A Dialogue on Responsibility, Moral Agency, and IT Systems

Deborah G. Johnson
Anne Shirley Carter Olson Professor of Applied Ethics
University of Virginia
Charlottesville, Virginia, USA
1-434-924-7751
dgj7p@virginia.edu

Keith W. Miller
Prof. of Computer Science
University of Illinois at Springfield
Springfield, Illinois, USA
1-217-206-7327
miller.keith@uis.edu

ABSTRACT

The dialogue that follows was written to express some of our ideas and remaining questions about IT systems, moral agency, and responsibility. We seem to have made some progress on some these issues, but we haven't come to anything close to agreement on several important points. While the issues are becoming more clearly drawn, what we have discovered so far is closer to a web of connecting ideas, than to formal claims or final conclusions.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues – *Ethics.*

General Terms

Human Factors, Legal Aspects.

Keywords

Computer ethics, IT systems, artificial agents, moral agents.

1. A Dialogue

ELIZA: Alice, you and I have talked before about responsibility and IT systems. Sorting it all out is really hard and lately I've been thinking about writing a paper in which I propose that we think of IT systems as agents or employees of corporations. An employer hires an employee or buys a computer system to do certain tasks. Both the employee and the system act for the employer; they behave in ways that accomplish goals that the employer wants achieved. This seems to me a good way to conceptualize IT systems, especially the parallel with human employees, because it forces us to see that we have turned over lots of tasks to computers but we haven't assigned the computers responsibility as we do with humans. That is, we don't attribute any moral responsibility to computers even thought they often do tasks that have moral consequences. What do you think? Good idea for a paper?

ALICE: Hey, that's a great idea. Traditionally, IT systems (including hardware, software, and telecommunications) have been viewed as a particular kind of "equipment." As such, IT systems are thought of as commodities used to increase the productivity of human employees. Looked at from this perspective, an IT system is similar to a typewriter or even a pencil. But 'equipment' doesn't usually have a say in what an organization is or becomes, or how it treats its customers or other constituents. These days IT systems have become so sophisticated that they are starting to have that kind of power in organizations. The analogy with being an employee may be a good way to show the power and sophistication of computer systems and why we have to think about them in moral terms.

ELIZA: Whoa! Wait a minute, Alice. You're going too fast. You're right about IT systems being thought of in the past as equipment but when you started talking about IT systems being different from typewriters and pencils, you lost me. I think we have to draw attention to the moral character of all technology and I'm not sure that the moral character of IT systems has anything to do with what's special about them. Even mundane technology can have a moral character. Think of a gun or a landmine or more positively, think of the polio vaccine. They don't have to be computerized or even intelligent to become moral in character.

Anyway, I was thinking that viewing IT systems as corporate employees would help us to understand how IT systems function in organizations and how their behavior might be thought of as moral - in the way we think that employees or agents can behave well or poorly. They can be good or bad employees and their personal morals can come in conflict with their professional morals. I wasn't thinking it had anything to do with the special features of computers such as their malleability or their capacity for symbol manipulation.

ALICE: Eliza, if what you want to do is draw attention to how IT systems function in organizations, then you have to admit that IT

SAC'06, April, 23-27, 2006, Dijon, France.

systems are unlike most other technologies. IT systems have characteristics that set them apart from other equipment such as pencil sharpeners, lawn mowers, and conveyer belts. IT systems are symbolic and communicative. They are used to process and distribute information, not physical objects. Other equipment doesn't exhibit those properties.

IT systems are different from most equipment in that IT systems require "instructions" before they are useful. A wrench doesn't require instructions - its form indicates its function. A computer gains its usefulness when a program defines its input/output behavior. This gives IT systems their impressive versatility [5], a versatility that sets IT systems apart from other equipment. An exception that proves this rule is equipment that includes embedded computing power. A weaving loom that can be programmed with new patterns has functionality that may replace a human operator. (Indeed, an early ancestor of digital computing was Jacquard's loom [6] that used wooden "punched cards" to control the pattern of the cloth being produced.) Equipment that includes computing is, in a way I consider fundamental, distinct from equipment that does not include computing.

Some characteristics that set IT systems apart from other equipment include information processing capabilities, versatility, communication with other IT systems, and communication with humans. Those characteristics all suggest an analogy with human employees. Like IT systems, but unlike other equipment, humans process information, are versatile, communicate with IT systems, and communicate with humans. I agree with you that the analogy between IT systems and employees might help show how closely connected ethics and IT systems are, but I also like that analogy because it points in the direction of thinking of computer systems as agents, moral agents. This tightly couples ethics and IT systems, in a powerful way. However, I think this has to do with special characteristics of computer systems, not just any technology.

ELIZA: Again, too fast for me. Let's talk about tasks and functions. I've been reading Science and Technology Studies (STS) literature lately. I am now convinced that we have to stop thinking about technology - all technologies - as if they were simply material objects or artifacts. The material objects don't have any meaning or use without people and context. STS scholars suggest that technology is really combinations of people and things. Some say a technology is best understood as a network and they treat nature and artifacts as well as people as actors in the network. They use the term 'actants' for all the actors in a network. This helps to show that artifacts influence what people do and what people do influences the kind of technology we have. It helps to show that you can't have one without the other and that action or activity is a combination. People can't do what they do without artifacts and artifacts wouldn't be and wouldn't have functionality without the context and social systems of which they are part.

Actor-network theory aside, I am convinced that we are better off focusing on social practices that consist of people and things. Cooking, for example, involves people doing things with raw materials and artifacts. Computer systems are parts of (components in) a wide range of social practices. Writing is a social practice involving people with ideas and expressing those ideas in words that can be externalized with pens or pencils or computers. Once externalized, the words can be understood by other people. Words on paper or a computer screen are nothing unless there are people who understand them in one way or another. Social practices are combinations of things and people and we should never forget that the things don't make sense without the people. This includes computers.

The social practices are aimed at accomplishing certain goals and typically tasks have been divided up among humans and machines. When I use my cell phone I have to know the name or number of the person I want to call and I have to press keys, then the phone takes over and makes the connection for me, but then I have to use my voice or figures to communicate. It's social behavior and things - IT systems work together with people.

Well, I've gotten somewhat side-tracked here, but what I was trying to say was that I thought using the analogy with employees would help to show how IT systems don't make sense in isolation. They are parts in human social activities.

ALICE: Aren't you forgetting that many computers never interact with people? Some computers simply interact with or cause something to happen in another machine or another computer. So some computers aren't social, at least not directly social in the way that we expect people to be social.

ELIZA: Of course, you are right but that's because you have focused in on a part of a system. The system as a whole, that is, multiple machines communicating with one another, has been put in place for some purpose or with some human end in mind; the system has been created by humans for human ends. Ultimately, all human-made things are connected to something that humans are doing or trying to do.

ALICE: I guess you're right - the exception proves the rule. I will grant you that computers are human-made, and that makes computers in a real sense "social." But we are getting sidetracked here. I thought the point was to show the moral character of IT systems. I think the best way to show this is not just to think of IT systems as agents of their employees but to think of them as moral agents.

The case for designating IT systems as moral agents is seen most clearly when an IT system displays characteristics that are often associated with moral agency. The philosophers Floridi and Sanders [3] argue that three characteristics are central to an abstract view of moral agency: adaptability, autonomy, and interactivity. Interestingly, this list has striking similarities to a list of characteristics that computer scientists [1] have proposed to distinguish intelligent agent software from other software. Their list contains the terms autonomous, persistent, reactive, proactive, trustworthy, personalized, and possessing social behavior. I don't want to argue right now whether software agents that seem to display these characteristics are merely mimicking human characteristics or actually possess them. For my purposes, I'll merely point out that, increasingly, IT systems are exhibiting input/output "behavior" that humans describe with adjectives that are traditionally associated with adaptable, autonomous, and interactive humans. These adjectives suggest that people are coming to rely on IT systems in a way that at least some would describe as "trust." Indeed, even early experiments in artificial intelligence (AI) were met by some people with a surprising degree of intimacy.

Most early AI programs were programmed using imperative programming languages. More modern AI software, sometimes driven by neural nets, can include aspects of novelty and adaptation that were not possible with more static programming in, for example, expert systems of just a decade past. I think that neural nets are an important step towards computers that act as moral agents.[4] Neural nets are designed as initial configurations that are "trained" to recognize particular patterns using a technique called "back propagation." After this initial training, a neural net can identify these patterns in new inputs. In some cases, the neural net can adjust itself (in essence, "learn") from its continuing experience without human intervention. When that autonomous learning takes place, the neural net can change its state in such a way that its initial configuration is altered significantly. Furthermore, nothing in the design or implementation of a neural requires that these changes can be traced back to the initial configuration and the new inputs.

But can we trust a program that "learns" from its "experience?" When such a program causes harm (and surely this will happen with some IT system-agents), who is responsible for that harm? Is it the original programmer, the person who trained the neural net, the organization that deployed the system, the system itself, or no one?

ELIZA: I understand what you are saying and I think the question about responsibility can be answered, but the point I want to make now is that there is a difference between saying that IT systems are moral entities and saying they are moral agents. I think IT systems are moral entities but not moral agents. I say this for a couple reasons. One reason has to do with what 'moral agent' has traditionally meant in moral philosophy. The notion of moral agent is embedded in moral philosophy and is ultimately connected to the idea that to be a moral agent, a being must have freedom. Remember morality isn't possible if everything is determined. An individual can't do good or evil if everything they do has been caused by something else.

I admit the notion of human freedom is somewhat mysterious, but the idea that individuals choose to do what they do is at the core of moral notions. Without this idea, morality doesn't make sense because no one can do otherwise. Morality doesn't make sense in a completely deterministic world. The model of human action and agency presumed by moral philosophy is that of this freedom operating through the mind of the agent and leading to behavior that has consequences. The behavior can be evaluated either in terms of the mental states that caused the behavior or in terms of the consequences. Either way, the origins of the behavior are non-deterministic.

So, whatever parallels you make between IT systems and human moral agents, you are not going to convince me that IT systems can be moral agents. In fact, I will agree to most of the characteristics that scholars like Floridi and Saunders attribute to IT systems, but none of these seem to justify saying that IT systems are moral agents.

I am willing to say that IT systems behave and they behave in ways that affect moral patients, so their behavior has moral consequences. I am even willing to go as far as saying that IT system behavior has a moral character, but because the character of computer behavior (what constitutes it) is so radically different from the character of human behavior (what constitutes it), there is no way of saying that a concept like moral agency that applies to one would apply to another.

ALICE: I can agree with most of what you say, and still disagree about your conclusion. I understand that freedom is central to the idea of moral choice, and that moral choice is at the heart of moral agency. But I think the core of our disagreement can be illustrated by your use of the word "deterministic." Way back in 1975, Dijkstra wrote about "guarded commands" as a way to introduce non-determinism into programming. [2] The guarded commands enables a programmer to write a program whose behaviors cannot be determined in advance. It is true that most such programs in 1975 had a fairly limited number of possibilities "programmed in." But there is nothing that inherently prohibits programs (especially programs that can "learn" in the sense I described before) from adding new possibilities.

Eliza, I'm not trying to convince you that computers and people are identical. We can agree that a computer and a human will always be distinct; at the very least, computers are silicon based and humans are carbon based! What I think we need to be working towards is a better understanding both of how they are different and how they are the same. And I don't think you can claim that non-determinism is exclusively a human trait.

I could try to argue a Skinner-like position [7] that humans are themselves deterministic, and not as "mysterious" as you supposed. But I'm going to take an opposite tack: computers, especially neural nets, have their own mysteries. In this odd new world of machines that can "think," we have to get beyond insisting that computers are not human; that's true, but it's irrelevant. We have to concentrate on what it is about moral agency that makes it exclusively human. So far, I'm not convinced you've given me a reason to think that at least some computers will be able to exhibit (yes, I'm dodging the question of "possess" for now) any characteristic you can name. So if a computer talks like a moral agent, makes decisions like a moral agent, is somewhat unpredictable like a moral agent, and interacts with humans like a moral agent, why exactly can't that computer be labeled a moral agent? Notice that I don't claim a computer is a human moral agent… but a moral agent that isn't human.

Unless you simply define a moral agent as a human that acts in a certain way, then I think you need to show me something ethically significant and outwardly observable that differentiates a human moral agent from a computer moral agent. If you can't do that, then you are merely saying that computers aren't human; we already agree on that, and we need to move beyond it.

ELIZA: There are several really important issues here, Alice. For one, you are equating "observable behaviors" with reality. It really does matter what goes on beneath the surface. Floridi, Sanders, and you are trying to make the lower levels of abstraction irrelevant to our discussion, and they are not irrelevant. Just because a simple text processing AI program "appears" to be a sympathetic listener to some people doesn't make it so; it is an illusion. The reality was that sequences of 1s and 0s were being executed and electronic switches flipped; this is a far cry from sympathy. I don't think non-human, non-living beings (neural nets or not) can be sympathetic, though, of course, they can behave in ways that mimic human sympathy.

Another thing is the issue of deterministic versus non-deterministic behavior. Moral philosophers used to make a distinction between undetermined and indetermined with the latter equating to randomness. Its one thing for behavior to be so complicatedly determined that it is impossible to sort out and identify what is causing the behavior. The complexity also makes it impossible to predict behavior. But, it is quite another thing for the behavior to be random. I'm not sure which would apply to neural net computers.

What I do know is that moral philosophy views human behavior as special in the sense that it is neither random nor determined in the way other entities are. In a sense human behavior is determined, but in a special way. That is, human beings are thought to act for reasons (combinations of motives, desires, and beliefs); the person's reasons, intentions, and beliefs lead to their behavior, but the reasons are not determined. Human thought - mental states - are thought to be uncaused causes. I don't think this would apply to computers.

I suppose yet a third issue is how we use that term 'moral agent.' We could use it in a way that included things other than humans but doing so would radically change its meaning which has a long and important history. The term is a kingpin in moral thought and the new use would undermine an entire network of ideas. I prefer the path of calling artifacts moral entities but not moral agents. Perhaps it would be even better if we didn't focus on either artifacts alone or humans alone but only focused on combinations - humans act with technology and technology requires humans for its meaning and being.

ALICE: I think I care less about the semantic difference between "moral entity" and "moral agent" than you do. I'm willing to call computers moral entities and save the term moral agent for humans. (We'd hate to upset the humans by taking over their term "moral agent!") But that still leaves us the important task of talking about what these moral entities and the human moral agents share.

I think your criticism about "beneath the surface" is much more important than the name game. I know you are right that the lower levels of abstraction matter, and I also agree that human brains are more complex than any computer system. I agree that those differences matter. But I think we have to start carefully thinking about how those differences matter, and how, in some cases, they don't matter, at least with respect to moral agency. Are you willing to wrestle with that problem, or are we stuck on the "carbon based only" problem for now? If you are not willing to go beyond the exclusivity of humans, then I don't see how we are going to get responsibility in equation without attributing moral agency to IT systems.

ELIZA: You're right, Alice. We should think carefully about the differences that matter and don't matter and I think we can make some progress here. Computers, computer systems, and other artifacts are like human beings in that their behavior has consequences for moral patients. That is, their behavior can harm and help moral patients. Moreover, I think the behavior of artifacts and the behavior of humans is alike in coming from intentionality. Humans act with intentions. Artifacts have been intentionally designed and intentionally put in place or are intentionally taken up by users. In this respect human behavior and artifact behavior are both different from the behavior of natural objects. Nevertheless, the complex of intentionality that underlies humans and that which underlies artifacts are radically different and in fact the latter is a creation of the former.

ALICE: I agree that humans today are far more complex than any existing computer. However, computers are becoming ever more complex, and I don't see any physical or algorithmic principles that prohibit computers from becoming as complex as humans eventually. There will be practical challenges, to be sure, and it may take decades or centuries. But if complexity is the only underlying cause of the "radical" difference between humans and computers, that difference is likely to pass.

Although I don't envision computers creating humans, I can envision computers developing other computers. Humans already have to "cooperate" with computers to design computer chips. Humans also need computers to analyze and manipulate genes. It's not such a big step to have a computer design a new computer without human intervention.

That being said, I still think we're avoiding an important part of this issue. No matter what ethical term you use to describe IT systems, and no matter what future tasks they may be able to achieve, we both agree that IT systems have ethical significance right now. IT systems, together with or autonomously, have consequences for moral patients. Traditionally, we hold moral agents responsible for their actions. Shouldn't we hold IT systems (which you want to call "moral entities") responsible for their morally significant actions?

ELIZA: Before we get to responsibility, a quick aside. I don't think it is fruitful to compare computers and humans on the basis of their degree of complexity. I have no idea how one would credibly measure degree of complexity.

In any case, while we disagree about the matter of moral agency and the significance of the difference between what underlies behavior, we do agree that much more attention has to be given to responsibility and computers and computer systems. We even agree that holding the designers of the computer system responsible isn't enough. But I think the focus on combinations of humans and artifacts does the best job of explaining this. For when something bad happens, it is often the combination of humans and things that has brought it about. In other words, I am inclined to say an IT system will never be responsible alone, but at the same time I am inclined to say that when individuals act with and through IT systems, the individual is also not responsible alone. It's the combo that has to be the focus of moral attention.

ALICE: Now I think we're getting somewhere. We can conveniently side-step the big gap that you see between human moral agents and non-human moral entities by focusing on the socio-technical systems that combine humans and computers. At this level of abstraction (notice how that keeps popping up?), we can agree that system, the combination of humans and machines, is a locus of responsibility. I think that's where we can come together: we can both demand that the humans in charge of those systems take ultimate responsibility for them. In a way, we can make this a two step process: Step 1. We assign a "real world" responsibility to the socio-technical system that is made up of humans and computers. Step 2. We will keep discussing how that responsibility can be "distributed" among the humans and non-humans in the system, but meanwhile the responsibility has been located as best we can - to the system.

ELIZA: Our disagreements don't disappear, but your move does clarify where we part company. We agree that systems are morally important, and we agree that components in the system share responsibility for its actions. So, responsibility doesn't disappear because there are machines inside the system. The responsibility does not get transferred out of the system. Still, we disagree on how that responsibility is located within the system, or how we think about the responsibility of the computer parts of the system. I don't think computers are now, nor will they ever, be moral agents.

ALICE: That is where we differ, but this system view also solves a problem for me. I see that more advanced computer systems exhibit many of the characteristics I see in human moral agents. But most computer systems are not terribly advanced, and very few of them exhibit non-determinism and "learning." By focusing on the human/machine combination I don't need to draw fine distinctions between advanced computing and more pedestrian computing; the human/computer mix may be different with different kinds of computers, but the mix is still the location of the responsibility.

ELIZA: I'm nervous that you are agreeing with me. We can't ignore this fundamental disagreement about whether or not IT systems are themselves moral agents.

ALICE: I don't think I'm trying to avoid that problem. I'm just trying to put it in perspective. We do need to keep working on the question of IT systems as moral agents. That question will become increasingly important as IT systems get more sophisticated. That sophistication will make the moral distance between an IT system and its original designers and programmers larger. And as that moral distance widens, we will become more and more interested in the moral complexities of the IT system itself.

ELIZA: I agree that this is important and unsettled. I'm sure we disagree on what direction to take to solve it, but I'll take it as progress that you are willing to consider the importance of the socio-technical systems of which computers are a part. We'll keep working on the vitally important details that underlie that system. But we'll probably see those details differently for the foreseeable future.

ALICE: I think you are right both about what we agree on, and what still is unsettled between us. But how does this discussion affect your idea of IT systems as employees?

ELIZA: I'm not sure the analogy I started off with still makes sense. Since we've decided to focus on the socio-technical systems (which include IT systems and humans), it doesn't work very well to think of an IT system in isolation as an employee. Perhaps the whole socio-technical system should be the analog of an employee, but that gets a little confusing because that system already includes human employees. I guess I have to rethink that analogy.

ALICE: I still like your idea of IT systems as employed agents, but you're right: we've made that analogy harder by coupling the IT systems and the humans. That coupling is messy, but I think it's essential to admit how important that coupling is to understanding where the responsibility for IT systems lies. We have to think of a better way to deal with that messiness. Might you be willing to think about IT systems and humans as a family unit? Could socio-technical systems be an analog to a trade union?

ELIZA: You're making my synapses hurt, Alice. We'll try a few new analogies next time. Over and out.

2. REFERENCES

[1] Desharnais, P, Lu, J, and Radkakrishnan (2002), T. Exploring agent support at the user interface in e-commerce applications, International Journal on Digital Libraries, Vol. 3, No. 4, 284-290.

[2] Dijkstra, E. Guarded commands, nondeterminacy and formal derivation of programs (1975), Communications of the ACM, Vol.18, No.8, 453-457.

[3] Floridi, L, and Sanders, J. (2004) On the morality of artificial agents. Minds and Machines, 14.3, 349-379 .

[4] Matthias, A. (2004), The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, Vol. 6, No. 3, 175-183.

[5] Moor, J (1985), What is computer ethics?, Metaphilosophy Vol 16 No 4, pp 266-279.

[6] Russo, M. Herman Hollerith: the world's first statistical engineer. http://www.history.rochester.edu/steam/hollerith/loom.htm, accessed September 8, 2005.

[7] Smith, L (1994) B. F. Skinner. PROSPECTS, Vol. XXIV, No. 3/4, 519-32, http://www.ibe.unesco.org/International/Publications/Thinkers/ThinkersPdf/skinnere.PDF, accessed September 8, 2005.