

Super Ethics or Just Human?

Boström and others refer to the 'super-ethical' question as the question on how to create or program artificial moral agents that will be able to make ethical decisions in the future – and basically for the rest of time – that will be to the benefit of humans, or put humans first, if we don't know what the future or the technology of the future will be like.

Already philosophers such as Brundage, Asaro, and Boström caution against being too optimistic about notions of artificial morality, as human morality is to a huge extent still an unsolved riddle to moral philosophers. Added to this is the fundamental concern that artificial 'minds' cannot have morality as they do not have the capacity for phenomenological experience or semantic understanding of the consequences of their decisions. In this context the super ethical question seems simply impossible to even contemplate.

I caution here about a related concern that impacts both on the possibility of artificial morality and on a meaningful response to the super ethical question. It is not just that the riddle of human morality has not been solved yet. It is that humans are not the best moral agents we can imagine by a long shot. Is this really as good as we want future morality to be? Just because humans understand – all things being equal – the consequences of their actions on themselves and others? What is the best model for artificial or super ethics, really? What should it be?

Keywords

Artificial morality; super ethical question; moral reasoning