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

Ivan Ivanchei[1]

1 Ghent University

Potential competing interests: No potential competing interests to declare.

The paper is dedicated to the challenges of developing a recommendation system or an artificial assistant for ethically aligned decisions at the workplace. The author suggests a roadmap and a set of recommendations based on cognitive science literature. That is a timely work with the potential to influence the field of applied Artificial Intelligence.

However, I had some difficulties reading it and understanding its main message. To me, it was not clear what problem the author is trying to solve and what type of solution should the reader expect in the paper. It is not a computer science report, nor a cognitive psychology empirical or theoretical study.

Therefore, it would be helpful to see a clear framing of the paper from the beginning. That is especially important for an interdisciplinary journal. Such framing would include answering the following questions:

1) what is the problem you are trying to solve?

2) what is your approach to this problem and, relatedly, what is the field of study, within which you are going to look at the problem?

3) what is the shape of the expected solution – a working algorithm, a theory, a set of guidelines, or a set of recommendations for AI developers? I can see from the title, that the shape of the solution would be a “conceptual model“ and a “roadmap”, but that still does not make it clear who and how can apply this paper. Maybe it’s a good idea to add a sentence or a paragraph on who is the target audience of the article.

Without clear framing, it is difficult for me to understand why the author introduces affective processes as well as metacognition. Is the author going to model human-decision making with its biases? It is not clear to me why would you do that for a system that is aimed to avoid biases. At least, at the moment when affect and metacognition are introduced in the paper – it is still not clear why it is important for solving the problem. It may be helpful to provide a complete example of a model/system that you are aiming for. In other words, describe a concrete situation where your desired model could operate.

One minor comment: Sloman et al. (2005) – no such reference in the list.

Overall, I think this could be an interesting paper, but it needs to be better framed and targeted to its audience. Right now it is difficult to evaluate its significance and role in ethical AI development.