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

Martin Takac[1]

1 Comenius University in Bratislava

Potential competing interests: No potential competing interests to declare.

The paper brings into focus an important question of gaps/inconsistencies between good intentions and actual actions in humans. The author does a very good job reviewing relevant literature - symbolic-cognitive architectures as well as psychology of dual-process theories of decision making, human biases and metacognition. The main goal of the paper is to outline a roadmap towards an artificial decision-making support system (with different levels of autonomy) that would help to eliminate the gap between behavioural choices and set ethical values (and norms stemming from them). The author also tries to identify different scenarios and provides an example use case. While I laud such endeavour, I can also see potential problems with some steps of the roadmap. I detail them below - not to criticize the author, but to bring attention to potential weaknesses / stumbling blocks.

First of all, while the paper tries to cover different use cases (a fully autonomous AI system making moral decisions vs. human decision support system), it sometimes makes the argumentation less valid, because different arguments apply to different cases.

Although the author proposes to start with "broad and shallow" system design and stay on Marr's level 1, I believe it is necessary to think ahead also about levels 2 (algorithms) and 3 (implementation), because whether such a decision-making system will be implemented as a symbolic, connectionist or hybrid architecture, all these architectures have known problems that will have impact on the overall functioning of the system. The author correctly points out that large connectionist / machine learning systems trained on massive amounts of data often have a problem of perpetuating existing biases (such as racism, sexism etc.). That is why, at least for the start, the author suggests using a symbolic reasoning system - similar to expert systems (see e.g. Gupta & Nagpal, 2020) that were successfully used especially in the 1970-80s. Unlike distributed/connectionist systems, the domain knowledge could be specified in the form of explicit facts and rules, which also made it easier to generate justifications/explanations for the system's decision, often in the form of the whole chain of reasoning. However, these systems suffered from the following problems:
1. Lack of common sense knowledge (Brachman & Levesque, 2022): Common sense is (often unspoken and unwritten) knowledge that all/most humans possess and involve even when solving domain-specific problems. On the other hand, AI systems can only reason from the data given to them, which can give them significant blindspots.
2. Frame problem (see e.g. Pylyshyn, 1987): What gives an expert system its power is the ability to not only consider explicit facts, but also consequences of those facts and consequences of the consequences. However, deciding which

consequences are relevant and which are not is a hard problem.

I am afraid the same problems would also carry over to the decision-support system proposed in the paper. That is why I believe such systems, with the current state of knowledge, should not be deployed as fully autonomous decision makers in the areas where their reliability is critical. The solution for now is to keep humans in the loop. Stuart Russell (2019) writes about "humble machines" - in the presence of uncertainty about goals they "ask for permissions, do trial runs, accept corrections". The author suggests including humans in the loop in section 5.6 (model M4). This, however, requires natural language processing (NLP). The best state-of-the-art NLP systems are based on embeddings - distributed representation of semantics obtained from hidden layers of large neural networks. I just want to point out here that in that way they can inherit biases and known issues of these large-scale connectionist language models.

I see the main value of the proposed system in the 2nd proposed use case - as a dialogue partner that will assist humans in making informed decisions. Especially the ability of the system to criticize human-suggested options, but also to ask additional questions is crucial. I see the key value of such a system in the following:

1. *Making a human aware of (implicit/unintended) consequences of proposed actions and their connection to the ethical values:* This can be based not just on logical inference, but also on domain-specific simulations and crowd simulations (e.g. modeled by multi-agent systems) often revealing emergent effects in complex systems. For example, an individual decision (such as picking one flower in a national park, or throwing away a piece of plastic) can have a massive effect if taken by many individuals.
2. *Suggesting novel solutions:* I believe this aspect was not elaborated in the paper. People sometimes make morally questionable decisions not because of lack of awareness of the gap between values and deeds, but because of more pressing factors, such as too high a cost of an ethically better solution. For example, a person can be aware that flying is not a sustainable way of travelling, but the time constraints take precedence. If, however, the person was reminded of an option to buy a ticket with a carbon offset price, she might do it and mitigate the environmental effects to some extent. Another example of a novel solution in a situation where a hard criterion has to be met is described by Miller (2021, p. 218-219) During the 2008 financial crisis, a CEO of a company was facing a hard decision of downsizing the company to reduce costs. Instead of firing people, he asked everyone in the company to take a month of unpaid leave. The goal of reducing costs was met without people losing their jobs and the whole operation emphasized and promoted a corporate value of caring for each other. The system could give suggestions such as "if you must do A, at least do also B to mitigate the negative effects of A" (this, however, goes beyond deontological ethics, more towards a consequentialist one).
3. *Motivating people:* Again, this aspect was not sufficiently addressed in the paper. The proposed system would be more effective if it did not just provide a consistency check between normative ethics and the considered decisions, but also supported the users in wanting to accept the suggestions of the system and behave accordingly. Often the problem is not that people don't see the value-action gap, but that they don't want to see it. The paper describes the dual processing systems in the way that system 1 is quick and effective, but faulty and biased, while system 2 can rationally demask the biases and find a better solution. I want to suggest that if we really want to facilitate change, we need to take System 1

seriously (not just "direct the driver", but also "motivate the elephant"; see Heath & Heath, 2010). In this way a dialogic system asking questions and exploring what are the hidden motivation factors and what prevents the human from taking ethical decisions would be very beneficial.

In relation to the above, it was unclear to me what was the purpose of including affects in the model - just to show what effect they have on the decision? Does the author see any positive role of emotions for decision making or are they viewed just as a source of biases? The metacognitive element in the model seems interesting, but I believe it should do more than just comparing the model M2 (contaminated by biases coming from M1) with an unbiased normative model M0.

Having said all this, I still believe the presented article is a valuable contribution to the much needed discussion about the role of AI in ethical decision making and also the broader area of AI ethics and value-based design.

ps: There's a typo in section 5.4.1: "In order words" should read "In other words".

References:

- Itisha Gupta and Garima Nagpal (2020): Artificial Intelligence and Expert Systems, Mercury Learning and Information.
- Ronald J. Brachman, Hector J. Levesque (2022): Machines like Us: Toward AI with Common Sense, The MIT Press.
- Pylyshyn, Z.W. (ed.) (1987), The Robot's Dilemma: The Frame Problem in Artificial Intelligence, Ablex.
- Russell, Stuart (2019): Human Compatible: Artificial Intelligence and the Problem of Control, Viking.
- Miller, L. (2021): The Awakened Brain: The Psychology of Spirituality, Penguin Books.
- Heath, C. and Heath, D. (2010): Switch: How to Change Things When Change Is Hard, Broadway Books.