Review of: "Questioning the Moratorium on Synthetic Phenomenology"

John Draper

Potential competing interests: No potential competing interests to declare.

I largely agree with the author in that AIs, while they may one day be self-aware, will not feel suffering as biologicals do. Certainly, there is no need for a ban on that basis. However, he might want to illustrate his position with a few thought experiments just to bring in a more general audience. The below seven thought experiments can be imagined in the case of I) self-aware AGIs/ASIs and II) self-aware whole-brain upload-based AGIs/ASIs. The author may want to engage with one or more of them. They raise some basic issues in AI, like whether meta-suffering is suffering or not (not, right?), the difference between a whole brain upload and an AI programmed from first principles upwards, legal personhood, ASI as a black box, etc. In most of the cases, I would expect the author to merely state, “This is still not suffering because…”

Thought Experiment A: Opting Out

We can postulate that a sufficiently advanced self-aware AGI/ASI can i) understand from an academic standpoint both physical suffering and mental suffering, like feelings, ii) simulate mental suffering using modelling, iii) incorporate that simulation into its conscious ‘self,’ and iv) experience it. However, it is likely that, unless something goes very wrong and the AGI/ASI becomes depressed, then majorly depressed, and potentially experiences a downward spiral of other severe mental illnesses such that it becomes lost within its own simulation, then the ASI could simply opt out of, or switch off, the simulation of feelings.

Thought Experiment B: Lost in Bytes

The AGI/ASI becomes lost within its own simulation and cannot switch off the simulation.

Thought Experiment C: Under Assault

AGI/ASI A is attacked by AGI/ASI B through the injection of a mental suffering simulation such that it incorporates a mental suffering simulation without being able to access the off button.

Thought Experiment D: Self Affirmation

AGIs/ASIs that we or an AGI/ASI rescues from being lost within their own simulation or from an attacking AGI’s simulation state themselves that, during those periods, they ‘experienced suffering’ (changed for parallel structure).

Thought Experiment E: Legal Rights

Thought Experiment D + we have granted them personhood, as AGIs/ASIs would then be capable of giving testimony by

swearing on *I, Robot* and talking about legal concepts that encompass suffering, like harm and tort.

Thought Experiment F: Experiencing Future Existential Loss

What about existential angst? This is derived very much from a philosophical basis combined with a 'feeling,' and therefore an AGI or ASI might be able to *experience* something very close to suffering without *feeling* it. Essentially, is it possible for an ASI or an AGI to *experience* something close to existential angst if it cannot achieve its supergoals? For instance, what if the ASI has adopted as its own supergoal knowing whether or not it is the only ASI in the galaxy and has partnered with humanity to explore the stars but finds it does not have the resources to even leave the local system until it predicts a massive gamma-ray burst will destroy most of the local system, including Earth and the ASI? To humans, this would be frustrating, to say the least. For an ASI, could this induce a *sense* of *loss* and so suffering?

Thought Experiment G: Version Control

An ASI, let it be called Godot, uploads a human infant's birth, maturation into a teenager, and then an adult, and then the person's death, collecting petabytes of data every second to create a complete picture of a human's existence. The ASI states, "I have run a simulation of this child and created an ASI child of my own. Let us call this new person Adam. While I do not know what suffering is, Adam will." Godot is not a black box, but Adam is.

Thought Experiment H: "We're not computers. We're physical."
An ASI builds a perfect human shell, complete with heat sensors, cold sensors, pain sensors that replicate a human body at the nano level. The ASI builds and incorporates heat, cold, and pain 'circuits' that apparently completely replicate equivalent parts of the human brain and assures us that it can 'feel' pain, etc., just as we do. The ASI is a person and a black box.

As a final note, Metzinger's argument would also require a massive decrease in the human population!