Computers as Surrogate Agents

(forthcoming in *Moral Philosophy and Information Technology*, eds. J.van den Hoven and J. Weckert, Cambridge University Press)

Deborah G. Johnson
Thomas M. Powers*
University of Virginia

Computer ethicists have long been intrigued by the possibility that computers, computer programs, and robots might develop to a point at which they could be considered moral agents. In such a future, computers might be considered responsible for good and evil deeds, and people might even have moral qualms about disabling them. Generally, those who entertain this scenario seem to presume that the moral agency of computers can only be established by showing that computers have moral personhood and this, in turn, can only be the case if computers have attributes comparable to human intelligence, rationality, or consciousness. In this paper we want to redirect the discussion over agency by offering an alternative model for thinking about the moral agency of computers. We argue that human surrogate agency is a good model for understanding the moral agency of computers. Human surrogate agency is a form of agency in which individuals act as agents of others. Such agents take on a special kind of role morality when they are employed as surrogates. We will examine the structural parallels between human surrogate agents and computer systems to reveal the moral agency of computers.

Our comparison of human surrogate agents and computers is part of a larger project, a major thrust of which is to show that technological artifacts have a *kind of intentionality*, regardless of whether they are intelligent or conscious. By this we mean that technological artifacts are directed at the world of human capabilities and behaviors. It is in virtue of their intentionality that artifacts are poised to interact with, and change, a world inhabited by humans. Without being directed at or being about the world, how else could technological artifacts affect the world according to their designs? Insofar as these artifacts display this kind of intentionality and affect human interests and behaviors, the artifacts exhibit a *kind of*

* Authors are listed in alphabetical order.

moral agency. If our account of technology and agency is right, the search for the mysterious point at which computers become intelligent or conscious is unnecessary.

We will not rely, however, on our deeper account of the intentionality of technological artifacts here. In this paper our more narrow agenda is to show that computer systems have a kind of moral agency, and that this agency has the structural features found in human surrogate agency. Both human and computer surrogate agents affect human interests in performing their respective roles; the way they affect interests should be constrained by the special morality proper to the kind of surrogate agents they are. To reveal the moral agency of computer systems, we begin by discussing the "role morality" of human surrogate agency, and the nature of agency relationships (part 1). We then turn our attention to specifying more carefully the object of our attention: computers, computer programs, and robots (part 2). The next part of our account draws out the parallels between human surrogate agents and computer systems and maps the moral framework of human surrogate agency onto the agency of computer systems (part 3). This framework allows us to identify the kind of interests that both human and computer surrogate agents can have, and also leads to an account of the two ways in which surrogate agents can go wrong. Finally, we review the account we have given and assess its implications (part 4).

1. Human Surrogate Agency

In standard accounts of agency, moral agents are understood to be acting from a first-person point of view. A moral agent pursues personal desires and interests based on his or her beliefs about the world, and morality is a constraint on how those interests can be pursued, especially in light of the interests of others. In this context, human surrogate agency is an extension of standard moral agency. The surrogate agent acts from a point of view that can be characterized as a 'third-person perspective'. In acting, the surrogate agent considers not what he or she wants, but what the client wants. While still being constrained by standard morality in the guise of such notions as duty, right, and responsibility, human surrogate agents pursue a subset of the interests of a client. But now, they are also constrained by role morality, a system of conventions and expectations associated with a role.[1] Examples of

[1] See Alan Goldman's *The Moral Foundation of Professional Ethics* (Totowa, N.J.: Rowman and Littlefield, 1980) for a theory of role morality and the justification of special moral rights and responsibilities attached to professional roles.

surrogate agents are lawyers, tax accountants, estate executors, and managers of performers and entertainers. Typically the role morality entails responsibilities and imposes duties on the agents as they pursue the desires and interests of the client. For example, lawyers are not supposed to represent clients whose interests are in conflict with those of another client; tax accountants are not supposed to sign for their clients; and estate executors are not supposed to distribute the funds from an estate to whomever they wish.

To say that human surrogate agents pursue the interests of third parties is not to say that they have no first-person interest in their actions as agents. Rather, the surrogate agent has a personal interest in fulfilling the role well, and doing so involves acting on behalf of the client. Failure to fulfill the responsibilities of the role or to stay within its constraints can harm the interests of the surrogate agent insofar as it leads to a poor reputation or being sued for professional negligence. Conversely, success in fulfilling the responsibilities of the role can enhance the reputation and market worth of the surrogate agent and may, thereby, fulfill some of his or her own goals.

Though surrogate agents pursue the interests of their clients, they do much more than simply take directions from their clients. To some extent, stockbrokers are expected to implement the decisions of their clients, but they are also expected to provide advice and market information relevant to making informed decisions.[2] In addition, stockbrokers form investment strategies based on a client profile of risk aversion, liquidity, goals, etc., and not based on generic investment strategies. The surrogate role of tax accountants is not merely to calculate and file a client's taxes, but also to find the most advantageous way for the client to fulfill the requirements of the tax code.[3] Estate executors provide a unique case of surrogate agents because they must pursue the expressed wishes of their clients after the clients are deceased. The client's will is comparable to a closed-source program; it is a set of instructions to be implemented by the executor, and not to be second-guessed or improved upon.

Generalizing across roles and types of human surrogate agents, we find at least two

[2] The practice of electronic trading has changed or eliminated the moral relations between investors and stockbrokers to a large extent.
[3] Tax accountants are important in our analysis because they can be compared to software programs that individuals use in preparing their annual income taxes.

different ways that surrogate agents can do wrong. First, they can act incompetently and in so doing fail to further the interests of their clients. Imagine a stockbroker who forgets to execute a client's request to buy 500 shares of IBM stock at a target price; the stock soars and the client looses the opportunity to obtain the shares at an attractive price. Or imagine a tax accountant who, in preparing a client's annual income tax, fails to take advantage of a tax credit for which the client is fully qualified. Finally, imagine the estate executor who neglects to include the appropriate parties in the meeting to read the will and, as a result, the will is thrown into probate court. These are all cases in which the surrogate agent fails to implement an action or achieve an outcome because of incompetence on the part of the agent. Such failures are generally unintentional, but nonetheless the surrogate agent fails to do what is in the interest of the client.

Second, surrogate agents can do wrong by intentionally violating one of the constraints of the role. We will refer to this form of doing wrong as misbehavior. Imagine the stockbroker encouraging a client to buy shares in a company and lying about the last dividend paid by the company or the company's price-to-earnings ratio. Worse still, consider the real case in which a major investment firm advocated the purchase of stock in a troubled company in order to gain the investment banking business of the company—at the expense of the interests of their investors.[4] Imagine the tax accountant violating confidentiality by giving out information about a client to a philanthropic organization; or imagine the estate executor giving money to the client's children despite the fact that the client specified that they were to receive none. These are all cases in which the agent does wrong by violating the duties or constraints of the role. In most cases of misbehavior, the surrogate agent pursues someone else's interests, e.g., the agent's or other third parties, to the detriment of the interests of the client. In this way the surrogate agent intentionally fails to fulfill one crucial expectation associated with the role: to take the (third-person) perspective of the client.

2. Computers, computer programs and robots

[4] A class action lawsuit settled in May of 2002 involved Merrill Lynch & Co. and the state of New York. The settlement required the investment firm to pay $100 million for misleading investors by giving them "biased research on the stocks of the company's investment banking clients." (See "Merrill Lynch, NY reach $100M Settlement," Frank Schnaue, UPI: 05/21/02) The Merrill Lynch agreement was the basis for many other settlements of suits against investment houses that had done basically the same thing—trade off the interests of individual investors for the interests of their investment banking business.

Up until now, we have used the phrase 'computers, computer programs, and robots' to identify the object of our focus. Since our primary interest is with the activities engaged in by computers as they are deployed and their programs are implemented, it is best to refer to them as computer systems. Users deploy computer systems to engage in certain activities and they do so by providing inputs of various kinds to the system – turning the system on, modifying the software, setting parameters, assigning values to variables, and so on. The user's input combines with the computer system's functionality to produce an output. Further, every computer system manifests a physical outcome, even if it is the mere illumination of a pixel on an LCD screen.

In this context robots are distinctive only in the sense that they have mobility and sensors which allow them to perform complex tasks in an extended space. Typically robots are responsive to their physical environment, and they have motors that allow locomotion. This special functionality allows robots to engage in activities that most computer systems cannot achieve. Robots are, nevertheless, computer systems; they are a special type of computer system.

For quite some time computer enthusiasts and cognitive scientists have used the language of agency to talk about a subset of computer systems, primarily search engines, web crawlers, and other software "agents" sent out to look for information or undertake activities on the Internet. The term 'bot' has even been introduced for software utilities that search and destroy invading forms of software or "spyware" on a resident computer system. Hence, the idea that computer systems can be thought of as agents is not novel.[5] However, we are extending the idea that software utilities and robots are agents to include all computer systems as agents.

Because of the similarities of some computer system behaviors to human thinking, mobility, and action, the simile of computer surrogate agency may seem strikingly obvious. However, our argument does not turn on functional similarities of computers and humans, such as mobility, responsiveness, or even logical sophistication. Rather, we want to focus on the relationship of computer systems (when in use) to human interests; that is, we want to see

[5] Software agents have been seen as agents of commerce and contract, in the legal sense. See Ian R. Kerr, "Spirits in the Material World: Intelligent Agents as Intermediaries in Electronic Commerce,"(*Dalhousie Law Journal* Vol. 22, No. 2, Fall, 1999, pp. 189-249).

the relationship in its social and moral context. It is there where we locate the key to seeing computer systems as moral agents.

Computer systems are designed and deployed to do tasks assigned to them *by* humans. The search engine finds online information by taking over a task similar to the task humans used to undertake when they rummaged through card catalogues and walked through stacks. Computer systems also take over tasks that were previously assigned to other mechanical devices. For example, an automobile braking system used to work mechanically (though not reliably) to prevent the caliper from locking the pad onto the rotor, which causes the roadwheel to slide in a severe braking maneuver. Now the mechanics of the caliper, pad, and rotor are changed, and an added ABS computer vacillates the caliper pressure very quickly in order to prevent roadwheel lockup. Technically, the same outcome could have been achieved by the driver pumping the brake very rapidly in a panic braking situation. However, given the psychology of panic, the automated system is more reliable than the system of driver-and-mechanical-device.

In addition to aiding humans and machines in doing what was formerly possible but still difficult or tedious, computer systems often perform tasks that were not possible for individuals and purely mechanical devices to achieve. Before fuel injection computers, a driver in a carbureted automobile could not vary the air/fuel mixture to respond to momentary changes in air and engine temperature and barometric pressure. Now, computers make all of these adjustments, and many more, in internal combustion engines. Similarly, an individual could never edit a photograph by changing resolution, colors, and dimensions, and erase all of those changes if not desirable, without the aid of a computer program. Now a child can do these things to a digital image. In all of these cases, users deploy computers to engage in activities the user wants done.

The conception of computer systems as agents is perhaps obvious in the case of search engines deployed on behalf of their users to find information on a certain topic, and in the case of tax software that does more or less what tax accountants do. But the comparison really encompasses a continuum: some computer systems replace human surrogates, other systems do tasks that humans did not do for one another, and still other systems perform tasks that no human (unaided) could ever do. We intend that our account will work just as well for

the automotive tasks described above as for other activities such as word processing, data base management, and so on. The user deploys the system to accomplish certain tasks, and we now talk freely of the computer "doing things" for the user.

Is it not the case, however, that all technological artifacts do things for their users? Not only is this an accurate characterization of all technologies, it seems indeed to define technology.[6] Among technologies, computer systems are special, however, insofar as they accept syntactically structured and semantically rich instructions from humans, both in the programs they implement and in the input they accept from users. These instructions are related to human interests. Most other technological artifacts – think here of a shovel – can be directed by us only via physical force. We can use shovels to further our interests, but we cannot instruct shovels to do our bidding.

Depending on the system and the task, the computer system may do all or some of a task. Spreadsheet software, for example, does not gather data but displays and calculates. Word processors do not help generate ideas or words; nor do they help get the words from one's mind to one's fingers. But the word processing system does help get the words from someone's fingers to a visible medium, and it facilitates change and reconsideration. Insofar as they do things at all, computers act as agents on behalf of humans. Thus, it seems plausible to think of computer systems as agents of humans. Sometimes the computer system is deployed on behalf of an individual; at other times it is deployed on behalf of groups of individuals such as corporations, or other kinds of organizations. As suggested above, when computer systems are deployed on behalf of humans, the activities engaged in involve varying degrees of automation and human involvement. Outcomes are achieved through a combination of human performance and automation. This is what happens with the automobile braking systems, as described above, as humans move their bodies in various ways in relation to automobile levers. This is also what happens with search engines and spybots, where humans manipulate keyboards (or other input devices) and set computers in motion. Tasks are accomplished by combinations of human and machine activity, but in pursuit of the interests of humans.

[6] See Joseph C. Pitt *Thinking About Technology*, (New York: Seven Bridges Press, 2000) in which he argues for a definition of technology as "humanity at work."

3. Computer Systems As Surrogate Agents

Computer systems, like human surrogate agents, perform tasks on behalf of persons. They implement actions in pursuit of the interests of users. As a user interacts with a computer system, the system achieves some of the user's ends. Like the relationship of a human surrogate agent with a client, the relationship between a computer system and a user is comparable to a professional, service relationship. Clients **employ** lawyers, accountants, and estate executors to perform actions on their behalf, and users **deploy** computer systems of all kinds to perform actions on their behalf. We claim that the relationship between computer system and user, like the relationship between human surrogate and client, has a moral component. How is this possible?

Admittedly, human surrogate agents have a first-person perspective independent of their surrogacy role, while computer systems cannot have such a perspective. They do not *have* interests, properly speaking, nor do they have a self or a sense of self. It is not appropriate to describe the actions of computers in terms that imply that they have a psychology. This comparison of agents, interests, and perspectives helps to clarify one of the issues in the standard debate about the moral agency of computers. Those who argue against the agency of computers often base their arguments on the claim that computers do not (and cannot be said to) have desires and interests.[7] This claim is right insofar as it points to the fact that computer systems do not have **first-person** desires and interests. In this respect, they cannot be moral agents in the standard way we think of humans being as moral agents—as having a rich moral psychology that supports sympathy, regret, honor, and the like.

However, the special moral constraints that apply to human surrogate agents do not rely on their first-person perspective. While human surrogate agents do not step out of the realm of morality when they take on the role of lawyer or tax accountant or estate executor, they do become obligated within a system of role morality; they become obligated to take on the perspective of their clients and pursue the clients' interests. Human surrogate agents are both moral agents of a general and a special kind. They are moral agents as human beings with first-person desires and interests and the capacity to control their behavior in the light of

[7] A similar argument against the intelligence of search engines is used by Herbert Dreyfus in *On the Internet* (New York: Routledge, 2001)

its effects on others; they are a special kind of moral agent insofar as they act as the agent of clients and have a duty to pursue responsibly their clients' interests and to stay within the constraints of the particular role.

The latter, special kind of moral agency best describes the functioning of computer systems when they are turned on and deployed on behalf of a user. Surrogate agency, whether human or computer, is a special form of moral agency in which the agent has a **third-person perspective** and pursues what we will call **second-order interests**—those interests of clients or users.

What exactly are the second-order interests of a surrogate agent? By definition, they are interests in or about other interests. Human surrogate agents have second-order interests (not their personal interests) when they pursue, by contractual agreement, the interests of a client. Computer systems take on and pursue second-order interests when they pursue the interests of their users. Computer systems are designed to be able to represent the interests of their users. When the computer system receives input from a user about the interests that the user wants the system to pursue, the system is poised to perform certain tasks for that user. As such, when a computer system receives inputs from a user, it is able to pursue a second-order interest. [8]

Let us be clear that we are not anthropomorphizing computer systems in claiming that they pursue second-order interests, when put to use. Without being put to use, a computer system has no relation to human interests whatsoever. But when a user interacts with the system and assigns it certain tasks, the computer system takes up the interests of the user. These second-order interests can be identified from the behavior of the computer system. For example, when a user commands a browser to search for a map of a destination—the destination to which the user is *interested* in traveling—the browser calls up just that map, and not the map that some other human might want to see. When the browser searches for the map the user wants, the browser has a second-order interest in finding that map. That second-order interest is determined by the combination of the program and the input from the user; the interest cannot be pursued until the user "hires" the computer system to find the map.

[8] We are tacitly claiming what may seem to be an unlikely psychological thesis: that having first-order interests is not a necessary condition for pursuing second-order interests.

When a tax accountant has an interest in minimizing a client's income tax burden, the accountant has a second-order interest. As indicated earlier, the first-order interests of human surrogate agents are not eliminated when they act in role; rather, some of the first-order interests of the human surrogate agent become temporarily aligned (in a limited domain) with the interests of a client. In other words, the human surrogate agent has self-interested reasons for seeing that some subset of the client's interests are successfully pursued. The tax accountant wants the client's tax burden to be reduced, both for the good of the client and for his or her own good. There is a temporary alignment between some of the first- and second-order interests of the human surrogate agent.

This alignment between the surrogate's interests and the client's interests cannot be a feature of computer systems, since computer systems do not have first-order interests. This is one important difference between human surrogate agents and computer systems and we do not mean to underestimate the significance of this difference. In psychological terms, computer agents do not sympathize or empathize with their users or "identify" with their interests. But the differences between human and computer surrogate agents, while significant, do not go as deep as many would think. Consider, for instance, the issue of expertise. It is important to acknowledge that in many cases of human surrogate agency, the client may not fully understand what the agent does. But this is true of users and their computer systems too. Indeed, in the cases we have discussed, the client/user has deployed the agent because the client/user does not have the requisite expertise or does not want to engage in the activities necessary to achieve the desired outcome. In the case of hiring a tax accountant as well as the case of using an income tax software package, the client/user does not need to understand the details of the service that is implemented. The user of the software package need not understand the tax code or how computer systems work; the client of the tax accountant need not understand the tax code or how accountants do their work. In both cases the client/user desires an outcome and seeks the aid of an agent to achieve that outcome.

Our comparison of human surrogates and computer systems reveals that both kinds of agents have a third-person perspective and pursue second-order interests. We have pointed out that the primary difference between human and computer surrogate agents concerns psychology and not morality; human surrogate agents have first-order interests and a first-person perspective, while computer systems do not. Note, however, that when it comes to

moral evaluation of the surrogate agent's behavior qua surrogate agent, these first-order interests and first-person perspective are irrelevant. The primary issue is whether the agent is incompetent or misbehaves with respect to the *clients'* interests. In other words, does the surrogate agent stay within the constraints of the special role morality?

It will now be useful to look in more detail at the ways in which human and computer surrogate agents can go wrong and see how this account of the moral agency of computer systems plays out. We will have to do so, however, without recourse to a specific role morality. The particular constraints of the role morality will depend on just what role is under consideration, and so our discussion here is necessarily general and abstract. The moral constraints of a tax accountant, for instance, differ significantly from those of an estate executor. Likewise, if our account is correct, the moral constraints on personal gaming software will differ from those on software that runs radiation machines, or secures databases of medical information, or guides missile systems.[9]

3.1 Incompetence

Both income tax accountants and income tax software system can perform incompetently. Just as the incompetence of the accountant might derive from the accountant's lack of understanding of the tax code or lack of understanding of the client's situation, a computer system can inaccurately represent the income tax code or errantly manipulate the input from a user. In both cases, the surrogate agent may not have asked for the right input or may have misunderstood or errantly assigned the input provided.

The outcome or effect on the client/user is the same in both cases. That is, whether it is a human surrogate agent or a computer system, incompetence can result in the client/user's interest not being fully achieved or not achieved to the level the client/user reasonably expected. For example, the incompetence of agents of either kind may result in the client/user missing out on a filing option in the income tax code that would have reduced the client/user's taxes. These are some of the morally relevant effects to which we referred in the opening section.

[9] See, e.g., Nancy Leveson *Safeware: System Safety and Computers*, (Boston: Addison-Wesley, 1995) and Mary L. Cummings and Stephanie Guerlain, "The Tactical Tomahawk Conundrum: Designing Decision Support Systems for Revolutionary Domains", IEEE Systems, Man, and Cybernetics Society conference, Washington DC, October 2003.

Admittedly, ordinary language often treats the two cases differently; we say of the tax accountant that he or she was 'incompetent,' and we say of the software package that it was 'faulty' or 'buggy.' This linguistic convention acknowledges that the one is a person and the other a computer system. No doubt, critics will insist here that the wrong done to the user by the computer system is done by (or can be traced back to) the designers of the software package. Indeed, the *de facto* difference between the two cases is in the way the legal system addresses the incompetence in each case. A client sues a human surrogate agent for negligence, and draws on a body of law focused on standards of practice and standards of care. Software users can also sue, but they must use a different body of law; typically software users will sue a software company for defective (error-ridden) software and will do so only if the errors in the system go beyond what the software company disclaims in the licensing agreement.[10]

There is a special kind of incompetence in designing computer systems that goes beyond programming errors. Problems can emerge when otherwise good modules in software/hardware systems are combined in ways that bring out incompatibilities in the modules.[11] It is hard to say where exactly the error lies; parts of the system may have functioned perfectly well when they were in different systemic configurations. Software packages and computer system configuration are generally the result of complex team efforts, and these teams do not always work in concert. Error can be introduced in any number of places including in the design, programming, or documentation stage. Thus, it will take us too far afield to identify and address all the different causes leading to a computer system performing incompetently. But certainly there is some level of incompetence when a computer system is put together without testing the compatibility or interoperability of its components, through various state-changes in the system.

3.2 Misbehavior

The second way in which a human surrogate agent can go wrong is through

[10] Standard EULA agreements make it exceedingly difficult for users of software to get relief from the courts for faulty software. In this section, we suggest that one way for computer surrogate software to be faulty is for it to be incompetent in pursing the interests of the client/user. If our argument about the human-computer surrogacy parallel is correct, it should be no more difficult to win a suit against a computer than against a human surrogate agent. We should add here that the two cases are alike for IRS purposes in that the client/user is always responsible for errors in their tax returns. The comparison between the two cases is also complicated because currently most income tax accountants use software packages to prepare their clients' tax returns.

[11] We would like to thank David Gleason for bringing this special kind of incompetence to our attention.

misbehavior. Here the agent uses the role to further the interests of someone besides the client, and in a way that neglects or harms the interests of the client.[12] As already indicated, computer systems cannot take a first-person perspective. Hence, it would seem that computer systems can not misbehave. Indeed, it is for this reason that many individuals and corporations prefer computer systems to human agents; they prefer to have machines perform tasks rather than humans, believing that computers are programmed to pursue only the interests of the user. Of course, machines break down, but with machines the employer does not have to worry about the worker getting distracted, stealing, being lazy or going on strike. Computer systems do exactly what they are told (programmed) to do.

Computer systems cannot misbehave by pursuing *their* personal interests to the neglect or detriment of their users. On the other hand, while computers do not have (and hence cannot pursue) their own interests, computer systems can be designed in ways that serve the interests of people other than their users. They may even be designed in ways that conflict with or undermine the interests of their users. As indicated earlier, computer systems have a way of pursuing the interests of their users, or of other (third) parties. Misbehavior can occur when computer systems are designed in ways that pursue the interests of someone other than the user, and to the detriment of the interests of the user. Consider the case of an internet browser that is constructed so that it pursues the interests of other third parties. Most browsers support usage tracking programs, cookies, pop-ups or adware, keyloggers (which transmit data about your computer use to third parties), and other forms of spyware on a personal computer. Most of this noxious software is installed without the user's expressed, or at least informed, consent. Accordingly, we might say that an internet browser is the surrogate agent of the end-user (client) when it searches the Internet for information, but at the same time acts as the surrogate agent of other clients, such as advertisers, corporations, hackers, and government agencies, when it allows or supports invasions of privacy and usurpation of computer resources.

Such misbehavior is embodied in other parts of computer systems. In the mid-1990's, Microsoft marketed a version of their platform for personal computers that was advertised to

[12] If the computer system merely neglects, but does not harm, the interests of the user, and the user has paid for or rented the system in order to further his or her interests, then it is still reasonable to say that the user has born a cost to his or her interests. That is, both opportunity costs and real costs to interests will count as harms.

consumers as more flexible than it really was. Though Microsoft claimed, in particular, that their Internet Explorer version 4.0 would operate smoothly with other Sun JAVA ™ applications, in fact Microsoft had programmed IE version 4.0 and other types of software with a proprietary form of the JAVA code. [13] The non-proprietary JAVA programming technology was, per agreement, to be supported by Microsoft, and in exchange Microsoft could advertise that fact on its software products. Hence consumers thought that they were getting products that would be compatible with all or most of Sun JAVA applications when in fact they were getting software that was reliable only with proprietary versions of JAVA. The expectation of broader interoperability was bolstered by the very public nature of the agreement between Sun and Microsoft. Here the users' interests in interoperability were undermined by Microsoft's interests in getting users to use only Microsoft applications — the very applications that would work with the Microsoft proprietary JAVA. In the courts, it appeared as though the problem was just a legal one between Microsoft and Sun, but in the technology itself users were confronted with a system that would not support at least some of the users' interests in using non-Microsoft products. Not surprisingly, the users were not informed of Microsoft's use of proprietary JAVA code, but would discover this fact when they tried (unsuccessfully) to use some Sun JAVA programs with their Microsoft computing platforms.

There are many kinds of misbehavior to be found in the activities of human surrogate agents. Imagine hiring a lawyer to represent you and later finding that while the lawyer represents you well, he or she is selling information about you to fundraising or advertising organizations. Here the agent's activities introduce the possibility of conflict between the agent's interests and third-party interests. Consider also the case of the Arthur Anderson auditors who were suppose to ensure that Enron stayed within the laws of corporate accounting. They can be thought of as agents of Enron stockholders. They misbehaved by allowing their judgment on behalf of Enron to be distorted by their own (Arthur Anderson's) interests. In parallel, users deploy a computer system such as a browser to seek out

[13] Sun and Microsoft agreed that the latter would support JAVA in their operating system and applications in March of 1996. Subsequently, Microsoft seems to have reneged on the deal, but still advertised that their products were "Sun JAVA compatible." The case was settled in favor of Sun Microsystems. The complaint can be accessed at http://java.sun.com/lawsuit/complaint.html. This is one of many lawsuits initiated over JAVA by Sun, not all of which were successful.

information they desire, believing that the browser will serve their interests. The browser, however, has been designed not just to serve the interest of the user but also to serve the interests of the software producer, or advertisers, or even hackers. The information delivered by the browser may or may not serve the interest of the user. In the literature on conflict of interest, these cases can be seen as classic conflicts of interest in which the agent's judgment is tainted by the presence of a conflicting interest. The client believes the agent is acting on his or her behalf and discovers that the agent has interests that may interfere with that judgment. In the case of pop-ups, adware, etc., the user typically has no interest in the functions that have been added to the computer system. Hence, from the perspective of the user, these aspects of browsers are a kind of misbehavior or, at the least, a candidate for misbehavior.

When a surrogate agent is hired by a client, the agent is authorized to engage in a range of activities directed at achieving a positive outcome for the client. Similarly, computer agents are put into operation to engage in activities aimed at an outcome desired by the user, that is, the person who deployed the computer program. Not only are human surrogate agents and computer agents both directed towards the interest of their client/users, both are given information by their client/users and expected to behave within certain constraints. For human surrogate agents, there generally are social and legal understandings such that when the behavior of a surrogate agent falls below a certain standard of diligence, authority, or disclosure, the client can sue the agent and the agent can be found liable for his or her behavior. This suggests that standards of diligence and authority should be developed for computer agents, perhaps even before they are put into operation.

3.3 Differences between computer systems and human surrogate agents.

We want to be clear about the precise scope of the computers-as-surrogates simile that lies at the heart of our argument. The most fruitful part of the simile comes in the way it reveals moral relations between human surrogate agents and clients, on the one hand, and computers and users, on the other. But we are not claiming that all computer systems are like *known* surrogate agents. Likewise, not all human surrogate agents engage in activities that could be likened to the operation of a computer system. There may be some human surrogate

agents, for instance, who rely on certain cognitive abilities, in the performance or their roles, which are in no way similar to the computational abilities of computer systems. Many skeptics about the human-computer comparison rely on a particular dissimilarity: humans use judgment and intuition, while computers are mere algorithmic or heuristic "thinkers."

If the surrogacy role always and essentially depended on the agent exercising judgment or guiding the client by using intuition, then computers could not be surrogate agents because they lack these mental capacities. But what reasons do we have for thinking that human surrogate agents rely principally or exclusively on judgment or intuition, and not on codified rules of law and standard practice—rules a computer system can also follow? Certainly the rules of the federal taxing and investment authorities, like the IRS and the SEC in the U.S., the statutes concerning estates and probate, and other laws can be programmed into a computer system. The best computer surrogate agents, then, are likely to be expert systems, or perhaps even "artificially" intelligent computers, that can advise clients or users through a maze of complex rules, laws, and guidelines. For those roles where the human surrogate cannot define such formal components of the agency—roles such as 'educational mentor', 'spiritual guide', or 'corporate raider'—perhaps there will never be computer surrogates that might take over.

Of particular importance is the role of information in the proper functioning of a surrogate agent. An agent can properly act on behalf of a person only if the agent has accurate information relevant to the performance of the agent's duties. For human surrogate agents, the responsibility to gather and update information often lies with those agents. For computer agents, the adequacy of information seems to be a function of the program and the person whose interests are to be served. Of course, the privacy and security of this information, in digitized form, are well-known concerns for computer ethics. A new aspect of information privacy and security is raised by computer surrogate agency: can computer programs "know" when it is appropriate to give up information (perhaps to governments or marketing agencies) about their clients? Discovering the proper moral relations between computers and users may depend in part upon further inquiries in information science.

A complete account of the cognition and psychology of human surrogate agency is beyond the scope of this paper. In lieu of such an account, it should be enough to note that

there are many forms of human surrogate agency that pursue the interests of clients, are prone to the kinds of misbehavior and incompetence we described earlier, and do not rely on non-formalizable "judgment" or "intuition". Likewise, there are many computer systems that serve the same role for their users.

4. Conclusion: Issues of responsibility, liability, and blame

What, then, do we learn from thinking of computer systems as surrogate agents? The simile brings to light two aspects of computer systems that together provide the basis for claiming that computer systems have a kind of moral agency. Computer systems have a third-person perspective that allows them to take on second-order interests, and the way in which they do this has effects on human interests and can be evaluated in much the same way we scrutinize the pursuit of second-order interests by human surrogate agents. In such an evaluation, we are able to apply *to computer systems* the concept of morality as a set of constraints on behavior, based on the interests of others. As surrogate agents, computer systems pursue interests, but they can do so in ways that go beyond what morality allows.

Recognizing that computer systems have a third-person perspective allows us to evaluate systems in terms of the adequacy of their perspective. Just as we evaluate human surrogate agents in terms of whether they adequately understand and represent the point of view of their clients, we can evaluate computer systems in terms of how they represent and pursue the user's interests. Such an evaluation would involve many aspects of the system including what it allows users to input and how it goes about implementing the interests of the user. Consider the search engine surrogate that pursues a user's interest in websites on a particular topic. Whether the search engine lists websites in an order that reflects highest use or one that reflects how much the website owner has paid to be listed or one that reflects some other listing criteria can have moral implications.[14] Recognizing the third-person perspective allows us, then, to ask a variety of important questions about computer systems: Does the

[14] Introna, Lucas D., and Nissenbaum, Helen (2000) Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society* 16, no. 3: 169–185.

system act on the actual user's interests, or on a restricted conception of the user's interests? Does the system competently pursue the users' interests, without pursuing other, possibly illegitimate interests such as those of advertisers, computer hardware or software manufacturers, government spying agencies, and the like?

Throughout the paper we have provided a number of analyses that illustrate the kind of evaluation that can be made. Tax preparation programs perform like tax advisers; contract-writing programs perform some of the tasks of attorneys; Internet search engines seek and deliver information like information researchers or librarians. Other types of programs and computer systems serve the interests of clients, but there are no corresponding human surrogate agents with whom to compare them. Spyware programs uncover breeches in computer security, but when they do so for the user, they do not replace the tasks of a private detective or security analyst. Increasingly, our computers do more for us than human surrogates could do. This is why it is all the more important to have a framework for morally evaluating computer systems, especially a framework that acknowledges that computer systems can do an incompetent job of pursuing the interests of their users and can misbehave in their work on behalf of users.

While our conception of computer systems as surrogate agents has wide-ranging implications, we will conclude by briefly focusing on one particular area of concern that is likely to be raised by the human surrogate-computer surrogate comparison. Foremost in the traditional analysis of role moralities are questions about rights and responsibilities. Many professional societies, in writing professional codes of ethics, have struggled with articulating the rights and responsibilities of human surrogate agents and their clients. How far can surrogate agents go to achieve the wishes of the client? If the surrogate agent acts on behalf of a client and stays within the constraints of the role, is the agent absolved of responsibility for the outcomes of his/her actions on behalf of the client?

Thus, the implications of our thesis for issues of responsibility, liability, and blame seem important. Since we claim that computers systems have a kind of moral agency, a likely (and possibly objectionable) inference is that computer systems can be responsible, liable, and blameworthy. This inference is, however, not necessary and should not be made too quickly. There are two issues that need further exploration. First, we must come to grips with issues of

responsibility, liability, and blame in situations in which multiple and diverse agents are at work. In cases involving computer systems, there will typically be at least three agencies at work -- users, systems designers, and computer systems; and, second, we must fully understand the kind of agency we have identified for computer systems.

In addressing the first issue it is important to note that we have **not** argued that users or system designers are absolved of responsibility because computer systems have agency. We anticipate that the standard response to our argument will be that the attention of moral philosophers should remain on system designers. Of course, computer systems are made by human beings, and, hence, the source of error or misbehavior in a computer system can be traced back, in principle, to human beings who made decisions about the software design, reasonable or otherwise. Similarly, when lawyers consider legal accountability for harm involving a computer system, they focus on users or system designers (or the companies manufacturing the computer system). In making these claims, however, moral philosophers and lawyers push computer systems out of the picture, treating them as if they were insignificant. This seems a serious mistake. A virtue of our analysis is that it keeps a focus on the system itself.

To understand computer systems merely as designed products, without any kind of moral agency of their own is to fail to see that computer systems also behave, and their behavior can have effects on humans and can be morally appraised independently of an appraisal of their designers' behavior. What the designer does and what the computer does (in a particular context) are different, albeit closely related. To think that only human designers are subject to morality is to fail to recognize that technology and computer systems constrain, facilitate, and in general shape what humans do.

The point of emphasizing the moral character of computer systems is not to deflect responsibility away from system designers. Since computer system and system designer are conceptually distinct, there is no reason why both should not come in for moral scrutiny. Ultimately, the motivation to extend scrutiny to computer systems arises from the fact that computer systems perform tasks and the way they do so has moral consequences--consequences that affect human interests.

This brings us to the second issue: since computer systems have a kind of moral

agency, does it make sense to think of them as responsible, liable or blameworthy? We do not yet have the answer to this question though we have identified some pitfalls to avoid and a strategy for answering it. First, while we have argued that computer systems are moral agents, we have distinguished this moral agency from the moral agency of human beings. Hence, it is plausible that the moral agency of computer systems does not entail responsibility, liability, or blame. We have acknowledged all along that computer systems do not have the first-person perspective, nor the moral psychology or the freedom that are requisite for standard (human) moral agency. Computer systems are determined to do what programs tell them to do. Instead we have proposed a kind of moral agency that parallels human surrogate agency. Before proclaiming that notions of responsibility, liability, and blame can or cannot be used in relation to computer systems, we need a more complete analysis of human surrogate agency and responsibility, liability, and blameworthiness of individuals acting in such roles.

The surrogacy comparison should go some distance in helping us here. For example, in the case of a trained human surrogate agent, a failure of incompetence would reflect poorly on the source of the training. A professional school that trains accountants for the CPA license, for instance, would be accountable if it regularly taught improper accounting methods. The designer of a computer accounting system, on the other hand, would be to blame if the computer program used the wrong database in calculating a user's tax rate. But the professional school would not be accountable if its graduates regularly forgot the proper accounting method (a form of incompetence), or diverted funds from the client's account to his or her own (a form of misbehavior). Likewise, the designer of the computer system would not be to blame if an unpredictable power surge changed a few variables while the user was calculating the tax rate (still, an incompetence in the computer system). Nor would the designer be to blame for every bug in a very complex computer program, on the assumption that complex programs cannot be proven "correct" or bug-free within the lifetime of the user.[15] The possibility of bugs in tax-preparation software, like the chance of cognitive breakdowns in the CPA, must be assumed as a risk of hiring another to pursue one's proper interests. On the other hand, the designer would be to blame for deliberately programming a

[15] See Brian Cantwell Smith, "The limits of correctness in computers," CSLI 1985, reprinted in Johnson and Nissenbaum (eds.) *Computers, Ethics, and Social Values* (Saddle River, N.J.: Prentice Hall, 1995)

routine that sent e-mail summaries of one's tax returns to one's worst enemies—certainly a form of misbehavior.

We do not claim to have figured out whether or how to adjust notions of responsibility, liability, and blame to computer systems. We leave this daunting task as a further project. Our point here is merely to suggest that computer systems have a certain kind of moral agency and this agency and the role of this agency in morality should not be ignored. In other words, while we have not fully worked out the implications of our account for issues of responsibility, they are worth facing in light of the virtues of the account.

In some ways, the need to give a moral account of computer systems arises from the fact that they are becoming increasingly sophisticated, in both technical and social dimensions. Though they may have begun as simple utilities or "dumb" technologies to help humans connect phone calls, calculate bomb trajectories, and do arithmetic, they are increasingly taking over roles once occupied by human surrogate agents. This continuous change would suggest that, somewhere along the way, computer systems changed from mere tool to agent. Now, it can no longer be denied that computer systems have displaced humans—both in the manufacturing workforce, as has long been acknowledged, and more recently in the service industry. It would be peculiar, then, for users to recognize that computers have replaced human service workers who have always been supposed to have moral constraints on their behavior, but to avoid the ascription of similar moral constraints to computer systems.