Towards Enhancing Moral Agency through Subjective Moral Debiasing, Mark Herman

Commentary, Nina Atanasova

The central question of Herman's project is "How could moral rationality be improved?" The significance of this question can be recognized in the fact that moral irrationality "can frustrate one's capacity to act in accordance with one's morality", which is in turn constitutive of moral agency, the author points out.

Herman entertains the idea that an empirical research program for approaching the problem could be drawn from empirical research on non-moral rationality. Considering the success of cognitive debiasing programs, systematic errors in moral reasoning might be manageable with similar means, the author suggests. For this reason, he dedicates his paper to providing an account of moral error which would lay the groundwork for the addressing systematic mistakes in moral reasoning.

Herman's first step is to introduce a simple model of subjective moral error, according to which "An agent A's performing some action φ is a subjective moral error insofar as φ-ing violates A's morality – i.e. frustrates A's moral ends" (p.3). Next, he proposes a simple model of internal reason, according to which "A has an internal reason to φ iff A *possesses* some desire whose satisfaction will be served by φ-ing" (p.4). Both accounts are revised and result in more sophisticated "idealized agent" models where agent A is replaced by the idealized A+.

This move enables Herman to present a model in which simply mistaken beliefs will not render mistakes in moral reasoning.

The next step in the project is to articulate an idealized subject's morality which would preserve normativity but also its applicability to actual agents. This can be ensured by specifying endowments among which informational and moral-psychological. However, specifying the relevant endowments presents additional problems which, according to the author, could be solved by adopting "*Two-Tier* Model of Subjective Moral Errors", according to which "A's φ-ing is a subjective moral error insofar as φ-ing deviates from A's genuine morality – i.e., frustrates A+'s moral ends, wherein A+ is a counterfactual idealization of A upon whom is bestowed those endowments that A considers morally authoritative under ordinary optimal conditions" (p.12).

This is a very intriguing project so far as it promises to contribute to the development of techniques which could be utilized for reduction of subjective moral biases. I fully agree with Herman that such techniques, if successful, would improve moral rationality. As the author points out, additional work remains to be done in order for the model to be utilized in empirical research. I would like to invite him to elaborate on this point as well as his proposal to look for ideas into the "*deliberate model* of medical consultation".

I would be particularly interested in examples of potential applications of this model for purposes of the differentiation between failures in moral reasoning and pathologies in which moral reasoning may be reduced to absurdity. For example, sociopaths tend to be rather intelligent and, I would speculate, they could identify their "moral" goals and act accordingly to achieve them. However, it would go against common sense to consider murder, which sociopaths may commit in accordance with their values, to be a conventionally moral act. On this account, would such a situation fall under the category of moral error or something else?