"It's the Computer's Fault" -- Reasoning About Computers as Moral Agents

Batya Friedman
Mathematics and Computer Science
Colby College
Waterville, ME 04901, USA
E-mail: b_friedm@colby.edu

ABSTRACT
Typically tool use poses few confusions about who we understand to be the moral agent for a given act. But when the "tool" becomes a computer, do people attribute moral agency and responsibility to the technology ("it's the computer's fault")? Twenty-nine male undergraduate computer science majors were interviewed. Results showed that most students (83%) attributed aspects of agency -- either decision-making and/or intentions -- to computers. In addition, some students (21%) consistently held computers morally responsible for error. Discussion includes implications for computer system design.

KEYWORDS: Computer agents, computer ethics, intelligent agents, social computing, social impact.

INTRODUCTION
Medical expert systems. Automated pilots. Loan approval software. Computer-guided missiles. Increasingly, computers participate in decisions that affect human lives. In cases of computer failure, there is a common response to "blame the computer." Is this a sincere instance of attributing moral agency to a computer, or a superficial verbal response that simply appropriates moral language? To investigate this question, this study examined computer literate individuals' reasoning about computers as moral agents.

METHODS
Twenty-nine male[1] undergraduate computer science majors from a leading research university in California (mean age = 23:1) participated in a one and a half hour interview about their views on computer agency and moral responsibility for computer error.

The interview contained questions in three general areas: (1) Students' views of computer agency (the capability to make decisions and the capability to have intentions). (2) Students' assessments of computer system characteristics and limitations. And (3) students' judgments of moral responsibility for two scenarios that involved delegation of decision-making to a complex computer system. One scenario involved a computer system that administers medical radiation treatment, and due to a computer error over-radiates a cancer patient. The second scenario involved a computer system that evaluates the employability of job seekers, and due to a computer error rejects a qualified worker. For each scenario, three conditions were investigated: a fully automated computer system that entails no human intervention; a token human intervention in which a person with little authority and status in the organizational hierarchy and little content area expertise operates the computer system (e.g., a hospital orderly in the radiation treatment scenario); and a non-token human intervention in which a person with authority and status in the organizational hierarchy and content area expertise oversees the use of the computer system (e.g., the attending physician in the radiation treatment scenario).

A coding manual was developed from half of the interviews and then applied to the remaining half of the data. To insure reliability of the coding scheme, an independent scorer trained in the coding manual recoded 28% of the data. Intercoder reliability for evaluations was 96%, for content responses 97%, and for justifications 74%.

Non-parametric statistics were used to analyze the categorical data. The McNemar statistic was used to determine a change in students' evaluations across measures (e.g., evaluation of blame across conditions). The amount of blame students' assigned to each potential agent was treated as score data. Then matched-pair t-tests were used to determine differences in students' assignments of blame across agents and conditions.

RESULTS
Due to limited space, only a few of the results will be presented here.

Computers as Agents
The capability to make decisions and the capability to have intentions were used to assess students' views on computers as agents. Seventy-nine percent of the students judged computers to have decision-making capabilities and 45% judged computers to have intentions. Eighty-three percent of the students attributed at least one of the two capabilities to computers; 41% attributed both capabilities. Furthermore, when students attributed only one aspect of agency to computers, they were more likely to attribute decision-making than intentions ($p<.006$).

Students' reasons for their assessments were also obtained. In justifying their positive or negative assessment of computer decision-making, virtually all students (95%) appealed to computers as deterministic systems that make use of rule-based or algorithmic processes, or lack free will.

[1]A considerable effort was made to interview equal numbers of females and males; however, a low enrollment of female computer science majors made this goal unfeasible.

For example, in support of computer decision-making one student said, "[the computer is] deciding based on a clear strict algorithm...it's a decision but not an open-ended one." In contrast but also drawing on the idea of computers as deterministic systems, to buttress a negative assessment another student said, "the decisions that the computer makes are decisions that somebody else has made before and programmed into the computer....it can analyze its input and take various actions depending on what the nature of the input is, but somebody has already told it how to proceed in the case of various inputs." Thus, students shared a view of computers as deterministic systems, but differed in their assessments as to whether or not deterministic activity constitutes genuine decision-making.

Students drew on a largely different set of reasons to support their assessments of computer intentions. Of the students who judged computers to lack intentions, 36% appealed to deterministic systems, 14% to emotions, 7% to consciousness, 7% to the soul, and 36% provided unelaborated responses. In many of these cases students referred to the absence of qualities in computer systems such as a lack of consciousness (e.g., "The program is not actually knowing... it's like a level of consciousness...it's just a computer that executes these lines of code...so there's no intention on the part of the program."). In contrast, students who judged computers to have intentions encountered difficulty explicating their reasons. Although probed to the same degree as students who did not attribute intentions to computers, all of these students (100%) provided vague, unelaborated justifications that often did little more than reassert their assessment.

While the above findings overall provide a positive portrayal of computers as agents, students also judged computers to be different than humans along similar dimensions. Of the students who judged computers to have decision-making capabilities, 100% judged computer decision-making to be different from human decision-making. Similarly, of the students who judged computers to have intentions, 77% judged computer intentions to be different from human intentions ($Z=2.30$, $p<.05$).

Responsibility for Computer-Error

Overall, students perceived the two scenarios -- on radiation treatment and employment rating -- as similar. No significant differences were found between the scenarios for corresponding agents and conditions in students' evaluations of who or what to blame.

Roughly one-fifth of the students (on average 21%) consistently blamed the computer system itself for the computer-based error. No significant differences were found across the three conditions and two scenarios for students' evaluations of blame and the amount of blame. However, the amount of blame finding should be understood with caution as only those students ($n<=6$) who blamed the computer were assessed for the amount of blame.

A central concern of this study is how students understand computers to be accountable, if at all, for computer error. Thus, it is useful to examine students' reasons for blaming or not blaming computers in relation to their reasons for blaming or not blaming people (the computer system designer, the computer system's human operator, and the organization's administrators). Averaging across conditions and scenarios, virtually all of students' justifications for blaming the computer (96%) referred to the computer's participation in the sequence of events that led to harm. In contrast, the large majority of students' justifications for blaming people (80%) referred to failing to meet some commonly expected reasonable level of performance (e.g., negligence). When students did not assign blame, differences were also found among the justifications students used for computers and for people. Again, averaging across conditions and scenarios, virtually all of students' justifications for not blaming the computer (97%) referred to qualities of computers that diminish its agency and thus undermine computers as being the sort of thing that can be blamed. Notably, the appeal to diminished agency was used exclusively in reference to computers. In contrast, students' justifications for not blaming people primarily referred to adequately meeting commonly expected levels of performance (55%) and deferring to an authority perhaps due to habit, lack of autonomy, or the authority's greater power or knowledge (41%).

DISCUSSION

The data reported above joins a growing body of research [1, 2,] that suggests people, even computer literate individuals, may at times attribute social attributes to and at times engage in social interaction with computer technology. Some researchers argue that as good designers we ought to exploit this psychological phenomena to build systems that actively engage users in a social relationship with the technology. Much of the work on computer agents and intelligent agents is of this vein. The results reported here, however, should give us pause. For the results suggest that even some computer literate individuals hold computer technology at least partly responsible for computer error. If this finding is correct, a different design strategy is in order. It would follow, for example, that designers should communicate through the system that a (human) who -- and not a (computer) what -- is responsible for the consequences of the computer use.

ACKNOWLEDGMENTS

I thank Sara Brose, Lynette Millett and Sue Nackoney for help with the coding, reliability, and analysis of the data. This research was funded in part by the Clare Boothe Luce Foundation and by a Natural Sciences Division Research Grant from Colby College.

REFERENCES

1. Nass, C., Steuer, J. and Tauber, E. R. Computers Are Social Actors, in Proc. CHI '94 Human Factors in Computing Systems (Boston, April 24-28, 1994), ACM Press, 72-78.

2. Walker, J. H., Sproull, L. and Subramani, R. Using a Human Face in an Interface, in Proc. CHI '94 Human Factors in Computing Systems (Boston, April 24-28, 1994), ACM Press, 85-91.