A Case for Investigating the Ethics of Artificial Life

Inari Thiel  
School of ITEE, University of Queensland  
Phone: +61-7-33651652  
inari@itee.uq.edu.au  
Neil Bergmann  
School of ITEE, University of Queensland  
Phone: +61-7-33651182  
n.bergmann@itee.uq.edu.au  
William Grey  
Department of Philosophy, University of Queensland  
Phone: +61-7-33652620  
w.grey@mailbox.uq.edu.au

A major stream of Artificial Life (ALife) research aims to build synthetic life forms, operating in virtual worlds, implemented as computer programs. A clear long-term target for this research is the evolution of digital life-forms with a complexity of structure and behaviour analogous to biological life-forms, potentially exhibiting intelligence and self-awareness. The creation of intelligent, self-aware digital life-forms has clear ethical implications, but there is no current research into how these ethical issues might be addressed. This paper argues that such ethical research is needed. Furthermore, it describes our future research plans to build a solid philosophical foundation for the consideration of these ethical issues.

Keywords:  
Applied Ethics, Artificial Life, Complex Systems, Moral Philosophy
1 Introduction and Motivation

The potential for humans to create artificial life forms is a theme that emerges repeatedly in Western mythology, from the Greek god Hephaestus’ “golden handmaidens” [1] to the sixteenth-century Prague legend of the Golem [2]. Mary Shelley’s nineteenth-century allegory of Frankenstein, and Karel Capek’s early twentieth century play in which he coined the term “robots” [3], continue and develop this fascination, providing both warning and encouragement.

1.1 Synthetic and Real

Recent developments in complex computing systems supporting powerful mathematical modelling techniques, and increasing understanding of the ways in which natural biological systems develop and operate, have combined to make the creative ambition more achievable, though possibly more problematic, given its instantiation in cyber-reality rather than in tangible form.

While ALife research has its roots in the modeling of real world environments, many researchers have now moved beyond this purely instrumental approach and set themselves the goal of creating fully functional worlds populated with independently evolving digital creatures, which their creators can either manipulate directly or leave to develop undirected. For example, Steve Grand, lead programmer of Creatures software, writes:

By combining simple cybernetic building blocks … we can make something that is not only alive in the technical sense but also alive in the richer, more rounded sense too. These creatures are not very smart, but they do have individual little personalities. They live out their lives and behave in ways that I, as their creator, didn’t program them to and sometimes didn’t even expect. [4]

Although Creatures is designed as a game, and its communities remain confined to cyberspace, Grand’s current interest is in bringing “cyberlife” into real world applications — the ancient fables may be approaching a twenty-first century realisation.

1.2 Prudential Foresight

The prospect of creating completely synthetic intelligent life-forms raises ethical issues in itself, but the possibility of interaction between the cyber-realm and the external physical world raises additional normative and conceptual questions.

As has happened in other areas of innovation (such as genetic engineering), rapid technological progress has outstripped the ethical resources which are required for rational deliberation about the new range of technologically-generated choices. In the field of artificial life, we may have opportunity to consider these issues before the more ambitious goals of the project have been achieved; but the timeframe is uncertain, and it would be prudent to begin the process of rational deliberation now.
Some practitioners in the Artificial Intelligence (AI) and ALife fields have raised questions about whether the goals of these projects ought to be pursued at all [5, 6]; and some philosophers have discussed the moral status of naturally occurring non-sentient life forms, in the context of environmental ethics [7-9]. However, while the extension of moral discourse to a wider domain than has been commonly allowed is a continuing theme in environmental philosophy [10], there has been little philosophical discussion of whether either synthetic intellects or digital biota should be admitted to the moral community, or what the implications of such admission might be (but see [11]).

In their paper, “Open problems in artificial life”, Mark Bedau et al. have presented a range of questions they describe as “a clear and fruitful challenge” designed to stimulate the ALife research community [12]. The final challenge in the third set of issues — grouped under the question “How is life related to mind, machines, and culture?” — is “establish ethical principles for artificial life”.

Here, Bedau et al. identify four areas of ethical concern, three of which — “(a) the sanctity of the biosphere, (b) the sanctity of human life, [and] (d) the risks of exploitation of artificial life” — relate to the potential effects of ALife on the world beyond its cyber-reality. It is the other area, “(c) the responsible treatment of newly generated life forms”, that we take up here, with a view to clarifying some of the ethical issues that are relevant.

2. ALife: Current Research and Trajectories of Development

While some writers [6, 13] have named John von Neumann as the progenitor of ALife, Mark Bedau et al. [12] credit Christopher Langton with coining the phrase “Artificial Life”. Langton has characterised Artificial Life as

… a field of study devoted to understanding life by attempting to abstract the fundamental dynamical principles underlying biological phenomena, and recreating these dynamics in other physical media — such as computers — making them accessible to new kinds of experimental manipulation and testing. [14]

The published Proceedings of successive Artificial Life and ECAL conferences have shown a range of ALife research areas, including the origins of life, evolutionary dynamics, learning, and some of the philosophical issues relating to functionalism and emergence. Many of the researchers have seen their work as contributing to an understanding of how life may have actually arisen and developed.

However, in the work cited above, Langton quickly moves from his characterisation of ALife as modeling and simulation of “life-as-we-know-it” to a Prometheus vision of broader possibilities:

… Artificial Life is not only about studying existing life, but also about the possibility of synthesizing new life, within computers or other “artificial” media. The life that is realized in these alternative media will force us to broaden our understanding of the proper domain of biology to include self-organizing, evolving, and even “living” machines, regardless of the specific physical stuff of which they are constituted, or whether or not they are based
on the same chemical and physical principles as the life that has evolved here on Earth. [15]

This development, hailed by some as potentially the next significant advance in evolution [6], opens up a range of empirical and philosophical problems not previously considered.

2.1 Philosophy and ALife

In the mid to late twentieth century, digital computing in general, and AI research in particular, attracted the interest of analytic philosophers working in the areas of Epistemology and Philosophy of Mind, providing an arena for extending already long-running discussions of such issues as intentionality, consciousness, and the “mind-body problem”. [15, 16]

The ALife field has revitalised debates about the definition of “life” or “living organism”, and the distinction between “simulation” and “realisation” of complex structures [17], while providing a tool for empirically testing the evolutionary effectiveness of certain theories [18].

The question of how to discern the emergence of real life from a simulation of it is one of the most contested in this field. ALife researchers like Langton tend to adopt an optimistic tone, confident that a sufficiently accurate transcription of the formal principles of biological organisation into a synthetic domain will generate truly living organisms, even if they are not embodied in forms currently recognised in our experience. He writes, for example:

"The life that is realized in these alternative media will force us to broaden our understanding of the proper domain of biology to include self-organizing, evolving, and even “living” machines, regardless of the specific physical stuff of which they are constituted, or whether or not they are based on the same chemical and physical principles as the life that has evolved here on Earth." [14]

Others are more sceptical. Theoretical biologist, Claus Emmeche, while acknowledging that non-carbon based life forms are conceivable, strongly denies that any computer simulation could cross over into the domain of actual life [19]. His principal objection is that life requires some material basis, whereas ALife consists of merely formal processes. Another is that all ALife simulations rely on computational models of biological systems, models that are both limited and fallible, being constructed on the basis of human interpretations shaped by theories about the nature of biological life. There are echoes here of Joseph Weizenbaum’s critique of AI researchers’ enthusiasm, over a generation ago [20].

3. Ethics and Ethical Theories

Ethics is the branch of philosophy concerned with the evaluation of conduct and the criteria for moral assessment of actions as permissible or impermissible; right or wrong. Within this field, the question of what makes something an appropriate object
of moral concern, that is, the criteria for moral considerability, and the requirements for moral agency, also arise.

Theories of normative ethics, which provide frameworks for ethical decision-making and moral evaluation, fall into three broad groups: deontological, consequentialist, and virtue ethics.

3.1 Deontological Ethics

Deontological ethics posit a framework of rule-governed principles which prescribe certain duties which moral agents ought to perform and rights which must be respected, irrespective of the consequences. These rules typically include the duty to keep promises that are freely made, and the right to restrict access to one’s own body. These duties and rights prescribe action irrespective of the advantage gained or disadvantage suffered by the agent or by those affected by the action.

3.2 Consequentialist Ethics

Consequentialist (also called 'utilitarian' or 'teleological') ethics, in contrast, evaluates actions morally by reference to the outcomes they produce. A best action is the one which produces the greatest surplus of positive over negative outcomes. If breaking a promise yields a positive outcome on balance, then that is what an agent should do. [22]

3.3 Virtue Ethics

Virtue ethics is a theory in which the focus is on neither the action nor the outcome but on the character of the agent. A virtuous agent is one disposed to perform good actions; a vicious agent, bad ones; therefore, one ought to develop one’s character by cultivating virtuous habits (such as courage and self-control) and avoiding vicious habits [23, 24].

4. Ethical Issues in ALife

While these differing ethical approaches notoriously yield different evaluations in particular cases, all agree that there should be a framework of action-guiding considerations which inform the decisions and choices of moral agents. But who (or what) are moral agents? What are the requirements for moral agency? And to what sorts of entities are moral considerations owed?

We believe it is necessary to examine a number of significant questions raised by ALife developments, in light of each of these ethical theories. Some of the problems are common to all moral theories, though there may be relevant points of difference as indicated below.
4.1 Counting the Costs

As noted in section 1.2 above, there are those who would also raise the question of whether ALife research should be pursued at all; and an anonymous reviewer of this paper suggested that there is a pressing need for some consideration of how to weigh the costs and benefits of such research.

Mark Bedau [12, 25] has rightly noted that there are two aspects to this set of questions. One is grounded in concern for the potential effects of ALife on human well-being, and even the continued existence of the species homo sapiens. Do we have a moral obligation to refrain from pursuing lines of research that have the potential to produce our evolutionary successors and which may lead to our species’ extinction? Bedau further refines this concern in a later paper[12]. We have not taken up that branch of enquiry here, though it is touched on elsewhere [26].

The other reflects a concern for the wellbeing of the digital biota themselves. Ought we refrain from bringing into being entities capable of suffering, in worlds in which we can reasonably expect that some suffering will befall them? How might we determine the threshold at which the costs outweigh the benefits for such creatures in such worlds? This second set of issues is contingent on acknowledgement that ALife creatures are appropriate for granting the status of moral considerability.

4.2 The Question of Moral Status

A foundational issue which any ethical theory must address is the criterion (or criteria) for inclusion in the moral community as an object of moral concern. For example, we commonly set increasingly stringent constraints on the treatment of the mice used in drug testing, but show no concern for the well-being of the bacteria used in microbiology experiments. (Our inclusion of mice in the research community contrasts with their treatment as agricultural pests; clearly their membership of the agricultural production community is more precarious. However the social consistency of moral judgements is not part of our present concern.)

Clearly, this question has been answered differently in different times and cultures, and it may be useful to consider briefly a selection of these responses and the reasons that inform them.

Some communities have drawn a moral distinction between themselves and those of different tribes, kinship groups, or ethnicities; so, for example, some eighteenth and nineteenth century caucasians bought and sold other human beings, constraining their freedom of movement and even their reproductive pairings, treating them like agricultural livestock [27]. Here, the principal criterion for assessment of moral status has been perceived similarity of form and appearance between those being evaluated and those doing the evaluating. The perception of physical difference seems to have generated presumptions about potential for rationality and moral character, which also contributed to the assignment of different (usually lower) moral status.

Some cultures have provided for the exclusion of members who breach significant norms, branding them as “outlaws”, beyond the moral protection of the culture’s laws. There are occasional resurgences of the same impulse in our own culture, when people call for more stringent conditions in prisons, especially for those convicted of par-
ticularly heinous offences. This illustrates the fragility of moral status — even those clearly within the moral community can be evicted, the operative criterion here centering on fidelity to some form of social contract.

Throughout the history of Western thought, there has been a range of attitudes to the proper moral status of non-human animals, from the medieval practice of putting them on trial in courts of law (suggesting that they were accountable for their actions) to the early nineteenth-century treatment of animals as little more than biological machines (a view challenged by dissenters such as Jeremy Bentham) and the late twentieth-century “animal liberation” movement which called into question the exploitation of non-human species over a wide area.

Bentham recognised that to draw the bounds of the moral community precisely around the human community was to make an apparently arbitrary choice: even such attributes as a capacity for reason or conversation are not universal among humans, and may not be completely absent from the higher mammals. He proposed a criterion for inclusion in the following terms: “[T]he question is not, Can they reason? nor, Can they talk? but, Can they suffer?” Note that this famous criterion of sentience does not require self-conscious intelligence, though the development of digital creatures with such capacities is a long-term goal of some ALife researchers. One might reasonably object to the gratuitous physical torment of a cane toad, for example, without supposing that the toad has any self-awareness in the sense of understanding itself as a self.

More recently, some philosophers have argued that not only sentient beings, but also naturally occurring ecosystems might be included among the objects of moral concern, on the grounds that they have interests that can be furthered or harmed by the choices that [human] moral agents make [7]. This claim is currently rather more controversial than that for the inclusion of animals. [8, 28, 29]

4.2.1 Genesis and Worth
In most of the cases considered above, there is a common restriction of morally significant entities to naturally occurring living things, or networks in the case of ecosystems. Among the questions one might ask in this context are: Why is natural origin morally significant? What distinguishes living from non-living entities in morally significant ways?

Philosophers have, in general, been much more willing to argue for the moral significance of wombats, butterflies, or forests than to champion the interests of chess-playing computers, industrial robots or power stations. While wombats may share an apparent kinship which elicits sympathy (cf.[30]), it’s difficult to claim the same for the forest, or even a butterfly, so we would need to base a difference in moral standing on other grounds.

Andrew Brennan [29] argues for the moral standing of inanimate natural objects on the ground of “their lack of intrinsic function”, a feature they have in common with plants and animals, but which is not shared by human artefacts, which Brennan sees as having a function by virtue of their origin and placement in human symbolic systems. However, Brennan briefly considers one of Stanislaw Lem’s artificial world scenarios, and concedes that some artefacts may develop in ways that place them in the same realm of moral considerability as natural things.
The appeal of Lem’s cyber-persons is grounded principally in their representation as perfect simulations of human beings, with all the attendant human qualities, including capacities for physical and psychological suffering. As a character in one of these stories says, “…when the imitator is perfect, so must be the imitation, and the semblance becomes the truth, the pretense a reality!” [31]. So we are invited to put aside the question of origins and consider the matter from the cyber point of view, as it were, a point of view which is in morally relevant aspects indistinguishable from our own.

Though some may see this as a clever authorial sleight-of-hand, the issue may arise in the development of sophisticated simulations of living systems as experimental subjects in lieu of conducting potentially damaging experiments on “real” systems. At some point, when the virtual subject responds precisely as a real subject would, Lem’s challenge may well confront us.

While many of the arguments for the ethical treatment of non-human animals have been structured around their physical well-being [32], the question of ethical treatment for “digital biota” raises a new range of issues which must be evaluated on different grounds, such as their capacity for “psychological” or “mental” suffering. This is not an entirely novel idea, resonating as it does with medieval theologians’ consideration of the suffering of non-physical beings (evil angels) by virtue of the “saddening” of their wills [33].

4.3 “Did I request thee, Maker, from my Clay to mould me Man…?”

If, as Langton and others have indicated, a goal of ALife research is the development of a virtual realm of autonomous entities, and if digital biota or “animats” are granted membership of the moral community, there may be grounds for evaluating the ethical standard not only of researchers’ interactions with the realms they have created but even of the care they have taken in initially setting up those realms. The literature of western culture, from the biblical story of Job to Milton’s Paradise Lost and Mary Shelley’s Frankenstein, has repeatedly raised the issue of the responsibility of life-creators for the lives they create. In each of these stories, a creature challenges the creator who has placed it in a position of unnecessary suffering [34]. Conversely, Christian theology has developed a range of theodicies in an attempt to deflect these and similar challenges.

A common element of both the challenges and responses is a recognition that one who presumes to embark on the creation of worlds or the sentient beings that inhabit them has a moral obligation to avoid or at least minimise the suffering attendant on such creations. In the case of world-creators, traditional arguments flow around whether there is a responsibility to create the worlds inhabited by sentient beings in such a way that those beings are given the opportunity to avoid meaningless or unnecessary suffering; and responses tend to find justification in some sort of optimisation thesis: it is impossible to have a richly complex world, permitting the benefits of individual autonomy, in which there isn’t also at least the potential for attendant evils. “You have to take the bad with the good,” as it were. Artificial life research provides scope for testing some of these arguments empirically.
While these concerns overlap with ethical issues about the creation and modification of new life forms through genetic manipulation (see [35, 36]) they also pose novel ethical questions which need to be addressed. Although raising such issues may now seem speculative in the extreme, they are in keeping with the directions already established by current developments and goals in the field of ALife research. As Bedau et al. write:

Artificial life’s ethical issues somewhat resemble those concerning animal experimentation, genetic engineering, and artificial intelligence. The extensive literature on the ethical issues raised in those three fields may provide some guidance for exploring the ethical issues in artificial life. On the other hand, creating novel forms of life and interacting with them in novel ways will place us in increasingly uncharted ethical terrain. [12]

4.4 Implications for ALife Researchers

There are two importantly distinct categories of ethical issues which arise in connection with ALife. First there are ethical issues which arise in connection with any category of research: what are the impacts which such research will have (or may have) on the community? Examples of these issues are those (reasonably) familiar questions raised in relation to the use of expert systems; e.g., who should bear responsibility for harms caused by the use of such systems? However as Bedau et al. suggest, there are further questions that arise when the gap between the creator’s intention and the creation’s behaviour widens, and the created entity, albeit a synthetic artefact, acquires in very real sense a "life of its own". At this point we need to address the ethical status of ALife entities which are (or may be, or may become) in Kant's terms, "ends in themselves" and therefore "something whose existence has in itself an absolute value".

If it can be shown that there are good reasons for taking the well-being or interests of ALife creations seriously, then some forms of ALife research may need to be reviewed by the ethics committees of the institutions in which they are conducted, as now happens routinely with projects using human or animal subjects. In addition, there may be new issues that arise as we chart the emerging terrain in this area. For example, a current precept in animal research is "replacement"; that is, an injunction to use, where possible, non-animal models in preference to animals in biological experiments. However what if the models themselves were to cross a threshold of complexity and become morally considerable, albeit synthetic, entities?

It may be, however, that investigation will show that there are not yet any compelling reasons to admit ALife creations into the realm of moral subjects in their own right, in which case, ethicists may appropriately adopt a “watching brief” in conjunction with ALife researchers, periodically reviewing progress in the field to evaluate whether there have been any salient developments.
5. Summary and Conclusions

A clear long-term goal of ALife research is the development of highly complex, intelligent artificial life forms. Equally clearly there are many ethical questions that arise and should be addressed before such ALife forms become a reality. Many of these potential ethical questions have already been posed in popular science fiction. It is perhaps surprising that there has been very little solid academic research into the ethical issues involved in ALife research. Ethical research must be distinguished from the ruminations of popular science fiction writers by seeking to build an ethical framework based on solid philosophical foundations.

Ethical research in ALife should combine the skills of computer science and philosophy by building a team with experts from both. Computer Science expertise can be used to predict the likely capabilities and actions of digital biota, and to understand the limits and possibilities of the artificial worlds that such biota will inhabit. Philosophy and Applied Ethics expertise can be used to build a solid philosophical framework from which to view these actions, capabilities, limits and possibilities in terms of ethical considerations.

However the significance of this research is not just restricted to this pragmatic requirement. As is the case for much artificial life research, results obtained from research in artificial worlds often reflect back to give new insights and understandings about the nature of real life. Research into the philosophy of artificial life is no different. We expect that the deliberation about ethical implications of ALife will reflect back as new insights about ethical implications in a variety of current real-life situations. Indeed, we expect that this research will throw up many more research questions which span the disciplines of philosophy and complex systems, and we hope to use this project as the springboard for a subsequent broader program of research in this area.

5.1 Our Roadmap for Research

The above outline of the history and current goals of ALife research, together with a review of philosophers’ current engagement with that field, shows clearly that there is both scope and a need for further research. Our future aims in this area are:

• To sketch an appropriate set of ethical guidelines for current and future ALife research.
• To develop proposals for the threshold of moral considerability for artificial life.
• To explore the need for a possible set of ethical constraints which might circumscribe permissible developments in ALife research
• To further understanding of philosophical aspects of ALife.
• To promote cross-disciplinary discussion between the IT and Philosophy communities.

An initial survey of the field of ALife research and a critical review of relevant philosophical literature are currently being undertaken in order to catalogue more fully the potential ethical issues that arise in this area.
The next stage of our research will consist of an examination of philosophical frameworks relating to the threshold conditions for the classification of an entity as “living”, and further conditions for the inclusion of living entities within the community of moral concern. This investigation will take as its context current and plausible future developments in artificial life.

Another research direction will be to examine existing guidelines for research ethics review committees in the biological and behavioural sciences and catalogue the normative ethical theories and effective decision-making strategies embedded in them. In light of these findings, consideration can be given to their applicability in relation to ALife research.

Finally, results of the investigations in all previous stages will be drawn together to develop a position on the scope of the responsibility that ALife researchers may have towards their creations, and to propose guidelines for the ethical conduct of further research in this field.

This paper represents a snapshot of work in progress, in terms of the exploration of the ethical issues involved in ALife research. Its main contribution is to argue the case for why such research needs to be undertaken in conjunction with technical research in ALife. As such it provides more questions than answers, but it does describe at least one pathway for seeking to explore these ethical issues.

References

33. Aquinas, T.: Summa Theologica 64.4.