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

Valeria Seidita[1]

1 University of Palermo

Potential competing interests: No potential competing interests to declare.

The aim of this paper is to present a conceptual model for the use of agents for ethical support systems.

The argument is very interesting that nowadays the complexity of current software systems is so high that there is an urgent need for a proper development phase considering serious complex aspects such as norms, rules, ethics, trust, etc. There is no disciplined approach in the literature to bridge the gap between the complex requirements and the actual implementation, and there is a lot of ongoing research on this topic.

The work presented in this paper would like to fall into this research area. Unfortunately, the work is very difficult to understand. The motivations for the work are clear, but the author fails to illustrate and really clarify what is being proposed. I think the main issue in this regard is that there are many sentences throughout the paper that seem to be disjointed, causing the reader to lose the flow of the paper. In addition, much of the information is unsubstantiated and unrelated to the topic of the paper, e.g., how and when (in the modeling phase) are options and arguments generated by the agent? How and when are ethical values and norms established? Do they depend on the scenario or does the author propose a general approach that can be used in different scenarios? The latter is the situation I expect from a scientific paper, I do not need an application for a specific scenario, but a way to model ethical agents that provide support in any situation where intelligent agents are supposed to support users ethically. Unfortunately, I can not find anything about that in this paper.

The "Related Work" section needs to be greatly improved by the addition of work related to the present work. The literature should be discussed from an ethical point of view, and it should be detailed why other works have similar goals or are of some interest.

The same applies to background; it is necessary to connect the discourse on affective models with ethical concerns.

I have two main concerns about this paper. The first relates to the notion of ethics. It seems that it is a normative fact. In a sense, ethics is based on norms, but the fact that the paper talks about ethics does not affect what is illustrated, except for the fact that some norms are considered, and I think that is reductive.

The second point is the concept of agent-based modeling and simulation. ABM is a tool for modeling and designing agents that exhibit complex emergent behavior, it is not used for modeling and designing multi-agent systems in general. It is not clear in the paper why simulation and ABM are used. It seems, and there is no evidence or proof, to be used only for simulation of agents that in a given scenario assist the user in some way. I am not convinced that the author is using the concept of agent and ABM in the right way. In my opinion, the theoretical basis of this work is very thin. Nowhere in the paper is it said which agent abstractions (and how) are used for the cognitive-affective models. Belief, desire, intentions, goals, and the others mentioned are the basic building blocks for constructing a multi-agent system, but nothing is said about which of them are related to ethics (or to norms, or to anything that might be used for expounding ethical explanatory activities). How they are used. Nothing more than the basic literature is mentioned, what further advancements?

It would be a lot to know how agents are used, then the real progress in the state of the art would be perfect. The roadmap listed is not progress in the state of the art. I think the paper should be more than just a list of actions to do to leverage what is in the literature.