Is there a continuity between man and machine?

Johnny Hartz Søraker[1]

Abstract. The principle of formal equality, one of the most fundamental and undisputed principles in ethics, states that a difference in treatment or value between two kinds of entities can only be justified on the basis of a relevant and significant difference between the two. Accordingly, when it comes to the question of what kind of moral claim an intelligent or autonomous machine might have, one way to answer this is by way of comparison with humans: Is there a fundamental difference between humans and machines that justifies unequal treatment, or will the two become increasingly continuous, thus making it increasingly dubious whether unequal treatment is justified? This question is inherently imprecise, however, because it presupposes a stance on what it means for two types of entities to be sufficiently similar, as well as which types of properties that are relevant to compare. In this paper, I will sketch a formal characterization of what it means for two types of entities to be continuous in this sense, discuss what it implies for two different types of entities to be (dis-)continuous with regard to both ethics and science, and discuss a dramatic difference in how two previously discontinuous entities might become continuous.

1 INTRODUCTION[2]

The concept of 'continuity' has been championed by MIT historian Bruce Mazlish, who claims that progress in science and technology will inevitably result in a fourth continuity between man and machine [1]. According to Mazlish, there have been three dramatic scientific revolutions in the history of mankind, and these revolutions are best described as the establishment of continuities; mankind has generally come to acknowledge that there is no sharp discontinuity between our planet and the rest of the universe (Copernican revolution), between humans and animals (Darwinian revolution), nor between rational and irrational humans (Freudian revolution). Mazlish argues that we should also overcome what he terms the fourth discontinuity; that there is no sharp discontinuity between humans and machines.

There are a number of problems with Mazlish's account, however. First, it is difficult to discern precise criteria for what it means for something to become continuous, which means that we seemingly operate with inconsistent conceptions of continuity. A clear example of this can be seen with regard to animal experiments. On the one hand, animal researchers presuppose a continuity between humans and animals – if not, the results would not be relevant to humans. On the other hand, they also presuppose a discontinuity – if not, the experiments would be unethical. It is clear from this example that we often regard two types of entities as continuous in one respect and discontinuous in another. Thus, we need to clarify what these different 'respects' are, and how they relate to each other.

Mazlish seems to claim that a continuity is determined by whether or not the inner workings of two entities can be explained within the same scientific framework, such as computationalism being able to explain both computers and the human brain. Although there are numerous problems with Mazlish's approach, which I will return to below, one if its advantages is that it takes an epistemological rather than ontological approach to the moral status debate – an approach that in many ways mirrors Alan Turing's well-known approach to the question of whether machines can be intelligent [2].

2 FOLLOWING TURING'S LEAD

For the present audience, I presuppose that it is not necessary to explain the Turing test as a means of judging whether a computer is intelligent enough, but in short Turing argues that a computer is to be regarded as intelligent if a human judge cannot reliably distinguish the computer from the human in an imitation game. What is important here is that Turing turns the question of intelligence from an ontological to an epistemological one. That is, Turing does not ask which properties a computer must possess in order to be deemed intelligent (which is an ontological question), but rather how an intelligent observer judges its behaviour. The latter is a type of epistemological question, where we are really asking what kind of explanatory framework we need to presuppose in order to understand a particular type of behaviour. If a computer were to pass the Turing test, this means that the judge had to explain its behaviour as coming from an intelligent being, which says nothing about which properties that being must have (other than being able to display the behaviour in question). Notice that this approach is radically different from the typical approach to questions of moral status and the like, where we typically discuss which properties an entity must possess in order to be regarded as a moral person (e.g. sentience[3], conception of one's own life [4], or having a will to live [5]).[3]

In a similar manner, Mazlish argues (indirectly) that two types of entities should be regarded as continuous if they do not require different scientific frameworks; if the same framework of scientific concepts and models can adequately explain the phenomenon under study. On this background, the Copernican revolution was really a realization that we do not need different scientific frameworks for the earth and the heavens (as was the case with the Aristotelian framework), the Darwinian revolution was a realization that we do not need different scientific frameworks for humans and other animals, and the Freudian revolution was a realization that we do not need different scientific frameworks for the mentally ill and the mentally healthy. Mazlish's prophesized fourth continuity, then, is the realization that we do not need different scientific frameworks for computers and humans either. Thus, all of these continuities amount to radical changes in how to explain different types of

[1] Dept. of Philosophy, University of Twente, Postbox 217, 7500 AE Enschede, Netherlands. Email: j.h.soraker@utwente.nl

[2] Allow me to emphasize that this paper, in line with the IACAP tradition, is work-in-progress, and is presented for the purpose of receiving peer feedback before being developed further. As such, I admit that this paper still lacks the precision and clarity that is to be expected from a finished paper. I hope the reader will apply the principle of charity, and regard this as a starting point for further discussion rather than a fully worked out standpoint.

[3] Cf. Søraker [6] for an overview of the discussion of moral status from an ontological, property-based perspective.

entities (epistemological), rather than saying that all the entities have the same properties and/or mode of existence (ontological). Just like Turing thought an epistemological approach to the question of machine intelligence was more fruitful than an ontological one, I will take a similar approach to the question of continuity in the remainder of this paper. In doing so, I first need to make some important changes to Mazlish's approach, which despite its advantages gives rise to some fundamental problems.

3 PROBLEMS WITH A SINGLE-LEVEL APPROACH TO 'CONTINUITY'

As mentioned, Mazlish seems to claim that a continuity is determined by whether or not the inner workings of two entities can be explained within the same scientific framework, for instance the same physics being able to explain both the earth and the heavens, behaviourism being able to explain both humans and other animals, psychoanalysis being able to explain both mental health and illness, and computationalism being able to explain both computers and the human brain. This is what I refer to as a single-level approach, for reasons I will explain in more detail below.

This approach is problematic for two related reasons. First, anything can be explained within the same scientific framework. Disregarding supernatural and/or substance dualist accounts, it is probably possible in principle to explain the workings of the human brain and a computer by physics alone – and if we believe in scientific progress, our ability to do so will increase in time with progress in physics (I will return to this below).

Second, anything can be explained as if it is an intentional agent, as also argued by Daniel Dennett, who refers to this as taking an intentional stance [7]. Since Mazlish does not specify how strict we need to be when claiming that the same scientific framework can explain two types of entities, his approach becomes inherently imprecise. If it is sufficient that it is in principle possible to explain something within the same framework, then every existing entity is continuous as long as there is no phenomena that cannot in principle be explained by some kind of physics. This would entail that humans are continuous with light bulbs, supernovae and clouds, which leaves the concept of little use. It seems more reasonable, then, to refer to some kind of pragmatism where it must not only be in-principle possible but also pragmatically feasible to explain two entities within the same framework. But, this would require some kind of measure for what it means to be pragmatically feasible. Is it, for instance, pragmatically feasible to explain the brain fully in terms of physical processes, or do we (also) need to invoke chemistry?

Some of these problems are difficult to escape, since it is hard to provide objective criteria for when a particular scientific framework ceases to be feasible. However, we can try to remedy the problem of explanations at different levels by explicitly invoking this into the conception of continuity. In the following section, I will sketch such a multi-level account of continuity.

Before outlining this multi-level account allow me to emphasize that my main concern in this paper is to discuss the formal nature of these continuities, so it is important to emphasize that the levels of explanation that I will use as examples below are to be seen as mere placeholders and the reader will inevitably find some of them problematic and/or imprecise. My goal is to first work out the formal schematics, and then the more substantial content should be worked out in more detail. This will, among other things, require a defence of a particular type of realism and view on scientific progress, both of which fall well beyond the scope of this paper.

4 A MULTI-LEVEL APPROACH TO 'CONTINUITY'

Rather than asking whether two types of entities can be explained within the same scientific framework, I believe it is better to approach this in terms of sets of scientific frameworks – or what I will refer to as sets of scientific levels of explanation. That is, rather than asking whether two entities can be explained within the same scientific framework, we should ask whether two entities require the same set of scientific levels of explanation. As Nagel [8] rightly argues, there are (at least) four fundamentally different types of explanation – deductive, probabilistic, teleological, and genetic – which makes it difficult to precisely define what a scientific level of explanation is. For present purposes, I will simply use the term in the more generic sense of a more or less coherent and mutually supportive set of principles, concepts and models that attempt to provide an account of the relationships between cause and effect.[4]

As mentioned, one of the problems with Mazlish's single-level approach, where continuity is established on the basis of a notion of sharing one scientific framework, is that we often choose different levels of explanation (or, to use Luciano Floridi's term, levels of abstraction [9]) depending on what it is that we seek to explain. Even if it is in-principle possible to explain human behaviour by physics alone, we typically employ higher-level explanations instead. For instance, at a behaviourist level of explanation, we employ concepts like stimulus and response to explain behaviour, without involving physics or chemistry. Even for entirely physicalist phenomena, such as an object moving through space, we often employ heuristics instead of explaining what is "really" going on. As such, the science of ballistics can be seen as a form of higher-level heuristics for explaining how an object moves through space without talking about the complex interplay between electrons and force fields. This, along with the other problems with a single-level approach mentioned above, entails that we cannot define a continuity in terms of a shared scientific framework. A much more promising approach is to define continuity in terms of having a particular set of scientific levels of explanation in common.

To simplify things, if we take a single-celled organism, we may be able to explain its functioning entirely in terms of physics.[5] As we get to more complex forms of life, however, such explanations quickly become untenable. At some point, the chemistry involved becomes too complex to be described in physics terms alone. At even higher levels of complexity, chemistry also fails to provide a full explanation and we need to start talking about biological processes and leave the actual

[4] I am very aware that this is far from precise, but it should be sufficient for establishing the more formal nature of continuities, which is the limited purpose of this paper. As Mieke Boon has pointed out to me, it is probably better to speak of 'practices of explanation', but this paper was due before this could be properly incorporated.

[5] For such an attempt, see Princeton University's Laboratory for the Physics of Life (http://tglab.princeton.edu/).

physical and chemical processes out of our explanations. At even higher levels, we may need to involve the environment and cognitive processes to a much higher extent, and start using principles from, say, behaviourism and comparative psychology. With humans, as evidenced by the widespread criticism and dismissal of radical behaviourism in the mid-1900s, we could also make the case that we need some kind of mental, phenomenological or folk-psychological level of explanation that cannot be reduced to any of the other levels. At even higher supraindividual levels, we may also require social, cultural and other value-laden levels of explanation – and we find ourselves far away from the original physicalist level. Which levels we may need in order to adequately explain a given entity is clearly controversial, and not my concern in this paper, but only the most radical and optimistic scientists maintain that we will in the foreseeable future be able to explain everything by means of one unified theory. On the basis of all this, it seems evident that if we are to define continuity in terms of which type of explanation is required, we must talk about sets of levels of explanation (multi-level) instead of Mazlish's single-level scientific frameworks.

On this background, we can stipulate the preliminary hypothesis: two types of entities are continuous if and only if an adequate understanding of their nature and properties require the same set of scientific levels of explanation; two types of entities are discontinuous if and only if an adequate understanding of their nature and properties does not require the same set of scientific levels of explanation. To illustrate, humans and other animals are continuous if and only if a full understanding of their nature and properties require the same set of scientific levels of explanation (LoE).[6] These definitions still lack a lot of precision, however, and we need to first specify what is meant by 'required'.

4 EPISTEMOLOGICAL VS ONTOLOGICAL CONTINUITY

There are two radically different ways in which a LoE may be required for an adequate understanding. On the one hand, we could for instance argue that the human brain works in such a way that we cannot adequately understand its functioning without employing a chemical level of explanation. Perhaps the chemical properties of neurotransmitters and hormones function in a way that cannot possibly be accounted for by means of more mechanistic explanations – which would be an anti-reductionist view of chemistry. If such a chemical LoE is required because of the brain's unique mode of existence, then that LoE is required for ontological reasons.

On the other hand, we could argue that the human brain works in such a way that it is much more convenient or tractable to use a chemical LoE, even if such an explanation can in principle be reduced to a more basic LoE. If we, despite this in-principle possibility, do require a chemical LoE for pragmatic reasons, then that LoE is required for epistemological reasons.

In light of the above, we can already differentiate between an ontological and epistemological continuity:

Ontological continuity: two types of entities are ontologically continuous if and only if an adequate understanding of their nature and properties require the same set of scientific levels of explanation in principle, due to their mode of existence.

Epistemological continuity: two types of entities are epistemologically continuous if and only if an adequate understanding of their nature and properties require the same set of scientific levels of explanation in practice.

To illustrate, humans and other animals are ontologically continuous if and only if a full understanding of their nature and properties require the same set of scientific levels of explanation in principle. Humans and other animals are epistemologically continuous if and only if a full understanding of their nature and properties require the same set of scientific levels of explanation in practice. It is far from uncontroversial which LoEs are ontologically or epistemologically necessary for an adequate understanding (as well as what is to be meant by 'adequate'), and it is far beyond the scope of this paper to discuss this for different types of entities, but this question maps directly on to the reductionism debates present in the different disciplines. In philosophy of mind, a property dualist would hold that consciousness is somehow ontologically irreducible to neurobiology and physics – which means that a "higher" LoE is ontologically necessary for a full understanding of a conscious being. Eliminiative materialism, on the other hand, holds that consciousness can and should be explained at a neuroscientific LoE, thus claiming that higher LoEs (folk psychology, in particular) are not ontologically necessary for a full understanding of conscious beings. If we compare humans and other animals, the former would hold that conscious animals are ontologically discontinuous from non-conscious animals, whereas the latter would hold that conscious animals are ontologically continuous with non-conscious animals. Non-reductive physicalism, however, holds that conscious states really are the same as physical states and that the former can in principle be explained by the latter – but not in practice. According to such a view conscious beings would be epistemologically discontinuous from non-conscious beings.

5 THE SCHEMATICS OF CONTINUITIES

In light of the considerations above, we can now attempt to schematize what a continuity might look like, according to this multi-level approach. Consider the following schematic:

Type of entity / Required LoE	Humans	Other animals
Psychological	X	
Behaviorist	X	X
Physical	X	X

In this example, humans require a physical, behaviourist and a psychological LoE for a full understanding, whereas other animals can be fully understood by physical and behaviourist LoEs alone. If this is the case, then humans would be discontinuous with other animals. If the psychological LoE is required in principle this is an ontological discontinuity, if required only in practice this is an epistemological discontinuity. To more clearly show the difference with Mazlish's single-level approach, consider the following:

[6] For the remainder of this paper, I will use 'LoE' as shorthand for '(scientific) Levels of Explanation'.

Type of entity / Required LoE	Humans	Intelligent machines
Psychological	X	
Computational	X	X
Physical	X	X

In this example, humans are discontinuous with machines because they require a psychological LoE, whereas machines can be fully understood without. Again, this would be an ontological discontinuity if the computational and physical LoEs are not sufficient for a full understanding of humans (which entails some type of dualism). It would be an epistemological discontinuity if the psychological LoE is only required for pragmatic reasons (which entails some type of non-reductive materialism). Note that Mazlish's single-level approach is unable to account for this, and would treat humans and machines, in this example, as continuous as long as the computational LoE somehow explains both.

Putting this together, the multi-level account of continuity ultimately suggests some kind of hierarchy when it comes to discontinuities:

Type of entity / Required LoE	Humans	Other animals	Intelligent machines	Other inanimate objects
Psychological	X			
Behaviorist	X	X		
Computational	X	X	X	
Physical	X	X	X	X

Now we are able to describe the aforementioned problem of animal experimentation seemingly being an inconsistent practice since it presupposes both a radical similarity (scientific validity) and radical difference (ethical justifiability) between humans and other animals. The scientific validity of such experiments can be grounded in the fact that the LoEs that are relevant for the scientific validity are shared, whereas the LoEs that are relevant for the ethical justifiability are not shared.

This further illustrates how one purpose of establishing discontinuities in this manner is to map their required LoE onto a classification of moral status (or moral worth). That is, there are different ways to harm entities corresponding to their required LoE. In a manner of speaking, the more LoEs that are required for understanding an entity, the more ways there are to harm that entity. At a physical level, we may speak of a minimal harm in terms of entropy, at a computational level we may be able to speak of a minimal harm to self-sustainability, at a behaviourist level we are dealing with harms in terms of rewards and punishment, i.e. infliction of pain and pleasure, and at a psychological LoE it should be evident that the harms become much more complex, including things related to offense, liberty, dignity, privacy, self-actualization, and so forth.

Another important purpose of this schematization is to include Freud's notion of scientific progress changing our conception of ourselves in dramatic ways: "the universal narcissism of men, their self-love, has up to the present suffered three severe blows from the researches of science" [10]. I will refer to such blows as downgrading as opposed to upgrading continuities, which also further illustrates what is meant by LoEs being required in principle and in practice.

6 DOWNGRADING VS UPGRADING CONTINUITIES

It follows from the notion of LoEs being only epistemologically necessary that scientific progress will bring about changes in which levels that are necessary to explain a given entity – which is reflected in the scientific ideals of parsimony, unification and reduction. This means that two types of entities previously seen as discontinuous may become continuous. That is, two types of entities that previously required different sets of LoE, at some point may end up requiring the same set of LoE. According to these schematics, this can come about in two different ways – which correspond to two radically different ways in which science may change our worldview, and where we can more precisely conceptualize Freud's notion of blows to the self-esteem of mankind.

First, we may come to realize that a type of entity no longer requires an LoE that we previously thought to be necessary – for instance when we are able to successfully reduce one scientific LoE to a more fundamental one. When two types of entities come to share the same set of LoE because one type loses a LoE, this amounts to a downgrading continuity. More schematically:

Type of entity / Required LoE	Humans	Other animals
Psychological	~~X~~	
Behaviorist	X	X
Physical	X	X

This was precisely the concern when Skinner's radical behaviourism aspired to explain both humans and other animals entirely in terms of behaviourist principles. This would, according to this line of reasoning, entail a continuity between humans and other animals because humans would no longer require an additional LoE. In a manner of speaking, this would downgrade humans to the level of animals. We can see the same concern when it comes to intelligent machines:

Type of entity / Required LoE	Humans	Intelligent machines
Psychological	- X	
Computational	X	X
Physical	X	X

In this case, humans would become continuous with intelligent machines because they come to share the same set of LoEs due to the loss of one LoE. In a manner of speaking, this does not only amount to humans and computers being "the same", but that "humans are nothing but machines".

There is a converse way of becoming continuous, however. Consider the following:

Type of entity / Required LoE	Humans	Intelligent machines
Psychological	X	+ X
Behaviorist	X	+ X
Computational	X	X
Physical	X	X

In this case, humans and intelligent machines become continuous because the latter attain new LoEs. That is, intelligent machines might become so complex that we can no longer explain their

function by means of computational principles alone. At some point we may need to adopt psychological principles to explain intelligent machines as well, not only metaphorically but as an in-practice (epistemological) or in-principle (ontological) requirement for explaining intelligent machine behaviour. This is what I refer to as an upgrading continuity, where two types of entities come to share the same set of LoEs due to one gaining a new LoE. Indeed, we can say that artificial intelligence has at least taken one important step towards making the sets similar, since highly advanced neural networks now require behaviourist notions in order to be explained. That is, if we want to explain exactly how a successful, complex neural network functions, we have to do so in terms of how the network was subjected to conditioning – a purely computational account of the weights of the nodes etc will often be incapable of explaining exactly how the network actually generates its output. Thus, even if there are good reasons to maintain a discontinuity between humans and machines, the necessity of a behaviourist LoE for explaining highly complex computers, neural networks and embedded systems in particular, entails that we can already now speak of an (epistemological) continuity between machines and non-human animals; they have come to share the same set of LoEs due to computers now requiring a behaviorist LoE

7 PROBLEMS WITH THE APPROACH

Needless to say, this approach is fraught with problems. Allow me to repeat that my only concern in this paper has been to sketch one possible formalization of 'continuity' and a lot of this has to be augmented by a particular conception of what a scientific (level of) explanation is, which will among other things have to rest on a particular stance in the realism debate. Indeed, the account sketched above presupposes some notion of scientific realism (I am leaning towards some form of structural realism) but should also be compatible with more pragmaticist notions of science as well. It also presupposes some idea of scientific progress – i.e. that science, through the development and elimination of LoEs, is providing us with an increasingly accurate picture of reality. That said, I certainly do not rule out the possibility of dramatic paradigm shift, but I think this can be accounted for within this conception of continuity as well. Indeed, a continuity between intelligent machines and humans may require a paradigm shift that obliterates our current LoEs – for instance if we arrive at some quantum mechanical LoE that allows us to explain consciousness and build conscious machines.

There are numerous other problems, primarily related to lack of precision, when it comes to most of the terms involved. This concerns, among other things, what it means to 'adequately' explain something, how to delineate precise levels of explanation, as well as some threshold for when a level ought to be seen as epistemologically and/or ontologically necessary. I hope to develop all of this further as soon as the formal schematics are in place, with the help of peer feedback from different backgrounds.

8 CONCLUDING REMARKS

I can only hope that this paper was read in the spirit intended – as an initial, exploratory and formal account of what it means for two types of entities to be (dis-)continuous. There is no doubt that that the details, if we can even agree on the formal nature, will require a lot of clarification. My only hope for this paper was that the reader, like myself, will on occasion find the notion of continuity an intuitively helpful concept – along with the distinctions between epistemological vs ontological and downgrading vs. upgrading continuities. I am also certain that the reader, like myself, will not be satisfied with the current level of precision, and I would certainly appreciate any help towards improving this. Judging from experience, the IACAP crowd is an excellent starting point to this effect.

ACKNOWLEDGEMENTS

Since this is an idea that has resisted precision, hence publication, for more than 10 years, I can no longer thank everyone that has given my advice over the years. Most importantly among them, my then-supervisor, Magne Dybvig certainly had an important role to play in the initial development. More recently, I am indebted to the helpful comments coming out of my own department's annual research meeting, in particular from Marianne Boenink, Mieke Boon, Mark Coeckelbergh, and Pak Hang Wong. Clearly, none of these are responsible for the lack of precision inherent to this topic and reflected in this paper.

REFERENCES

1. Mazlish, B., The Fourth Discontinuity. 1993, New Haven and London: Yale University Press.
2. Turing, A., Computing Machinery and Intelligence. Mind 1950(236): p. 433–460.
3. Singer, P., Animal Liberation. 2nd ed. 1990, London: Thorsons.
4. Regan, T., The Case for Animal Rights. 2004, Berkeley, CA: University of California Press.
5. Wetlesen, J., The Moral Status of Beings who are not Persons: A Casuistic Argument. Environmental Values, 1999. **8**: p. 287-323.
6. Søraker, J.H., The Moral Status of Information and Information Technologies – a relational theory of moral status, in Information Technology Ethics: Cultural Perspectives, S. Hongladarom and C. Ess, Editors. 2007, Idea Group Publishing.: Hershey, PA. p. 1-19.
7. Dennett, D.C., The Intentional Stance. 1989, Cambridge, MA: MIT Press. 388.
8. Nagel, E., The structure of science: Problems in the logic of scientific explanation. 1979, Indianapolis, IN: Hackett Publishing.
9. Floridi, L., The Method of Levels of Abstraction. Minds and machines, 2008. **18**(3): p. 303-329.
10. Freud, S., A Difficulty in the Path of Psycho-Analysis [Eine Schwierigkeit der Psychoanalyse], in The standard edition of the complete psychological works, J. Strachey, Editor. 1953, Hogarth: London. p. 135-145.