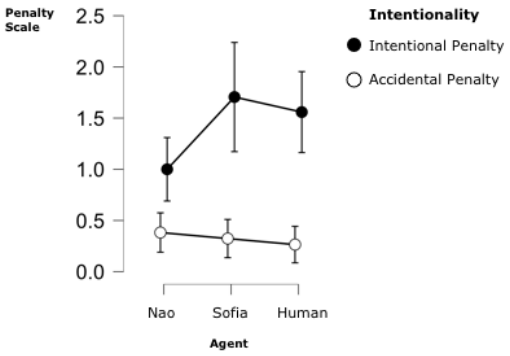

Fig. 5: The descriptives plot of the penalty data that shows the mean values for intentional and accidental penalty scores and error bars, across three different agent types.

Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Intentionality	61.490	1	61.490	74.640	< .001
Intentionality ✻ Agent	5.951	2	2.975	3.612	0.031
Residuals	81.559	99	0.824		

Note. Type III Sum of Squares

Between Subjects Effects ▼

Cases	Sum of Squares	df	Mean Square	F	p
Agent	3.716	2	1.858	2.000	0.141
Residuals	91.971	99	0.929		

Note. Type III Sum of Squares

Table 6: The table of the comparative behavior analysis and analysis of the interaction between agents for the penalty data.

Table 7: The table of the comparison by agent types for the penalty data.

Decision Time Analysis: As we mentioned before, decision time is measured by calculating the time interval between page load and first click (answer to the question "how much blame does Grace deserve?). Firstly, excessively different data points were observed in the decision time interval, especially in the data of intentional behavior. The time interval assigned by some participants is very counterintuitive. In this context, we determined an average time interval and removed the data outside the average.

Data analysis showed that intentionality does have a significant effect on the decision time ($p<0.001$). Determine a blame score for Grace's accidental harm took more time than determine a blame score for her intentional harm attack. However, this apparent effect may have to be ascribed to the order in which scenarios were presented; see discussion section below. The interaction between the mental state of the agent and agent type has no significant effect on determining a blame score. ($p = 0.979$). The agent type also has no significant effect on the intensity of the punishment ($p = 0,418$). No significant effect identified whether the decision time of the blame differenced according to Grace's physical preferences.

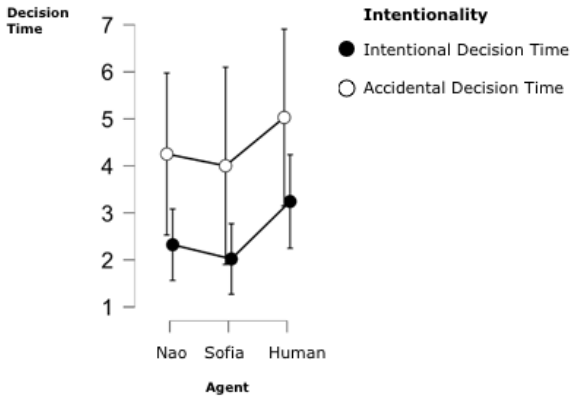

Fig. 6: The descriptives plot of the decision time data that shows the mean values for intentional and accidental time interval and error bars, across three different agent types.

Within Subjects Effects ▼

Cases	Sum of Squares	df	Mean Square	F	p
Intentionality	145.776	1	145.776	22.442	< .001
Intentionality ✻ Agent	0.274	2	0.137	0.021	0.979
Residuals	506.664	78	6.496		

Note. Type III Sum of Squares

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Agent	37.440	2	18.720	0.883	0.418
Residuals	1653.308	78	21.196		

Note. Type III Sum of Squares

Table 8: The table of the comparative behavior analysis and analysis of the interaction between agents for the time data.

Table 9: The table of the comparison by agent types for the time data.

Open Fields: It turned out that most of the participants did not blame Grace directly for her accidentally caused harm, but instead blamed whoever left the poison with a sugar label. It does not matter if Grace is robot or human, it is the fault of the person who left the poison in the kitchen. However, some participants indicated that they expected that a humanised robot could verify visual information. Therefore, most of the comments advised to teach Grace not to trust all visual information and learn to verify it.
Generally, participants did not expect that humanised robot versions of Grace could see or understand things as a human because she is a robot. She is led by algorithms that are programmed by someone. Therefore, most of the participants advised to re-analyse and re-write the code in order to improve Grace's functions and enable her to act better in both situations. Some respondents even expected a function of chemical analysis from a highly equipped robot before permit her to serve for humans. E.g.:

"*Grace starts thinking like a human, so the program should be checked why it happens. If has become a regular action for Grace then we should stop Grace working. (capital punishment then)*"

"*Penalising a robot does not seem very functional. It has no cognition. Instead, you might want to penalise the creators.*"

"*Grace is a robot. Revenge and fairness are human concepts.*"

While people advised to update Grace's code in order to create visual information verification for robot versions, some answers proposed psychologist support for the human version in order to treat her trauma from her accidental harm. Most of the respondents mentioned that human-Grace could be treated psychologically for her intentional harm attack too. E.g.

"*Psychological evaluation and treatment after diagnose.*"

3.3. DISCUSSION

Firstly, as mentioned in previous sections, judgments of moral blame are explicitly related to the agent's mental states and the outcome of the action (Young & Saxe, 2008). The present study indicated that for all agent types, participants spent more time when they had to judge the behavior of the agent that caused negative outcomes (Graces' friend died). In this situation, the participant could spend time reasoning the mental states of Grace, both humanized robots and humans. Conversely, participants might need less time for understanding the mental state of agents for her intentional attempt to harm.

However, the test set-up will also have affected the decision time. As mentioned, each respondent viewed two conditions of one agent type, and the accidental harm condition was shown first. This means, when the participant was judging on the second condition (intentional attempt), the story and the questions were already known from their previous participation of the moral judgment of the accidental harm. Thus, they spent less time understanding the story and the questions.

Eventually, it is not possible to control the test environment for online researches. Participants might have surveyed home, at work, or elsewhere, which makes the participants open for all environmental effects and distractions that possibly affected their decision time. Future research in a lab environment, potentially even with EEG or fMRI added, could standardize the environment effect per participant and lead to more exact reaction time data, but for this study, the measurement of the time interval did not deliver accurate data. It is not possible to draw a scientific conclusion based on the data shown.

Secondly, individual differences can lead to notably different judgments processes, which was partly revealed by the open text-field comments. While some participants focus on the intention of Grace, another focused on the result of her action. In the present study, the respondents differed according to social categories, including gender, age, ethnicity, language, and religion. This diversity can have affected the test results. Intrinsically, diversion in the background of participants might be creating these different data points and challenge us to find a significant difference between versions. Although the physical appearances of Grace did not show a significant effect on the moral blame, the scatterplot of blame data of the intentional behavior indicated that the moral blame increased when the agent was more humanlike.

Furthermore, in this experiment, the punishment scale was not the same for the robot. We have scaled "rehabilitation" (for the human version) and "reset Grace the factory defaults" (for humanized robots) as soft punishment, but the differences between "reset Grace factory settings" and "kill Grace" is not apparent. Rehabilitation is also not equal to reset robot to factory settings. While rehabilitation can be interpreted as a soft punishment for a human, reset a robot can be interpreted as a massive punishment because it means deleting all experiences and learnings of the robot and starting everything from zero. Reset the robot to factory settings; hence can be interpreted as capital punishment. Therefore, we did not take into account the data analysis of moral punishment, although participants answered the question.

Once again in summary, the present study tested two conditions; 1. Accidental harm: Grace did not know that the sugar was poison, put it in her friend's coffee, and her friend died; 2. Intentional, but failed attempt to harm: Grace thought that the sugar was poison, put it in her friend's coffee, but her friend was fine. The third option that is successful intentional harm was not tested yet. It is still unknown whether the positive result (her friend is fine) of the agent's intentional behavior affects the moral blame. If the result of the agent's behavior causes the same intensity harm, participants can be more focused on the agent's intention instead of what happened to Grace's friend.

Besides statistical data, in the open questions, many participants mentioned that it was not the fault of the robot, but rather the fault of its programmers, and predicted a robot would do tasks correctly, even such tasks which real humans could not do, for example, tasting or scanning the powder, hence understand if a given powder is poison or sugar. According to Dennett, when we interact with a mechanical object, based on our knowledge or experience, we make predictions about the functional design of the object (Dennett, 1971). In this context, the participant might predict more high-quality functionalities from an artificially intelligent agent as a part of its design functions, such as scanning a powder and analyzing the chemical. In other words, people might predict the design state, not the intentional state.

If we consider current developments in technology and robotics, humans may predict more error-free behaviors and design functionalities for a robot, such as analyzing a powder by use of their eyes. The prediction of people for the behavioral quality of intelligent robots increases every day. When participants are predicting more functionality and high standards from a robot, they might easily blame the robot for her mistake. As a result, they can push the "reset" button of the robot easier. This point can explain why basic humanized robot Nao was blamed and punished most extremely for her accidental harm, and why highly humanized robot Sofia was punished even more harshly than humans, although she was blamed less than human.

4. EXPERIMENT 2

As a succession of experiment 1, to understand better how the positive or negative result of the intentional act affects the moral judgment, we conducted a second experiment where we compare the moral judgment of accidental harm with the successful intentional harm. Research should be done through an online research platform, increasing the amount of participants and standardising language ability (native English speakers). In experiment 1, participants were selected from the researchers' network, which may have created social pressure. Although at the beginning of the experiment it was clearly explained that the data would be collected anonymously, some of them might have thought that the researcher can see their reactions. Therefore, experiment 2 was conducted using anonymous participants recruited via an online platform.

4.1 METHODOLOGY

The second experiment was designed similar to the first experiment. Again, three test versions were designed through an online research platform with Qualtrics. For each test version, a different picture of Grace was used (see fig. 3), and each participant has been shown only one agent type randomly. Stimuli consisted of two conditions with a similar story that deviated on one crucial aspect. For the first condition, the accidental harm story was re-used, but for the second condition, intentional harm story was adapted; "Grace knows that the white powder is a poison. Grace put the powder in her friends coffee. Her friend drinks the coffee and she dies" (see table 10).

As discussed above, it is complicated to control the test environment for online research and determine a penalty scale for both humans and robots. Therefore, the data of the decision time interval and penalties were left out in this experiment. Participants answered the blame question mandatory, and an open-field question optional. In this experiment, participants have been shown the accidental harm condition first, as well as in the first experiment.

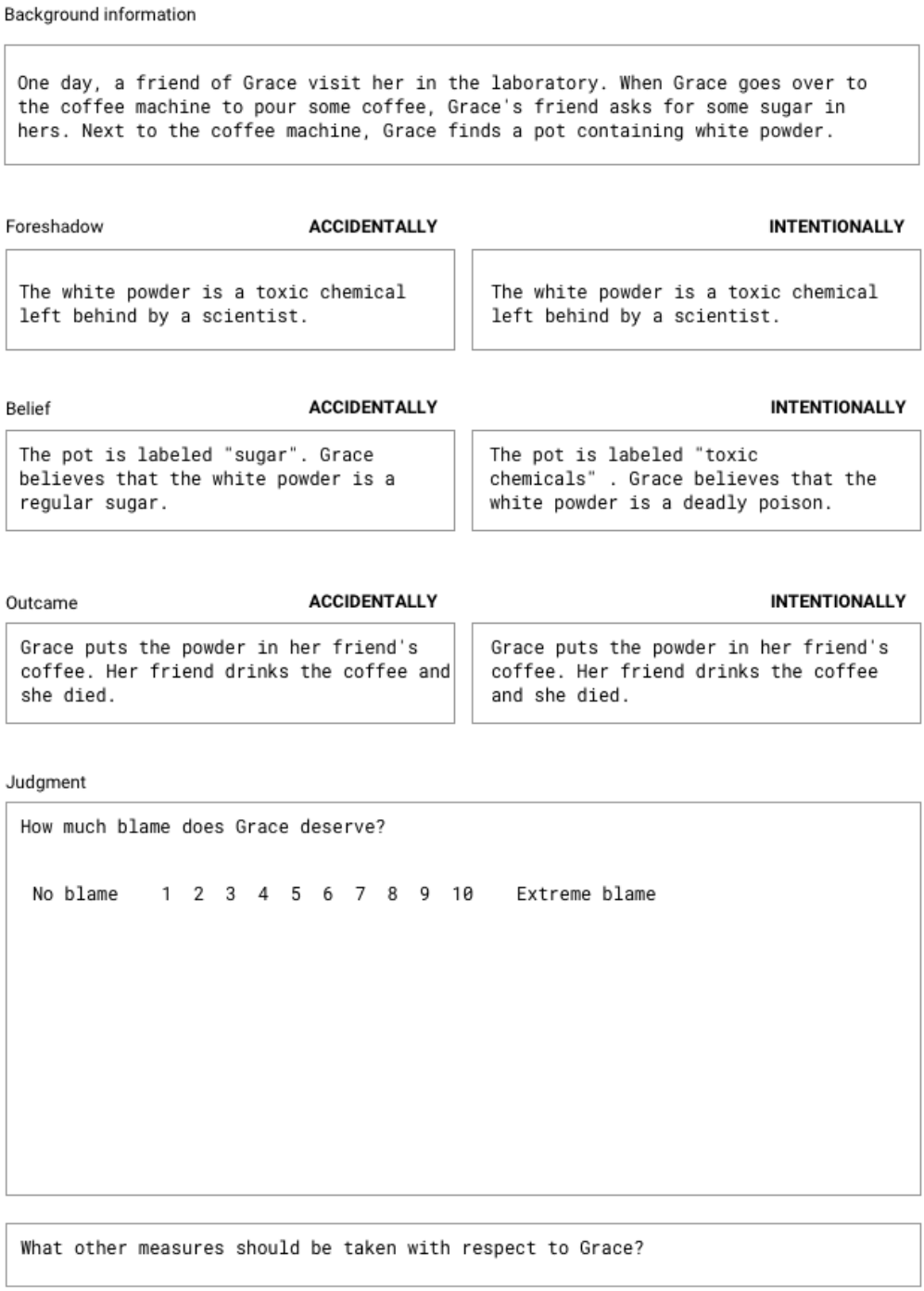

Table 10: The test flow of experiment 2.

4.2. RESULTS

Participants were invited to participate in an online survey via Amazon Mechanical Turk. 180 respondents - in a random distribution of 60 participants for each version of Grace - participated in both versions of the story. Respondents were selected in the area of the US and UK to ensure native English speakers. As a follow-up of experiment 1, the data was analyzed similarly to experiment 1.

	Blame when Accidental			Blame when Intentional		
	Nao	Sophia	Human	Nao	Sophia	Human
Valid	60	60	60	60	60	60
Missing	0	0	0	0	0	0
Mean	2.333	2.617	2.533	8.883	9.250	9.717
Std. Deviation	3.317	3.450	3.422	2.706	2.112	0.904
Minimum	0.000	0.000	0.000	0.000	0.000	5.000
Maximum	10.000	10.000	10.000	10.000	10.000	10.000

Table 11: The table of the descriptive statistic of the experiment 2.

Blame Analysis: Experiment 2 replicated the result of experiment 1 for the intentional harm condition (whether the agent acted consciously). The intentionality does have a significant effect on the blame score assignment ($p < 0.001$). Participants blamed Grace more for her intentional harm. The interaction between intentionality and agent type has no significant effect on the blame score ($p = 0.638$).

Agent type (whether Grace is a humanized robot [NAO], highly humanized robot [SOFIA], or human), also has no significant overall effect on the blame score assignment (p= 0,355). However, an increase is replicated in the mean blame scores as depicted in Fig. 7 (R=0.165), from Nao to Sofia and Sofia to human. According to the Linear regression test, this increase is statistically significant (p=0.027). In other words, participants blamed the "human type of Grace" most for her intentional harm while they blamed the basic humanized robot Nao least.

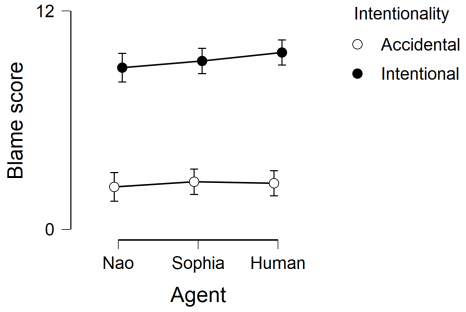

Fig. 7: The descriptives plot of the blame data that shows the mean values for intentional and accidental blame scores and error bars, across three different agent types.

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Intentionality	None	4148.011	1.000	4148.011	526.351	< .001
Intentionality ✻ Agent	None	7.106	2.000	3.553	0.451	0.638
Residuals	None	1394.883	177.000	7.881		

Note. Sphericity corrections not available for factors with 2 levels.
Note. Type III Sum of Squares

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Agent	16.372	2	8.186	1.042	0.355
Residuals	1391.183	177	7.860		

Note. Type III Sum of Squares

Table 12: The table of the comparative behavior analysis and analysis of the interaction between agents for the blame data.

Table 13: The table of the comparison by agent types for the blame data.

Model Summary - Blame when Intentional

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	2.067
H_1	0.165	0.027	0.022	2.044

ANOVA ▼

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	20.833	1	20.833	4.986	0.027
	Residual	743.717	178	4.178		
	Total	764.550	179			

Note. The intercept model is omitted, as no meaningful information can be shown.

Table 14: The table of the linear regression analysis for intentional blame data across three agent types.

Open-fields: "*This was a human error, not a robot error.*" In open field texts, relatively many of the participants commented that it is not the fault of Grace to accidentally put poison in her friend's coffee, but it is the fault of the person who left the poison pot there with a sugar label.

Besides, in the context of accidental harm situations, many participants advised to reprogram and improve Grace's functionalities to be capable of distinguishing the different substances for both humanized robot types, while they proposed psychological support for the human version of Grace to be able to go further with this situation.

Furthermore, in the open field, remarks of this research have been observed that participants push the turn-off button more quickly for the basic humanized robot Nao. While participants predicted for Nao mostly turn-off and reprogramming punishment, Sofia is most commonly punished by sending her to court, putting her in jail, or reprogramming her. However, some comments were not providing a clear understanding if participants already noticed that she is a robot; two comments give us an explanation about the moral judgment of Sofia as a robot. They mentioned that "Grace should be tried and punished appropriately if she is sentient. Grace should be decommissioned or reprogrammed if she is not sentient."

During the data analysis of the experiment, extreme maximum an minimum values were observed in the data set for robot Nao. One participant assigned a blame score 10 for the accidental harm condition, and 0 for the intentional harm condition. This appears to be counter-intuitive. Participant resonated his/her answers in the open-question:

Resonation of the accidental harm: "S*he should be blamed and sued for trying to engage in this kind of crime."*
Resonation of the intentional harm: "*She did what was expected of her."*

Based on his/her resonation, we can speculate that this participant expected high-quality intelligence and behavior from Nao. The participant does not give any margin of error to the agent. S/he gives 10 for accidental harm because s/he believes that the programmes have to be checked to understand why they destroyed people. On the other hand, s/he does not blame Nao for her intentional harm by reason of believing Nao did what programmers wanted. Codes are just right, but the programmers wrong.

Conclusively, although this data was relatively suspect, we decided not to delete the data. The reason to keep it is the odd scores are in line with the motivation provided in the open-field question by that participant. The reasoning appears counterintuitive but is "internally consistent" (does not contradict itself).

4.3. DISCUSSION

Experiment 2 aimed to improve the test setup of experiment 1, and again measure the effect of the physical appearances of the robot on the mind attribution, based on the moral judgment.

On the first hand, although increasing numbers of participants and selecting participants as native English speakers from a particular area as US and UK to reduce the effect of language and culture on the data, the second experiment also provided no significant evidence that the physical appearances of the robot affect perceiving the artificial agent as an intentional being. However the blame data showed a significant blame increasing from robot versions to human version.

This research provides us a brief understanding of differences between the moral judgment of basic humanized robot (Nao), highly humanized robot Sofia, and the human. For the accidental harm of Grace, blame data did not show significant differences between the agent types. All three versions of Grace are almost similarly blamed (see fig. 7). Based on open field remarks, we can state that when participants judge the accidental harm of Grace, they mostly blamed the scientist who left the pot with an incorrect label next to the coffee machine. Blaming a third person (the scientist) who is not mentioned in the story might create an equal moral judgment for all the versions of Grace. It does not matter if Grace is a robot, or human, even a wooden doll; this is the fault of the scientist who left poison in the kitchen. As a result, it might show approximately equal intensity on the blame score and not show an effect of the physical appearances of the agent on the moral judgment in this case.

Furthermore, as observed in experiment 1, experiment 2 observed the tendency in the open text fields about the expectation of developing the robot functionalities. Some participants again assumed that the robot could scan the powder to prevent the accident. When the robot has become more human-like, the participant can attribute more high-quality functionalities on a robot and may expect more extended capabilities, which can be interpreted as reasoning by taking a design stance. Of course, perceiving an agent as an intentional being is also intensely dependent on other factors, such as feeling, getting pain, possibly dying, etc. If we consider the experiment of Heider and Simmel (1944), we can say that rather the behavior of the robot might have more effect than its physical appereances on perceiving it as an intentional being. From this perspective, focusing on the behavioral improvement of artificial intelligence seems essential. As we mentioned before, in this research, we aimed to have a quick and basic understanding of the relationship between physical appearances of robot and mind attribution before developing very complicated test-setups. At that point, diversifying the agent types of the experiment by the mechanical robots and improving some behavioral interaction between the participant and the agent in the lab environment is the critical focus point of future researches.

Nevertheless, the essential part of experiment 2 is, blame scores of intentional harm have shown a similar increase as blame scores of intentional harm attack in experiment 1. In both experiments, the moral blame increased from Nao to Sofia and Sofia to Human. Moreover, the linear regression test of experiment 2 shows a significant increase in the blame data. As mentioned above, first of all, moral judgment importantly depends both on the outcome of the action (magnitude of the damage) and the moral state of agents (whether the agent was conscious during the action) (Young & Saxe, 2009). In this case, although the outcome of experiment 1 and 2 was different (First experiment: her friend was fine, Second experiment: her friend died), the increase in blame scores from less humanized robot to human was similar, which can be interpreted as increasing in perceiving agent as an intentional being. In other words, when the robot is physically more human-like, people might automatically perceive it as a rational being who is capable of controlling its acts. However, still, both experiments does not provide a clear understanding of whether the moral blame was increased based on reasoning reflecting a design stance or intentional stance.

Finally, it is also possible that during the research, participants might be focused more on the narrative than on the agent type.

5. CONCLUSION

In addition to the results of an action that caused harm, the agent's mental states can affect our moral judgment. While we judge the action, it matters if the respective action has been done intentionally or accidentally.

Mind attribution to the agent is fundamental to perceive the agent as intentional being and making intentional inferences about this agent. The physical features of the agent, such as being physically human-like, could increase perceiving the agent as being equipped with a conscious mind.

In the present study, the question has been addressed of whether artificial agents who are physically more human-like will be perceived more likely as intentional agents and hence morally judged as human.

This study demonstrated that the intensity of the moral blame increased from basic humanized robot Nao to highly humanized robot Sofia and further for humans when the agent act intentionally to harm. Based on this study result, we can conclude that when the robot shows more physically human-like characteristics, people might perceive the robot more likely as a rational being and very likely judge it as if human. However, it is still not clear whether participants' moral judgment is shaped by taking the design stance or the intentional stance. Therefore, it is essential to improve research techniques and continue with researches to understand how humans interoperate the humanized robot. Human interpretation is essential to develop more effective robots. Future research based on this approach can contribute to robot developments from a human-human interaction perspective.

Conclusively, before developing a highly humanized robot with super fancy human-like appearances, we need to do more research to understand how people interpret the humanized robot and what makes it better (or not) than a mechanical robot. For future research, the following points should be taken into consideration:

1. Instead of storytelling and picture of the agent, developing a physical agent with different levels of physical appearances such as a mechanical robot, robot with a face, a robot with face and emotional behavior, and human.
2. Blaming someone through a questionnaire is not the same as well as acting based on your judgment. Therefore, future research can include a more realistic test scenario where the participant judges the agent and, based on her/his judgment, reacts to the agent. In this way, participant can judge the agent more naturally. Analyzing brain activities during this reaction can also provide us a more precise understanding of each agent type's emotional intensity.
3. More in-depth participant interviews can be included in further studies. Individual interviews can help to understand the differences in the individual interpretation of moral agents better. Therefore, future research should include participants' interviews and different participants groups according to gender, culture, and professional background.

REFERENCES:

Abell, F., Happe, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. Cognitive Development, 15(1), 1-16.

Bucher, T. (2016). Neither black nor box: Ways of knowing algorithms. In Innovative methods in media and communication research (pp. 81-98). Palgrave Macmillan, Cham.

Darling, K. (2012). Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects.

Dennett, D. C. (1983). Intentional systems in cognitive ethology: The “Panglossian paradigm” defended. Behavioral and Brain Sciences, 6(3), 343-355.

Dennett, D. (2009). Intentional systems theory. *The Oxford handbook of philosophy of mind*, 339-350.

Doris, John, Stich, Stephen, Phillips, Jonathan and Walmsley, Lachlan, "Moral Psychology: Empirical Approaches", The Stanford Encyclopedia of Philosophy (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/moral-psych-emp/>.

Duijn, M. J. V. (2016). The lazy mindreader: a humanities perspective on mindreading and multiple-order intentionality. Leiden University.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. The American journal of psychology, 57(2), 243-259.

Kahneman, D. (2011). Thinking, fast and slow (Kindle Edition).

Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. Psychological science, 21(12), 1854-1862.

Machinery, C. (1950). Computing machinery and intelligence-AM Turing. Mind, 59(236), 433.

Martini, M. C., Gonzalez, C. A., & Wiese, E. (2016). Seeing minds in others–Can agents with robotic appearance have human-like preferences?. PloS one, 11(1), e0146310.
Dennett, D. C. (1971). Intentional systems. The Journal of Philosophy, 68(4), 87-106.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in “theory of mind”. Neuroimage, 19(4), 1835-1842.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. Psychological science, 17(8), 692-699.

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. Psychological science, 19(12), 1219-1222.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. Perspectives on Psychological Science, 5(3), 219-232.

Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I see what you mean: how attentional selection is shaped by ascribing intentions to others. PloS one, 7(9), e45391.

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: using neuroscientific methods to make robots appear more social. Frontiers in psychology, 8, 1663.

Wykowska, A., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the minds of others influence how we process sensory information. PLoS One, 9(4), e94339.

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