Towards a coherent account of moral agency

Fabio Tollon[1]

[1] Stellenbosch University, Stellenbosch Central, 7602
fabiotollon@gmail.com

Johnson and Noorman's (2014) account of agency provides us with an understanding of how we may have agents without mental states, and therefore gives a more nuanced account of the moral *roles* that Artificial Agents (AAs) may perform. However, this account only provides a truncated understanding of the potential for Artificial Moral Agents (AMAs). In order for an entity to be considered a real moral agent, on their account, it must still have some form of *intentionality*, and it is the specific sense of intentionality argued for by Johnson (2006) that I claim is at issue. Johnson uses human beings as the paradigmatic case of moral agents, which seems to be a fair claim. We consider most human beings to be moral agents, and so any account of moral agency that contradicts this claim would seem to seriously undermine our conventional ways of holding each other responsible. The issue here, however, is that in Johnson's account she makes it clear that the buck stops and starts with human intentionality: the only intentionality an AA can ever have would be derived from a human being, with the implication that the human being remains the sole locus of responsibility. This way of understanding matters is incapable of accounting for a future in which AAs become increasingly autonomous, as it does not allow for them (AAs) to be considered as genuine AMAs.

My argument in this paper will proceed in three main steps: firstly, I will evaluate Johnsons (2006) argument that because machines do not have "intendings to act" they should not be considered moral agents. Secondly, I will introduce the notion of *functional* moral agency, as an alternative to the *intentional* moral agency argued for by Johnson and show how the former account of moral attributability does not have the same shortcomings as the latter. Lastly, then, I will provide a positive account of functional moral agency, drawing on the work of Floridi and Sanders (2004).

Johnson's main objection to the idea of AMAs turns on the fact that she denies that artificial entities can be moral agents as they do not have the requisite mental states that could be the cause of the events that we would understand as their actions. The thrust of this argument turns on the fact that these intentional states combine to form reasons for acting (in Johnsons terminology "intendings to act"), and these reasons are internal to the entity in question (2006). For Johnson, this *independence* is essential, as since AAs have the essential feature of having been designed by paradigmatic moral agents (human beings), this ontogenetic fact undermines their ability to be considered "real" moral agents. What this implies for present my purposes is that "real" intentionality is reserved for human beings, and AAs can only have a derived type of intentionality (Grodzinsky, Miller and Wolf, 2008; Powers, 2013). Moreover, this "derived" kind of intentionality is insufficient as a criterion for grounding moral agency. Key to this distinction is that human beings have *internal* mental states, which AAs lack. A key issue which arises at this point is the problem of other minds.

The way in which we go about attributing internal mental states to one another rests on external behaviouristic cues: the only evidence we have to go on when attempting to figure out what caused another human being to behave in a specific way is evidence which is expressed in some external way in order for us to be able to evaluate it (Powers, 2013: 232). While it may be possible in principle to have some kind of access to an agent's internal mental states, in practice this possibility is not guaranteed. For now, if we have the choice between abstract metaphysical speculation (inferring the presence and nature of others' internal mental states) and using clearly identifiable, observable cues, it seems obvious that we should prefer the latter method of investigation. The latter provides us with greater epistemic security than the former, in the form of verifiable data points which we can use to make relatively sound judgements. If all that we reliably have to go on when judging whether human beings are moral agents are external cues that they provide, then positing the verifiable existence of internal mental states as a necessary requirement for moral behaviour is a condition that cannot even be met by humans. Johnsons stipulation that agents must have internal, mental states, therefore, precludes us form having a coherent account of how *human beings* are moral agents. This is surely not a desirable outcome and is tethered to an assumption that any theory of moral agency that we endorse should preserve the status of human beings as moral agents.

Following this critique of Johnson, I will show how a general functionalist account of moral agency, which preserves our current understanding of human beings as moral agents, can provide a better theoretical foundation than intentional accounts. A functionalist account of moral agency does not look at internal mental states, but rather towards the role that an entity plays in the production of certain events. The basic functionalist claim is that the level of abstraction proposed by intentional accounts of moral agency is too low, and by raising it we can come to understand less anthropocentric perspectives on which entities qualify as moral agents. Moreover, I suggest that it might even be possible to hold AMAs morally responsible for their actions. This would involve a re-examination of the role that "punishability" plays in our moral assessments.

An AMA may not "feel bad" when punished, but this need not stop us from potentially holding them morally responsible. They are still moral agents "deserving" of "punishment", even if that means we just turn them off or reprogram them for the benefit of society. This introduces the possibility that there is a case to be made for a sense of moral responsibility that is independent of whether the agent in question is "punishable" or not. This sense of moral responsibility would be anchored to the pragmatic benefits such an ascription might confer to society. Examples of such benefits could be the resolution of "responsibility-gaps" (Champagne and Tonkens, 2013; Nyholm, 2017), cases in which it is unclear whether a human being was responsible for a given morally significant action while it is clear that an AA was causally involved in the action.

Reference List

1. Champagne, M. and Tonkens, R. Bridging the Responsibility Gap. *Philosophy and Technology* 28(1), 125–137 (2013)
2. Floridi, L. and Sanders, J. W. On the Morality of Artificial Agents. *Minds and Machines* 14, 349–379 (2004)
3. Grodzinsky, F. S., Miller, K. W. and Wolf, M. J. The Ethics of Designing Artificial Agents. *Ethics and Information Technology* 10, 115–121 (2008)
4. Johnson, D. G. Computer systems: Moral Entities but not Moral Agents. *Ethics and Information Technology* 8, 195–204 (2006)
5. Johnson, D. G. and Noorman, M. Artefactual Agency and Artefactual Moral Agency, in Kroes, P. and Verbeek, P.-P. (eds) *The Moral Status of Technical Artefacts*. New York: Springer, 143–158 (2014)
6. Nyholm, S. Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci, *Science and Engineering Ethics*, 1–19 (2017)
7. Powers, T. M. On the Moral Agency of Computers. *Topoi* 32(2), 227–236 (2013)