Moral Responsibility for Computing Artifacts: Five Rules

Preamble

The complexity of computing artifacts does not absolve those who design, develop, or deploy systems from responsibility for the technologies they introduce. We propose five rules as a guide to moral responsibility for computing artifacts.

This is a collaborative document, and it does not include everything each of us believes about this subject. However, each signer of this document supports what is written here.

A Working Definition of "Moral Responsibility"

The following is from the Stanford Encyclopedia of Philosophy [1]:

> When a person performs or fails to perform a morally significant action, we sometimes think that a particular kind of response is warranted. Praise and blame are perhaps the most obvious forms this reaction might take. … To regard such agents as worthy of one of these reactions is to ascribe moral responsibility to them on the basis of what they have done or left undone. … Thus, to be morally responsible for something, say an action, is to be worthy of a particular kind of reaction—praise, blame, or something akin to these—for having performed it.

For our limited purposes, we will not make statements about legal liability, concentrating instead on issues of ethical accountability. The particular kind of accountability we discuss here will be identified using the phrase "moral responsibility" throughout.

A Working Definition of "Computing Artifacts"

An artifact is an object made or shaped by humans [2]. We use "computing artifact" for any object that includes an executing computer program as part of the object. We intend to include software applications running on a general purpose computer, programs burned into hardware and embedded in mechanical devices, robots, webbots, programs distributed across more than one machine, and many other configurations.

A Working Definition of "Sociotechnical Systems"

We place each computing artifact into the context of "sociotechnical systems." A sociotechnical system includes people, relationships between people, other artifacts, physical surroundings, customs, assumptions, procedures and protocols. [3]

We acknowledge the importance of sociotechnical systems to the issue of moral responsibility for computing artifacts. As a straightforward example, a GPS navigator is a computing artifact, but in isolation from the satellites it uses for ascertaining location, it cannot perform its function. People, commercial enterprises, governments and artifacts were necessary to design, develop, and deploy the satellite system and the navigator. The people who make the device encourage and discourage different uses by the navigator's design. The people who buy the navigators choose to use it in different ways. A communications protocol for communicating with those satellites had to be negotiated between stakeholders. The methods by which people have agreed to identify places on the earth's surface form another part of the sociotechnical system without which an automated navigator is incomprehensible.

The significance of sociotechnical systems complicates any discussion of moral responsibility for computing artifacts. On one hand, ignoring the sociotechnical systems in which a computing artifact is embedded is folly. On the other hand, including all relevant sociotechnical systems components in every discussion of moral responsibility involving a computing artifact will make it impractical to assign meaningful responsibility to the humans most directly involved with that specific artifact. In order to negotiate this tension, we first discuss moral responsibility for computing artifacts in a more focused sense (Rules 1, 2 and 3), and then place this discussion into a perspective that explicitly includes sociotechnical systems (Rules 4 and 5).

Rule 1: **The people who design, develop, and deploy a computing artifact are morally responsible for that artifact, and for the foreseeable effects of that artifact. This responsibility is shared with other people who design, develop, deploy and knowingly use the artifact as part of a sociotechnical system.**

Rule 2: **The shared responsibility of computing artifacts is not a zero-sum game. The responsibility of an individual is not automatically reduced as more people become involved in designing, developing, deploying and using the artifact. Instead, a person's responsibility includes his or her accountability for the behaviors of the artifact and for the artifact's effects after deployment, to the degree to which these effects are reasonably foreseeable by that person.**

When humans design, develop and deploy computing artifacts, they do so consciously and intentionally. This intentionality is important when discussing moral responsibility. Rules 1 and 2 are meant to clarify the moral responsibility of the people most directly accountable for a specific computing artifact and its effects on others.

Rule 2 is meant to address the problem of many hands, in which responsibility that is broadly shared is sometimes considered insignificant for any one individual. See [4] for more detail on this issue.

The words "reasonably," "foreseeable" and "knowingly" add complexity, subtlety and ambiguity to Rules 1 and 2. This is unfortunate, as we'd like the rules to be as straightforward as possible. However, in order to make the rules realistic, we don't immediately see a way to make them simpler than this.

By using the word "foreseeable," we acknowledge that the people who design, develop and deploy artifacts cannot reasonably be expected to foresee *all* the effects of the artifacts, for all time. However, implicit in our use of this word is the expectation that these people make a good faith effort to predict the uses and effects of the deployment, and to monitor them after deployment. Willful ignorance, or cursory thought, is not sufficient to meet the ethical challenges of Rules 1 and 2.

Furthermore, if people design an artifact in such a way that it is not possible to reasonably predict its future behaviors, then those people are particularly responsible for the unpredictable, and potentially harmful, results. We insist that machines that are designed to adapt over time, "learn" without human supervision, or self-modify their own code, are machines for which the people who launch them are *more* rather than *less* responsible than people who launch more predictable machines. We assert that a machine's unpredictability increases people's responsibility for anticipating problems and safeguarding against them. People who recognize their responsibilities in this way are likely to make their machines simpler and more predictable in order to make them safer and more reliable; we would welcome this outcome.

Another caution about Rules 1 and 2 concerns a decision to *not* create or deploy a computing artifact. This decision also has consequences. For example, if it is reasonably foreseeable that the creation and deployment of an artifact is likely to have a *good* effect, and the decision is made to *not* create and deploy that artifact, that decision has ethical significance, and those who made the decision are responsible for the consequences.

Rule 3: **People who knowingly use a particular computing artifact are morally responsible for that use.**

The word "knowingly" is problematic in Rule 3, but we think it is, on balance, appropriate. People who "use" a particular computing artifact may or may not be aware of this use. For example, the driver of a car may not have any knowledge of a computing artifact embedded in the car, an artifact that records data for analysis in case of a crash. It seems to us counter-intuitive to assign moral responsibility to the driver for the use of that artifact. However, when someone knowingly, intentionally, uses a particular computing artifact, that person takes on moral responsibility attached to that use. A dramatic example is when someone launches a cruise missile at an enemy target; a more mundane example is when someone searches the web for information about a prospective employee. The moral responsibility of a user includes an obligation to learn enough about the computing artifact's effect to make an informed judgment about its use for a particular application.

It is not our intent to absolve the users of computing artifacts from any moral responsibility if they are willfully ignorant about artifacts or their effects. Rule 3 could be misused in this way. We acknowledge this problem, but we judge that the possibility of this abuse does not negate that there are practical and ethically significant differences in the way people interact with computing artifacts. For example, "users" of computing artifacts cannot be reasonably held accountable if the use is hidden from them. (The hidden nature of the artifact may be intentional or incidental.) However, people should not seek, or even allow themselves, to be ignorant about technology and its effects in order to avoid responsibility for their use of technology. As with Rules 1 and 2, Rule 3 applies to people who consciously decide *not* to use a computing

artifact. In order to place Rules 1, 2 and 3 into perspective, we assert two more rules:

Rule 4: **People who design, develop, deploy, and knowingly use a computing artifact can only do so responsibly when they make a reasonable effort to take into account the sociotechnical systems in which the artifact is embedded.**

Sociotechnical systems are increasingly powerful. If people thoughtlessly produce and adopt these systems, they are, in our opinion, being morally irresponsible. Ignorance is not a justification for harms associated with sociotechnical systems and the computing artifacts imbedded in those systems.

Rule 4 is intended to be a progressively heavy burden. It requires an honest effort to identify and understand relevant systems, commensurate with one's ability and one's depth of involvement with the artifact. Thus, the burden is heavier for those with more expertise and more influence over the artifact's effects. Those in design and development cannot shift their burden to the users (see Rule 2), and users cannot shift the burden to developers when users' local knowledge is critical to appropriate ethical action. The sociotechnical systems in which an artifact would be embedded should be considered even when the decision is to not design, develop, deploy or use a computing artifact.

Rule 4 expands the effect of Rules 1, 2 and 3, since people who obey Rule 4 will know more about the effects of a computing artifact they produce and/or use.

Rule 5: **People who design, develop, deploy and promote a computing artifact should not explicitly or implicitly deceive users about the artifact or its effects, or about the sociotechnical systems in which the artifact is embedded.**

Morally responsible use of computing artifacts and sociotechnical systems requires reliable information about the artifacts and systems. People who design, develop, deploy and promote a computing artifact should provide honest, reliable, and understandable information about the artifact, its effects, and to the extent foreseeable, about the sociotechnical systems in which they assume the artifact will be embedded.

Computing Artifacts that are Not Exceptions to the Rules

No matter how sophisticated computing artifacts become, the rules hold. For example, if an artifact uses a neural net, and the designers subsequently are surprised by the artifact's effects, the rules hold. If a computing artifact is self-modifying, and eventually becomes quite different from the original artifact, the rules still hold. If a computing artifact is a distributed system or an emerging system, the rules still hold for the humans associated with the pieces that are distributed, for the humans associated with the organization of the overall system, and for the humans responsible for the system from which the new system emerged. If a computing artifact A is launched to build another computing artifact B, and artifact B builds computing artifact C, then the rules are applied repeatedly so that the humans moral responsible for A are also responsible for B and C. (That is, if you are not willing to accept moral responsibility for A, B, and C, then you should not launch A.)

As described in the discussions of the Rules above, there are responsibilities associated with *not* launching any computing artifact. However, when the predictability of an artifact's future behavior is in serious doubt, we maintain that the precautionary principle [5] should be applied, which will require a particularly serious effort to weigh (uncertain) benefits against (unpredictable) costs. We recognize that this is a heavy burden on those advocating launching such artifacts, and we contend that this is appropriate.

References

[1] Eshleman, Andrew, "Moral Responsibility", *The Stanford Encyclopedia of Philosophy (Winter 2009 Edition)*, Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/win2009/entries/moral-responsibility/, accessed 17 May 2010.

[2] Wiktionary entry for "Artifact." URL = http://en.wiktionary.org/wiki/artifact, accessed 9 March 2010.

[3] Chuck Huff. Why a Sociotechnical System? URL = http://computingcases.org/general_tools/sia/socio_tech_system.html, accessed 9 March 2010.

[4] Nissenbaum, H. 1994. Computing and Accountability. *Communications of the ACM 37*, 1 (Jan. 1994), 72-80.

[5] Som, C., Hilty, L. M. & Ruddy, T. F. (2004). The Precautionary Principle in the Information Society. *Human and Ecological Risk Assessment, 10* (5), 787-799.

Signed by:

Colin Allen, Indiana Univ.
Joe Herkert, Arizona State
Chuck Huff, St. Olaf
Deborah Johnson, Univ. of Virginia
Keith Miller, Univ. of Illinois Springfield
James Moor, Dartmouth
Ken Pimple, Indiana Univ.
Noel Sharkey, Univ. of Sheffield
Wendell Wallach, Yale

Meta-Rules (rules about changing the text of this document):

Meta-rule 0. In this document, “we” refers to people who have signed on to this document. The “coordinator” is one of the signers, and is in charge of version control. The coordinator is currently Keith Miller, miller.keith@uis.edu.

Meta-rule 1. Anyone we invite can sign on to the document.

Meta-rule 2. Anyone we invite can suggest changes in the document.

Meta-rule 3. Any proposed changes should be emailed to the coordinator. The coordinator emails all the signers with proposed changes. If there are no objections sent to the coordinator within 10 days after the coordinator sends out a proposed change, the proposed change is accepted and a new version is emailed out.

This document has been discussed at:

Michael S. Pritchard and Joseph R. Herkert, (2010), "Robots and Ethical Responsibility," *The Forum on Philosophy, Engineering, and Technology*, May 9-10, Golden, CO.