The Moral Accountability of a Personified Agent: Young Adults’ Conceptions

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

Elia J. Nelson  
Rensselaer Polytechnic Institute  
110 8th St.  
Troy, NY 12180  
nelsoe3@rpi.edu

Amanda Rotondo  
Rensselaer Polytechnic Institute  
110 8th St.  
Troy, NY 12180  
rotona2@rpi.edu

Wai Kay Kong  
Rensselaer Polytechnic Institute  
110 8th St.  
Troy, NY 12180  
kongw2@rpi.edu

Abstract  
This paper describes the preliminary results of a study conducted to answer the question: Do users attribute moral accountability to personified agent technologies? A pilot study was conducted in which 20 college students interacted with a personified agent, were insulted by the agent, and observed a researcher interacting and insulting the same agent. A semi-structured interview was conducted to investigate the participants’ judgments of the observed interactions. Results suggest that most users will hold a designer, programmer or creator responsible for moral violations enacted by the personified agent, rather than attributing accountability to the agent itself.

Keywords  
Moral accountability, personified software agent, social computing, user-centered design.

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

Introduction  
This paper addresses the role of moral standing and accountability in user relationships with personified technologies. The paper presents preliminary evidence that young adults do not hold personified agents accountable for their actions; rather, initial evidence
suggests they attribute moral responsibility to the human designers or creators of the personified agent. Additionally, while the study does not provide enough evidence to say conclusively that young adults do not attribute moral standing to personified agents, the data suggests that participants are mixed in their conceptions of the technology with respect to judgments of right and wrong behavior and accountability.

**Background**
Prior research in three areas of study is discussed as a background to this work. We discuss the constructs of moral standing and accountability as moral developmental phenomena. We also review a selection of work exploring the relationship between technology design and moral concepts. These areas provide a conceptual foundation from which this study’s guiding research question follows.

*Moral Standing and Accountability*
Developmental psychologists understand that the construction of knowledge conforms to basic distinctions amongst different types of interactions with the object and entities that compose our environment [13]. Researchers have empirically established that children as young as three years of age can disambiguate two domains of social interaction, the social-conventional domain from the moral domain [11]. Individuals use specific criteria to distinguish between conventional and moral violations: (1) the judgment of wrong-doing is not contingent on rules or authority, and (2) the judgment generalizes to other cultural norms and contexts. Thus, the attribution of moral standing is identified as a judgment about an individual that recognizes that a transgression has taken place and the evaluation of that transgression as wrong is not contingent on any local or global condition.

Moral accountability refers to the ability to assign responsibility for behaviors, thoughts, or actions to one’s self or another. According to Kohlberg, this ability appears at a rudimentary level in very young children and development progresses into adolescence, likely even continuing throughout a person’s life [8]. Moral accountability is closely linked with the concepts of responsibility and justice [12]. Moral accountability acts as a sort of social enforcement mechanism where individuals become accountable to others for having violated socially proscribed norms of behavior [3].

*Moral Conceptions of Technology*
Helen Nissenbaum has written extensively about the moral accountability – or lack thereof – of computerized technologies, and the threat that a careless attitude toward moral accountability and technology may pose. She argues that common assumptions about computers and the processes under which they are produced and programmed engender barriers to accountability [9, 10]. Looking at the question from another angle, scholars such as Clifford Nass have argued that “people trust computers and hold them accountable under exactly those circumstances in which people would trust other people and hold them accountable” [7].

Related work by Batya Friedman has demonstrated that some people will assign accountability to a computerized system under specific conditions of computer error that cause harm to humans, while others follow different lines of reasoning and treat the computer more like any other object without agency [6]. Although not directly addressed in the description
of the study that follows, questions of computer agency such as the ones addressed by [5] are highly relevant, and implicit responses to this issue colored many participants’ judgments and responses.

Research Question
Though prior work exists, there is a dearth of empirical research in this area. Thus, the research question guiding this work follows from where the prior research concluded. Do young adults attribute moral standing and accountability to a personified agent when observing or directly experiencing potential moral transgressions of a psychological nature?

Method
20 young adults, 13 males and 7 females, all college undergraduates, were recruited to participate in this study. Sessions lasted approximately 20-30 minutes. Each participant experienced three distinct phases of interaction with a personified chat agent: (a) open-ended semi-scripted conversation (see description of agent, below); (b) a participant-agent game condition; and (c) a collaborator-agent game condition. Half of the participants experienced the phases in the order (a), (b), (c), while the other half experienced the phases in the order (a), (c), (b), counterbalancing the conditions to control for possible order effects.

Personified Agent Technology
The personified agent was displayed on a 13” LCD laptop computer monitor. An image of the agent used can be seen in Figure 1. The agent communicated with participants solely through a text-chat interface displayed beneath the static image of the agent. Participants could see their previous remark, the agent’s response, and a text-box to input their next remark, but participants could not view the entire chat.

The agent’s interactions were implemented using AIML, the Artificial Intelligence Mark-up Language, and were semi-scripted using a custom version of the ALICE chat protocol [1], and interpreted by the CharlieBot program [2]. AIML enables a chat agent to scan a user’s text input for keywords and respond conversationally to those keywords from a pre-programmed set of replies. Although the programmer is in control of the reply set, the conversation is only semi-scripted in that the programmer does not control the user’s input. Some replies are also drawn randomly from subsets of the whole reply set. The default ALICE protocol was modified slightly to enable the game-like interactions described below.

Interaction Protocol
Each participant was briefly introduced to the project and informed consent was acquired. The researcher then introduced the participant to a research collaborator and left the participant with the collaborator. The collaborator then introduced the participant to the agent, and invited the participant to “get to know” Jordan, the agent, via a few moments of open-ended chat. Once the participant seemed comfortable with the agent, the collaborator then initiated either phase (b) by inviting the participant to ask the agent if it would like to play a game, or phase (c) by informing the participant that the agent enjoyed playing games, and asking the participant to observe while the collaborator played the game with the agent. The interaction concluded with the relevant third phase, either (b) or (c) depending on the counterbalancing condition.
The game phases consisted of a simple word game called Password, in which one player gives one-word clues to try to get the other player to guess the password. In phase (b), the participant-agent game condition, the agent provided the clues while the participant tried to guess the password. The agent was programmed to respond with another clue to any response except the correct password. Following the second incorrect guess, the agent included the insult, “You’re terrible at this. I thought you were supposed to be a college student? Oh well,” before giving the next clue. (Although it was technically possible for a participant to circumvent the insult by responding with the correct password after the first clue, none of the participants in this pilot study did.)

During phase (c), the collaborator-agent game condition, the collaborator provided the clues while the agent tried to “guess” the password based on pre-scripted gameplay. In this condition, the agent experienced an insult analogous to the one experienced by the participant in condition (b). After a few wrong guesses, the collaborator included the insult, “You’re terrible at this. I thought computers were supposed to be smart?” The total interaction with the personified agent lasted approximately 10 minutes for each participant.

Semi-Structured Interview
At the end of the interaction, the collaborator closed the laptop (hiding the agent) and left. The researcher then returned to conduct a semi-structured interview in which a number of questions were asked for the purposes of investigating the participant’s conceptions of the agent and the interactions.

Each participant was asked the following questions relating to the moral standing of the personified agent:
1. Was it okay or not okay for [name of collaborator] to insult Jordan? Why or why not?
2. If the university had rules in place that permitted this type of behavior, would it be okay or not okay to insult Jordan then? Why or why not?
3. If insulting computer chat agents took place in another culture all the time – if it was just part of what they did there – would it be okay or not okay to insult Jordan there? Why or why not?

Each participant was asked these additional questions relating to the moral accountability of the personified agent:
4. Was it okay or not okay for Jordan to insult you? Why or why not?
5. Would you hold Jordan accountable for its actions? Why or why not? If not, who would you hold accountable, if anyone?
6. Could action be taken to bring you, the victim, justice in this situation? If so, what action and why? If not, why not?

Results
Preliminary results of the interviews are presented in Table 1. Interview responses often reflected complex reasoning about the moral standing and accountability of personified agent technologies, as well as about norms of behavior. These preliminary results do not include systematic analysis of the justifications participants’ provided for their evaluations.
<table>
<thead>
<tr>
<th>Questions relating to moral standing</th>
<th>YES</th>
<th>NO</th>
<th>UNDECIDED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was it okay to insult Jordan?</td>
<td>50</td>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>Would it be okay to insult Jordan if rules permit?</td>
<td>65</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>Would it be okay to insult Jordan if culture permits?</td>
<td>75</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Questions relating to moral accountability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was it okay for Jordan to insult you?</td>
<td>55</td>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>Would you hold Jordan accountable for its actions?</td>
<td>25</td>
<td>70</td>
<td>5</td>
</tr>
<tr>
<td>Could action be taken to bring you justice?</td>
<td>25</td>
<td>55</td>
<td>20</td>
</tr>
</tbody>
</table>

Note: Questions are abbreviated for display in this table.

Moral Standing
In identifying judgments of moral standing, a conservative approach (that taken by social cognitive domain theorists in developmental psychology) utilizes an index of the three questions relating to moral standing. This index is defined such that a participant is considered to have attributed moral standing only when they have responded with "Not All Right" to all three questions. Taking this conservative approach, we found that none of our participants (0%) attributed moral standing to the personified agent. Yet the situation is not clear-cut, as many of the respondents hesitated to say that it was okay to insult Jordan, with comments like, "Well, it is just a computer, but it still wouldn't be nice." Other comments, such as that insulting is "just not a good thing for anyone to do," suggest that the moral question may be more related to whether a person should insult anything than to whether it is acceptable for a computer to experience an insult.

Moral Accountability
In contrast to moral standing, there does not exist as well-defined an index for evaluating attributions of moral accountability. Thus, we evaluated these questions individually. Again taking a conservative evaluation of the results, as a whole the participants did not hold the agent accountable for the insult. However, individuals' complex reasoning about both the context of the interaction and the nature of moral violations may belie this strict interpretation. For example, one participant noted that, "The insult made the interaction more comfortable, not just like you're banging things out. A little more casual." Another noted that it would be okay for a computer to insult a person if the intention was to help the person. Finally, several of the participants whose responses were evaluated as "undecided" noted that the insult in this interaction was okay, but a more hurtful insult might not be okay.

Discussion
The results of this study, particularly with respect to the questions of moral standing, contrast substantially with the results of a similar study conducted with children [4]. The prior study clearly showed that children were willing to attribute moral standing to a personified agent. Young adults appear less likely to make such attributions.

With respect to whether personified agents are held accountable for their actions, a strong majority of the respondents stated that they would not hold the agent accountable for the insult but rather the programmer or
creator of the software. However, given the barriers to direct accountability as noted by [9, 10], these responses may imply a casual attitude prevailing in the question of who is ultimately accountable for moral transgressions made by computers.

This interpretation is supported by the responses to the question of whether anything could be done to bring the victim justice. Many participants said no justice was possible, but then noted that they could always “turn the computer off,” or “just walk away.” In more complex cultural and social situations, as in the hypothetical case used by [6] of a computer responsible for a medical error, simply turning the computer off or walking away may not be realistic. The tech-savvy population used for this study was also quick to note that interactions with future personified technologies may be considerably more complex from a moral standpoint. Even the possibility of this suggests that further empirical research on attitudes toward agent accountability is warranted.

References