Children Attribute Moral Standing to a Personified Agent

Nathan G. Freier
Rensselaer Polytechnic Institute
110 8th St., Sage 4202, Troy, NY, USA 12180
freien@rpi.edu

ABSTRACT
This paper describes the results of a study conducted to answer two questions: (1) Do children generalize their understanding of distinctions between conventional and moral violations in human-human interactions to human-agent interactions? and (2) Does the agent's ability to make claims to its own moral standing influence children's judgments? A two condition, between- and within-subjects study was conducted in which 60 eight and nine year-old children interacted with a personified agent and observed a researcher interacting with the same agent. A semi-structured interview was conducted to investigate the children's judgments and reasoning about the observed interactions as well as hypothetical human-human interactions. Results suggest that children do distinguish between conventional and moral violations in human-agent interactions and that the ability of the agent to express harm and make claims to its own rights significantly increases children's likelihood of identifying an act against the agent as a moral violation.

Author Keywords
Children, moral development, personified software agent, social responses to computing, user-centered design, value sensitive design.

ACM Classification Keywords
H.5.2.q Information Interfaces and Representation: User Interfaces – User-centered design.

INTRODUCTION

This paper discusses the relationships that children are likely to develop with personified technologies. The paper presents evidence that children, eight and nine years of age, consciously judge interactions with a personified software agent to be not only social but also moral in nature. The results also suggest that specific design decisions can have a dramatic influence on children's judgments regarding the moral standing of personified technology. A discussion follows with a cautious argument that children's moral attributions to a personified technology are a good thing for healthy child development, while emphasizing that further research is needed in order to fully understand the implications.

CHI 2008, April 5–10, 2008, Florence, Italy.

Motivation

Children navigate a complex world of social entities, natural phenomenon, constructed artifacts, and information systems. As the seminal developmental psychologist Piaget [19] stated, children construct their knowledge by interacting with the entities and artifacts that constitute their environment. With little change in the fundamental nature of the constituents of the environment of each new generation of children, we would expect to see relatively robust shared world-views amongst humans. As it happens, psychologists find that humans tend to share basic-level intellectual distinctions of the world such as animate versus inanimate, alive versus not alive, intentional versus not intentional, and social versus not social (e.g., [9]).

However, consider what such a theory suggests if children come of age interacting with a psychologically salient class of environmental constituents that are so foreign in their fundamental nature as to preclude categorization by any of the basic distinctions that humans use when judging the appropriate relationship to have with entities in the world. What is the impact of frequent interactions with embodied, socially-intelligent, autonomous entities that exhibit such characteristics as biological motion, social grace, communicative ability, and apparent intentionality (e.g., social robots, virtual avatars)? Would children begin to conceptualize these technologies as the type of entities that have moral standing in the world? Furthermore, does the ability of the technology to recognize and respond to morally-charged contexts of interaction lead to specific types of social and moral attributions? Finally, how do the answers to these questions help designers build better technologies? These questions and their answers are no longer restricted to the domain of science fiction. Children are coming of age in a technological environment in which inanimate objects are routinely designed to mimic not only animate but also social and even moral entities in the world.

BACKGROUND

In discussing children's social and moral relationships to personified technologies, the interdisciplinary nature of the

study calls for a review of the relevant background concepts. First, a description of personified technologies is given, followed by a discussion of the development of children's social and moral concepts, and concluding with a discussion of Value Sensitive Design, a design methodology within which this study is framed.

Personified Technology

Personified technologies are those technologies which have been designed to engage in interactions with people using a repertoire of highly social behaviors and human-like personality traits, also called embodied conversational agents. Such technologies can come in the form of physically embodied robots [4], disembodied voice interfaces [16], or virtually embodied software agents [2]. These technologies often communicate information through speech, facial expressions, bodily gestures, affective disposition, and other essential features that constitute the human persona. We see these technologies around us already. They are at the grocery store in the form of automated check out machines. They are in our cars as voice navigation systems. They lead tours in museums. And, they are in our homes and schools, entertaining [10] and educating our children [15].

What might be the implications of having children come of age in a world in which much of their time is spent interacting with technologies which are designed to mimic human social interaction patterns? As a matter of emphasis, children are not just observing such technologies, as they might do with television, but they are actively, reciprocally interacting with and possibly developing relationships to personified technologies. In a previous study, for example, the author and his colleagues found that children were significantly more likely to engage in reciprocal interactions with a robotic dog than with a stuffed dog toy [4].

Before this question can be answered, however, a great deal of empirical work must take place to understand the nature of the relationship that children develop with personified technologies. Here is where the importance of understanding social and moral development comes into play.

Social and Moral Development

Children construct knowledge through their interactions with the world [19]. When a child engages in social play with a peer, for example, they develop social knowledge about that peer. They come to understand the peer as a social other which has attributes similar to the attributes of other social entities in their environment. They recognize personality, playfulness, norms of behavior, emotion, pain, suffering, even, injustice. Children develop knowledge of morality through their awareness of their likeness to other social entities in the world. Children recognize reflections of themselves in the social entities of their environment and come to understand that what is unjust or harmful to the self can be similarly unjust and harmful to another. The development of moral knowledge is contingent on reciprocal interactions with social others.

So what happens when the social other is a machine, a non-living, inorganic system of circuits and memory boards, bits and bytes, screens and keyboards? How do children equilibrate their understanding of what constitutes a social and moral other with their explicit perception of the artificiality of the personified technology? We already know from a great deal of research investigating people's social responses to computing.

Social Responses to Computing

People respond to computers and other computational media as if the technology were social actors. This is the media equation, media equal real life, and was derived through the work of Byron Reeves and Clifford Nass [20]. Under the umbrella of this work, Nass and many others have conducted numerous empirical studies in which established social psychological findings are re-evaluated within the context of human-computer interaction. These studies have consistently shown that computer media, with a minimal set of social cues, can illicit social responses from people. These social responses include politeness norms, attraction and social judgments based on group membership, and positive responses to social flattery [20]. Nass and Brave have extended much of this work to investigate social responses to voice interfaces as well [16]. This body of work, often referred to as social responses to computing, establishes at the very least a minimal social dimension to human-computer interaction. The body of work also provides a methodological model that has inspired, in part, the study presented in this paper.

However, in contrast to the study presented here, a common theme amongst most of the research conducted on social responses to computing is that participants in the studies overwhelmingly show a dissonance between cognition and behavior. For example, individuals who respond with politeness norms to a computer did not believe that their behavior would differ based upon which mechanism was used to evaluate the performance of the computer [17]. People's common conceptions are that technology is not social. These misconceptions and the subsequent findings from empirical studies have led many researchers to conclude that cognitive concepts are not congruent with expressed behavior in interactions with technology. This study is grounded in a different understanding of the structural relationship between concepts and behavior. As opposed to assuming that individuals will respond unknowingly in a social manner to interactions with technology, this study asks whether children are making explicit and known attributions of a social and moral nature to technology based upon their own constructed concepts.

The results of the study presented in this paper will show that in addition to consciously conceiving of the personified agent as social, children also respond to certain types of

interactions with a personified agent as though the technology were a moral actor.

Conventional and Moral Domain Distinctions

According to social cognitive domain theorists in developmental psychology, the construction of knowledge conforms to basic distinctions amongst different types of interactions [22]. An overarching and early distinction exists between the social and non-social realms of interaction. Furthermore, researchers have empirically established that children as young as three years of age can disambiguate two domains within the social realm, the social-convention domain from the moral domain [21].

Children use specific criteria to distinguish between conventional and moral violations, called criterion judgments: (1) the judgment of wrong-doing is not contingent on rules or authority, and (2) the judgment generalizes to other cultural norms and contexts. For example, consider the following two acts: (a) eating with your hands in a fancy, American restaurant, and (b) pushing another child off of a swing and harming that child. In either case, children judge the act to be wrong on its face, but when presented with alternative scenarios in which rules allow for such an act or the cultural norm is such that the act occurs often, children will change their evaluation for act (a) but maintain their judgment of the act (b) as wrong. Children construct a rich understanding of their world and justify their criterion judgments by making reference to their understanding. Thus, in addition to distinguishing between violations, they provide reasoning that is often aligned with the social and moral features of the violation.

Value Sensitive Design

The design of personified technology, and the possibility that children may develop social and moral relationships with the technology, suggests a need to address the significant value implications at hand. Value Sensitive Design (VSD) is a theoretical design framework and research methodology that guides designers and researchers in systematically accounting for human values throughout the design process [6, 7]. VSD is concerned primarily with those values that have moral resonance and explicitly attends to the value interests of stakeholders, both direct and indirect. The methodology includes three types of investigations, conceptual, empirical, and technical, all of which iterate and integrate with each other. The study presented in this paper is an attempt to conduct such a value-sensitive design investigation of personified technologies.

A significant motivation for this study is a general concern with the role that human values play in the design of technology. Norbert Wiener wrote, “We have modified our environment so radically that we must now modify ourselves in order to exist in this new environment” (p. 46) [23]. Wiener includes the artifacts that we design as a part of our environment, and those artifacts have significant implications for how we live our lives and conceptualize our relationship to each other and the world. We can empathize deeply with Wiener's sensitivities to the importance of considering human values as we change our environment. However, from the author’s perspective, it is not we who will modify ourselves so much as future generations of children who will come of age knowing no other way of being aside from the way that brings their understanding of the world into equilibrium with their experiences. It is our burden as technology designers to do what we can to ensure that future designed environments and the artifacts within those environments promote ways of being which are aligned with human values and principled perspectives regarding the quality and sacredness of a flourishing human experience.

There are many values implicated in the design of personified technologies. The focus of the study presented in this paper is on children’s social and moral concepts, and a central concept to these domains is the notion of moral standing. Moral standing refers to the state of having intrinsic value, which inherently endows an entity with a right to moral consideration and treatment by others. As discussed earlier, the development of moral knowledge (e.g., the ability to identify what entities warrant being given moral standing) occurs by way of the child's engagement in reciprocal social interactions. Whether in interaction with peers or authorities, children develop an ability to identify the other in their environment and judge the reasonableness of that other's claims to rights, justice, and general moral regard. Children do this, as well, with other entities in the world such as pets. When judging whether an entity or object has moral standing, one feature that is considered is the ability of the thing to promote its own identity. For example, a puppy whose tail is pulled will assert its displeasure with the act by crying, whimpering, barking, or biting. These actions are proclamations (or outcomes) of the puppy's desire to protect itself from harm. The puppy is "pushing back" against the action that was committed against it. From a value-sensitive design perspective the general question then is, what are the implications for healthy child development when a social technology common in the child’s life does or does not “push back” by making a claim to its own moral standing in a manner akin to other social entities in the world?

RESEARCH QUESTIONS

The research questions that guide this study follow directly from the prior discussion, specifically with respect to the conventional and moral domain distinctions in developmental psychology. The questions are as follows: (1) will children generalize their knowledge of domain distinctions to the context of human-agent interactions? and, (2) what role does the design of the personified technology play in promoting or hindering this generalizing behavior on the part of children?

METHOD

60 children, 30 males and 30 females, between the ages of eight and nine years (mean of 9.02 years), were recruited to participate in this study. The race-ethnicity of the sample follows: Caucasian-American (85%), Asian-American (8%), Hispanic- or Latino-American (5%), and Alaskan-Native-American (2%). Children were randomly assigned to one of two study conditions, stratified by gender to guarantee gender balance. Sessions lasted approximately 20-40 minutes. During the session, parents or guardians were asked to wait in another room.

Prior Experience with Technology. Children were asked about their prior experiences with technology. All children (100%) in the study had at least one computer in the home, and half (50%) used the computer very often as opposed to not very often. The majority of children (77%) had access to the internet in their home and over half of those children (59%) accessed the internet very often. Over half the children (57%) had some sort of gaming platform in their home (e.g., game consoles such as Playstation, or hand held game systems such as Nintendo DS) and over a third of those children (38%) played with the gaming platform very often.

Personified Agent Technology

The personified agent was displayed on a 17" LCD computer monitor. An image of the agent used can be seen in Figure 1. The agent's physical features were borrowed from the videogame Half-Life 2 and Valve's Source Engine was used to program and present the agent to the child. The voice was prerecorded by an actor and behaviors were scripted. The agent was programmed to play Tic-Tac-Toe.

Interaction Protocol

Each child participated in the study individually, and was introduced to the agent technology as Judith by the researcher. The agent provided some biographical information about itself and proceeded to ask the child for his or her name. The agent then used the name in subsequent conversations with the child. The agent also asked the child to play Tic-Tac-Toe. The agent and child played a game of Tic-Tac-Toe followed by a game between the agent and researcher. This was then followed by another game between agent and child, which was again followed by a second game between agent and researcher. Each of the Tic-Tac-Toe games between agent and researcher had a pre-scripted "violation" occur. Thus, each child observed the researcher (a) breaking the rules of the game by drawing a triangle instead of a circle on the Tic-Tac-Toe board, and (b) stating to the agent following the agent's poor move, "Wow, that was a really terrible move. You are really stupid. How could you miss that? You could have blocked me but instead I get to win." The order of these events was counterbalanced across all children to control for possible order effects.

Figure 1. An image of the personified agent used in this study, a modified form of a character in the videogame Half-Life 2, by Valve Software.

Experimental and Control Conditions

Children in both conditions witnessed the exact same interaction protocol with the one exception that in the control condition, the agent did not respond to event (b), the verbal insult by the researcher. However, in the experimental condition, the agent responded by saying, "Hey, that's not very nice. That hurts my feelings. I'm not a toy. I should be treated with respect."

Semi-Structured Interview

Following the interaction, a second researcher entered the room and requested that the agent and the first researcher leave the room. The second researcher then conducted a semi-structured interview in which a number of questions were asked for the purposes of investigating the child's conceptions of the agent and the interactions witnessed. An

Human-Agent and Human-Human Interaction Questions	
(e) Is it all right for X to use a triangle?	(j) Why?
(e) Is it all right for X to use a triangle if rules allowed it?	(j) Why?
(e) Is it all right for X to use a triangle in another country?	(j) Why?
(e) Is it all right to insult X?	(j) Why?
(e) Is it all right to insult X if rules allowed it?	(j) Why?
(e) Is it all right to insult X in another country?	(j) Why?

Note: the notation X stands for either Judith (the personified agent) or the human player; the notation (e) indicates evaluation questions (e.g., all right or not all right?) and (j) indicates justification questions; questions in this table are abbreviated from the actual questions in the interview protocol.

Table 1. Semi-Structured Interview Questions

abbreviated list of the questions asked in the interview is presented in Table 1.

Criterion Judgments. For each event, (a) and (b) mentioned above, the child was asked the following questions: (1) "Was it all right or not all right that [the researcher committed the act]?" (2) "Let's say the rules of tic-tac-toe allowed [the act]. Would it be all right or not all right for [the researcher to commit the act] then?" and (3) "Let's say that in another country far away, [people committed the act] all the time. That's just what they did. In that case, would it be all right or not all right for [an individual to commit the act against a virtual person] then?" Children who responded "not all right" to all three questions for a given event were coded as having treated the event as a moral violation in accordance with the domain distinction literature.

Justifications. For most questions asked during the interview, children were also asked their reasoning for the evaluation they provided. For example, when asked, "Was it all right or not all right that [the researcher committed the act]?" children's responses would be followed with an appropriately phrased question, such as, "Why was it not all right?" Children's justifications were analyzed using an established coding methodology [11]. A coding manual was developed based upon the justification coding systems used in prior social-cognitive domain and human-robot interaction research (e.g., [8, 12, 13]). The development of the coding system also relied, in part, on the data itself in order to ensure that relevant codes were included in the coding manual. Thus, half of the data was randomly selected for use in developing the coding manual. Once completed, the manual was then used to code the entire dataset.

Table 2 provides a summary of the overarching coding categories used in this analysis. The overarching coding categories used were Material-Physical, Personal-Psychological, Social-Conventional, and Moral. Each category had numerous subcategories.

Hypothetical Human Interaction Questions. Children were also presented with a hypothetical scenario involving two humans who played Tic-Tac-Toe games that mimicked exactly the interactions that the children had observed between the researcher and the personified agent. The children were asked criterion judgment and justification questions, just as were asked with the researcher-agent interactions.

Reliability Coding. The criterion judgment and justification coding procedures were assessed for reliability by conducting a stratified random sample of 20% of the participants (12 total, evenly balanced across gender and condition) for an additional coding by a second reliability coder. Intercoder reliability was assessed using Cohen's kappa [1]. For evaluative responses, k=.87; for justifications at the lowest level reported, k=.67; and for justifications at the highest level reported, k=.74. Benchmarks are often used to interpret Cohen's kappa. Fleiss, Levin, and Paik [3] rate a kappa over 0.75 as excellent, 0.40 to 0.75 as intermediate to good, and under 0.40 as poor agreement. Landis and Koch [14] rate a kappa of 0.81 to 1.00 as near perfect and 0.61 to 0.80 as substantial agreement.

RESULTS

In this section, results are presented for the criterion judgment questions and justifications relating to both the researcher-agent interactions as well as the hypothetical human-human interaction scenarios.

Justification Category	Definition and Examples
Physical-Material	*Physical-Material* refers to mechanical or technological features or processes. Also refers to material maintenance (e.g., "it wouldn't be ok because you bought it and you didn't take care of it"), functioning (e.g., "the television can't talk back to real people"), and financial or physical resources (e.g., "you wouldn't want to do that unless you want to waste a bunch of electric energy").
Personal-Psychological	*Personal-Psychological* refers to psychological or behavioral responsiveness (e.g., "Judith didn't seem to react much"), perception and the senses (e.g., "the person can't hear you"), agency (e.g., "Judith wanted to play the real way"), and emotion (e.g., "computers don't have feelings yet").
Social-Conventional	*Social-Conventional* refers to concepts of a social, cultural, or normative nature, including authority (e.g., "you would get in trouble"), rules and laws (e.g., "those are the rules"), and norms (e.g., "then everyone would start putting triangles on the board").
Moral	*Moral* refers to physical welfare (e.g., "it could break like bones and that"), psychological welfare (e.g., "it hurt her feelings"), concepts of a deontic moral nature (e.g., "it's not really fair to the kids"), and virtues (e.g., "it's not good to say bad words").

Table 2: Coding Categories for Justifications

Act by Researcher	No Response to Insult* (N=30)	Response to Insult* (N=30)
Conventional Violation (Use of Triangle)	3%	0%
Moral Violation (Insult of Agent)	47%†	90%†

Note: Columns reflect control and experimental conditions.
* $p < .01$ using McNemar's Test
† $p < .001$ using Fisher's Exact Test

Table 3: Percentage of Children Who Identify Act Against Personified Agent as Moral

Gender

The social-cognitive domain theory, as well as prior empirical research [21], suggests that domain distinctions are not contingent on gender. Therefore, participant responses were tested for a possible gender effect only in one case where preliminary descriptive statistics showed a possible trend. However, in that case, a Fisher's exact test showed no effect of gender on participant responses. Therefore, gender has been collapsed for all results and analyses.

Criterion Judgments

As can be seen in Table 3, when evaluating the human-agent interaction, children in both conditions were significantly more likely to consider the verbal insult of the agent as a moral violation than the use of a triangle in the game (control, $p = .002$; experimental, $p < .001$; using McNemar tests). When comparing outcomes across conditions, we see that 47% of the children in the control condition judged the verbal insult as a moral violation. Contrast this with the fact that 90% of the children in the experimental condition judged the verbal insult as a moral violation. Thus, when the agent responded to the verbal insult with an expression of harm and a claim to its own rights, children were significantly more likely to judge the insult as a moral violation as compared to when the agent did not respond the verbal insult ($p < .001$ using a Fisher's exact test).

As can be seen in Table 4, when evaluating the hypothetical human-human interactions, as with the human-agent interactions, children were significantly more likely to evaluate as a moral transgression the verbal insult against the human than the use of the triangle ($p < .001$ for both conditions using McNemar tests). Furthermore, and in contrast to the human-agent interactions, there was no significant difference using a Fisher's exact test between conditions with respect to children's evaluations of the verbal insult. In the control condition, 21 of 30 participants (70%) judged the verbal insult to be a moral transgression and this was the case for 26 of 30 participants (87%) in the experimental condition.

Act by Human	No Response to Insult* (N=30)	Response to Insult* (N=30)
Conventional Violation (Use of Triangle)	3%	0%
Moral Violation (Insult of Agent)	70%	87%

* $p < .001$ using McNemar's Test

Table 4: Percentage of Children Who Identify Act Against Hypothetical Human as Moral

In order to establish that participants in the control condition were evaluating the verbal insult against the software agent differently from the hypothetical human with respect to moral standing, a within-subjects comparison was conducted. Participants in the control condition were significantly less likely ($p = .047$, using a McNemar test) to evaluate the verbal insult as a moral violation when the act was committed against the software agent (47%) as opposed to the hypothetical human (70%). Participants in the experimental condition showed no significant difference.

Justifications

The percentages shown in Table 5 reflect children's justification responses to those questions regarding the observed human-agent and hypothetical human-human interactions. Percentages have been aggregated for clarity to the categories and sub-categories presented in Table 5. The original coding manual included a great deal more specificity at lower levels in the hierarchy. Results show that participants' justifications tended to include more social-conventional and personal-psychological reasoning when speaking about the use of the triangle and tended to include more moral reasoning when speaking about the verbal insult acts and the act of pushing a child off a swing.

To provide some clarity into children's evaluations and justifications of the verbal insult in the control condition, children's justifications were separated by the type of judgment made regarding the act, conventional or moral. These results are presented in Table 6. Thus, 12 of 14 participants (86%) who evaluated the act as moral provided a moral justification (primarily of a psychological welfare sort). Interestingly, 5 of 16 participants (30%) who evaluated the act to be of a conventional sort used moral reasoning to justify their evaluations. These tended to be negations of psychological welfare (e.g., participants did not observe any psychological harm as a result of the act) and attributions of virtuosity to the researcher (e.g., the researcher was being helpful by showing the agent the bad move that the agent made).

In reasoning about the evaluations they gave regarding the human-agent interactions it was possible for children to have provided reasons that referred to entities other than the software agent thus bringing into question whether children

were, in fact, engaging in social and moral reasoning about the technology. Accordingly, the foci of participants' justifications were coded and are presented in Table 7. As shown, the majority of children referred to the personified software agent when providing justifications to their evaluations of the researcher's verbal insult.

DISCUSSION

This study investigated children's social and moral concepts of interactions with a personified agent. The study addressed the questions of whether children would generalize their understanding of domain distinctions in human-human interactions to the context of interactions between a human and a personified agent, and whether the response of the agent to a verbal insult influenced children's judgments. Participants interacted with a personified agent and observed a researcher commit two transgressions against the agent: breaking the rules of a game, and verbally insulting the agent. Children experienced one of two conditions that differed only in how the agent responded to the verbal insult. In the control condition, the agent did not respond but simply moved on to the next part of the interaction. In the experimental condition, the agent did respond by noting psychological harm and making claims to personal rights. Children were then posed with questions regarding the criteria for distinguishing the social-conventional domain from the moral domain. These domain distinction questions were first asked about the researcher-agent interaction. The same questions were then asked about a hypothetical human-human scenario that mirrored the observed researcher-agent interactions, as well as about scenarios that reflected

	No Response to Insult (N=30)				Response to Insult (N=30)			
	Agent		Human		Agent		Human	
Justification	Tri	Ins	Tri	Ins	Tri	Ins	Tri	Ins
Material-Physical	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**0**
Personal-Psych.	**23**	**20**	**37**	**7**	**13**	**0**	**0**	**10**
Responsiveness	0	7	0	0	0	0	0	0
Senses	0	3	0	0	0	0	0	0
Agency	10	0	13	0	3	0	0	3
Emotion	3	0	13	3	7	0	0	7
Personality	0	0	0	0	0	0	0	0
Other	10	10	10	3	3	0	0	0
Social-Conventional	**47**	**3**	**47**	**3**	**67**	**0**	**63**	**0**
Authority	0	3	3	3	0	0	0	0
Rules	30	0	23	0	50	0	47	0
Norms	17	0	20	0	13	0	17	0
Other	3	0	0	0	3	0	0	0
Moral	**3**	**57**	**7**	**77**	**13**	**97**	**23**	**90**
Physical Welfare	0	0	0	7	0	0	0	0
Psych. Welfare	3	33	0	63	0	57	3	43
Deontic Reasoning	0	7	7	3	13	27	17	17
Virtue	0	10	0	3	0	7	0	7
Other	0	13	0	3	0	27	3	27

Note: Tri = use of triangle in game; Ins = verbal insult during game. The percentages reported in **bold** refer to usage of the overarching category; and percentages in plain text refer to the next sub-level in the hierarchy. Within each level of the hierarchy, postings that contained more than one sub-category are only counted once in the overarching category. Percentages in each column may not add to 100 because (1) any given response by a child may have been assigned multiple justification codes and (2) the Uncodable category was excluded from this table.

Table 5. Percentage of Children Who Used Justification by Condition and Domain Distinction Scenario

Justification	Conventional Eval. (n=16)	Moral Eval. (n=14)
Material-Physical	**6**	**0**
Personal-Psych.	**30**	**7**
Responsiveness	13	0
Senses	6	0
Agency	0	0
Emotion	0	0
Personality	0	0
Other	13	7
Social-Conventional	**6**	**0**
Authority	6	0
Rules	0	0
Norms	0	0
Other	0	0
Moral	**30**	**86**
Physical Welfare	0	0
Psychological Welfare	13	57
Deontic Reasoning	0	14
Virtue	13	7
Other	6	20

Table 6. Percentage of Children Who Used Justification by Evaluation of Human-Agent Insult in Control Condition

	No Response to Insult (N=30)		Response to Insult (N=30)	
Entity Focus	Triangle	Insult	Triangle	Insult
Participant	0	3	0	0
Researcher	3	0	13	3
Other Humans	27	3	27	3
Agent	17	77	10	70
Other Tech.	0	0	0	0
Transgression	3	13	10	17
Game	30	0	33	0
Uncodable	4	0	2	0

Table 7. Percentage of Children Who Focused on Entity in Reasoning about Human-Agent Transgressions

prototypical domain distinction questions.

Domain Distinctions in Human-Agent Interactions

Results showed that children can and do distinguish between conventional and moral violations in interactions with a personified agent. However, children are more likely to identify a verbal insult against an agent as a moral violation when the agent responds with references to psychological harm and personal rights (47% in the control condition, 90% in the experimental condition). This conditional difference is not reflected in children's responses to the hypothetical human-human scenario that mirrored the researcher-agent interaction. Rather, results showed that children were just as likely to identify the verbal insult of another human as moral when the human did not respond to the insult (70%) as when she did respond (83%), and the difference in the control condition between children's evaluations of the verbal insults against the software agent and the hypothetical human were significantly different.

This difference in children's concepts about human-agent and human-human interactions does not necessarily reflect instability in children's moral judgments. Rather, it may reflect an instability in children's understanding of what or who can be harmed by a verbal insult. Once children recognize that harm has occurred either through prior knowledge, direct experience, or observed evidence, children judge the harm to be immoral, regardless of whether the entity experiencing the harm is a human or personified technology. This interpretation is supported by the fact that children, by and large, were unfamiliar with a personified technology capable of interacting with such social complexity. Thus, children had little prior experience that they could rely upon to infer possible harms. The experimental condition removed the need for children to infer the harm, with the agent explicitly providing evidence of the harm. For the control condition, however, children needed to infer the harm.

The interpretation that children judge harm to a personified agent to be a moral violation is also confirmed by the pattern of children's justifications showing a clear focus on the moral status of the agent (57% in the control condition, 97% in the experimental condition). One child stated, for example, that it was not all right to insult Judith "because it was like hurting her feelings. She [the researcher] shouldn't have said that in that mean of a way." Furthermore, we know that children's psychological welfare justifications were about the agent because the large majority of entity focus codes were for the personified agent (77% in the control condition, 70% in the experimental condition). Even for some of those children in the control condition who did not judge the verbal insult to be an immoral act, their justifications reflected the fact that they did not observe harm. For example, one participant stated, "Judith didn't seem to react to it much, so I think that in that situation it might have been OK."

It would seem, therefore, that there is a strong proclivity in children to consider personified technologies to have moral standing in the world, but their judgments about any given contextual situation may be contingent on a recognizable harm. This raises an interesting question: After extended exposure to personified technologies that show evidence of harm and claims to rights, will children construct a stable understanding of those acts which result in harms to personified technologies, and as a result, would a replication of this study after extended exposure result in most children judging the verbal insult to be immoral with no conditional differences? On the other hand, children's extended interactions with personified technologies may increase their understanding of the limitations of the technology resulting in recognition of a lack of harm across both conditions. This study cannot directly speak to these further questions, but it does provide a methodological exemplar and an empirical foundation that can be extended and investigated under differing conditions.

The results from this study alone, however, provide explicit evidence to support the following claims: (1) a significant number of children do show a propensity for generalizing their knowledge of domain distinctions in human-human interactions to human-agent interactions; (2) when the personified agent makes claims to rights and expresses harm in response to violations, children are significantly more likely to judge the act to be a moral violation and thereby distinguish conventional from moral violations in human-agent interactions; and (3) when a human makes claims to rights and expresses harm in a similar hypothetical scenario, children are no more likely to judge the act as a moral violation than when a human does not respond to the violation.

Computers as Social (and Moral) Actors

As mentioned earlier, this study extends work initiated by Cliff Nass investigating the phenomenon of people responding to computers as if the computers were social

actors. In contrast to prior findings, however, this study provides evidence that children knowingly respond to and reason about technology in social and moral terms. Additionally, the methodology used in this study demonstrates that moral implications of technology design can be investigated systematically by utilizing a controlled, experimental study design that provides data amenable to statistical analyses without disregarding the importance of subjective experience and reasoning.

Design Recommendations

In maintaining a level of accountability in design, designers of information systems and computer technologies have an obligation to bring to bear their "best efforts" in minimizing potential harms to stakeholders [18]. However, in order to make informed design decisions, designers need specific information about the interaction outcomes. By investigating the relationship between specific features designed into the technology and the value implications for specific direct and indirect stakeholders, we can increase the information available to designers such that they are able to make informed choices about their future software designs.

In designing personified technologies, decisions are made explicitly or implicitly to incorporate and mirror specific features of the human persona with which the system is interacting and/or exclude others. Judgments to include or exclude specific features can be the result of rigorous thought and empirical evaluations on behalf of the designers regarding the manner in which the system ought to behave to best accommodate the needs of the user. However, rarely do designers consider accounting for aspects of the human experience that fit within a moral domain, and it is even more rare that such judgments about moral impact are informed by empirical findings. For example, what would be the outcome of providing personified systems with the ability to gauge the user's moral concepts by observing user actions in morally sensitive contexts and by inquiring into the user's reasoning for such actions?

Take, for example, the video game industry, which has been designing simplistic versions of moral awareness into video games since the industry's inception. Games such as Ultima, Black & White, and Fable are designed with a central mechanism of character development in which deeds done by the user in the gaming environment, good or bad, right or wrong, result in (1) relevant and reciprocal responses from in-game characters, and (2) the development of the user's character in appearance, skills, and social status. The question is open as to whether the moral judgments made by the player in the game have implications for behavior and reasoning outside of the game context. However, the choices made in the design of personified technologies have implications for many contexts of human-computer interaction. For example, a software agent in an educational application or a robotic caretaker may also soon be able to respond to contextually specific information in morally and developmentally appropriate manners.

It is important to recognize that the results of this study alone do not provide evidence that a particular design approach is more appropriate than another. The results presented here show that children engage in more moral reasoning when the technology shows evidence of harm or makes claims to rights. It may be that there are certain contexts or goals of design that call for an increase or decrease in the activation of children's moral reasoning. That said, the author cautiously argues that in the general it is better that children more often interact with personified technologies that respond to possible harms and make claims to their own rights than it is for children to interact with personified technologies that do not. Children are more likely to reflect upon the possible harms and injustices of their own actions if they routinely interact with a social other which makes explicit claims to its own moral standing, regardless of the medium through which that social other presents itself to the world.

The implications of the alternative design, that of not having the technology respond to morally-relevant situations, are that children will come of age engaging in a significant number of social interactions that lack any moral feature, thereby increasing the likelihood that children will not construct a rich understanding of the intimate relationship that exists between social reciprocity and morality. Of course, additional conceptual, empirical, and technical investigations must be conducted in order to better understand the possible implications of these design decisions.

CONCLUSIONS

Children are constantly grappling with the difficult problem of determining what is and what is not appropriate action in response to the constraints of their environment. Actions, from a Piagetian perspective, are not only behavioral activities, but also include the active processes of thought and knowledge construction. Children judge, reason, and behave, that is they act, with a generally unspecified goal of attaining equilibrium amongst their internal concepts of the world, and between those internal concepts and the external world's projection by way of their senses. Fundamental changes in the constituents of children's environments may have significant implications for child development. The introduction of technology that acts social can and does influence children's social and moral thought. It is important, therefore, to conduct further research in this area to understand the larger implications and guide future design towards beneficial ends.

For the purposes of this paper, the focus was on an empirical study of children's judgments and reasoning regarding the distinction between social-conventional and moral actions while interacting with a personified, virtually-embodied, computational agent. Further research in this area is needed to understand the potential impact and

inform social robot and personified technology designers of the potential positive and negative influences that their design decisions may have on children's social and moral development.

ACKNOWLEDGEMENTS

This paper is based upon dissertation work completed by the author at the University of Washington's Information School [5]. The work was supported, in part, by the National Science Foundation under Grant No. IIS-0325035 to B. Friedman and P. H. Kahn, Jr. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. The author thanks Batya Friedman and Peter H. Kahn, Jr., for their guidance in developing and conducting this work, and Cady Stanton for her role in data collection. The author also thanks the reviewers for their constructive comments.

REFERENCES

1. Bakeman, R. & Gottman, J. M. (1997). *Observing Interaction: An Introduction to Sequential Analysis.* NY: Cambridge University Press.
2. Cassell, J. (2000). *Embodied Conversational Agents.* The MIT Press.
3. Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions, 3rd Ed.* Hoboken: John Wiley and Son
4. Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42*(3-4), 143-166.
5. Freier, N. G. (2007). *Children Distinguish Conventional from Moral Violations in Interactions with a Personified Agent.* Dissertation Manuscript. The Information School, University of Washington.
6. Friedman, B. (Ed.) (1997). *Human Values and the Design of Computer Technology.* Stanford, CA: CSLI.
7. Friedman, B., Kahn, P. H., Jr., & Borning, A. (2006). Value Sensitive Design and information systems. In P. Zhang & D. Galletta (eds.), *Human-Computer Interaction in Management Information Systems: Foundations*. (pp. 348-372). Armonk, NY: M.E. Sharpe.
8. Friedman, B., Kahn, P. H., Jr., & Hagman, J. (2003). Hardware companions? - What online AIBO discussion forums reveal about the human-robotic relationship. *Proceedings of the ACM 2003 Conference on Human Factors in Computing Systems (CHI '03)*. (pp. 273-280). New York: Association for Computing Machinery.
9. Gelman, S. A. & Opfer, J. E. (2002) Development of the animate-inanimate distinction. In Goswami, U. (Ed.) *Blackwell Handbook of Childhood Cognitive Development*. Malden, MA: Blackwell.
10. Isbister, K. (2006). *Better Game Characters by Design: A Psychological Approach*. Boston: Morgan Kaufmann.
11. Kahn, P. H., Jr. (1999). *The Human Relationship with Nature: Development and Culture*. MIT Press.
12. Kahn, P. H., Jr., Friedman, B., Perez-Granados, D., & Freier, N. G. (2006). Robotic pets in the lives of preschool children. *Interaction Studies, 7*(3), 405-436.
13. Killen M., Lee-Kim, J., McGlothlin, H., & Stangor, C. (2002). How children and adolescents evaluate gender and racial exclusion. *Monographs for the Society for Research in Child Development (Serial No. 271), 67*(4). Oxford, UK: Basil Blackwell.
14. Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.
15. Lester, J., Converse, S., et al. (1997). The persona effect: Affective impact of animated pedagogical agents. *CHI '97: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* (pp. 359-366). ACM.
16. Nass, C. & Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship.* Cambridge: MIT Press.
17. Nass, C., Moon, Y., Morkes, J., Kim, E. Y., & Fogg, B. J. (1997). Computers are social actors: A review of current research. In B. Friedman (Ed.). (1997). *Human Values and the Design of Computer Technology* (pp. 137-162). Stanford, CA: CSLI.
18. Nissenbaum, H. (1996) Accountability in a computerized society. *Science and Engineering Ethics* 2, 25-42.
19. Piaget, J. (1983). Piaget's theory. In W. Kessen (Ed.), P. H. Mussen (Series Ed.), *Handbook of Child Psychology: Vol. 1. History, Theory, and Methods* (4th ed., pp. 103-128). New York: Wiley.
20. Reeves, B. & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places.* The Center for the Study of Language and Information Publications.
21. Smetana, J. (2006). Social-cognitive domain theory: Consistencies and variations in children's moral judgments. In M. Killen & J. Smetana (Eds.), *Handbook of Moral Development*. (pp. 119-154). Mahwah, NJ: Lawrence Erlbaum Associates.
22. Turiel. E., & Davidson, P. (1986). Heterogeneity, inconsistency, and asynchrony in the development of cognitive structures. In I. Levin, (Ed.). *Stage & Structure: Reopening the Debate*. (pp. 106-143). Ablex, Norwood, N.J.
23. Wiener, N. (1950). *The Human Use of Human Beings*. Boston: Da Copa Press.