# Review of: "Metacognitive Agents for Ethical Decision Support: Conceptual Model and Research Roadmap"

Ben Kenward[1]

1 Oxford Brookes University

Potential competing interests: No potential competing interests to declare.

This paper has now received a number of reviews focussing on various technical aspects of the proposals, and I'm anyway not expert in the technical aspects of these ideas. This review therefore focuses on only two questions which arise from the ideas proposed in this paper, and which need consideration because they have implications for whether the proposed ideas are advisable, and how they might or might not work. The questions are: (1) how might people react to suggestions from an artificial agent that are framed as moral, and (2) what implications might these reactions have for whether such a system could result in moral progress in human behaviour.

Consider how most or many people now have an intellectual understanding that (1) those who eat only plants tend to produce appreciably lower emissions, and (2) producing appreciable lower emissions is a major moral improvement. A simple syllogism (also intellectually understood by most) implies these people are engaging in morally better behaviour. And yet, many people hate being told by such people to be more like them, and will find many rationalisations that allow them to ignore the most important moral issues and derogate the sanctimonious vegans (Kurz et al., 2020). This single example illustrates the complexity of attempting to achieve improvements in moral behaviour by giving advice. The current paper focuses very much on issues of cognitive bias that lead to errors in moral judgement, but in many cases (including the most important case facing humanity, climate change) the problem is not that individuals are incapable of the the correct moral judgements, but that we have psychological defence mechanisms which allow us to sideline them (Lamb et al., 2020; Norgaard, 2006).

So we need to think carefully about how our words will be received, before we deliver words intended to improve moral behaviour, or design machines intended to deliver such words. In a recent paper (Kenward & Sinclair 2021), we argued that artificial agents that act as moral advisors are between a rock and a hard place. On the one hand, if such agents offer advice to humans that is appreciably more progressive than currently common human norms, there is a risk the agents are even more severely derogated than other humans would be in the circumstances. On the other hand, if the agents are in fact designed to offer advice that is roughly consistent with currently common human norms, they will be doing nothing to achieve the moral progress that humans so desperately need. The human-agent interaction proposed in step 4 of the currently proposed roadmap does not appear designed to avoid such problems. I suggest the author and others who consider designing such systems give further thought to the psychological processes likely to occur when humans interact with the machines.

References

Kenward, B., & Sinclair, T. (2021). Machine morality, moral progress, and the looming environmental disaster. Cognitive Computation and Systems, n/a(n/a). https://doi.org/https://doi.org/10.1049/ccs2.12027

Kurz, T., Prosser, A. M. B., Rabinovich, A., & O'Neill, S. (2020). Could Vegans and Lycra Cyclists be Bad for the Planet? Theorizing the Role of Moralized Minority Practice Identities in Processes of Societal-Level Change. Journal of Social Issues, 76(1), 86-100. https://doi.org/10.1111/josi.12366

Lamb, W. F., Mattioli, G., Levi, S., Roberts, J. T., Capstick, S., Creutzig, F., . . . Steinberger, J. K. (2020). Discourses of climate delay. Global Sustainability, 3, e17, Article e17. https://doi.org/10.1017/sus.2020.13

Norgaard, K. M. (2006). “People Want to Protect Themselves a Little Bit”: Emotions, Denial, and Social Movement Nonparticipation. Sociological Inquiry, 76(3), 372-396. https://doi.org/10.1111/j.1475-682X.2006.00160.x