# Towards a digital ethics of agential devices

written by Jeremy Huggett | 23/05/2019




Image by Rawpixel CC0 1.0 via Creative Commons

Discussion of digital ethics is very much on trend: for example, the *Proceedings of the IEEE* special issue on 'Ethical Considerations in the Design of Autonomous Systems' has just been published (Volume 107 Issue 3), and the *Philosophical Transactions of the Royal Society A* published a special issue on 'Governing Artificial Intelligence – ethical, legal and technical opportunities and challenges' late in 2018. In that issue, Corinne Cath (2018, 3) draws attention to the growing body of literature surrounding AI and ethical frameworks, debates over laws governing AI and robotics across the world and points to an explosion of activity in 2018 with a dozen national strategies published and billions in government grants allocated. She also notes the way that many of the leaders in both debates and the technologies are based in the USA which itself presents an ethical issue in terms of the extent to which AI systems mirror the US culture rather than socio-cultural systems elsewhere around the world (Cath 2018, 4).

Agential devices, whether software or hardware, essentially extend the human mind by scaffolding or supporting our cognition. This broad definition therefore runs the gamut of digital tools and technologies, from digital cameras to survey devices (e.g. Huggett 2017), through software supporting data-driven meta-analyses and their incorporation in machine-learning tools, to remotely controlled terrestrial and aerial drones, remotely operated vehicles, autonomous surface and underwater vehicles, and lab-based robotic devices and semi-autonomous bio-mimetic or anthropomorphic robots. Many of these devices augment archaeological practice, reducing routinised and repetitive work in the office environment and in the field. Others augment work by developing data-driven methods which represent, store, and manipulate information in order to undertake tasks previously thought to be uncomputable or incapable of being automated. In the process, each raises ethical issues of various kinds. Whether agency can be associated with such devices can be questioned on the basis that they have no intent, responsibility or liability, but I would simply suggest that anything we ascribe agency to acquires agency, especially bearing in mind the human tendency to anthropomorphize our tools and devices. What I am not suggesting, however, is that these systems have a mind or consciousness themselves, which represents a whole different ethical set of questions.

The question of digital ethics can be approached from three directions: the ethics of the designers

and developers of the device, the ethics of the device itself, and the ethics of the users and decision-makers using the device. In reality, of course, it will involve a combination of any or all of these. After all, the users of systems will be influenced by the ethical decisions taken by the developers, while developers may implicitly or explicitly incorporate ethical perspectives in the devices themselves. There is also the question of whether the devices can be full ethical agents, which would require them to be able to make explicit ethical judgements and be able to justify them (Moor 2006, 20), although this is likely to be a question for some point in the distant future.

Robertson *et al* (2019, Table 1) illustrate the alignment of ethical agency and autonomy in relation to cars (see Figure 1). In an archaeological context we might suggest that a line is drawn between implicit and explicit agency, emphasising the role of the device as one of support, assisting and complementing the human user. According to Moor (2006, 19), implicit ethical agency in devices avoids unethical outcomes by implicitly incorporating ethical behaviour in the software (for example, in decision support systems, autopilots etc.), ensuring in some way their 'correct' behaviour without having ethical judgements encoded in them. Do we wish to conceive of a situation where the device has explicit ethical agency and acts autonomously with limited reference (if any) to the archaeologist? Certainly Juan Barceló has proposed a specialized automated archaeologist capable of learning through experience to associate archaeological observations with explanations, and using them to solve archaeological problems (see Barceló 2009, for example). This seems likely to be a distant proposition given challenges associated with the physical interaction with archaeological spaces as well as the knowledge-based and explanatory components.

| Level of robot moral agency | Sheridan's autonomy | SAE Car autonomy |
|---|---|---|
| No moral agency | Computer offers no assistance; human does it all | Level Zero: No Automation<br>You drive it. |
| No moral agency | Computer offers a complete set of action alternatives | Level One: Driver Assistance<br>Hands on the wheel. |
| Implicit moral agent | Computer narrows the selection down to a few choices | |
| Implicit moral agent | Computer suggests a single action | |
| Implicit moral agent | Computer executes that action if human approves | Level Two: Partial Automation<br>Hands off the wheel, eyes on the road. |
| Implicit moral agent | Computer allows the human limited time to veto before automatic execution | |
| Explicit moral agent | Computer executes automatically then necessarily informs the human | Level Three: Conditional Automation<br>Hands off the wheel, eyes off the road - sometimes. |
| Fully moral agent | Computer informs human after automatic execution only if human asks | |
| Fully moral agent | Computer informs human after automatic execution only if it decides to | Level Four: High Automation<br>Hands, off, eyes off, mind off - sometimes. |
| Fully moral agent | Computer decides everything and acts autonomously, ignoring the human | Level Five: Full Automation<br>Steering wheel is optional. |

Figure 1: Alignment of Ethical Agency and Autonomy in vehicles. The red line/arrow has been added to highlight the boundary between implicit and explicit agency (Robertson *et al* 2019 Table 1).

However, the table in Figure 1 focuses on the ethical agency of the device (the car), and its relationship with the user. What it does not incorporate is any question of the ethical agency of the designers or developers who created these devices in the first place and who instilled implicit or explicit ethical agency in them. Contrary to the table, I'd argue that a device cannot be without ethical agency since its developers themselves will have applied implicit or explicit ethics in situating their product. The intentions of the developers, designers and programmers will be key in establishing the extent of ethical behaviour of the device, and while the end-user will still bear ethical responsibility for its eventual application, the developer side of the ethical equation cannot be underestimated.  Indeed, establishing human responsibility is a feature of proposed legal arrangements concerning robotics and artificial intelligence in the European Union, for instance: the need to ensure the visibility of the makers, designers, data scientists, suppliers and companies responsible for creating artificial agents, as well as all the other actors who interact with and use them, such as workers, employers, consumers, patients, users and trainers (Del Castillo 2017, 9-10). This recognises that the ethical imperative lies with the human elements, ensuring that humans – archaeologists – cannot avoid responsibility by devolving it onto the device. As Olivier Penel has recently observed:

> Algorithms do not have ethics, moral, values or ideologies, but people do. Questions about the ethics of AI are questions about the ethics of the people who make it and put it to use (Penel 2019)

We can identify a number of ethical concerns in relation to archaeology. These may include, for example:

**Control and oversight**

This emphasises the importance of retaining archaeological supervision of the devices we use, and not devolving responsibility for action to the system itself. This would suggest resistance to a fully automated autonomous approach to archaeological agential devices. While we might expect that our archaeological motor skills will remain largely unchanged in the medium term, devolving decisions to machines erodes our critical abilities, as is the case in everything from aircraft and vehicle automation to satellite navigation. At the same time, retaining appropriate levels of supervision requires knowledge of the origins, assumptions, methods and operation of these devices if we are to be able to use them properly. Ethical control and oversight sets the bar for knowledge higher than is frequently the case at present, when we frequently employ tools without a full understanding of them. It also raises the question of where this oversight lies and with whom.

**Augmentation and replacement**

Digital devices are heavily implicated in questions of technology replacement, threatening the displacement of individuals who perform mental as well as physical labour. This may seem a remote threat for archaeologists, not least because of the reliance on complex sensorimotor skills which are as yet problematic to implement in automated devices and the challenge of AI operating outside discrete, well defined and limited application areas (Huggett 2018a). Nevertheless, the question of replacement needs to be considered: are we content to see aspects of the craft and profession of archaeology replaced by digital devices? If so, it again underlines the significance of retaining

oversight. Augmentation is something we are more familiar with: digital, semi-automated assistance with routinised and repetitive work, and devices which operate in inaccessible or unattractive locations, for instance. Retaining emphasis on the human actor at the centre of the process as an active participant rather than an observer – an archaeologist-in-the-loop – should make it possible to avoid the overly scientistic, postitivistic or instrumentalist perspectives on the past.

**Algorithmic opacity and hidden machine bias**

Algorithms have been frequently characterised as black-boxing procedures and processes, and consequently making them inscrutable. Leaving aside the undesirability of not knowing how something has been arrived at, this inscrutability can disguise a range of hidden biases which, while they may ultimately have their origins in the human biases of the original creators, ultimately impact on the system outputs in the form of discrimination or cultural bias, as has famously been demonstrated in recent examples of facial recognition software failures, for instance. Of course, the key issue here is that these are hidden, obscured given the black boxed processes which can perpetuate and reinforce them, and make them difficult to surface.

**Explainability**

It is often said that explanations may be unnecessary where the decisions are not crucial or where there are no unacceptable consequences (e.g. Doshi-Velez and Kim 2017, 3). However, for the reasons just described, it seems unwise to accept the implementation of black-boxed systems without some degree of explainability built into them, although the means by which this is achieved may be open to debate. Simply put, we should not black-box archaeological systems that classify or categorise data without requiring some understanding of the basis on which they draw their conclusions. This is crucial with machine learning systems, but equally with more basic analytical tools, although the locus for the explanation will change accordingly (see Huggett 2019).

**Reproducibility**

The issue of reproducibility has come to the fore in the context of open science, but is equally relevant in this context. With basic analytical tools, we can generally assume that a process is reproducible, in that given the same inputs the same outputs will result given unchanged functionality. However, with AI and machine learning systems, it is not necessarily the case as they adapt to new data and new inputs and so may provide different conclusions. Reproducibility in this context is therefore problematic without detailed information about the internal pipelines applied in each case which could in principle be used to recreate the specific sequence adopted at any stage.

**Trust/Authority**

Trust evidently entails knowledge of the inputs and processes in order to have trust in the outputs. However, this is extended to include trust in the actions and the ability to direct human action as well as verify outcomes in place of human intervention. This should make their transparency and explainability all the more important; however, studies show that devices are frequently used without real consideration and their authority is accepted without question (Huggett 2018b). For example, satnavs are notorious for taking our navigational cognitive load upon themselves and consequently leading drivers who are insufficiently aware of their surroundings into undesirable, even dangerous situations.

**Automation bias**

Linked to questions of authority is automation bias: the increasingly routinised use of devices can lead to them being taken for granted, with the devices simply seen as means to ends and their outputs accepted unquestioningly because they derive from the device rather than from another person. This is related to the kind of fetishization, habituation and enchantment associated with our expectations and use of these devices. For example, in a recent study looking at the adoption of algorithms it was found that simply knowing that other people were using it made it more than twice as likely to be adopted, even in the face of knowledge that it gave imperfect advice (Alexander et al. 2018).

These are just some areas where digital ethical issues within archaeology need to be considered, and apply across the range of agential devices we use. Indeed, over recent years archaeology has been transitioning towards more computerised, automated practices, but our consideration of the ethical implications of this has lagged behind. This is compounded by the way in which digital devices are increasingly moving into areas we might previously have considered un-computable: for example, the combination of big data approaches and machine learning has enabled computers to perform tasks that might have been thought to require cognitive ability and to improve themselves with little or no human intervention. In this context, therefore, the need to debate the nature of the ethics associated with these tools becomes all the more important.

[This post is based on a paper presented at CAA Krakow in April 2019 in the *Ethics in Digital Archaeology: Concerns, Implementations and Successes* session organised by Meghan Dennis and chaired by Cat Cooper]

## References


Alexander, V., Blinder, C. and Zak, P. 2018 'Why trust an algorithm? Performance, cognition, and neurophysiology', *Computers in Human Behaviour* 89, 279-288. https://doi.org/10.1016/j.chb.2018.07.026

Barceló, J. 2009 *Computational Intelligence in Archaeology* (Information Science Reference, Hershey PA).

Cath, C. 2018 'Governing artificial intelligence: ethical, legal and technical opportunities and challenges', *Philosophical Transactions of the Royal Society* A376. https://doi.org/10.1098/rsta.2018.0080

Del Castillo, A. 2017 'A law on robotics and artificial intelligence in the EU?', *Foresight Brief* 2. https://www.etui.org/About-Etui/Foresight-unit2/Foresight-Brief

Doshi-Velez, F. and Kim, B. 2017 'Towards a Rigorous Science of Interpretable Machine Learning'. arXiv:1702.08608v2 [stat.ML] https://arxiv.org/abs/1702.08608v2

Huggett, J. 2017 'The Apparatus of Digital Archaeology', *Internet Archaeology* 44. https://doi.org/10.11141/ia.44.7

Huggett, J. 2018a 'Artificial Archaeologies', *Introspective Digital Archaeology* Oct 22.

https://introspectivedigitalarchaeology.com/2018/10/22/artificial-archaeologies/

Huggett, J. 2018b ‘Digital Place, Cognitive Space’, *Introspective Digital Archaeology* Nov 7. https://introspectivedigitalarchaeology.com/2018/11/07/digital-place-cognitive-space/

Huggett, J. 2019 ‘Explainability in digital systems’, *Introspective Digital Archaeology* Jan 22. https://introspectivedigitalarchaeology.com/2019/01/22/explainability-in-digital-systems/

Moor, J. 2006 ‘The Nature, Importance, and Difficulty of Machine Ethics’, *IEEE Intelligent Systems* 21 (4), 18-21. https://doi.org/10.1109/MIS.2006.80

Penel, O. 2019 ‘Ethics, the new frontier of technology’, *Towards Data Science* April 24. https://towardsdatascience.com/ethics-the-new-frontier-of-technology-815454f0d158

Robertson, L, Abbas, R., Alici, G., Munoz, A. and Michael, K 2019 ‘Engineering-Based Design Methodology for Embedding Ethics in Autonomous Robots’, *Proceedings of the IEEE* 107 (3), 582-599. https://doi.org/10.1109/JPROC.2018.2889678