Machine Self-Sacrifice

C. J. Reynolds and A. Cassinelli
University of Tokyo
Engineering Building 6, Room 230
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-8656, Japan
Phone: +81 3-5841-6937
carson@k2.t.u-tokyo.ac.jp

Abstract

The concept of self-sacrifice as it relates to artificial entities is defined. Illustrative anecdotes drawn from computer science, robotics, and microprocessor architecture are then provided. Building upon this we will argue for the utility of self-sacrifice in existing biological phenomena such as kin altruism. We will conclude by making a counter-intuitive claim from the standpoint of information ethics: that information should have the capability to destroy itself.

Keywords: ethics, artificial moral agents, artificial life, artificial death, self-sacrifice

SELF-HARM, WEAPONIZATION, SUICIDE AND SELF-SACRIFICE

Consider a machine that is capable of self-destruction. There are many genres of self-destruction with attendant causes, effects, harms, and benefits. Let us provide a simple taxonomy. Of self-destructing machines physical harm that is a byproduct of the process of destruction can either be limited to the machine or vented upon the external world. A machine that destroys itself, and does so with the intent of harming others in the external world is one variety of a weapon. For instance, a land mine is an artificial machine with sensors and mechanically or digitally encoded logic, which was built with the intent of harming some and thus clearly a weapon.

If we exclude weapons, then we are left with the set of machines that self-destruct but limit physical harm to the artificial self. One example of such a system is the MIPS-X microprocessor, which included a special machine instruction **hsc** (Chow, 1986). The programmer's manual for the processor documents the **hsc** instruction as follows;

> 4.58. hsc - Halt and Spontaneously Combust
>
> ...
>
> The processor stops fetching instructions and self destructs.
>
> Note that the contents of Reg(31) are actually lost.
>
> ...
>
> This is executed by the processor when a protection violation is detected. It is a privileged instruction available only on the -NSA versions of the processor.

This microprocessor is able to disable itself in response to a special command that can be sent by programmers. The destruction however is limited to the processor and does not (by design) seek to cause physical harm outside the processor.

Among non-weapon artificial machines with the capability of self-harm we can further distinguish between systems which are artificially suicidal and those that perform artificial self-sacrifice. One might argue that the distinction between suicide and self-sacrifice is a matter of perspective (as some political groups might label an individual a martyr or suicidal terrorist depending on affiliation). However, we will side step this deeply politicized argument with the following stipulation. When an artificial entity intends to self-destruct to induce psychological harm then we will define it as artificial suicide. This obviously has only a limited correspondence with suicide as it is defined in the human domain, which may stem from various intentions as evidenced through actions like euthanasia and running amok. Alternatively, when an entity intends to self-destruct to induce physical or psychological benefit to others then we will define it as artificial self-sacrifice.

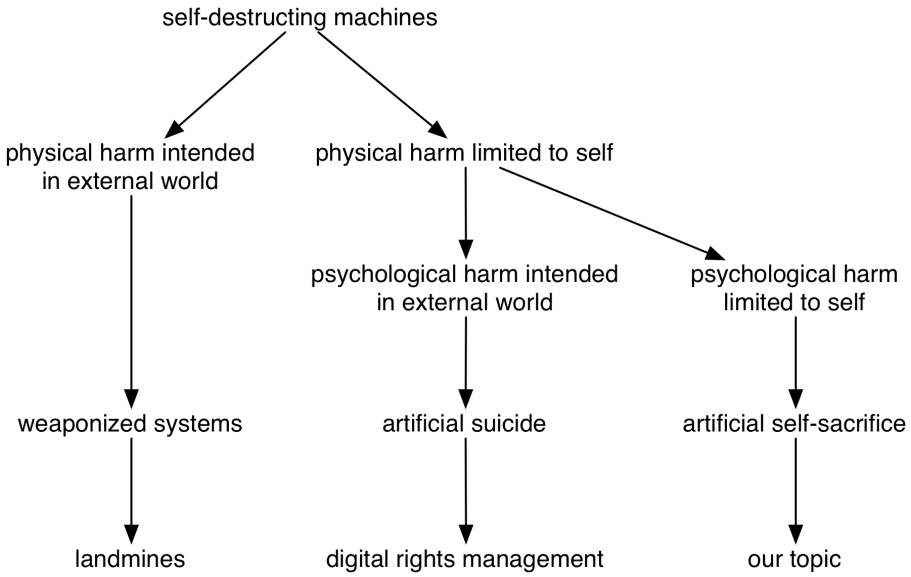

Figure 1: taxonomy of self-destructing machines

The taxonomy employed in this paper is diagrammed in figure 1. From it one can see that we consider digital rights management to be a variety of artificial suicide. Digital rights management systems may render digital content useless after a certain condition is met. For instance, the video game Spore includes a system, which allows the game to be installed on a computer three times before disabling itself. We believe this self-destruction is intended to induce psychological harm among would-be pirates of the video game.

ARTIFICIAL DEATH

Although the topics of artificial self-sacrifice and artificial suicide are somewhat morose, one can view them as liminal topics in the area of Artificial Life (Langton, 1986). If Artificial Life researchers are seeking to provide a definition of life, as evidenced in artificial systems, then it seems that by extension that they must grapple with under which conditions an artificial system dies. It is relevant to note that the relationship between artificial moral agents and Artificial Life has already been discussed by Sullins (2005).

So here we will directly discuss the notion of artificial death. With robot systems, death is something that is ascribed analogically to human death. When a robotic system loses the ability to move or process then it is colloquial to call it "dead." Further evidence of this analogical thinking can be observed in the naming of "kill switches" which stop the movement of mechanical systems. A recent art installation, "Shockbot Corehulio," illustrates some ideas about robotic death:

> At the point of contact a short-circuit occurs creating a fault current... As the damage to the computer increases, there is a proportional rise of dysfunction to the control signal. This overload of errors ends in a total collapse of the system (Andel & Gütze, 2005).

But then robots can resuscitate by repairing broken components and restoring processing function. A more final variety of robot death would be a case in which the robot is physically destroyed and the ability to resuscitate the robot is no longer possible.

The case of death for digital or software systems is tricky. Digital systems can under normal circumstances be ceaselessly duplicated. So when we "kill" a process in the UNIX operating system, we are only deactivating one of a nearly infinite supply of clones. There is little consequence to this action within the machine since a new process can be quickly activated. However, if we destroy the only remaining copy of some digital system or software then something more serious has happened. The system cannot be easily re-created. Taking this idea further, one might argue that in order for a digital system to truly experience death, it would have take the form of the destruction of a unique non-reproducible system.

Artificial death is not only limited robotic and software systems. Consider for instance a recent variety of self-erasing paper invented at Xerox that clears itself of printed information in "16-24 hours" (Ramplel, 2008). Genetically engineered systems can also self-destruct. Recent work on "fail-safe mechanisms to terminate (gene) therapy" consists of biological suicide switches that can be activated (MacCorkle et al., 1998).

There are fictitious accounts of machine suicide and sacrifice; HAL 9000 appearing in Clarke's Space Odyssey series comes immediately to mind. In 2010: odyssey two the computer sacrifices itself to save crew members. Galatea 2.2 is a retelling of the Pygmalion myth in which the narrator teaches an artificial intelligence that increasingly becomes overwhelmed and decides to shut down (Powers, 2004).

What each of the artificial systems we have thus far described does have in common in some element of information that is destroyed. Floridi has previously argued that information objects have "intrinsic value" (Floridi, 2002). We will argue that there is value in the ability of information to destroy itself in varieties of artificial self-sacrifice.

A SHREDDER IN THE LIBRARY OF BABEL

Let us consider a universe in which information is never destroyed. *Prima facie* this appears an advantageous state of affairs. Any piece of information that comes into existence in this universe is guaranteed to always exist. A Library of Alexandria would not be destroyable and its texts would remain extant. Putting speculation aside we know that both the Library of Alexandria's texts but also an accurate account of their destruction have receded into history. A present-day observer quipped: "It's inherently difficult to get reliable information about an event that consisted of the destruction of all recorded information" (Stephenson, 1996).

Clearly some would like to look in the lost texts of the Library of Alexandria and still others would prefer to know precisely how the books came to be destroyed. But in a universe free of information destruction both could not exist. The Library of Alexandria would remain unburned but the cost would be the loss of all non-fictitious information about destroyed information.

"The Library of Babel" hints at a world where information is lossless (Borges, 1999). If all possible texts exist in Borges's library then it would include purely random gibberish, dead languages, and all possible future texts. We can similarly imagine the collection of all finite sequences that under some encoding would then produce every finite text that has or ever will exist.

One problem with the Library of Babel or our collection of all finite sequences is how one quickly finds anything useful. How do I separate out gibberish and languages unknown to me from information that is pertinent? Searching these universes is hard.

For information to be meaningful it must be observable or readable by at least one entity. A permanently inaccessible set of information is as useful as no information at all. Alternatively, a universe in which no information is destroyed (no matter how useless to the observing entities) becomes rapidly cluttered.

Consider the following dilemma: you are the only reader in a finite variety of Borges's Library of Babel. You carry a shredder formed into a backpack. After examining a book you have the choice to replace it back where you found it or to shred it. If you shred the books you believe to be gibberish then you can more rapidly find texts that you can read. Of course you risk shredding a book you would only understand the value of later after some Flowers for Algernon type transformation or just some hard time spent learning to read a language like Mandarin Chinese. What would you make of a text composed of random numbers arduously composed to ensure randomness? Such a text exists (RAND, 2002) and is even economically valuable enough that statisticians and engineers will pay actual money to obtain a copy.

LIMITED RESOURCES, KIN ALTURISM AND SMALLPOX

From these contrivances we can assert that destruction only becomes an issue in the presence of limited resources. In the above cases, the limited resource is the speed with which the reader can sequentially access texts in the library. If the reader could access in parallel simultaneously all of the texts in the library, then shredding books becomes a vapid activity. The shredder dilemma also becomes interesting when there is limited shelf space in the library or storage in our information system.

Another situation in which resources are limited and fiercely contested is the biological world. Food, sexual partners, sunlight, soil, and water are among the (re)sources of conflict both within and between species. Let us think analogically: a biological system can be equated with information. This might be done by simply transcribing the genome of a particular creature to a set of symbols. Loss of both the symbols and the original creature is akin to information death. We see in the biological world a bewildering array of mechanisms for culling organisms. Interestingly there are many defined varieties of biological death (Schneider and Matakas, 1971):

- necrobiosis: individual cell death (but not necessarily of hosting multi-cellular organism)
- necrosis: death of a group of cells such as an organ
- apotosis: programmed death of a cell within an organism
- brain death: total necrosis of the central nervous system
- extinction: death of a species

In the case of apotosis: "most, if not all animal cells have the ability to self-destruct by activation of an intrinsic cell suicide program when they are no longer needed or have become seriously damaged" (Steller, 1995). Death is pervasive in biological systems both within organisms and between organisms.

One edge case is that of *Turritopsis nutricula*, a species of jellyfish that can reverse the life-cycle typical of jellyfish. If not killed by a predator, the jellyfish can repeat the process of returning to a polyp and again becoming sexually mature indefinitely. This leads to one of the few examples in the biological world of a degree of immortality (Piraino, 1996). However, the topic of this article is self-sacrifice so we will now turn our attention to biological self-sacrifice.

One variety of self-sacrifice is kin altruism in which an individual performs actions or undertakes a strategy to their own detriment but to the benefit of genetic relatives. Kin selection is exemplified in sterile ants, bees, wasps, and termites that spend their lives protecting and feeding related offspring (Griffin and West, 2002).

One spectacle of kin selection is the forming of rafts, plugs, ovens, walls, and bridges by ants and beetles. Such assemblies may have a variety of immediate uses for the reproducing members of the kin groups for instance: "mantle shields against rain," "thermoregulatory clusters," "swarms (against desiccation)," or "rafts (against flooding)." In some of these assemblies the organisms give up their lives to benefit the survival of the group (Anderson et al., 2002).

Within evolutionary biology Hamilton's rule is used to model cases in which kin altruism will cause genes to propagate in the genome of a species.

> When J. B. S. Haldane remarked, "I will jump into the river to save two brothers or eight cousins," he anticipated what became later known as Hamilton's rule...This ingenious idea is that natural selection can favor cooperation if the donor and the recipient of an altruistic act are genetic relatives (Nowak, 2006).

Hamilton's rule is stated as follows:

$$r > \frac{c}{b}$$

Here *r* is the probability of relatedness, *c* is the cost (reproductively) of performing an altruistic action, and b is the benefit (reproductively) to the recipient of the action. If relatedness outweighs the ratio of cost to benefit then according to Hamilton's rule an action should be performed and potentially inherited by offspring. Would it be worthwhile if such a rule existed for information entities?

Another interesting biological example related to the intrinsic value of information is the debate surrounding the destruction of smallpox stocks (see both Jolick, et al., 1993 and Mahy et al., 1993). One group argues that the smallpox virus stocks should be saved for research purposes (or as information for future generations) while a second group argues that the potential for weaponization or accidents outweighs the benefit of preserving the biological specimens. We speak in the next section to provide a frame for arguments concerning destruction of information and its benefit.

INTRINSIC VALUE AND SOCIETAL VALUE

Floridi argues for the moral worth of information objects (2002). We are taking the position that in some cases the value of destroying information outweighs its intrinsic moral value. Indeed, Floridi later hints at this possibility with the following quotation:

> Nobody would ever argue that this is equivalent to saying that a spider's and a human life are equally worthy of respect. Culling, for example, is an ethical duty in environmental ethics (Floridi, 2008).

Extending from arguments in the section above it is easy to see the biological systems benefit from self-sacrifice. The soldier ant queen is able to continue the colony because she rides upon drowning workers during the flood.

Himma (2004) has previously critiqued "the moral value of things *qua* information objects." We will make a different claim that some types of information systems do not function without constant information destruction. A slight extension of this is the claim that it is worthwhile for information to destroy itself. Responding to the needs of a society of information entities or biological entities may in some cases enjoin a duty to destroy.

Imagine the following predicament: the grey goo meme. This is a piece of information that duplicates exponentially. An information system into which grey goo is input begins to broadcast the grey goo to other information systems. Such a piece of information would behave as a computer worm does.

One way in which packet switching networks, such as Internet Protocol networks limit their vulnerability to such rouge information is assigning a *time to live* for information encapsulated. Another concept *packet filtering* seeks to discard data packets according to certain criteria (Chang, 2002). Internet routers kill packets of stale or unwanted information with a prodigious efficiency. How might we square intrinsic moral value of information at the same time with networking protocol designer's propensity to define ways of destroying information?

A counterpoint to consider is that even malignant worms and viruses that exist in information systems are not entirely worthless. A script kiddy may achieve social esteem by successfully attacking and co-opting personal computers. Indeed, Robert Tappan Morris, the author of the synonymous worm has even been appointed a professorship perhaps in part for his knowledge of such distributed systems. The creators of anti-social exploits may be financially rewarded by groups seeking botnets for nefarious activities. Security researchers catalog and intentionally entrap such pieces of information to write software to protect and harden networks and computing systems.

Our argument may be reduced to the following statement: it is good that information is capable of being destroyed. Let us revisit cases in which information is duplicated without check: computer worms, runaway cell growth, and the grey goo doomsday scenario (Pheonix and Drexler, 2004). Some types of information entities produce negative outcomes for the (eco)systems in which they reside.

We are not arguing that all instances of such information should be destroyed. Obviously cancer researchers benefit from examining cancer cells and a policy of constantly destroying the cells wholesale is ludicrous. We are taking the more moderate view that it would be better if information entities in such systems could knowingly self-sacrifice themselves when the surrounding society, environment, or situation warrants this course of action. While wholesale destruction of all variants of an information entity is not advocated, selective self-destruction of information entities to maintain the function of a society is advocated.

THE SIGNAL TO NOISE RATIO OF INFORMATION ENTITIES

We have examined some more and less absurd situations in which non-destruction of information has negative consequences. Conversely, we've examined some cases (often drawn from biology) in which self-sacrifice has a positive effect on the biological community.

As themes, artificial death and artificial self-sacrifice raise many interesting philosophical and design questions. For instance: should robots destroy themselves in certain circumstances? Also: should the robot have control over the choice to destroy itself? As information entities, they may have intrinsic moral worth, but they may also have duties. Is the act of negating an information entities own existence among these duties?

REFERENCES

C. Anderson, et al. (2002). `Self-assemblages in insect societies'. Insectes Sociaux 49(2):99-110.

E. Andel and C. Gütze (2005). 5VOLTCORE - [Shockbot] "Corejulio".

http://5voltcore.com/typolight/typolight257/index.php?id=1&articles=1

J. L. Borges (1999). Collected Fictions. Penguin.

R. K. C. Chang (2002). `Defending against flooding-based distributed denial-of-service attacks: a tutorial'. Communications Magazine, IEEE **40**(10):42-51.

P. Chow (1986). `MIPS-X instruction set and programmers manual'. Technical Report, Stanford, CA, USA. ftp://reports.stanford.edu/pub/cstr/reports/csl/tr/86/289/CSL-TR-86-289.pdf

A. C. Clarke (1984). 2010: Odyssey Two. Del Rey, first edition.

L. Floridi (2002). `On the intrinsic value of information objects and the infosphere'. Ethics and Information Technology **4**(4):287-304.

L. Floridi (2008). `Information ethics: a reappraisal'. Ethics and Information Technology **10(**2&3):189-204.

A. S. Griffin & S. S. West (2002). `Kin selection: fact and fiction'. Trends in Ecology & Evolution **17**(1):15-21.

K. Himma (2004). `There's something about Mary: The moral value of things qua information objects'. Ethics and Information Technology **6**(3):145-159.

W. K. Joklik, et al. (1993). `Why the smallpox virus stocks should not be destroyed'. Science (New York, N.Y.) **262**(5137):1225-1226.

C. G. Langton (1986). `Studying artificial life with cellular automata'. Phys. D **2**(1-3):120-149.

R. A. MacCorkle, et al. (1998). `Synthetic activation of caspases: artificial death switches.' Proceedings of the National Academy of Sciences of the United States of America **95**(7):3655-3660.

B. W. Mahy, et al. (1993). `The remaining stocks of smallpox virus should be destroyed'. Science (New York, N.Y.) **262**(5137):1223-1224.

M. A. Nowak (2006). `Five rules for the evolution of cooperation.'. Science **314**(5805):1560-1563.

C. Phoenix & E. Drexler (2004). `Safe exponential manufacturing'. Nanotechnology **15**(8):869-872.

S. Piraino, et al. (1996). `Reversing the Life Cycle: Medusae Transforming into Polyps and Cell Transdifferentiation in Turritopsis nutricula (Cnidaria, Hydrozoa)'. Biological Bulletin **190**(3):302-312.

R. Powers (2004). Galatea 2.2: A Novel. Picador.

C. Ramplell (2008). ‘Xerox Creates Self-Erasing Paper’. The Chronicle for Higher Education Wired Campus Newsletter. http://chronicle.com/wiredcampus/article/2956/xerox-creates-self-erasing-paper

RAND Corporation (2002) A Million Random Digits with 100,000 Normal Deviates. American Book Publishers.

H. Schneider & F. Matakas (1971). `Pathological changes of the spinal cord after brain death'. Acta Neuropathologica **18**(3):234-247.

H. Steller (1995). `Mechanisms and genes of cellular suicide'. Science **267**(5203):1445-1449.

N. Stephenson (1996) 'Mother Earth Mother Board’. Wired **4**(12). http://www.wired.com/wired/archive/4.12/ffglass.html

J. Sullins (2005). `Ethics and Artificial life: From Modeling to Moral Agents'. Ethics and Information Technology **7**(3):139-148.

BIOGRAPHIES

Carson Reynolds is a project assistant professor in the Department of Creative Informatics of the University of Tokyo. He holds a Doctor of Philosophy and Master of Science from the Massachusetts Institute of Technology upon recommendation by the Program in Media Arts and Sciences in the School of Architecture and Planning. His research work there was performed at the Media Laboratory in the Affective Computing Group. Carson also holds a Bachelor of Science in Technical Communication with a Minor in Philosophy from the University of Washington at Seattle.

Alvaro Cassinelli was born in Montevideo (Uruguay) in 1972. In 1990 he obtained both French and Uruguayan Bachelor degree, and a grant to pursue his studies in France. In 1996 he obtained a Graduate Engineering diploma from the Ecole Nationale Superieure des Telecommunications (ENST), in Paris. He completed the same year a Doctoral Qualifying Degree (DEA) in physics (laser and matter interaction) from the University of Paris-XI/ENST/Ecole Polytechnique. In 2000 he received a Ph.D degree from the University of Paris-XI Orsay for his work on optoelectronic stochastic parallel processors for image processing. Since 2001 he has been working as a Research Fellow, Research Assistant and since 2006 as Assistant Professor at the Ishikawa-Komuro Laboratory, where he is actively involved in creation and development of the new Meta-Perception Group. He has been awarded several prizes as a Media Artist, including the Grand Prize [Art Division] at the 9th Japan Media Art Festival and an Honorary Mention at Ars Electronica 2006.