Anna Rogers: A sanity check on emergent properties

One of the frequent points in the mainstream narrative about large language models is that they have "emergent properties" (sometimes even dangerous enough to be considered existential risk to mankind). However, there is much disagreement about even the very definition of such properties. If they are understood as a kind of generalization beyond training data - as something that a model does without being explicitly trained for it - I argue that we have not in fact established the existence of any such properties, and at the moment we do not even have the methodology for doing so.