Artefactual ethics as opportunity for rethinking "natural" ethics

Joel Parthemore*& Blay Whitby†

Abstract

This paper serves as introduction to a significantly longer paper in progress. It argues that, within the ethics community, the wider philosophical establishment and society in general, people have been far too lax about what to accept as morally "right" behaviour – far too quick to let themselves and, all too often, each other off the hook. By drawing comparisons to artefactual behaviour and the objections people raise to calling that behaviour the morally acceptable behaviour of authentic moral agents, this paper lays out a framework by which human ethics and meta-ethics can more fruitfully be approached. An earlier paper of ours (Parthemore and Whitby, 2014) argued that, for an action to be morally right, one must have a convergence of the right motivations, the right means, and the right consequences. The underlying insight is that deontological, virtue-ethics-based, and consequentialist accounts all have their necessary role to play, but each tends to get too focused on itself and its merits to the loss of the bigger picture; while utilitarian accounts, as perhaps the most prominent division within consequentialism, face the further problem of failing to allow for those occasions where the needs of the few, or the one, outweigh the needs of the many, as Ursula K. LeGuin (1973) so devastatingly addressed. Although the requirement to align motivations, means, and consequences may seem impossibly onerous, it need not be, provided one is prepared to allow that moral behaviour is far more difficult to achieve, either for artefacts or human beings, than it might seem at first glance. Mistakes will be made. Perhaps it matters more to take responsibility for those mistakes than to assure oneself, despite reasonable argument to the contrary, that one has avoided them. It is time to hold artefactual and natural agent alike to a higher standard.

1 Introduction: Human beings, artefactual agents, and the responsibility game

For purposes of this paper, we will take moral agency as the capacity to take responsibility, and be held responsible, for one's actions.[1] Intimately wrapped up in all matters moral is the responsibility question: who individually has, and who collectively have, responsibility for any given action or set of events.[2] People have been attributing all manner of agency to virtual and physical artefacts – including moral agency – at least since the advent of Eliza. With the advent of "self-driving" cars and "autonomous" battlefield robots (see, e.g., Sharkey 2011a,b), the responsibility question has only grown. In attempting to answer it, researchers interested in artefactual moral agency (e.g., Wallach et al., 2011; Allen et al., 2000) have tended to focus more-or-less equally on what artefacts do and what they fail to do – what morally relevant "choices" they make or fail to make – and here the standard objection is that existing artefacts either do the "wrong" things (e.g., battlefield robots producing "friendly fire") or fail to do the "right" ones (say, making no response on seeing someone in danger, as with the Uber car that failed to brake for the pedestrian in Arizona). Setting aside whether it provides an adequate litmus test for moral agency – as it is surely attempting to do – Colin Allen and colleagues' (2000) proposed Moral Turing Test[3] sets a standard that, it would seem, no existing artefact could pass. Over-attribution of moral agency is, seemingly, met by bald under-performance.

*Adjunct researcher, University of Skövde, Sweden; joel.parthemore@his.se

†Visiting lecturer, University of Sussex, UK; B.R.Whitby@sussex.ac.uk

[1]This is to make the traditional distinction from *moral patienthood,* which may usefully be described as an entity having certain moral responsibilities attached to it on the part of moral agents (see, e.g., Pluhar, 1988).

[2]Needless to say, neither individual nor collective responsibility excludes the other.

[3]In brief, a purported agent is a moral agent if it takes what they consider the morally "right" decision a sufficiently high percentage of the time.

1.1 Action, inaction, and intention

Somewhat by contrast, psychology tells us that people are, *ceteris paribus,* far more willing to excuse inaction in themselves or others – a failure to act – than to excuse actions they consider morally problematic.[4] To fail to save someone's life – to allow that death to happen – is generally considered less morally wrong than to take a life, even if the two circumstances are, in all other relevant aspects, the same. At the same time, it seems difficult how one might logically justify how the passive vs. active nature of the behaviour could make the necessary difference – as, e.g., Sisela Bok (1999) has pointed out in discussing the nature of lies. How is a *lie of omission* (what I fail to tell you) any less a lie than a *lie of commission* (what I tell you wrongly)? If the one is morally problematic, then so is the other.

Along similar lines, the *Doctrine of Double Effect* (DDE) – often invoked to uphold Roman Catholic thinking on abortion – holds that knowing that something otherwise morally unacceptable will happen as the unintended consequence of one's actions (or inactions) is at least sometimes acceptable whereas intending that same thing to happen would not. The doctrine is necessary for reconciling moral absolutes (killing of human beings is always wrong; human foetuses are human beings; therefore abortion is always wrong) with real-world cases that would otherwise pose problems for those moral absolutes. (What if allowing the pregnancy to go to term – not performing an abortion – would kill the mother or both the mother and the child? Many defenders of the DDE would argue that that is morally preferable because the death of the child, though foreseeable and unfortunate, is not intended; whereas abortion is always an intentional act.) The trolley problem, as originally formulated – quite succinctly![5] – raises difficulties here, as the DDE can equally be used to argue for saving the life of the one person on the one track (with the unintended consequence of killing five on the other) or for saving the lives of the five at the cost of the one: it all depends on one's intentions, which Foot (rightly, we believe) declares unacceptable.[6] For Foot, intention is important but insufficient; means matter; and, clearly, she takes a utilitarian-inspired interest in numbers in favouring the lives of the five over that of the one. For Foot, the outcome *must* be weighed along with the means and intention. For all her sympathy with those who oppose abortion and support the DDE, she sees merit not only in saving the mother's life at the deliberate loss of the child's – i.e., via abortion – when both would otherwise be certain to die; but also in pursuing abortion in cases where only one or the other might be saved. Foot rescues a version of the DDE at the loss of the possibility of absolute moral principles; but one might see this as a good thing. Claims to absolute moral principles may serve to excuse behaviour, as that by persons inclined to take a dogmatic position on abortion, that perhaps should not be excused. If artefacts are not allowed resort to sophistry – whether we think them capable of genuine sophistry or not – then neither should people be.

1.2 Hard-and-fast rules, rules of thumb, and ground rules

Much ink has been spilled within the machine ethics community on what rules to hardwire into artefactual moral agents, and much effort has been made to draw inspiration from Isaac Asimov's *Three Laws of Robotics* – despite the many times, in his stories, where Asimov showed just what impossible conundrums those rules created: a rule intended to anticipate every possible circumstance rarely if ever can. Such rules set a bar so high that not even those who clearly qualify as moral agents can reach it, never mind those whose moral agency may be considered in dispute. If artefacts are ever to be considered candidates for moral agency, then they should be held to no higher a standard than what human beings can achieve.

Rules of thumb might fare better. First-order predicate logic may rely on universal quantification, but the *lifeworld* (Husserl, 1970)with which people engage on a daily basis has a habit of throwing up exceptions. That said, if Foot is right – and we think she is – then *any* strictly rule-based approach will fail. Perhaps the lesson to be learned from present-day artefacts, and the reason so few are willing to grant them moral

[4]If one objects that no one could excuse the human equivalent of the Uber case, the authors have personally encountered it more than once.

[5]"... It may be supposed that [the man] is the driver of a runaway tram which he can steer from one narrow track onto another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed" (Foot, 1967).

[6]"A certain event may be desired under one of its descriptions, unwanted under another, but we cannot treat these as two different events, one of which is aimed at and the other not. And even if it can be argued that there are here two different events... the two are obviously much too close for an application of the doctrine of double effect" (Foot, 1967).

agency,[7] is not that they lack the right rules with which to make the right decisions; rather it is that they lack the capacity to make decisions or take responsibility for them in the first place – in any but the most loosely metaphorical of senses. Remember that, by our definition, moral agency requires the capacity to take responsibility: something that – in company with newborn infants and certain among the mentally infirm[8] – present artefacts would appear to lack. Most infants and at least some mentally infirm persons can be expected to outgrow their present conditions; by contrast, no amount of time and patience will change present-day artefacts or their close kin into moral agents.

This is not to say that one can or should avoid hard-and-fast rules altogether. At least at first blush, the principle that what is acknowledged as morally wrong should never simultaneously be accepted as morally right seems like a suitable candidate. Indeed, if one accepts that moral right and wrong are mutually exclusive, then it follows of logical necessity. Yet "lesser of two evils" arguments, widely used, require that the "lesser" evil is, at the least, morally acceptable if not strictly speaking "right"; and "just war" accounts – to take one example – critically depend on such arguments. The evil action (or inaction) becomes the good because, it is said, there is no alternative. Jean-Paul Sartre showed that, on nearly every occasion where people claim a lack of alternatives, there *are* alternatives; the problem is either that we fail to see or that we fail to acknowledge them. If people would not accept "lesser of two evils" arguments to excuse artefactual behaviour – and we believe that few would – then they likewise should not accept them to excuse their own.

The solution posed by the full paper is to let go of moral absolutes – few things indeed are *always* morally right or wrong – and to embrace personal responsibility, as Sartre (1946) has challenged us all to do: taking responsibility and acknowledging both when we believe that we *have* done right, despite all evidence and arguments to the contrary, with a willingness and ability to defend the reasoning that led us there; and when we know we have done wrong, either because we could not see an alternative or lacked the courage to embrace it. The proper response to the high standards imposed on moral agency for artefacts is not to lower those standards on artefacts but use them to raise the bar for ourselves.

Section Two of the intended full paper, currently a work in progress, will examine the ethical theory that serves as the foundation for this extended abstract – one that calls for a convergence of the "right" motivations, the "right" means, and the "right" outcomes – and consider how it can be made to work.[9] Section Three will consider the consequences of applying that framework to purported artefactual agents. Section Four will offer three case studies: one from the field of autonomous vehicles, one from aviation, and one from medicine. Section Five will bring the artefactual lessons back to the human case and offer prescriptions on the way forward.

References

Allen, C., Varner, G., and Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3):251–261.

Bok, S. (1999). *Lying: Moral choice in public and private life*. Vintage. First published 1978.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15. Available online from https://philpapers.org/archive/footpo-2.pdf (accessed 26 January 2020).

Husserl, E. (1970). *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*. Northwestern University Press. tr. David Carr. First published (in German) 1954.

LeGuin, U. K. (1973). The ones who walk away from omelas. In Silverberg, R., editor, *New Dimensions*, volume 3, pages 1–8. Doubleday.

Parthemore, J. and Whitby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness*, 6(2):1–21.

[7] . . . Despite the haste with which others would do so!

[8] . . . Who nevertheless qualify as moral patients!

[9] A reviewer suggested basing that discussion around climate ethics but that, to our mind, would be a different paper. For better or worse, we have chosen to return to the artefactual question we addressed in our earlier papers.

Pluhar, E. (1988). Moral agents and moral patients. *Between the Species*, 4(1):32–45.

Sartre, J.-P. (1946). The flies. In *The Flies and In Camera*. Hamish Hamilton. tr. Stuart Gilbert.

Sharkey, N. (2011a). Automating warfare: Lessons learned from the drones. *Journal of Law Information and Scienc*, 21:140–154.

Sharkey, N. (2011b). Killing made easy: From joysticks to politics. In Lin, P., Abney, K., and Bekey, G. A., editors, *Robot Ethics: The Ethical and Social Implications of Robotics*, chapter 7, pages 111–128. MIT Press.

Wallach, W., Allen, C., and Franklin, S. (2011). Consciousness and ethics: Artificially conscious moral agents. *International Journal of Machine Consciousness*, 3(1):177–192.