

Ethical Issues In A-Life: Cyber Gods As Moral Monsters?

Inari Thiel is a Research Officer in the School of Information Technology and Electrical Engineering, University of Queensland.

Neil Bergmann is a Professor of Embedded Systems in the School of Information Technology and Electrical Engineering, University of Queensland.

William Grey is an Associate Professor in the Department of Philosophy, University of Queensland.

Abstract

The authors have undertaken an exploration of some significant social and ethical issues that arise in relation to the emerging field of Artificial Life (A-life). These issues have been approached from a philosophical perspective, taking into account reports of current developments in A-life research, and the application of A-life software in elementary school education. It has been suggested that the use of such systems may impact on the development of moral character in children, and illuminate that of adults. In addition, it is argued that if A-life researchers achieve their aims and evolve digital biota that are both intelligent and autonomous, they may be responsible to their creations for the quality of the worlds in which they live. The authors conclude that, given the stated aims and current progress of A-life researchers, there is a clear need for further consideration of the potential social and ethical implications of these technologies.

Keywords: Artificial life, ethics, moral considerability

Introduction

Artificial Life, or A-Life, is a new research area which includes the creation of novel life forms that inhabit synthetic environments implemented as computer programs. Some A-Life researchers use computers to model high-level characteristics of life (such as ants cooperating on a task, or birds travelling in a flock), in an attempt to understand these characteristics. The more ambitious branch of A-Life seeks to create synthetic environments, with their own artificial chemistry and physics, and to seed these with self-replicating structures. The aim is then to develop artificial life forms, or digital biota, that 'live' and operate in the synthetic environment. These biota should then evolve over time with increasingly complex structure and behaviour. There is a clear, long-term target of creating intelligent artificial life (Grand 2001; Langton 1991). This intelligent A-Life may exist only in the synthetic world, observed from the real world. Alternatively the digital biota may have sensors and actuators that allow them to interact with the real world, effectively implementing intelligent, self-aware robots.

This paper offers a preliminary investigation of some of the social and ethical implications of the development of such projects, drawing on the writings of both philosophers and A-life researchers, as well as a wider tradition of parables and stories in which these developments have been imaginatively foreshadowed.

Western cultures have long been fascinated with, and alarmed about, the possibility of creating synthetic life. It provides a recurrent theme in our mythology, from the classical Greek stories of Daedalus (whose statues reputedly had to be tied down to prevent their running away) and Hephaestus (who created a range of robots and cyborg assistants) to the sixteenth-century Prague legend of the Golem. Mary Shelley's nineteenth-century allegory of Frankenstein, and Karel Capek's early twentieth century story about intelligent and affective robots, continue and develop this ambivalent fascination which has always provided as much warning as encouragement (Capek 1923; Shelley [1818] 1983).

The fictional Golem, for example, was a giant humanoid fashioned from clay and animated by a fragment of kabbalistic text, created to serve a human community; but the creature escapes from its designer's control and runs amok until he manages to disable it (Meyrink 1928). The story presents a tension between the hope of assistance for humanity through the enhancement of human capabilities by a semi-autonomous artefact, and the fear that the creators of such artefacts might unleash unforeseen catastrophe.

Similarly, Frankenstein's creature, a composite of human body-parts animated by electrical energy, is set loose, unsocialised, and eventually inflicts revenge on its creator by destroying those he loves most (Shelley [1818] 1983). Again, there are themes of optimism about the human capacity to create new life forms, and pessimism about our lack of foresight in doing so. More than the Golem fable, Mary Shelley's story raises the question of what responsibilities the creator of such an artefact might have towards the creature itself, as well as to those likely to be affected by its actions.

Issues Being Raised Within The A-Life Research Community

Although philosophers have taken up some of the questions in epistemology and metaphysics raised by Artificial Intelligence (AI) and A-life, there has yet been little or no discussion of the potential ethical implications of either field. However, some researchers within the community developing AI and A-life systems have begun raising issues of moral concern.

For example, Hugo de Garis asks: 'Who or what is to be dominant species on this planet – human beings or artefacts (artificial intellects)?' and suggests that moral philosophers need to consider the problems of whether artefacts should be created at all and whether, if they are created, they should be permitted to evolve freely (de Garis 1990: 131-8). Steve Grand reflects on the relationship between an A-life creator and the creatures created and destroyed in the cyber-realm (Grand no date). Peta Wyeth and Helen Purchase have used A-life simulation software in developing high-order thinking skills among primary school students (7-8 yr olds). They don't discuss ethics, but report comments such as 'I like seeing ZOIDS die' and 'I liked the SPOTS getting eaten' as evidence of students' enjoyment and engagement (Wyeth and Purchase 2001: 239).

Through these issues, there is a common thread of questions about the moral responsibility of A-life creators or controllers for the beings they create. De Garis's concern is with the responsibility of technologists to the wider society. He foresees that, should A-life researchers succeed in their quest, synthetic intelligences, unless deliberately impaired by their creators, are likely to evolve rapidly to 'a state of sophistication beyond human comprehension' and escape all human control, attaining a position of intellectual dominance from which they may regard human beings with no more respect than we commonly accord to insects or animals we consider our inferiors.

The question we face, then, is: Should we attempt to place a moratorium on this kind of research, at least until we are certain about our safety as a species? One response, acknowledged by de Garis, is that such a proposal would be practically unachievable, as we have seen recently in the matter of human cloning. Whether cloned human beings have actually been produced is less relevant in this context than the fact that the attempt to persuade a world-wide scientific community to effect a self-imposed moratorium on such research has clearly failed. Another kind of response is that of Pamela McCorduck, who reflected on similar fears in terms of the common human propensity to project onto others our own least desirable traits. She writes: 'Will an intelligent artefact grasp for power over us? We're into the Coyote problem here, ready to assume that the worst about ourselves is also true of others. AI researchers dismiss as perfervid science fiction the notion that a will to power is necessarily concomitant with intelligence; this just happens to be so in the human species. AI can conceive of intelligences where power simply isn't at issue.' (McCorduck 1979: 327)

McCorduck raises an interesting question. One could suggest that, since twentieth-century AI was being created largely by analogy with human intelligence, it might reasonably be expected to reproduce the image and likeness of human intellects, possibly even specifically masculine intellects (Adam 2002). A-life, however, is less constructed than evolved, and might be more open to development along a significantly different trajectory; unless, of course, the optimal evolutionary path leads to just our kind of intellect.

While de Garis is concerned for how artefacts might impact on human beings, Steve Grand's reflection, though somewhat facetious in tone, takes up the issue of whether 'cyber-gods' are morally free to exercise unfettered power over their creations. The basic question is not new, having been debated extensively in Judeo-Christian theology and philosophy, from the biblical story of Job to more contemporary academic writing (Felt 1984). This debate has been conducted from the perspective of the creature attempting to understand the reasoning of a creator who is generally not a participant in the discussion; Grand, however, finds himself in the position of an effectively omnipotent power seeking to justify his provision of a less than perfect virtual world for his virtual creatures. Here the common resort to inscrutable

wisdom is not available, and Grand's closing position — 'if I can't have immortality, nor can they' — seems petty.

The use of A-life as a teaching aid in classroom projects raises yet another set of issues, this time relating to the area of character development in children. Although hardly equivalent to the Milgram experiments, these activities could, perhaps, foster moral insensitivity when conducted in the absence of any mitigating guidance or prior discussion by teachers or other classroom facilitators. According to Wyeth and Purchase, it is precisely because the children bring to their engagement with the CULTURE software their real world experience of survival and death among plants and animals that they are able to manipulate the simulation effectively. Their research did not investigate how the students conceived of the cyber-creatures they created and destroyed, or how the attitudes they expressed in this context correlated with their attitudes towards real-world life forms. Clearly, there is scope for further exploration of these issues, possibly in collaboration with researchers in education and developmental psychology.

Freedom And Responsibility, Real And Virtual

Our consideration of these issues does not develop in a vacuum; we are well practised in dealing with similar issues in other contexts. For example, our culture, while valuing scientific enquiry and technological development, already places constraints on what kinds of things researchers ought to create, using international treaties or conventions to restrict the generation of viruses or harmful organisms; and the international research community recently attempted to impose a voluntary moratorium on human cloning. This is not a decision taken lightly, as freedom of enquiry and priority of discovery are highly valued by researchers; and it is not always a successful strategy, but nor are our moral constraints on murder, lying, and the like.

The reasons advanced by the International Committee of the Red Cross (ICRC) in 1972 for proscribing the production of 'microbial or other biological agents, or toxins [...] of types and in quantities that have no justification for prophylactic, protective or other peaceful purposes' are grounded in a desire to minimise risk to humans, as are de Garis's concerns about the unfettered development of 'ultraintelligences'. A potentially significant difference may lie in the fact that a synthetic intelligence could offer considerable benefits to humankind along with the risks that de Garis identifies. However, we quite commonly weigh that kind of problem when we consider issues like the use of atomic energy or such individual choices as whether to drive a car.

We also have some experience in determining the responsibilities of designers and fabricators for the products of their creative activities. Hammurabi's code of direct responsibility for the design and construction of buildings has been refined and developed in modern jurisdictions, and there are similar clearly defined duties and obligations placed on those who build and operate such facilities as factories, water supplies or zoos. When there is a malfunction that results in damage to persons or the environment, the cause is investigated and responsibility is apportioned accordingly. There are some situations in which such investigations are difficult to resolve, when the systems involved are particularly complex and both the design and the continued operation are carried out by groups that are large in number and fluid in composition, but the underlying principle of human responsibility for human artefacts remains.

However, the development of synthetic intellects may challenge this principle. Although we would usually hold a person responsible for injury or damage caused by an animal they had

brought into existence, we do not generally consider parents liable for the wrongs committed by their adult offspring. To the extent that the offending product of human creative endeavour is an autonomous moral agent, the creator is absolved from liability for its actions. So if, as de Garis suggests, researchers are on the way to creating 'superintelligences', and if those intellects have a capacity to effect action in the real world, it will be necessary to determine who is responsible for the results of such action. Both Nick Bostrom and Eliezer Yudkowsky recommend that synthetic intellects should be constructed in such a way as to ensure that their highest value is 'friendliness' towards humankind, in order to minimise the risk of harm that might attend their action (Bostrom 2003; Yudkowsky 2001).

One area in which we have little prior experience is in determining the responsibilities of a creator to the product/s of creation. In the case of A-life, we encounter 'quality of life' as a potential design issue, an idea that has some resonance with earlier theological debates on whether a divine creator is obliged to create the best of all possible worlds (Felt 1984). However, there are also analogies in philosophical discussions around whether parents may have a duty to terminate a pregnancy that will result in the birth of a defective baby. David Benatar goes further, arguing that, because even healthy persons suffer hardships of various kinds, and because being deprived — by not coming into existence — of the opportunity to experience pleasure is not bad for anyone, it is 'morally desirable' that people refrain from bringing any children into existence (Benatar 1997: 353). This is, admittedly, an extreme view in relation to humans, but it is effectively the argument put by Michael laChat in relation to synthetic intellects when he writes: '... the ratio of risk to benefit is higher in AI than in human reproduction. [...] An AI experiment that aims at producing a self-reflexively conscious and communicative "person" is prima facie immoral.' (laChat in Dejoie 1991 et al :287)

The type of world in which A-life develops and exists is just one of the factors a 'cyber-god' will need to consider, as Steve Grand acknowledges in his reflections on whether he should have made his cyber-creatures immortal, given that he could have done so (Grand no date). He concludes that he is beyond such obligation: '...that's a god's privilege, never to be accused of making a mistake. Whatever he or she does is "right", more or less by definition.' This may be a debatable point; however, our treatment of these creatures during their lives is subject to moral evaluation. If, instead of simply setting up a system and allowing it to develop, we choose to intervene and our interventions prove deleterious, we will place ourselves under judgement by our own standards of justice and care. In this respect we are not, as Grand suggests, completely unfettered creators arbitrarily defining right and wrong; instead we are members of a moral community, or set of moral communities, and subject to the judgement of our peers.

Social Implications Of This Technology

As the discussion above reveals, the ongoing development of A-life provides a new context in which some long-standing issues in ethics and social responsibility can be explored and extended, and some of our existing principles of moral judgement can and should be applied. The novelty of having available to us a set of sophisticated simulations or replications of living and evolving worlds also facilitates the development or illumination of moral character. To date, there seem to be signs that ultimate power does, indeed, tend to corrupt those who wield it (Grand no date; Wyeth and Purchase 2001). These observations serve to highlight a need for the development of widely-agreed research/application guidelines similar to those that now apply to biotechnologies. Equally, since children whose moral characters are not yet fully formed are being introduced to the exercise of effectively unlimited power over

virtual creatures in virtual worlds, there is clearly a need for the incorporation of ethics awareness education into classroom use. If such safeguards are neglected, we may find that our cyber-gods are formed as moral monsters.

ⁱ For example, *Metaphilosophy* 33(1/2), January 2002, was a special issue devoted to “cyberphilosophy”.

ⁱⁱ See Eliezer Yudkowsky’s “AI Box” experiment <http://sysopmind.com/essays/aibox.html> (2002) for a purported illustration of this possibility.

References

- Adam, A. 2002, 'Gender/body/machine.' *Ratio (new series)* Vol 14, No 4, pp. 354-375.
- Benatar, D. 1997, 'Why it is better never to come into existence.' *American Philosophical Quarterly*, Vol 34, No 3, pp. 345-355
- Bostrom, N. 2003, 'Ethical issues in advanced Artificial Intelligence.' (unpublished paper).
- Capek, K. 1923, *R.U.R. (Rossum's Universal Robots): a play in three acts and an epilogue*. London, Oxford University Press.
- De Joie, R., Fowler, G. & Paradice, D. eds 1991, *Ethical Issues in Information Systems*, Boston, Boyd & Fraser. pp 278 - 294
- de Garis, H. 1990, 'Moral dilemmas concerning the ultra intelligent machine.' *Revue Internationale de Philosophie*, Vol 44, No 172, pp.131-138.
- Felt, J. W. 1984, 'God's choice: reflections on evil in a created world.' *Faith and Philosophy*, Vol 1, No 4, pp. 370-377.
- Grand, S. 2001, *Creation: life and how to make it*. Cambridge, Mass., Harvard University Press.
- Grand, S. No date, *Confessions of a cyber-god*. Accessed 31 July 2003, <http://www.cyberlife-research.com/articles/confessions.htm>.
- Langton, C. G. 1991, Preface. *Artificial life II : the proceedings of an interdisciplinary workshop on the synthesis and simulation of living systems held 1990 in Los Alamos, New Mexico*. C. G. Langton, C. Taylor, J. D. Farmer & S. Rasmussen. Redwood City, Calif., Addison-Wesley: xiii-xviii.
- McCorduck, P. 1979, *Machines Who Think*. San Francisco, W H Freeman and Company.
- Meyrink, G. 1928, *The Golem*. London, Gollancz.
- Shelley, M. [1818] 1983, *Frankenstein : or, the modern Prometheus*. New York, N.Y., Signet.
- Wyeth, P. & H. C. Purchase 2001, 'Exploring the learning potential of an artificial life simulation.' *International Journal of Continuing Engineering Education and Life-Long Learning*, Vol 11, No 3, pp 229-241.
- Yudkowsky, E. 2001, *Creating friendly AI 1.0: the analysis and design of benevolent goal architectures*. Accessed 31 July 2003, <http://www.singinst.org/CFAI/index.html>.