Deke Cainas Gould, dekegould@augustana.edu
27 August 2018
Philosophy Pub (BREW in the Village)

Creating Artificial Intelligence: The Ethics

Thesis: Given some anti-natalist considerations, we should not develop artificial moral agents (AMAs) that resemble us in some respects. However, given those same considerations, the development of some AMAs might be permissible, perhaps even obligatory.

Definitions:

- Artificial Intelligence (AI): the project of developing a machine that either thinks or acts either humanly, or that thinks or acts rationally (Russell and Norvig 2015).
- Artificial Moral Agents (AMAs): an artificially intelligent machine that qualifies (in some sense) as a moral agent (Allen, Varner, Zinser 2000; Floridi and Sanders 2004).
  - "full" AMAs: one end of a gradient scale possessing full autonomy and full ethical sensitivity (Wallach and Allen 2010)
- Anti-Natalism: the view that it is morally wrong to reproduce
  - "strong" anti-natalism: it is always wrong to reproduce human offspring because human existence is inherently not worth it (Benatar 2008)
  - "weak" anti-natalism: it is wrong to reproduce given the likely harms future generations will face (cf. Reider 2015)

Analogy to Procreation:

1. The decision to become a parent involves the introduction of a new conscious entity, and that involves the responsibility to ensure, to the best of one's ability, the well-being of that entity.
2. The decision, on the part of humanity, to create full AMAs involves similar responsibilities, to the extent that they are relevant.
3. If one flippantly decides to create a new human being, that person is acting immorally.
4. Therefore, we would likewise act immorally if we flippantly decided to create full AMAs.
   - the connection to parental ethics goes further…

Metzinger's argument:

1. The development of full AMAs will involve endowing those entities with consciousness.
2. If AMAs are endowed with consciousness, then AMAs will have a capacity to suffer. (2009, p195)
3. If an entity is created with the capacity to suffer, then there is a higher likelihood that overall suffering will be increased in the world.
4. We should do whatever we can to decrease overall suffering in the world.
5. Therefore, we should not develop full AMAs. (2003, p620)

Bostrom's "orthogonality thesis": "intelligence and final goals are orthogonal axes along which possible agents can freely vary." (2012 p73)

My argument:

1. An artificial (or even "superintelligent") will need not resemble our own, given varieties of combinations of motivation and intelligence (Bostrom 2012); it is possible to construct AMAs that resemble our own will, but it's also possible to construct many other varieties that do not.
2. If some version of anti-natalism is true, then if it is possible to create some agents who are likely to suffer in ways that we can, we shouldn't construct such agents.
3. Some version of anti-natalism is true: either the strong version or the weaker version gives a strong *prima facie* case.
4. If it is possible to construct many other varieties of AMA that do not resemble our own will, then it is not obligatory that we shouldn't construct AMAs at all.
5. Thus, we shouldn't create AMAs that resemble our own will, but (given careful consideration of the sort of will an AMA might have) it is permissible to create some forms of AMAs.

Consequences for Metzinger: Blissful AMAs?

- One might imagine that the gains in happiness by creating AMAs capable of positive affective states are so immeasurably high that even if the introduction of AMAs into the universe via academic research causes them to experience some suffering, it would still be optimific to bring them about.

Consequences for Anti-Natalism: AMAs with no (or severely reduced) suffering?

- Seems as though some forms of AMAs in nearby regions of Bostrom's orthogonal chart would not be subject to these considerations.
- Even stronger, this suggests that some forms of utilitarian thinking might urge that we are obligated to develop such, near-human, suffering-free AMAs.
- Perhaps these considerations should get us to consider "artificial replacement": hand over the earth to AMAs? (Shiller 2017)

Works Cited

1. Allen, Colin, Gary Varner, and Jason Zinser. (2000) "Prolegomena to Any Future Moral Agent." *Journal of Experimental and Theoretical Artificial Intelligence*. 12: 251-261.
2. Benetar, David. (2008) *Better Never to Have Been: the Harm of Coming into Existence*. Oxford University Press.
3. Bostrom, Nick. (2012) "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines*. 22: 71-85.
4. Floridi, Luciano and J.W. Sanders. (2004) "On the Morality of Artificial Agents." *Minds and Machines*. 14: 349-379.
5. Metzinger, Thomas. (2003) *Being No-One: The Self-Model Theory of Subjectivity*. MIT Press.
6. Metzinger, Thomas. (2009) *The Ego Tunnel: The Science of the Mind and the Myth of the Self.* New York: Basic Books.
7. Rieder, Travis N. (2015) "Procreation, Adoption and the Contours of Obligation." *Journal of Applied Philosophy*. 32.3: 293-309.
8. Russell, Stuart and Peter Norvig. (2015) *Artificial Intelligence: A Modern Approach, Third Edition*. Pearson.
9. Shiller, Derek. (2017 June) "In Defense of Artificial Replacement." *Bioethics*. 31.5. 393-399.
10. Wallach, Wendell and Colin Allen. (2010) *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.