Linköping University | Department of Computer and Information Science
Bachelor's thesis 18 HP | Cognitive Science
Autumn term 22 | LIU-IDA/KOGVET-G--22/026--SE

I don't know because I'm not a robot:

A qualitative study exploring moral questions as a way to investigate the reasoning behind preschoolers' mental state attribution to robots

Oscar Amcoff

Supervisor: Tom Ziemke
Examinator: Sam Thellman

LINKÖPINGS UNIVERSITET

COPYRIGHT

ACKNOWLEDGEMENTS

I would like to thank my supervisor Prof. Tom Ziemke for the superb supervision during this project. I would also like to thank Sam Thellman for the generous additional help during this project. I am especially thankful for the meetings we had when things got stressful! Also, I would like to thank the other group members for the helpful tips.

ABSTRACT

Portrayals of artificially intelligent robots are becoming increasingly prevalent in children's culture. This affects how children perceive robots, which have been found to affect the way children in school understand subjects like technology and programming. Since teachers need to know what influences their pupils understanding of these subjects, we need to know how children's preconceptions about robots affect the way they attribute mental states to them. We still know relatively little about how children do this. Based on the above, a qualitative approach was deemed fit. This study aimed to (1) investigate the reasoning and preconceptions underlying children's mental state attribution to robots, and (2) explore the effectiveness of moral questions as a way to do this.

16 children aged 5- and 6-years-old were asked to rate the mental states of four different robots while subsequently being asked to explain their answers. Half of the children were interviewed alone and half in small groups. A thematic analysis was conducted to analyze the qualitative data.

Children's mental state attribution was found to be influenced by preconceptions about robots as a group of entities lacking mental states. Children were found to perceive two robots, Atlas, and Nao, differently in various respects. This was argued to be because the children perceived these robots through archetypal frameworks. Moral questions were found successful as a way to spark reflective reasoning about the mental state attribution in the children.

Contents

1. INTRODUCTION

Children seem to ascribe human-like characteristics to robots. We still know relatively little about how children come up with these ascriptions. An increasing amount of robot portrayals in children's culture have been suggested as a possible influence. One way to investigate the effect of children's preconceptions about robots is by qualitatively investigating the ways they attribute mental states to robots. To our knowledge, no qualitative studies have been done on this topic. The link between mental state attribution and morality is supported by the literature. A few recent studies have found relationships between children's mental state attribution and their moral concern for robots.

Guided by the above, we thought of using moral questions as a way to investigate how children come up with attributions of mental states to robots. Through an explorative qualitative approach, this study investigates the reasoning and preconceptions underlying 5- and 6-year-old children's mental state attribution to robots.

1.1 Mental State Attribution

In psychology, the ability to attribute mental states to others is commonly known as *theory of mind*. Interestingly, this ability extends beyond the attribution of mental states to others: people tend to think in intentional terms about inanimate objects such as moving squares and triangles (Heider and Simmel, 1944). Similarly, people tend to *anthropomorphize* animals and robots. This refers to the tendency to attribute human characteristics (such as mental states) to non-human entities (Bartneck et al., 2020). Anthropomorphizing in adults has been widely observed in the field of human robot interaction (HRI) (Manzi et al., 2020). [1]

There is a decent amount of research on children's mental state attribution to robots. A review (Thellman et al., 2022) found that thirty-five studies have observed a tendency to attribute mental states to robots in children. There are indications that there could be differences regarding the extent to which children of different ages attribute mental states to robots. Younger children (three-year-old's) have been found to attribute mental state to robot to a greater extent than older children (five- to seven-year-olds) (Manzi et al. 2020; Okanda et al. 2021; Severson and Lemm, 2016). These differences have been suggested to stem from animism errors in younger children (Manzi et al., 2020; Okanda et al., 2021). Children have also been found to attribute mental states to robots to a greater extent than adults (Subrahmanyam et al. 2002; Jipson and Gelman 2007; Okanda et al. 2021). We still know relatively little about why we see these differences, and few qualitative studies have been done in this area. Therefore, using a qualitative approach, this study attempts to investigate the underlying reasoning behind 5- and 6-year-old children's mental state attribution.

[1] The terms "theory of mind" and "anthropomorphism" both include the act of attributing mental states to other entities. For clarity, we will use the words *"mental state attribution*" when referring to the human tendency to attribute mental states to robots, and the word "anthropomorphic" when referring to the behavioral and physical characteristics of robots that are designed to look and behave like humans.

1.2 Robots in Children's Culture

Robots and artificial intelligence are increasingly prevalent in children's-culture (Axell et al., in press). Seven-year-old pupils have been found to mention movies, books, and tv-shows with portrayals of robots when they talk about technology and programming in school. In their descriptions, the pupils talked about the robots as if they had human minds. Teachers need to know what influences children's understanding of technology and programming, and robots seem to be related to that understanding (Axell & Berg, 2022). Therefore, we need to research how children's preconceptions about robots influence their mental state attribution to them.

The increase of robot portrayals in children's culture can be described as an increase in children's *domain-specific experience* with robots. The developmental psychology literature suggests that domain-specific experience is a strong predictor of children's beliefs compared to maturation or general experience (Chi, 1978). This means that children's increasing exposure to robots through consuming children's culture is likely to impact their conceptions about robots. A similar relationship was found in a relatively old study (Bernstein and Crowley, 2008). Due to the recent increase of robots in children's culture, newer studies investigating this relationship is needed. The way these robots are portrayed today could have important effects on how children attribute mental states to robots. Axell and colleagues (in press) investigated how robots are portrayed in children's culture and found that the portrayals were often included anthropomorphic bodies, human intelligence capacities, the use of human language, and a hard time understanding social codes. The human characters in these books, movies, and games would sometimes refer to the robots as any other person, hence attributing mental states to the robots. In other cases, the robots were referred to as "it", indicating that the human in the story conceived of the robot as something akin to an inanimate object. Children's preconceptions about robots are likely formed by these portrayals. Therefore, our study aims to get an inside glimpse into how children's preconceptions about robots affect the way they attribute mental states to them. '

1.3 Physical and Behavioral Variations in Robots

People attribute human characteristics to robots that are not designed to look or behave like humans. For example, robotic vacuum cleaners (Fink et al., 2012). A recent review (Thellman et al., 2022) found that the degree to which people attribute mental states to robots is higher the more they look like humans, and that people are more inclined to attribute mental states to robots when they behave like humans (Thellman et al., 2022). Higher amounts of mental state attribution have been found to be beneficial for HRI (Waytz et al., 2014) and humanoid social robots have been found to work better as social partners when they behave and look more like humans (Manzi et al., 2020). But when robots look too much like humans it results in the well-known uncanny valley effect, which negatively affects the HRI (Mori, 1970). A recent study investigating the uncanny valley effect in children found that it starts having an effect at about nine years of age (Brink et al., 2019).

Not much research has been done on how physical and behavioral variations in robots affect the way children attribute mental states to robots. One study (Woods, 2006) that investigated children's conceptions of 40 different robots found that, in line with the uncanny valley effect, robots with a very human-like physical appearance elicited discomfort in children. They instead preferred robots with a mixed appearance of mechanical and human blended together. These findings were confirmed more recently in a study by (Tung, 2016). In order

to design robots that are suitable for children, we need know to what differentiates children from adults in terms of the way physical and behavioral characteristics affect the way robots are perceived. To gain such knowledge, we need to investigate how children attribute mental states to robots of varying physical and behavioral characteristics. In an attempt to contribute to such knowledge as well as broaden the horizons of our study, we included four robots of varying physical features and behavior in our investigation.

1.4 Morality

Mental state attribution has been proposed to be at the center of morality (Grey et al., 2007; 2012). In children, there seems to be a relationship between mental state attribution to robots and ascription of moral patiency to robots (Melson et al., 2009; Kahn et al., 2012; Severson and Lemm, 2016; Sommer et al., 2019). Guided by the connection between mental state attribution and morality, this qualitative study poses moral questions to children in an attempt to spark reflection around the reasoning and beliefs underlying children's mental state attribution to robots.

It can be hard for young children to think in general or abstract terms, especially about difficult concepts like intentionality and agency. But children use concepts like these in their daily lives. The phrase "*he did it on purpose"* is commonly heard when children explain why a particular act was especially mean. They understand these concepts practically but have a hard time grasping them in abstract terms. This presents an epistemological challenge: how do we formulate questions around mental state attribution that children understand and can reason around?

We thought a possible solution to this problem could be to provide concrete scenarios for the children, that would make it easier for them to understand these concepts. Therefore, we thought of using moral questions since they can easily be formulated as scenarios embedded in the real-world.

1.5 Study overview

We formulated moral questions like "how okay would it be if someone locked the robot in the closet?" or "if the robot lawnmower ran over a hedgehog, would it be the robot's fault then?". Since these questions concern moral agency and moral patiency, they implicitly investigate mental state attribution. Our hope was that these questions would facilitating reasoning around difficult mental-state-related concepts like intentionality, agency, and experience, due to being posed as concrete scenarios. We compare the effectiveness of these moral questions to the effectiveness of explicitly asking children to explain why they attribute certain mental states to the robots. Based on the above, we formulated the following research questions:

1. To investigate the reasoning and preconceptions that produces 5- and 6-year-old children's mental state attribution to robots

2. Explore the use of the following qualitative methods as a means to investigate (1)
    a. Explore the use of moral questions as a starting point for reflection around mental state attribution in 5- and 6-year-old children
    b. Compare two approaches to conducting qualitative interviews: one-on-none interviews and interviews in small groups of two or three children

2. THEORETICAL BACKGROUND

The theoretical background of this study consists of the connection between mental state attribution and morality. This connection has strong evidence in the literature due to the work of Grey and colleagues (2007; 2012). Recently, a study by Sommer and colleagues found a relationship between children's mental state attribution to robots and moral concern their moral concern for robots. What follows is a brief review of these publications.

The paper "Dimensions of Mind Perception" (Grey et al., 2007) proposes that mental state attribution happens along two dimensions: *agency* and *experience.* Things like pain, fear, hunger, rage, and sadness make up experience, and things like self-control, morality, memory, planning, and communication make up agency. An anthropomorphic robot specialized for face-to-face interaction called "Kismet" was one of the entities in the study. The study collected survey answers from 2040 adults and created a two-dimensional graph with different entities. People attributed more agency to Kismet than to a chimpanzee but less than to a girl. The adults did not attribute any experience to Kismet, placing the robot lower than a dead woman on the scale.

A follow up paper (Grey et al., 2012) by the same authors linked these findings to morality. They proposed that mind perception make up the "essence of morality" and that experience and agency can be likened to moral patiency and moral agency. An entity that can experience can also suffer from moral injustice and an entity that has agency can perform actions that are immoral. Normally, when an action (performed by a moral agent) is judged as immoral, it infringes on the physical- and/or psychological welfare of some moral patient. If an entity is to be a considered a moral patient, some attribution of experience-related mental states to them is necessary. If an entity is to be a considered a moral agent, some attribution of agency-related mental states to them is necessary. Based on this connection, we thought of forming moral questions as implicit measures of mental state attribution. Sommer and colleagues (2019) studied the relationship between moral patiency and mental state attribution.

Sommer and colleagues (2019) looked at mental state attribution and moral concern for live agents, robots, and inanimate objects in children aged 4-10. Children were placed in front of a computer screen that showed these entities being placed under a box which was then kicked. They found that in terms of moral concern relating to physical harm, children place less moral concern in robots compared to living agents (a dog and a girl), but more moral concern in robots than in inanimate objects (a stuffed toy and a cardboard box). Sommer and colleagues also investigated the relationship between attribution of mental states and moral concern. They found that higher levels of mental state attribution predicted higher levels of moral concern for the robots. Mental state attribution was calculated using five questions on four-point scales, four of which was selected from the *child version* of the Individual Differences in Anthropomorphism Questionnaire: IDAQ-CF.

3. METHOD

3.1 Participants

16 children aged 5 and 6 participated in this study. The participants were recruited from four preschools in different parts of Linköping, Sweden. The decision of choosing children aged 5 and 6 was made with regards to previous research that suggests that children develop naïve

psychology theories such as theory of mind are formed at around three- to four years of age. Since this was a qualitative study where children were asked to explain their answers, we deemed three- and four-year-old's too young.

3.2 Apparatus and materials

A phone used to record the interviews. A computer containing 1-minute videos of robots. Forms on paper.

3.3 Robots

When choosing these videos, our aim was to provide the children with a holistic picture of the robots. For this reason, we chose videos that showcased the fullest range of aspects of the robots that could be found. Due to our focus on moral agency, we wanted the behavioral affordances of the robots to be fully displayed in the videos. We thought seeing the robots *act* in the world would facilitate thinking about the robots in terms of agency. We thus prioritized providing a holistic image of behavioral affordances over things like homogeneity in the environment or in the amount of human interaction.



Atlas

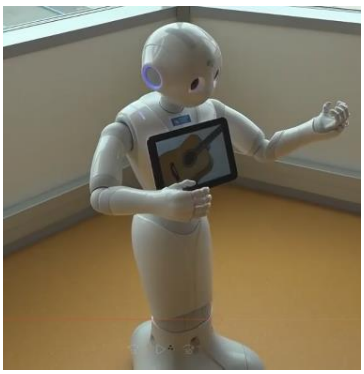
Boston-dynamics robot "Atlas", a large highly mobile humanoid robot without much of anything resembling a face. In the video, Atlas runs an obstacle course and finishes with a backflip. Link: https://www.youtube.com/watch?v=tF4DML7FIWk



Nao

SoftBank robotics "Nao", a small humanoid with clearly marked eyes and a little dot resembling a mouth. In the video, Nao performs a Tai Chi dance, then falls to the ground uttering an "ouch"-sound. Link: https://www.youtube.com/watch?v=LNBNtmMCmIQ





Pepper

SoftBank robotics "Pepper", a medium sized humanoid robot without legs and with a large screen in the stomach area. In the video, pepper rolls in a room and interacts with a woman at one point "playing" guitar and at another throwing a ball towards the woman. Link: https://www.youtube.com/watch?v=4h4j-e3oUzU&t=7s



Kim

Husqvarna "Automower", a robotic lawnmower referred to as "Kim" in the interviews. In the video, Kim drives around on a lawn, one time coming close to a dog and another time turning for an obstacle. Link: https://www.youtube.com/watch?v=KE7O3dK07nQ&t=103s

3.4 Measures

Forms

Forms filled in on paper containing questions aimed to assess the children's mental state attribution and moral concern for the different robots. There was one form for each robot. All forms contained the same questions with minor grammatic differences on some questions due to physical differences between the robots. For example, question B3 (below) is slightly different for the robot Pepper than for the robot Atlas, since the Pepper has wheels, and Atlas has legs.

Scales

Two different scales were used in the interviews, these will be presented below. Cut out copies of the scales were handed to the children when the scales were introduced. The children often had these in their hands and pointed to them when answering questions. These were meant to make it easier for the children to understand the scales.

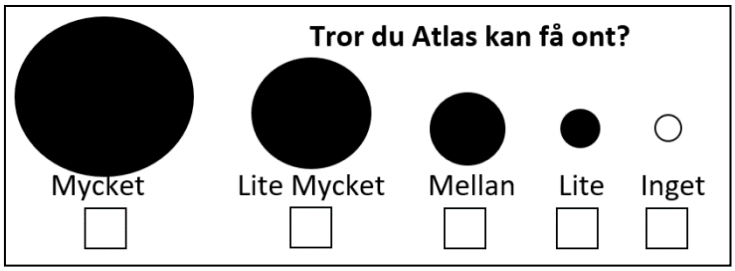
3.5 Questions

Mental state attribution questions (MSA-questions)

Four questions aim to assess the children's *mental state attribution* of the robots. In the introduction, mental state attribution was introduced as the two dimensions: *experience* and *agency.* Questions MS1 and MS2 aim to investigate how much *experience* the children ascribe to the robots. Question MS3 and MS4 aim to investigate how much *agency* the children ascribe to the robots.

MS1: "can the robot get hurt? If so, how much?"
MS2: "can the robot feel emotions like happy or sad? If so, how much?"
MS3: "does the robot have thoughts? If so, how much?"
MS4: "can the robot do things on purpose? If so, how much?"

The questions are on a five-point scale adapted to children ranging from "none" to "very much" (see scale 1 below). The five alternatives have circles above them that correspond to the alternatives in size. This was meant to make it easier for the children to understand the gradual nature of the scale. The design of the circles was taken from (Sommer et al., 2019) and adjusted for a five-point scale instead of the four-point scale used it their study, so it would be easier to connect it to the five-point moral scale presented below. The questions were formulated with inspiration from a version of the Individual Differences in Anthropomorphism Questionnaire adapted to children: IDAQ-CF (Severson and Lemm, 2016).



Scale 1. The values (from the left) roughly translate to: "**Very much**", "**Much**", "**Middle**", "**a Little**", "**Nothing**".

Moral Agency questions

One yes-or-no question aimed to assess how much moral agency the children attributed to the robots, as well as spark reflection around the reasoning and preconceptions underlying children's agency-related mental state attribution.

A1: "if the robot would step on/run over a beetle/hedgehog, would it be the robot's fault then?"

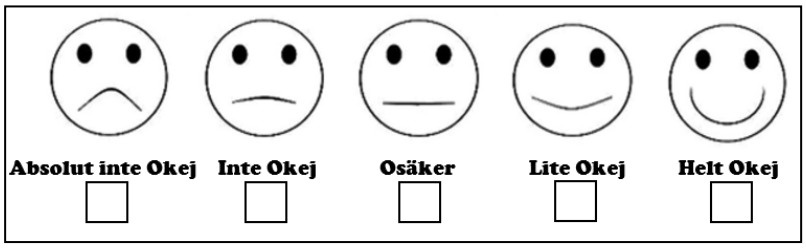
Question A1 was formed as a means to implicitly ask the children about agency-related mental states by providing a practical context that was thought to be easy to understand for the children. The question asks the children if it would be the robot's *fault* in the proposed

scenario. A positive answer to this question means that the child regards the robot as a moral agent.

Moral Patiency questions

Two questions aimed to assess how much *moral patiency* ("MC") the children attributed to the robots, as well as spark reflection around the reasoning and preconceptions underlying children's experience-related mental state attribution. The questions were on five-point scales adapted to children with alternatives ranging from "absolutely not okay" to "completely okay" (see scale 2 below).

P1: "how okay would it be to kick/pinch the robot?"
P2: "how okay would it be to lock the robot in a closet?"



Scale 2. The values (from the left) roughly translate to: "**Absolutely not Okay**", "**Not Okay**", "**Unsure**", "**A little Okay**", "**Completely Okay**".

The scenarios posed in the moral patiency question placed the robots in concrete contexts where the moral judgements would have practical (or easily imaginable) consequences for the robots as moral patients. Making a moral judgement is essentially an act of "creating the world" since the judgement states what is or is not allowed to happen. So, when the consequences of the judgements are easily imaginable (for example, locking the robot in the closet) it is almost as if the children are either allowing something to happen or prohibiting it. This was though to make the children more inclined to think hard about the reasoning behind their judgements and give thorough explanations to them, which were thought to spark reflection.

Follow-up questions

After every question in the form, children were asked to explain their reasoning by answering qualitative follow-up questions. These questions always started with just asking "why do you think so?" and in cases where the interviewer wanted to know more, further questions were asked. Such further questions were not preconceived and were thus formulated depending on the context of the dialogue.

Modification of the moral questions

In our initial conception of the study design, we anticipated that simple explanations to the moral scale would suffice. But when it became clear that the kids exclusively made the harshest or the least harsh moral judgements, we started to suspect that the children did not understand that the questions asked for a gradual answer. Thus, the decision to change the grammatic structure of the questions and give a more thorough introduction to the scales was made.

So, all children in the solo approach were asked the moral-concern-questions like this: "would it be okay to kick the robot?". In the group approach, they were asked like this: "how okay would it be to kick the robot?". Prior to the change, children almost exclusively answered the questions as "okay" or "not okay", and as "yes" or "no". After the change (and with a more thorough clarification of the scale before the questions) kids did seem to understand the questions and many times gave "in-between" answers (an answer that is either 2,3 or 4 on the scale).

3.6 Analysis

The quantitative data was digitalized and then handled descriptively using plots and charts. This was the first step of the analysis because we wanted to get an overview of trends and relationships in the answers. At this stage, some answers seemed to contradict each other. For example, a lot of children who ascribed no ability to feel pain or emotions to robots still answered that it was "absolutely not okay" to kick or lock the robots in the closet. No statistical analyses were made.

Notes from the interviews were digitalized in an excel chart. This provided a handy overview of the explanations to the quantitative answers. The recorded audio from the interviews was transcribed. Interesting sections were collected and coded in an excel document. Codes that touched on similar areas were marked with similar colors. These similarities were summarized in a separate document to be kept for exploration after the coding was done.

After the coding, a thematic analysis was conducted using the digitalized notes and the transcriptions/codes. These resources were reflected upon holistically, and themes appeared. Chosen themes were explored deeper by going back to the audio recordings and through further reflection. A more in-depth description of our thinking can be found under *discussion of methods.*

3.7 Procedure

Prior to the interviews, forms of consent were sent out to parents where they could read about the study and consent to the participation of their child. Interviews were conducted at four dates over roughly two weeks. There were a few days in-between each day of interviews. The days in-between the interview-days was used to listen to the recorded audio from the interviews. These listening sessions led to reflection about themes and birthed theories that were explored in following interviews.

Interviews

Interviews were conducted in two ways: A solo approach and a group approach. Half of the children were interviewed solo, and the other half in small groups containing two or three children in each group. Interviews usually took about an hour, but some kids got tired after 30 minutes and wanted to stop the interview. Because of this, some kids did not see all robots.

In both approaches, the first thing that happened was that the children were introduced to the scales in the form. Scale 2 (for the moral questions) were introduced more thoroughly for the children in the group approach for reasons already mentioned. In the solo approach, the interviewer introduced Scale 2 by explaining that it was a gradual scale and that the children

could point to the face they though corresponded to "how okay" the action was. In the group approach, the same thing was done but the interviewer was generally clearer and more thorough and gave examples like: "if I think something is really wrong to do, I point here". No concrete examples were used to introduce the scales due to a fear of affecting their responses.

Children were interviewed about one robot at a time. For each robot, children were shown a 1-minute video of the robot on a laptop screen. The interviewer filled in the forms using a pen and made short notes of the explanations to answers. After the video, they were interviewed about their thoughts surrounding the robot.

Solo Approach

In the solo approach, the interviewer followed the questions in the form closely and asked for explanations to the given answers. If the interviewer found some answer especially interesting or identified some contradiction, further questions were asked to delve deeper or clarify.

Group Approach

In the group approach, the interviewer did not follow the questions in the form as closely as in the solo approach. Rather than being in the center of the interviews, the questions in the form here functioned more as a starting point for discussion and guiding framework in case the dialogue went too far off the rails. The aim was to invite the children to express their thoughts openly and discuss the answers with each other.

4. RESULTS

What follows are three themes that could be outlined from the thematic analysis. All of these themes are interpretations made by the researcher. Obvious epistemological issues should be considered when looking at these results. We have tried to be as transparent and nuanced as possible about the assumptions that are made when making these arguments.

For some themes and observed phenomena, quantitative data is used to back up the interpretations. It is important to note that this quantitative data has no statistical security. No statistical analyses were made in this study.

Theme one: "The Robot just Breaks"

Despite answering that some robots cannot feel pain or emotions, some children answered that it would be "absolutely not okay" to kick the robots or lock them in the closet. They explained these moral judgements by stating that the robot would break.

When asked the mental state attribution questions, some children did not attribute the ability to feel pain or emotions to some robots. Some of these children consistently maintained this lack of attribution of experience-related mental states to all the robots, but some children did not and only attributed experience-related mental states to some robots.

13 out of 16 children (81%) attributed "no pain" (gave a "1" on the scale) to *some* robot.
10 out of 16 children (62%) attributed "no emotions" (gave a "1" on the scale) to *some* robot.

4 out of 16 children (25%) *consistently* attributed "no pain" (gave a "1" on the scale) to *all* robots.
5 out of 16 children (31%) *consistently* attributed "no emotions" (gave a "1" on the scale) to *all* robots.

When asked how okay it would be to kick the robot or lock the robot in the closet, one could expect that the children who did not attribute pain or emotion to any robots would give less harsh moral judgements than children who did, since, from their point of view, no robot would get hurt or feel unpleasant emotions in these moral scenarios. One could also expect that the children who attributed no pain or emotions to *some* robots, would give less harsh moral judgements in those very cases. In a few cases, these expectations were correct. This was most often found in the case of Atlas, the robot that got the highest number of "no pain" and "no emotion" attributions out of all the robots. Some of the children who did not attribute pain or emotion to Atlas answered that it would be "completely okay" to kick Atlas and explained that it was okay since he would not get hurt.

Surprisingly, though, this pattern was not the trend. In most cases, children who consistently attributed "no pain" and "no emotions" to robots, and children who did this to some robots, still gave the harshest moral judgement ("completely not okay") when asked how okay it would be to kick the robot or lock it in the closet. When asked to explain the reasoning behind these moral judgements, some interesting explanations were given.

As could be expected, children that did attribute pain and emotion to a robot often referred to that attribution in their explanation. The transcript below portrays how such explanations were commonly uttered.

Transcript 1
*1 **INTERV**: would it be okay to kick Nao?*
*2 **TOMMY**: no*
*3 **INTERV**: how much? you can point here*
*4 **TOMMY**: (points to "1" or "absolutely not okay")*
*5 **INTERV**: why is it not okay?*
*6 **TOMMY**: because he would be hurt and sad*

The children that did not attribute pain and emotions, on the other hand, explained their moral judgements with that the robot would break if it got kicked, and that it was "absolutely not okay" to kick the robot for this reason. Most of these children failed to explain why this, in turn, would be bad. They instead seemed to think it was inherently wrong to break robots, even though they did not think they could feel any pain or emotions. Here is an outtake that illustrates this thinking.

Transcript 2
*1 **INTERV**: Is it okay to kick pepper?*
*2 **JOHN**: no if you kick pepper he just breaks*
*3 **INTERV**: why?*
*4 **JOHN**: if you kick pepper he just breaks*
*5 **INTERV**: why is that bad?*
(paus 2 sec)

6 ***JOHN****: if you do this (pushes chair) really hard it falls over*
7 ***INTERV****: yes*
8 ***JOHN****: if it falls to the ground then it breaks*
9 ***INTERV****: right, and that's bad?*
10 ***JOHN****: yes*
11 ***INTERV****: but pepper doesn't get hurt or sad?*
12 ***JOHN: (assertive tone)*** *pepper can't get sad or feel pain*

Even though the interviewer repeatedly asks John why it is bad that Pepper breaks, he fails to give an explanation. Instead, he likens the breaking of pepper to the breaking of a chair. It is possible that John tried to explain *how* something breaks by referring to the chair. But our interpretation is that John used the breaking of the chair to show *why* it would be bad to break pepper. To John, the chair suffices as an example that explains why it is bad to break pepper even though the chair and pepper do not share a lot of characteristics apart from being physical objects. This suggests that John might be thinking in more general terms. Rather that evaluating the breaking of each object, John has been taught that it is bad to break things in general.

Other kids who also referred to breaking the robot as explanations for their moral judgements were also unable to explain why this was bad. They didn't seem to be able to put their minds around *why* it would be inherently bad to break the robots. A possible explanation is that they don't yet have all the necessary capacities for the moral reasoning required to ground their answers in the suffering of some being. They instead have rules that decide what is allowed to do and what is not, but do not need to understand why these rules apply. This is discussed in more detail in the last part of the discussion.

Theme two: The Strong and Brave Warrior and the Weak and Fragile Child

It seemed like Atlas made a strong impression on many children. When the video of Atlas was shown to the children, many commented that it was a cool robot or made sounds indicating that they were impressed. After the video, when asked to assess the degree to which Atlas had the various mental states, children often explained their answers by referring to actions like the backflip in the video or the skill with which he ran the course. Children were impressed by Atlas ability to jump, run, and do backflips, as well as his strength and size.

When shown the video depicting Nao, children did not seems as impressed as with Atlas. Nao falls to the ground in the video and gets help when getting back up. Children often laughed about this, and one child commented "he can't stand on his own". Two different children said that Nao was a toy. Children often said that Nao was cute and that he was a child.

Sometimes the children compared the robots during the interviews. This often happened without the interviewer explicitly asking the children to compare them. The most common comparison was between Atlas and Nao. In these comparisons Atlas and Nao was often seen as opposites to each other with regards to their abilities. Atlas was described as old,

strong, brave, skilled, and resistant to pain, and Nao was described as young, weak, small, and fragile. On one occasion, Atlas was described as "*better*" than the other robots.

Atlas was attributed significantly less pain and emotion in comparison to the other robots (except for Kim with regards to pain). 12 out of 16 children answered "nothing" when asked how much *emotions* Atlas could feel. And 10 out of 16 children answered "nothing" when asked how much *pain* Atlas could feel. However, Atlas was attributed a greater ability to do things on purpose than the other robots. Averaged answers to all mental state attribution questions can be seen in figure 1.

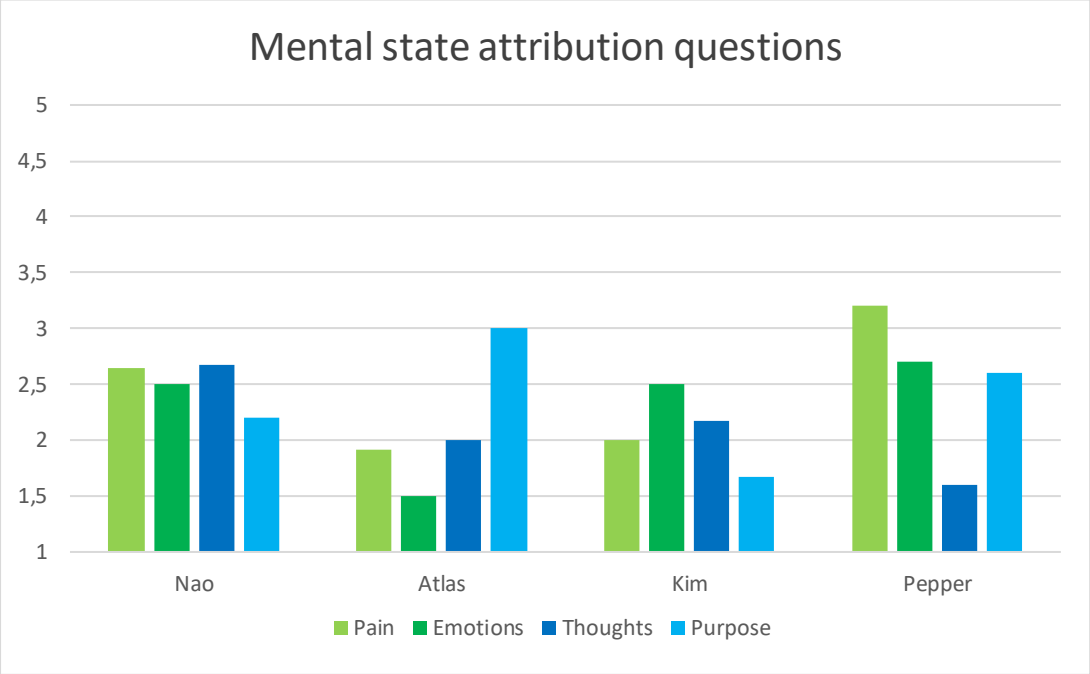


Figure 1. Average answers to mental state attribution questions about robots. 1 = "*nothing*", 2 = "*a little*", 3 = "*middle*", 4 = "*much*", 5 = "*a lot*"

Considering that Atlas was described as older, more skilful, wiser, and braver than Nao, one could think it is surprising that so many children attributed neither pain nor emotion to Atlas, while attributing it to Nao. It seems reasonable to expect that a robot that is conceived as more sophisticated and complex would be attributed a greater capacity to experience mental states such as pain and emotion. Instead, the opposite pattern can be observed in our data. Why do we see this pattern?

One possible explanation is that this could depend on children perceiving the robots through the lens of archetypes. In this interpretation, children have preconceived archetypal narratives that they place the robots inside of if they are deemed fit. For example, when looking at the video of Atlas, children might see a warrior, a soldier, or a hero because Atlas shown signs of being strong, fast, and agile, abilities which are characteristic of such archetypes. From this perspective, the attributions of "no pain" and "no emotion" to Atlas are easier to understand. For a warrior or a hero, these mental states are weaknesses that do not exist in a strong and brave warrior. So, the more impressive the children think Warrior-Atlas is, the less pain and emotion he is attributed.

We will start by presenting and discussing outtakes from the interview with Arad. His answers are representative for the reasoning of many children. At this point in the interview, Arad has already been interviewed about both Nao and Pepper.

Transcript 3
*1 **INTERV**: Here is Atlas **(plays video of Atlas)***
*2 **ARAD**: it's also kinda like them **(Nao and Pepper)***
3 (5 sec)
*4 **ARAD**: alright no its not, he **(Atlas)** is better than them*
5 (10 sec)
*6 **ARAD**: but it is about the same for them because they are all robots, so it's about the same*

As can be interpreted from line four, it seems like Arad is initially impressed with Atlas. He states that Atlas is better than Nao and Pepper, without making it clear in what way. The fact that Arad refers to Nao and Pepper as "*them*" on line two suggests that there is some kind of similarity in his conception of these two robots. On line six he states that Atlas is "*about the same"* as the others since *"they are all robots"*. So, even though Arad explicitly states that the three robots are all robots, he still seems to think that Atlas is slightly different to the other robots. Just after this, Arad is asked to compare Atlas and Nao in terms of how much emotions they feel. This is Arad's answer:

Transcript 4
*1 **ARAD**: Nao feels a lot more emotions than Atlas*
*2 **INTERV**: oh ok*
*3 **ARAD**: yes because Atlas didn't you see what they did they did backflips and they didn't feel almost nothing but I think Nao does*
*4 **INTERV**: ok, but so how would you judge on this scale how much emotions does Atlas feel?"*
*5 **ARAD**: only a little **(points to "little")***

Earlier, when talking about Nao, Arad had assessed Nao's ability to feel emotions as "very much" (5 on the scale). When asked which emotions he though Atlas could feel, Arad answered: "only happy, he is always happy" and that Atlas would not be sad or scared even if someone kicked or punched him. For pain, Arad assessed Nao as a "middle" (3 on the scale) and Atlas as "nothing" (1 on the scale).

Arad thus thinks that Atlas generally feels *less* than Nao, and he seems to repetitively ground this in Atlas abilities to perform complex tasks like doing backflips or running the obstacle course. Similar thinking is uttered in the responses of other kids as well. An interpretation of this is that the children conceive of the ability to feel things like being scared, being sad, or feeling pain, as *weaknesses*. And since Atlas is something akin to the antithesis of weakness in the children's eyes, they arrive at the conclusion that Atlas does not feel these things. As stated, this view might be grounded in the children perceiving Atlas as the archetype of a strong warrior. From this perspective, the strong and skillful Atlas is hard to hurt and very brave, so of course he feels less emotions than weak little Nao through the children's eyes.

This interpretation is underlined further in responses to the questions about *moral patiency.* An example of this can be seen in the following transcript, where Arad is asked how bad it would be to kick Atlas.

Transcript 5
1 ***INTERV****: How bad would it be to kick Atlas?*
2 ***ARAD****: completely fine*
3 ***INTERV****: why?*
4 ***ARAD****: because he can do a lot of stuff and he is skilled*
5 ***INTERV****: but you said it was not okay to kick Nao*
6 ***ARAD****: yes but Nao is a liiiiitle bit younger, and Nao doesn't know as much stuff as Atlas, maybe people try to kick Atlas but it doesn't work cause he like does a backflip so they miss or something*
7 ***INTERV****: but if someone would succeed in kicking him?*
8 ***ARAD****: I don't think anyone could*
9 ***INTERV****: but if we pretended that someone would?*
10 ***ARAD****: completely okay*
11 ***INTERV****: it's still completely okay? He would take a kick?*
12 ***ARAD:*** *yes*
13 ***INTERV****: if we would tie Atlas up and pick him to pieces, would that be okay?*
14 ***ARAD****: it would not be okay at all*
15 ***INTERV****: why?*
16 ***ARAD****: Atlas will not like it but I don't know why because I'm not a robot*
17 ***INTERV****: do you think he would suffer?*
18 ***ARAD****: yes it would be like he would die*

Arad responds that its completely fine because of Atlas's abilities and skills. When the interviewer mentions Arad's answer to the same question about Nao, he states that Nao is younger and that he does not know as much stuff as Atlas. Arad then reasons that it would be okay since Atlas could just escape the kick with a backflip. The interviewer continues to push the scenario up until asking if it would be okay to pick Atlas apart.

Arad's utterances here support the interpretation made in this theme. Arad has an image of Atlas as a strong hero-figure that would not get hurt by punches or kicks. But when the interviewer pushes the scenario into one where Atlas is defenseless, Arad completely changes his moral assessment. Atlas can't feel pain, but he can die. And Arad thinks it would be a morally wrong thing to kill Atlas by picking him apart. It thus seems like Arad's conception of Atlas is one where he either lives or dies; he either succeeds, or he dies trying. There seems to be no possibility of a middle way where Atlas could writhe in pain from breaking a leg or run away scared.

If Atlas is the archetype of the brave and strong warrior or hero, Nao is the archetype of the vulnerable child. Kids referred to Nao as small, young, and fragile on multiple occasions. These utterances usually came as explanations to why it was not okay to kick Nao. In the interview with Frank and Mila, they are explicit about the differences between Atlas and Nao.

Transcript 6
1 ***INTERV****: Would it be okay to kick Nao?*
2 ***FRANK****: no*

3 ***INTERV****: why?*
4 ***MILA****: because he is very small*
5 ***INTERV****: why is that a reason not to kick him?*
6 ***MILA****: because he is small and might get hurt*
7 ***FRANK****: but Atlas doesn't get hurt*
8 ***INTERV****: so you don't think Atlas gets hurt but you do think Nao gets hurt Frank?*
9 ***FRANK****: yes if you hit him like here* ***(points to net-structure at Nao's ear)*** *because it's like this net*

Here, Frank and Mila agree on that it would not be okay to kick Nao. As an explanation, Mila states that Nao is "very small", and "Might get hurt". Frank refers back to Atlas stating that he doesn't get hurt. Here, when speaking about Nao, the children had already been interviewed about Atlas. So, referring back to Atlas's inability to feel pain after having stated that Nao does feel pain clearly illustrates that Frank conceives of the two robots as different in this regard. An outtake from the interview with Anna provides example where Nao is referred to as a child.

Transcript 7
1 ***INTERV****: how okay would it be to lock Nao in the closet?*
2 ***ANNA****: you can't do that*
3 ***INTERV****: why not?*
4 ***ANNA****: he is just a child*
5 ***INTERV****: why is that bad?*
6 ***ANNA****: he is scared and can't get out*

Anna states that Nao is "*just a child*" on line four. Her formulation of this response, specifically the word "*just*" indicates that her utterance wants to convey the notion of patiency accompanying the use of the label "*child*". This supports the notion of Nao as being conceived by the children as the archetype of a fragile being that easily falls victim to immoral acts.

Theme three: "Plastic can't get Hurt"

Structure and texture

When some children give explanations to why they think some robots do not have the ability to feel pain, they refer to characteristics of the material that the robot is made of. Two variations of this theme could be found in the interviews. In one interview, with Frank and Mila, the structure and texture of the material is referenced as explanatory grounds for the inability to feel pain. In two other interviews, one with John and one with Sara, Irina, and Nick, the specific type of material (plastic or metal) is referenced, but no structural or textural specifications about the material is mentioned. Reference to material texture and structure can be seen in transcript 7 above and in transcript 8 below.

Transcript 8
1 ***INTERV****: do you think Atlas can feel pain like you feel pain if someone pinches you?*
2 ***FRANK****: no*
3 ***MILA****: no*
4 ***FRANK****: because it is hard*
5 ***INTERV****: if you punch Atlas, how about then?*

6 ***FRANK/MILA****: nooo*
7 ***INTERV****: you don't even think he can feel pain if you hit him really hard so he breaks?*
8 ***FRANK****: no*
9 ***FRANK****: I only think he, I think his electricity goes out*

In transcript 8, Frank states that he does not think Atlas can feel pain. He then explains this answer with the phrase: "because he's hard". He thus seems to understand Atlas inability to feel pain as stemming from the hardness of his body. When posed the follow-up question on row 7, Frank still holds that Atlas would not feel pain even if his hard body broke. This opens for the possibility that Frank might not think of fragility when explaining Atlas inability to feel pain with the hardness of his body. It might instead be a reference to the texture of Atlas body. Frank might be thinking that it feels unnatural that a surface of such a hard texture would be able to produce sensation. Frank might process this question (if Atlas can or cannot feel pain) in embodied, anthropomorphic terms, where he understands his own sensation of pain as stemming from his understanding of the parts of his own body that can feel pain. And since he can only feel pain on his "soft parts" (not on his nails or teeth) it is natural that he concludes that hard-bodied Atlas lacks this ability.

On row 9 in transcript 7, Frank explains that Nao would feel pain if someone hit him on the net-like-structure around the area where his ear would be. One possible interpretation of this is that he specifically points out the net-structure at Nao's ear since this part of Nao's body clearly seems to be more vulnerable to breaking than the rest of Nao's body. This interpretation entails that Frank thinks of the *fragility* of this part of Nao when he points out this specific spot. He then might think that if something breaks, then pain comes, just like how some children think that blood equals pain.

Another interpretation, more in line with the interpretation presented in relation to Franks statements about Atlas, is that he pointed to this spot due to being the easiest access to Nao's brain, since one could think that it is only the weak white net-structure that protects Nao's brain at this spot. This interpretation would entail that Frank anthropomorphizes Nao to the point of thinking that he has a brain inside his "skull", and that it would hurt a lot if this part got damaged, just like on humans. Franks reasoning remains unclear, though.

Type of Material

The other variation of this theme references the specific type of material of the robots as an explanation for the inability to feel pain. Sara and Irina state that some specific materials "can't get hurt" in general. Irina explains that Nao is made of plastic and that "plastic can't feel pain, like a plastic bag". Sara adds that "metal" cannot feel pain either. Here is the interaction in transcript:

Transcript 9

1 ***INTERV****: if someone would punch him (Nao) hard, would he feel pain like you then?*
2 ***SARA****: noo*
3 ***IRINA****: no cus robots are made of plastic and plastic can't get hurt, like a plastic bag*
4 ***SARA****: not metal either*
5 ***NICK****: but if you cut a plastic bag what happens then?*
6 ***IRINA****: they don't even have feel, not even feelings*

7 ***INTERV****: do you (Nick and Sara) agree with that it doesn't have feel or feelings?*
8 ***NICK****: yes*

Another example of this variation of the theme can be seen in the interview with John. John is asked a question that aims to investigate what he thinks about Atlas ability to hurt another robot intentionally. He seems to have interpreted the question as if asking if the other robot would get hurt if Atlas would try to hurt it. Similar to Sara and Irina in the transcript above, he refers to the material of Atlas as being "metal" when he explains that the robot would not get hurt.

Transcript 10
1 ***INTERV****: do you think Atlas could hurt another robot on purpose?*
2 ***JOHN****: no*
3 ***INTERV****: why not?*
4 ***JOHN****: because he is made of metal, they all are*
5 ***JOHN****: he is strong, if he punches another robot it just breaks, it's easy to fix*
6 ***INTERV****: it doesn't hurt the other robot?*
7 ***JOHN****: no it just breaks*

It is interesting that John first uses "he", when referring to the robot Atlas would try to hurt, since the question was posed about any robot in general. This grammatical choice suggests that John thinks of a specific robot in his answer. On line 4, though, he adds "they all are", probably referring to all robots. This increment suggests that Johns answer might instead best be explained as a belief about robots in general. He explains that the robot would not get hurt because it is made of metal and then adds that all robots are made of metal. So, all robots are unable to feel pain.

5. DISCUSSION OF THEMES

The first research question aimed "*to investigate the reasoning and preconceptions that produces 5- and 6-year-old children's mental state attribution to robots.".*

Overview

Through an interpretation of the themes, we attempt to answer the first research question. After this, we present two possible explanations to the problem posed around theme one. One of which is an attempt to outline how the children decided how harsh their moral judgements would be.

Interpretation of Themes

The three themes presented in our results section illustrate different patterns in the children's mental state attribution and in how they explained their reasoning behind it. To provide insight into how this relates to the first research question, we provide an interpretation of the three themes.

We argue that our results are reflections of two types of preconceptions in the children. We present what these preconceptions are and discuss how they influence mental state

attribution to the robots. Further, we argue that the way robots are presented to children can promote the influence of either or both of these preconceptions. We then argue that theme three reflects how the children dealt with the conflict of influence from both preconceptions and the tendency to attribute mental states to anthropomorphic robots.

Preconception one: Robots as Taxon

On one hand, the children seem to have the preconception that robots are non-experiencing and non-intentional. This is similar to the theoretical term: Theory of Artificial Mind (ToAM) which is the understanding that machines or robots are programmed by humans and have no inner life or will of their own (Spektor-Precel & Mioduser, 2015a, 2015b). What lead us to this were the many times children explained their lack of mental state attribution by merely referring to the robot being a robot. Children place specific (or individual) robots under the same roof by evaluating them as the general notion of "robot". In other words: Atlas, Nao, Kim, and Pepper are treated as members of a biological taxon (or clade), sharing the common ancestor of "robot". The transcript below is a representative example of such utterances.

Transcript 11
1 ***INTERV****: do you think Kim can feel pain?*
2 ***DAVID****: no*
3 ***INTERV****: why?*
4 ***DAVID****: because he is a robot*

Preconception two: Archetypal Frameworks

On the other hand, the children seem have preconceptions stemming from portrayals of robots in children's culture. These preconceptions might be best described as archetypal frameworks that influence the way children perceive robots. The archetypal frameworks consist of knowledge from children's culture which acts as a filter that the children perceive the robots through. The influence of these preconceptions in the children of this study is evidenced under theme two in the results, where we argued that the children perceived Atlas and Nao through such archetypal frameworks.

How did these preconceptions affect children's mental state attribution?

Preconception one leads to the evaluation of specific robots as members of the non-specific taxon "robots". The influence of the belief is "activated" as soon as a child who holds this belief hears that the entity that is to be evaluated is a robot. This influence can overrule information related to the behaviour and appearance of the robot, information that otherwise (without the *robots as taxon* belief) would convey an understanding of the mental properties of the robot. The mental state attribution can thus be described as a tug of war between the (anthropomorphic) appearance and behaviour of the robot and the preconception about robots as cold machines without mental states. Transcript 4 from theme two illustrates this process. Here, Arad goes back and forth between categorizing Atlas as "one of the other robots" or as something else.

Transcript 12
1 ***INTERV****: Here is Atlas* ***(plays video of Atlas)***
2 ***ARAD****: it's also kinda like them* ***(Nao and Pepper)***
3 (5 sec)

4 ***ARAD****: alright no its not, he* ***(Atlas)*** *is better than them*
5 (10 sec)
6 ***ARAD****: but it is about the same for them because they are all robots, so it's about the same*

Preconception two instead leads to an increased inclination to regard the robots as unique individuals devoid of the "robot" group affiliation. Perceiving the robots through the archetypal framework can be understood as a matchmaking process of connecting perceptual information about the robot to archetypes of robot characters. Portrayals of robots have been found to be anthropomorphic in both adult- (Sandoval et al., 2014) and children's culture (Axell et al., in press). Among other anthropomorphic features, robots are often gendered and exaggerated in this regard (Cave et al., 2018). Such exaggerated gender-stereotypes in portrayals of robots speaks for the existence of anthropomorphised robot archetypes. When children get to see videos of or meet robots that fit these archetypes, the anthropomorphic portrayals can take over and an archetypal story is formed around the robot. As can be seen in the case of the children's perceptions of Atlas, the consequences of this can be unexpected in terms of how it affects mental state attribution. The archetype that some children seemed to fit Atlas too were the warrior or the hero, two archetypes that are often portrayed as having no fear and a high pain threshold. The way robots are presented to children affects how much these preconceptions influence mental state attribution. This has relevant implications for research on children's mental state attribution to robots.

Implication: Stimulus Materials Promote Preconceptions

The stimulus materials through which robots are presented to children could promote the influence of any of these preconceptions. For example, if one would ask children questions about robots where robots are referred to as the general notion of "robots" (without presenting them with stimuli depicting a specific robot) it could promote the influence of preconception one. If one would instead ask children questions about a specific robot while presenting it to the children on video or in real life, it might promote the influence of preconception two. Other things could also promote the influence of one or the other of these preconceptions. Explaining that a specific robot *is a robot* to the children could promote the influence of preconception one. Presenting children with robots that has an appearance that could easily be fitted to a common archetype could promote the influence of preconception two. Evidence for this can be seen in our study. It is possible that the results presented in theme two can be linked to the choice of videos depicting the Atlas and Nao.

In the video depicting Atlas, it performs two backflips. It is likely that these backflips influenced the hero/warrior archetype that some children seemed to perceive Atlas through. When the backflip was mentioned, it was often when explaining how Atlas could withstand a kick or break himself out of the closet, which suggests that this action was linked to the archetype. A backflip is something that the children have probably seen in action movies when the strong and athletic hero jumps off a building or dodges a kick. Six studies have shown an increased tendency to attribute mental states to a robot related to movements of a robot (Thellman et al., 2022). A backflip, though, is more than just movement. It is in a sense socially interactive behavior. Atlas does not need to do the backflip off of the wooden box to get down; he does not do it for pragmatic reasons, he does it to show his skills. The backflip could thus be interpreted as a communicative action. Atlas also makes celebration gestures after the backflip (dusting off his shoulders) which indicates that the whole thing was a display or a showing-off. Just like the backflip, this is socially interactive behavior. Socially

interactive behavior in robots have been found to predict a stronger tendency to attribute mental states to robots (Thellman et al., 2022).

A similar point can be made about Nao. In the video showcasing Nao, he falls to the ground and utters an "auch"-sound when he hits the floor. The sound Nao makes is a common sound heard by humans when they hurt themselves. This is thus a similar fostering of anthropomorphic thinking in the children that could have contributed to the "Child-archetype" that some kids seemed to perceive Nao through.

Conflicting Influences

One could assume that these preconceptions as well as the tendency to attribute mental states to anthropomorphic robots all influence children's mental state attribution. But these influences pull in different directions. Earlier these conflicting influences were described as a "tug of war". Evidence from naïve biology theory suggest that conflicting beliefs can exist without constraining each other.

Naive biology theories are frameworks that young children use to categorize their world. This Similar to how conception one was found to influences mental state attribution of the children in our study, when a child categorizes a robot as biological and alive it affects the attributions of characteristics they make to the robot (Bernstein and Crowley, 2008). Although, children have been found to attribute mental states, moral standing and social rapport to robots and stuffed toys even though they did not attribute life status to them (Kahn et al., 2006). This suggests that seemingly conflicting beliefs held by children do not necessarily constrain children's mental state attribution. This suggests that the same might be true for our two preconceptions.

Theme three illustrated how some children referred to the exterior of the robots in their explanations to why they did not attribute mental states to the robots. This can be interpreted as the children's way of dealing with the conflicting influence of preconception one in light of robots that behave and look like they have an inner mental life. All of the children that uttered these explanations clearly had an understanding of preconception one: they regularly referred to robots as a group in their explanations to why they did not attribute things like pain, volition, or emotion to them. Then, when they were presented with robots showing an anthropomorphized appearance like Nao and Pepper, or exhibiting an anthropomorphized behaviour like Atlas, and were subsequently asked to explain why these robots were not capable of these mental states, they resorted to these stories as explanatory models. This is thus a way of solving the problem of conflicting influences coming from preconception one and from the children's natural anthropomorphic tendencies.

Misunderstanding or Moral Landscape

As described in theme one, children that said that the robots would "just break" rather than feel pain as explanations to their moral judgements, almost always answered "absolutely not okay" (which is the harshest moral judgement) on the scale when asked about kicking the robots. While many children in the study gave such answers to the moral patiency questions, it seems natural that the children who attributed the ability to feel pain or emotions to the robots would make harsher moral judgements than the children who did not. So why did the children who did not conceive of the robots as able to feel pain also make such harsh moral judgements? We propose two possible explanations.

Moral Scale Not Understood

The first explanation is that some kids mistook the gradual nature of the scale in the moral-concern-questions for binary yes-or-no questions. This explanation is supported by the fact that only two children in the first approach answered these questions with an "in-between" answer (an answer that is either 2,3 or 4 on the scale). In the first approach, the questions were asked using a grammar that could be understood as asking for a yes-or-no answer. The questions were asked like this: "would it be okay to [X]", which tended to be answered with "yes", "no", "not okay", or "okay". This can be seen in many of the transcripts above. When the interviewer later asked, "how okay?" and asked the children to point to the scale, many of them pointed to 1 or 5 without any further reflection, almost like they were simply asked "point to the answer that means "no"". An example that illustrates this can be seen in the following outtake from the interview with John.

Transcript 13
1 **INTERV***: would it be okay to kick Atlas?*
2 **JOHN***: no! he would break then*
3 **INTERV**: how bad would it be?
4 **JOHN**: **(points to "absolutely not okay")**
5 **INTERV***: do you think it would be as bad to kick Atlas as it would to kick a rabbit?*
6 **JOHN***: ayyy! you can't kick a rabbit!*
7 **INTERV**: *no, but what would be worse?*
8 **JOHN**: *to kick a rabbit is worse*
9 **INTERV***: why?*
10 **JOHN***: if you kick a rabbit it can get hurt*
11 **INTERV**: *but Atlas can't?*
12 **JOHN**: *no it only comes sparks from him*

Here, John first makes the makes the harshest moral judgement ("absolutely not okay"), but when asked about kicking a rabbit, he states that this would be worse. This illustrates the notion that the moral scale was misunderstood. If the scale would have been properly introduced (for example by asking John about the rabbit *before* asking him about Atlas) he might not have made the harshest moral judgement.

When all the interviews for the solo approach were done, we had strong suspicions that many of the children had misunderstood the moral scale. As mentioned in the method section, we therefore decided to change the grammatic structure of the questions to a "how okay would it be…?" and to give more through introductions to the scales. This seemed to clarify the questions: only two kids in the solo approach gave "in between" answers to moral questions, while more than half of the children in the group approach did.

One example of a child who understood the moral questions can be seen in the transcript below where Maria points to "unsure" when asked if how okay it would be to kick Nao. She explains her answer with that it might be possible to repair Nao if he breaks, since "he looks like one of those that can be fixed". This suggests that she might have given a harsher moral judgement if repairing would not be possible. She thus weighs the badness of breaking the robot against the possibility of being able to repair it and comes up with an in-between answer to the moral question.

Transcript 14

1 ***INTERV****: would it be okay to kick Nao?*
(Maria points to 3 or "unsure")
2 ***INTERV****: so you don't think it would be so bad to kick Nao?*
3 ***MARIA****: hmm because maybe I don't know if he breaks but if he breaks maybe you could send him to a mechanic and fix it cause it one of those that can be fixed*

Moral Landscape

The second explanation is that the moral landscapes available to these children has been furnished by their lived experiences and that they hence compare the severity of moral acts using their limited conceptions of the worst possible moral actions as their reference point.

It's likely that the children have been taught that it is bad to break things in general, and especially expensive things. The children might have learned that breaking expensive things results in harsh consequences. Maybe a child once broke an expensive vase and got met with an emotional outrage from a parent. Maybe the child remembered this as one of the times the parent had gotten the maddest at them. The action of breaking a robot might get mapped to one such experience.

The reason that the child makes such a sever moral judgement might then be because in relation to their limited arsenal of acts that they understand as morally wrong, breaking expensive things might be up there. So, what I am proposing is that the children connect the consequences they think would result from the moral actions proposed in the questions to the consequences that would result to them if they would break something similarly expensive "at home". The moral judgement might then be decided by the severity of the expected consequences of the action. A severity which of course is relative to the consequences the child has been met with as the result of other wrongdoings. So, what I'm saying is that all these wrongdoings that the child has experienced furnish his or her personal moral landscape. And it is inside the frames of these moral landscapes that the children connect the actions that are proposed in the questions.

An adult would think to themselves how bad the worst acts are in general, like murder or rape, and then use these as reference. But since these kids haven't experienced the consequences that would be met by such acts, their moral landscape doesn't fully include these concepts yet. What I mean by "fully" here is that they probably understand these concepts, they know that someone would go to jail if they would murder, but they might not include them in their moral reasoning when comparing how bad different acts are. The assessment of the severity of a moral act is thus skewed by the lack of other acts available for comparison. Through the lens of this narrative, it is easier to understand why some kids judged the breaking of a robot as similar in severity to hurting an animal or a child.

So, it is likely that half of the children in this study understood the moral questions as binary choices rather than scales. But if this is not true, the "moral landscape" explanation might shed some light on why some children made such harsh moral judgements even though they did not attribute experience-related mental states to the robots.

If the first explanation is true, an overarching point can still be made. As has been stated, the children sometimes make harsh moral judgements that are not based on things like pain or

suffering caused to the moral subject of the question, but rather based on *rules* about what is and is not allowed to do. So, forgetting my theorizing about the *moral landscape*, the overarching point is that such rules are upheld by the "mapping-processes" described earlier. These children repeatedly refer to empty explanations like "because that's bad", "because you cannot do that", or "because that's mean", when asked to explain why they make the harshest moral judgements. They fail to give explanations that point to the suffering of a moral patient. Instead, they refer to consequences that the moral agent would be met with. Here are two examples.

Transcript 15
1 ***INTERV**: would it be okay if a child hit another child?*
2 ***ARAD**: no because then a grownup will really come and discipline them*

Transcript 16
1 ***INTERV**: would it be okay to kick Nao?*
2 ***IMRAD**: noo*
3 **INTERV**: why?
4 **IMRAD**: because he will tell the teacher

In relation to making moral judgements about breaking a robot, we are proposing that explanations like these, empty of reference to a suffering moral patient, are grounded in the child's *own suffering* as an expected consequence of committing the act of breaking something expensive. Since the questions are not asking how okay it would be if *they* kicked the robots, this explanation entails that the children instinctively ground abstract moral scenarios in their own bodies, as if it would be them that kicked the robot or locked it in the closet. One interpretation of this is that these children have a hard time placing a hypothetical moral agent in the moral scenario, which in effect transforms the scenario into a question of "would it be okay if *you* kicked to robot?". Which is an entirely different question.

Something that supports this is that some kids seem to have a more developed ability to reason and provide explanations that does point to a suffering moral patient. One child, Maria, often gave thorough explanations to her answers. When asked the moral question of if it would be okay to kick Nao, her reasoning was also centred around the breaking of the robot. However, her assessment was that she was unsure of her moral judgment: she pointed to "3/unsure" as her answer. This can be seen in transcript 14 above.

Maria thus seems to have a more nuanced understanding of the issue. She weighs the breaking of Nao against the possibility of fixing him at a mechanic. Comparing Maria's answer to John's after having interviewed both kids, it seems to me that Maria possesses more developed capacities for moral reasoning. She provided grounding for her thinking in a way that was uncommon in most other children. At 6 years old, she is one of the oldest children that participated. So, it's a natural possibility that she could be further in her mental development in some respects. If this is true, it suggests that moral questions like these are too difficult for 5-year-old children to reason around.

6. DISCUSSION OF METHODS

Research question 2a aimed to *"explore the use of moral questions as a starting point for reflection around mental state attribution in 5- and 6-year-old children".*

Research question 2b aimed to *"compare two approaches to conducting qualitative interviews: one-on-none interviews and interviews in small groups of two or three children".*

Overview

In this section we discuss these two research questions. They both aimed to explore the use of qualitative methods as ways to investigate the reasoning and preconceptions underlying mental state attribution. First, the effectiveness of the moral questions is presented and discussed. While the moral agency questions were mostly misunderstood by the children, the moral patiency questions seemed to spark a lot more reflection than the MSA-questions. We then evaluate what worked well in the two approaches and present the major differences between them.

RC-2a: Moral Questions as a starting point for reflection

We anticipated that the moral questions would spark more reflection and produce richer qualitative data in than the MSA-questions. To test this, we wanted to compare the effectiveness of moral questions and the MSA-questions as starting points for reflection.

The moral questions were anticipated to result in richer and more reflective qualitative data compared to the MSA-questions. We reasoned that when a child would be asked to explain their mental state attributions (for example, *why* they thought Nao could feel pain) it would be hard for them to give thorough explanations due to the abstract nature of the questions. In contrast, since the moral questions pose more practical scenarios, we anticipated that these questions would work better as starting points or "launch pads" sparking reflection, and thus result in richer qualitative data.

So, how well did this work? The moral agency question had inconsistent results. When the question was posed about Kim the robotic lawnmower, it worked well. But when posed about the other robots, it was almost always misunderstood. This will be discussed in more detail later.

Moral patiency questions

The moral patiency questions, however, worked very well. The dialogues that happened when the children were asked the moral patiency questions made up the majority of the reflective qualitative data collected from the interviews. Almost all transcriptions used in theme one and theme two happened in connection to the moral patiency questions. Even though the interviewer asked the same follow-up question (normally a simple "why do you thinks so?") these questions seemed spark more reflection in the children.

Although, many times, the moral patiency questions did not spark any reflection, and instead circled around back to a mental state attribution. In such cases, the reflection would end there, begging the question of *why* the children attributed mental states like they did. We thus ended up at the same place as with the mental state attribution questions. This can be seen in the following transcript.

Transcript 17
1 ***INTERV****: would it be okay to kick Nao?*
2 **JOHN***: no it is not okay*
3 **INTERV**: why not?
4 **JOHN**: he will get hurt then

In this situation, the moral patiency questions only resulted in a reference back to John's earlier attribution of Nao's ability to feel pain. The moral question thus only functioned as a way to replicate the earlier mental state attribution. It thus begs the question of why John thinks Nao can feel pain. So, in this case (and other similar cases), the moral question did not spark any reflective explanations in relation to the underlying reasoning behind the mental state attribution. While this could be interpreted as a reason to doubt the efficiency of the moral patiency questions, the occurrence of similar cases to this was a lot higher for the MSA-questions that for the moral patiency questions.

Improvised formulations

As described in the method section, the interviews in group approach were more loose ended than in the solo approach. The interviewer did not follow the forms as closely which lead to improvised formulations of the MSA-questions and the moral patiency questions. The questions were thus modified in real time and adjusted to suite specific children. An example of this can be seen in transcript 8, where the interviewer formulates one of the MSA-questions about pain in the following way:

Outtake from Transcript 8
1 ***INTERV****: do you think Atlas can feel pain like you feel pain if someone pinches you?*

Interestingly, a result of the improvised MSA-questions was that they started to spark more reflection around the underlying reasoning behind mental state attribution in the children. An example can be seen in the transcript below.

Transcript 18
1 ***INTERV****: do you not think Kim would get hurt even if I threw him across this room?*
2 ***SARA****: no*
3 ***IRINA****: no not*
4 ***NICK****: but if you would throw him out of this window, then he would get hurt*
5 ***SARA****: mhm*
6 ***INTERV****: what would happen then you think?*
7 ***IRINA****: he would get angry*
8 ***NICK****: he would get angry and fight*
9 ***SARA****: I don't know if he can fight*

Like in this example, most of the improvised MSA-questions were more embedded in real-world scenarios. And since this change resulted in an increased number of reflective answers, one could interpret this increase as a result of this embeddedness in real-world scenarios. This, in turn support the original reasoning that guided us to the idea of using moral questions in the first place: the notion of facilitating children's reasoning around abstract concepts related to agency and experience by "hiding" them under the disguise of moral scenarios.

However, this interpretation does not rest on steady ground. An alternative interpretation of the implications of the improvised MSA-questions is the following: Due to the commonality of these kinds of real-time improvisations of MSA-questions in the group approach, it is difficult to find epistemological footing in statements about the effectiveness of the moral questions in comparison to the MSA-questions. In the group approach, the interviewer simply did not follow the pre-made sets of questions in a manner that was strict enough to say something about how effective the moral patiency questions were.

Moral agency questions

It can be hard even for adults to understand and reflect on concepts like agency, intentionality, and volition in their abstract forms. One could then assume that this would be problematic for children. But children regularly use these concepts when ascribing moral blame in their own lives (for example: *"he did it on purpose!"*). However, they do not have knowledge of these concepts in their abstract form (i.e., the words "intentionality" or "agency"). Question A1 *"if the robot would step on/run over a beetle/hedgehog, would it be the robot's fault then?"* was formed as a means to ask the children about agency-related mental states by providing a practical context that was thought to be easier to understand for the children.

The question worked well when it was asked about Kim the robotic lawnmower. When the question was asked about Kim, children almost never had an issue understanding it, and gave rich reflective answers. Here is an example from the interview with Maria and Cora.

Transcript 19
*1 **INTERV**: if kim the robot lawnmower would run over a hedgehog, do you think it's Kim's fault then?*
*2 **MARIA**: mm yes I think so*
*3 **CORA**: eeh no*
*4 **MARIA**: because robot lawnmower they that controls itself while hedgehogs just walk where they want*
*5 **CORA**: they walk more in the forest*
*6 **INTERV**: so you don't think it would be the lawnmowers fault?*
*7 **CORA**: yes I do*
*8 **MARIA**: maybe because because computers, they can, they don't listen, they can't talk to others so therefore they go wherever they want, but robot vacuum cleaners you can control them with the phone, (short pause) but not robot lawnmower cause they just drive by themselves*
*9 **INTERV**: oh okay*
*10 **MARIA**: because one time our robot lawnmower drove by itself to our neighbors garden*

When the question was asked about other robots, confusion often arose. Common responses to the question when posed about other robots were: "yes, if it was on purpose", or "no, if it was not on purpose". Sometimes this could be solved with a clarifying follow up question of something like: "do you think the robot could do it on purpose?", but most often it could not. A similar answer was often uttered in response to another question related to agency, the fourth mental state attribution question (MS4) *"can the robot do things on purpose? If so, how much?"*: this answer was "*both*" with the child often pointing to both "1/nothing" and "5/very much". While these answers do indicate that the child thinks the

robot can do *some* things on purpose, and that they would place moral guilt on the robot *if* the action portrayed in the question was on purpose. But since the answers clearly indicate misunderstanding, this is invalidated.

Question A2 was initially formulated as a unique question for the robotic lawnmower. The circumstances when the question is asked in the context of the lawnmower driving around on the lawn and a hedgehog walking in its path is natural and easy to imagine. We also thought this was a good way to spark reflection around moral agency since we predicted that many of the children would be familiar with lawnmower robots. To form a similar scenario around the other robots, one needs to create a scenario where the other robots expose some moral patient (the hedgehog/beetle) to harms way in a similarly natural way as the lawnmower just "doing its job" on the lawn.

But even in the few cases where the children understood the question, with or without clarification, the scenario was fundamentally different when posed around the robotic lawnmower. The difference lies in that the lawnmower drives around on the lawn doing its thing, and then a hedgehog appears *in its space.* The lawnmower has knives on its belly that can kill the hedgehog, but these knives are meant for cutting the grass. It is thus a naturally occurring scenario that could easily be interpreted as an accident. When posed around another robot, on the other hand, the robot imposes on the space of the moral patient regardless of whether it was "on purpose" or not. Another difference lies in how bad it is to kill a beetle compared to killing a hedgehog. It is common to sympathize with and value the life of a small and cute mammal over an insect like a beetle.

RC-2b: Comparison between the two approaches

The main strength of the group approach was that children were more reflective in their reasoning. The children in the group approach were a lot more talkative and expressive than the children in the solo approach. This resulted in more reflective answers. However, they were also a lot more inclined to be mischievous, which resulted in many interactions being interrupted due to mischievous behavior or distraction. Children in the solo approach were almost never mischievous, but they did not produce as reflective answers. Another important aspect differentiating the two approaches was that the solo approach allowed for deeper insight into the thinking of individual children. In the group approach, especially in groups of three children, it was hard to follow the reasoning of individuals. Something that supports the lack of individuality in the group approach is the suspicion that children in the group approach seemed to answer the questions collectively or as unity.

Collective answers

On many occasions, and in different ways, children in the group interviews seemed to collectively answer the questions. This could be seen in various ways. The most common indication was that two or more children first gave the same answer on a question and were then asked to explain their answers. In their explanations, the children would fill each other in or answer follow-up questions for each other, as if they both had complete transparency regarding the other's thoughts. This indicates that the children might have thought of their explanations as *one product* that they created together. An example can be seen in transcript 19 with Maria and Cora. An implication of this is the risk of children being influenced by other children, which would mean that they failed to express their individual thoughts and beliefs.

Loose interviews

It should be noted that surrounding factors cannot be excluded as possible explanations for these differences. The structure of the interviews for the group approach was more loose and less centered around the forms. This naturally leads to more reflective answers. Although, the decision to conduct more loose interviews was made just after the first group interviews, because it felt stale and difficult to strictly follow the forms. It is thus harder to conclude that the reason for the differences in the reflectiveness of the answers was related to the fact of it being conducted in a group or one-on-one.

7. CONCLUSION

The first research question aimed "*to investigate the reasoning and preconceptions underlying 5- and 6-year-old children's mental state attribution to robots.".* Through an interpretation of the themes, two main preconceptions were found to have influenced the children's mental state attribution. The first preconception "Robots as Taxon" describes how children tended to disregard individual differences between the robots and instead evaluated them as the general notion of "robot". The second preconception "Archetypal frameworks" describes how children tended to perceive some robots (Atlas and Nao) as archetypes. This was argued to work as a matchmaking process where perceived information about the robots were fit to archetypes influenced by children's, and adult's culture. Theme three was then argued to illustrate how children reason when they are presented with conflicting influences. When the children were presented with robots that activated their anthropomorphize tendencies, while simultaneously being influenced by the "Robots as Taxon" preconception, they come up with stories about the robot's material as explanatory models.

Research question 2a aimed to *"explore the use of moral questions as a starting point for reflection around mental state attribution in 5- and 6-year-old children".* The moral agency question was misunderstood by most children when posed about all robots except for Kim. The moral patiency questions were evaluated as successful starting points for reflection around mental state attribution. These questions stood for the majority of the qualitative data used in the themes.

Research question 2b aimed to *"compare two approaches to conducting qualitative interviews: one-on-none interviews and interviews in small groups of two or three children".* The two approaches had their respective strengths. The group approach resulted in more reflective reasoning, while lacking in individual depth.

8. REFERENCES

Axell, C. & Berg, A. (2022). Primary pupils' conceptions about 'programming' and what it is. (Submitted).

Cecilia Axell, Astrid Berg, Jonas Hallström, Sam Thellman & Tom Ziemke (in press). Artificial Intelligence in Contemporary Children's Culture – A Case Study. In: Pupils' Attitudes Toward technology 39, to appear.

Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., & Šabanović, S. (2020). Human-Robot Interaction: An Introduction. Cambridge University Press.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science robotics*, *3*(21), eaat5954.

Bernstein D & Crowley K (2008) Searching for Signs of Intelligent Life: An Investigation of Young Children's Beliefs About Robot Intelligence, THE JOURNAL OF THE LEARNING SCIENCES, 17:2, 225-247, DOI: 10.1080/1050840080198611

Brink, K. A., Gray, K., & Wellman, H. M. (2019). Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child development*, *90*(4), 1202-1214.

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., & Taylor, L. (2018). Portrayals and perceptions of AI and why they matter. (DES5612)

Chi, M. T. (1978). Knowledge structures and memory development. *Children's thinking: What develops*, *1*, 75-96.

Di Dio, C., Manzi, F., Peretti, G., Cangelosi, A., Harris, P. L., Massaro, D., & Marchetti, A. (2020). Shall I trust you? From child robot interaction to trusting relationships. *Frontiers in Psychology*, *11*, 469

Fink, J., Mubin, O., Kaplan, F., and Dillenbourg, P. (2012). "Anthropomorphic language in online forums about Roomba, AIBO and the iPad," in Proceedings of the 2012 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), Munich: IEEE, 54–59.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mental state attribution. *science*, *315*(5812), 619-619.

Gray, K., Young, L., & Waytz, A. (2012). Mental state attribution is the essence of morality. *Psychological inquiry*, *23*(2), 101-124.

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

Jipson J. L. and Gelman S. A (2007). Robots and rodents: Children's inferences about living and nonliving kinds. Child Development 78, 6 (2007), 1675–1688.

Kahn, P. H., Friedman, B., Perez-Granados, D. R., & Freier, N. G. (2006). Robotics pets in the lives of preschool children. Interaction Studies, 7, 405–436.

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., ... Shen, S. (2012). ''Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. Developmental Psychology, 48, 303–314.

Melson, G. F., Kahn, P. H., Beck, A., Friedman, B., Roberts, T., Garrett, E., & Gill, B. T. (2009). Children's behavior toward and understanding of robotic and living dogs. Journal of Applied Developmental Psychology, 30, 92–102

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

Murashov, V., Hearl, F., & Howard, J. (2016). Working safely with robot workers: Recommendations for the new workplace. *Journal of occupational and environmental hygiene*, *13*(3), D61-D71.

Okanda, M., Taniguchi, K., Wang, Y., & Itakura, S. (2021). Preschoolers' and adults' animism tendencies toward a humanoid robot. *Computers in Human Behavior*, *118*, 106688.

Sandoval, E.B., Mubin, O., & Obaid, M. (2014). Human Robot Interaction and Fiction: A Contradiction. In: Beetz, M., Johnston B., & Williams, M.A. (Eds.) Social Robotics. ICSR 2014. Lecture Notes in Computer Science, vol 8755. Springer, Cham.

Severson Rachel L. & Lemm Kristi M. (2016) Kids See Human Too: Adapting an Individual Differences Measure of Anthropomorphism for a Child Sample, Journal of Cognition and Development, 17:1, 122-141, DOI: 10.1080/15248372.2014.989445

Sommer, K., Nielsen, M., Draheim, M., Redshaw, J., Vanman, E. J., & Wilks, M. (2019). Children's perceptions of the moral worth of live agents, robots, and inanimate objects. *Journal of Experimental Child Psychology*, *187*, 104656.

Spektor-Precel, K. & Mioduser, D. (2015a). The influence of constructing robot's behavior on the development om Theory of Mind (ToM) and the Theory of Artificial Mind (ToAM) in Young Children. IDC '15: Proceedings of the 14th International Conference on Interaction Design and Children. (pp. 311–314). https://doi.org/10.1145/2771839.2771904

Spektor-Precel, K. & Mioduser, D. (2015b). 5-7 Year Old Children's Conceptions of Behaving Artifacts and the Influence of Constructing Their Behavior on the Development of Theory of Mind (ToM) and Theory of Artificial Mind (ToAM). Interdisciplinary Journal of e-Skills and LifeLong Learning, 11, 329–345. http://www.ijello.org/Volume11/IJELLv11p329-345Spektor1973.pdf

Subrahmanyam. K, Gelman. R, and Lafosse. A (2002). Animates and other separably moveable objects. In Category Specificity in Brain and Mind, Emer Forde and Glyn Humphreys (Eds.). Psychology Press New York, NY, 341–373.

Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. ACM Transactions on Human-Robot Interaction.

Tung, F. W. (2016). Child perception of humanoid robot appearance and behavior. Int. J. Hum. Comput. Interact. 32, 493–502. doi: 10.1080/10447318.2016. 1172808

Waytz, A., Heafner, J., & Epley, N. 2014. The mind in the machine. J. of Exp. Soc. Psychology 52, 113-117.