Commentary: Modeling the Social Dynamics of Moral Enhancement While Illustrating Some Basic Divergences in the Enhancement Debate

SØREN HOLM

The article by Anders Sandberg and Joao Fabiano in this issue of *Cambridge Quarterly of Healthcare Ethics* is a very important addition to the enhancement literature.[1] It reports an important simulation study of the stability of prosociality in society and at the same time implicitly illustrates some of the enduring uncertainties and divisions in the philosophical enhancement debate that often lead to participants in that debate talking at cross purposes. These come to the surface when reading Sandberg and Fabiano's article, because specifying the mathematical model requires that terms be defined and operationalized to a degree that is unusual in the philosophical literature. In this brief comment I will outline three such uncertainties or divisions that come to the surface in Sandberg and Fabiano's article.

The first of these uncertainties is about the meaning of terms such as "moral enhancement" and "prosociality" and how we should think about them. Sandberg and Fabiano mention the problems in defining moral enhancement and in distinguishing it from psychological changes that may lead to morally better actions in certain circumstances, such as prosociality, but that may lead to problematic outcomes in other circumstances. However, there is an additional issue concerning how one should think about and model an increase in prosociality or any other psychological characteristic or character trait. Sandberg and Fabiano model changes in social value orientations (SVO) such as a change toward or away from prosociality on the SVO Ring Measure; however, that seems very restrictive compared with how possible changes are usually conceived of in the enhancement literature. In the enhancement literature, and especially among transhumanist writers, we are consistently being told that our powers of valuation will be enhanced and that we will be able to value things and to access welfare levels that are currently unavailable to us.[2] If I take this approach to enhancement of prosociality, then enhancement does not change my position on the SVO Ring Measure to bring me closer to optimal prosociality, it takes me beyond the ring to experiences of welfare in prosocial outcomes that are not accessible to the unenhanced. Now, both conceptions of enhancement of prosociality are valid conceptions, but they are radically different and unless I were to specify (as Sandberg and Fabiano very explicitly do) which conception I am talking about, there is a huge potential for equivocation.

The second division also relates to the meaning of moral enhancement, but it relates more to the issue of how the enhancement label gets attached to individual agents. On one side of the moral enhancement debate moral enhancement is seen as a solution to a social problem: we need to be morally enhanced or we will

Cambridge Quarterly of Healthcare Ethics (2017), **26**, 446–448.

doi:10.1017/S0963180116001110

become extinct.[3] If this is the justification for moral enhancement, it does not matter whether moral enhancement is good for everyone, or even most of those who have been enhanced, what matters is the total outcome. I will call this the "societal account" of moral enhancement. On the other side of the debate, moral enhancement has to be good for those who are enhanced, or at least likely to be good for them at the point of choice of enhancement, to count as a true enhancement.[4] I will call this the "personal account." Sandberg and Fabiano's modeling shows that a society with a predominant SVO of "sadomasochism"; that is, a negative evaluation of positive outcomes for both a particular person and for the agent with whom that person interacts, can be a stable society given certain initial parameter choices. If this society is less likely to become extinct than our current society, then the move to the sadomasochistic SVO will count as an enhancement on the "societal account," but presumably not on the "personal account" of moral enhancement. This very nicely illustrates the possibilities of equivocation and vagueness when using the term "moral enhancement" in an unspecified way.

The third and perhaps most fundamental uncertainty is about the scope of enhancements that one should assume to be possible when analyzing a future with enhancement. Sandberg and Fabiano's modelling of the social stability of prosociality explicitly assumes that agents interact one to one, that they are equal in power,[5] and that they can only change their evaluations to another point on the Social Value Orientations Ring. However, all these assumptions are incompatible with the more radical enhancement scenario they outline in the Conclusion "… once we break free from its [our human being shaped by evolution] chains we are overwhelmed by possibilities; we no longer have a set of common traits and preferences that must necessarily be held and whereby all other choices can be evaluated." But how are we supposed to model or even think about such a future when agents can interact in ways that we may not be able to imagine, have valuational structures that are beyond our current comprehension (partly because we may not know what features of the world it is that they value), and have powers of action far beyond our ken in addition to society being structured along lines that are equally unfathomable from our current standpoint? This problem of truly unimaginable complexity caused by transhuman diversity can be handled in several ways, but none of them are really satisfactory. One can simplify and try to remove all of this transhuman diversity by arguing that there will only really be one or a few types of transhuman agents, as Nick Bostrom does in his recent book on superintelligence when he argues that only one superintelligence will emerge thereby obviating the need to analyze a world with many superintelligent agents.[6] We can make our task seem easier by implicitly (and illicitly) importing assumptions drawn from human anthropology[7]; or we can retreat into the mysticism of "the Singularity."[8] All of the approaches make modeling and analysis easier, but at great cost, because none of them are likely to provide us with a realistic assessment of a future predicted to be radically different.

Notes

1. Fabiano J, Sandberg A. Modelling the social dynamics of moral enhancement: social strategies sold over-the-counter and the stability of society. *Cambridge Quarterly of Healthcare Ethics* This issue
2. Bostrom N. Letter from Utopia. *Journal of Evolution and Technology* 2008; 19(1):67–72.
3. Persson I, Savulescu J. *Unfit for the Future: The Need for Moral Enhancement*. Oxford: Oxford University Press; 2012.

4. Harris J. *How to be Good*. Oxford: Oxford University Press; 2016.
5. There may be inherent inequalities in the game played in each encounter, but because position in the game is randomly allocated, this does not lead to systematic differences in power.
6. Bostrom N. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press; 2014.
7. Holm S Naturalness and anthropology in modern bioethics, with a special view to trans- and post-humanism. In: Kragh H, ed. *Theology and Science - Issues for Future Dialogue*. Aarhus: University of Aarhus; 2007:17–29.
8. Holm S. Evaluating the posthuman future–some philosophical problems. *European Review* 2016;1–9.