# A conversation with Brian Tomasik, October 6, 2016

## Participants

- Brian Tomasik – Research Lead, Foundational Research Institute (FRI)
- Luke Muehlhauser – Research Analyst, Open Philanthropy Project

**Note**: These notes were compiled by the Open Philanthropy Project and give an overview of the major points made by Mr. Tomasik.

## Summary

The Open Philanthropy Project spoke with Mr. Tomasik of FRI as part of its investigation into which types of beings should be of moral concern, and thus a potential target for the Open Philanthropy Project's grantmaking. This conversation focused on one particular factor plausibly relevant to whether a being should be of moral concern or not — namely, whether that being is phenomenally conscious, and what the character of its conscious experience is. The conversation covered a wide range of topics, focusing on the arguments for and implications of Mr. Tomasik's brand of panpsychism. See our full report on consciousness and moral patienthood for context.

## Overview of Luke and Mr. Tomasik's views

### Points of agreement

Luke and Mr. Tomasik found that they agreed about the following:

- Physicalism and functionalism about consciousness.
- Specifically, Mr. Tomasik endorses "Type A" physicalism, as described in his article "Is There a Hard Problem of Consciousness?" Luke isn't certain he endorses Type A physicalism as defined in that article, but he thinks his views are much closer to "Type A" physicalism than to "Type B" physicalism.
- Consciousness will likely turn out to be polymorphic, without a sharp dividing line between conscious and non-conscious systems, just like (say) the line between what does and doesn't count as "face recognition software."
- Consciousness will likely vary along a great many dimensions, and Luke and Mr. Tomasik both suspect they would have different degrees of moral caring for different types of conscious systems, depending on how each particular system scores along each of these dimensions.

### A core disagreement

In Luke's view, a system needs to have certain features interacting in the right way in order to qualify as having non-zero consciousness and non-zero moral weight (if one assumes consciousness is necessary for moral patienthood).

In Mr. Tomasik's view, various potential features (e.g. ability to do reinforcement learning or meta-cognition) contribute different amounts to a system's degree of consciousness, because they increase that system's fit with the "consciousness"

concept, but *all* things have non-zero fit with the “consciousness” concept.

Luke suggested that this core disagreement stems from the principle described in Mr. Tomasik's "Flavors of Computation are Flavors of Consciousness":

> It's unsurprising that a type-A physicalist should attribute nonzero consciousness to all systems. After all, "consciousness" is a concept — a "cluster in thingspace" — and all points in thingspace are less than infinitely far away from the centroid of the "consciousness" cluster. By a similar argument, we might say that *any* system displays nonzero similarity to *any* concept (except maybe for strictly partitioned concepts that map onto the universe's fundamental ontology, like the difference between matter vs. antimatter). Panpsychism on consciousness is just one particular example of that principle.

For example, Luke's view would deny that an "electron" is, even to some very small degree, "furniture" — not because the concept of "furniture" is sharply defined, but because the fuzzy borders of the "furniture" concept become indistinguishable from zero very far in concept-space from "electron” (at least, for the purposes of moral caring).

On Mr. Tomasik's view, the "degree of fit" that a given entity has with a concept corresponds to the inverse distance between it and the concept's "centroid" in a many-dimensional concept-space. For example, a human might have a 99% fit with “conscious” and 1% fit with “non-conscious,” whereas an electron may fit "conscious" on the order of .01% (i.e. very close to a prototypical “non-conscious” thing). Similarly, a turtle might have .1% fit with the concept of "mammal." In Mr. Tomasik's view, every concept is a finite distance from every other concept's centroid; i.e., every concept "fits" every other concept to some non-zero degree. Therefore, if one assigns moral weight to an entity based on the degree to which it fits the "consciousness" concept, one will end up assigning non-zero moral weight to every entity.

This would remain the case if moral concern were based on some other criterion; e.g., if one's pre-theoretical moral intuitions valued "life" (rather than "consciousness") as the primary criterion for moral patienthood, Mr. Tomasik's position also claims that electrons fit the concept of "life" to some tiny degree.

**Example of contemporary brutalist architecture in Ukraine vs. sea slugs**

Luke finds this way of using concepts strange, in part because it would be unusual to use concepts this way in other domains.

For example, Luke notes that it would be unusual to suggest to someone with an affinity for “contemporary brutalist architecture in Ukraine” (hereafter, "CBAU") that sea slugs *are* CBAU "at least a little bit," and that therefore they ought to have at least some affinity for sea slugs. This would be unusual both as a way of using words and as a way of choosing how to spread one’s affinity over things in the world. Luke’s moral intuitions operate similarly to his intuitions about useful talk and action in other domains; just as Luke would not extend an affinity for CBAU to sea slugs merely because they can be modeled as being non-infinitely separated in a multi-

dimensional concept-space, he does not spread his moral concern for conscious beings to electrons.

Mr. Tomasik's view, Luke suggests, amounts to pansychism about consciousness as an uninformative special case of "pan-everythingism about everything."

**Relevance of entities with low fit to a concept in the case of maximizers**

Mr. Tomasik agrees that, in ordinary communication, we refer to objects using only the nearest concept centroid (or thereabouts). However, Mr. Tomasik thinks that "edge cases" of entities with very low (but non-zero) fit with a particular concept become more relevant when considering the decision-making processes of "maximizing" agents. This is particularly the case in situations where the extremely large quantity of some entity (e.g. electrons) might counterbalance, in a maximizer's weighting function, the low degree to which that entity fits a particular concept. For example, a table maximizer would need some way of deciding to what degree an electron "counts" as a table. If an electron counts as a table even to some extremely small degree, the maximizer would need to decide whether the abundance of electrons, relative to all of the ordinary tables it could possibly create, leads it to prefer the proliferation of electrons to ordinary tables.

This is also the sense in which Mr. Tomasik thinks that electrons have a very small amount of moral weight for him (though, he does not have a numerical estimate). While a single electron would never be relevant for any practical moral issue, the combined moral weight of the roughly 10^80 electrons in the universe might be morally relevant.

Mr. Tomasik suggests that a CBAU-maximizer might favor sea slugs if there happened to be a massive number of real or potential sea slugs in the universe. Mr. Tomasik also claims that a CBAU-maximizer would put more weight on sea slugs than on things further from CBAU in concept-space than sea slugs (e.g. electrons), depending on its algorithm for carving up concept-space and maximizing a utility function that refers to its concepts.

While we would not expect a table maximizer to care about things that are conceptually very far from tables, Mr. Tomasik suggests that this is only because the maximizer is not faced with an extreme trade-off. If a table maximizer had to choose between one table and a universe of $3\uparrow\uparrow\uparrow 3$ electrons (see: Donald Knuth's up-arrow notation), it is less clear which it should favor, since the tiny amount of "table-ness" of each electron could, in aggregate, outweigh the single ordinary table. In reality, such an extreme trade-off might not arise.

**Shape of Mr. Tomasik's "moral care function"**

Mr. Tomasik's moral intuitions seem to roughly fit a "moral care function" that gives more weight to particular types of complex systems. Depending on how steeply the moral weight that Mr. Tomasik's view assigns to an entity drops off as complexity decreases, it is possible that Mr. Tomasik's model might produce object-level views similar to Luke's, particularly if it were truncated wherever it fell below some very

small amount (rather than failing to ever reach zero). Where Luke's model assigns zero moral weight, Mr. Tomasik's might assign moral weight low enough to be treated as zero for practical purposes.

## Difficulty of adjudicating moral intuitions

Moral intuitions are difficult to adjudicate because there may be no experiment or investigation that could discriminate between conflicting views. However, if Luke and Mr. Tomasik were both conceptualizing their reflective moral judgments in terms of their “idealized” or “extrapolated” values, and if they were able to perform some piece of an agreed-upon extrapolation procedure (e.g. learning many more true facts) and observe the direction that their intuitions shifted, that might provide a small amount of evidence that could discriminate between their conflicting moral intuitions.

## Potential methods for estimating the moral weight of a program

Mr. Tomasik is not sure what the best method is for assessing how much moral concern to have for a given computer program, on his view. One potential metric is the size of the program. For example, one might compare a program's complexity in terms of lines of code or number of steps to the complexity required to implement a neuromorphic general-purpose artificial intelligence system. This ratio could be used as a very rough estimate of the program’s relative degree of consciousness. Qualitative shifts in the program's function would also be relevant.

Similarly, on Mr. Tomasik’s view, comparing the number of neurons (or the square root of the number of neurons) in insects and humans may be one rough way of estimating their relative degree of consciousness.

In Luke's view, the functional operations of neurons or steps in a program are critical. For example, Luke would not morally care about a system that used lots of cycles and computation to use a giant look-up table to approximate the input and output behavior of a human brain, despite it involving far more computational steps than a human brain, and displaying the same external behavior. Similarly, Luke would need a better functional understanding of an insect brain to know whether his moral concern extends to insects.

## "Parliamentary method" for uncertainty among theories

Mr. Tomasik puts some weight on theories of moral concern besides the one he defended in this conversation. He finds the possibility that an unusual concern like suffering in fundamental physics could dominate his moral considerations intuitively unappealing, and to some extent he would like to be able to limit his concern to, e.g., complex vertebrates. For practical purposes, Mr. Tomasik balances his uncertainty using something like Professor Nick Bostrom’s "parliamentary method" of compromise between theories, assigning some weight to the different intuitions produced by different arguments or intuition pumps.

*All Open Philanthropy Project conversations are available at*
*http://www.openphilanthropy.org/research/conversations*