WHEN SHOULD TWO MINDS BE CONSIDERED VERSIONS OF ONE ANOTHER?

BEN GOERTZEL

Novamente LLC
1405 Bernerd Place, Rockville MD 20851
ben@goertzel.org

Received 21/ 12 / 2012

What does it mean for one mind to be a different version of another one, or a natural continuation of another one? Or put differently: when can two minds sensibly be considered versions of one another? This question occurs in relation to mind uploading, where one wants to be able to assess whether an approximate upload constitutes a genuine continuation of the uploaded mind or not. It also occurs in the context of the rapid mental growth that is likely to follow mind uploading, at least in some cases – here the question is, when is growth so rapid or discontinuous as to cause the new state of the mind to no longer be sensibly considerable as a continuation of the previous one? Provisional answers to these questions are sketched, using mathematical tools drawn from category theory and probability theory. It is argued that if a mind's growth is "approximately smooth", in a certain sense, then there will be "continuity of self" and the mind will have a rough comprehension of its growth and change process as it occurs. The treatment is somewhat abstract, and intended to point a direction for ongoing research rather than as a definitive practical solution. These ideas may have practical value in future, however, for those whose values favor neither strict self-preservation nor unrestricted growth, but rather growth that is constrained to be at least quasi-comprehensible to the minds doing the growing.

Keywords: mind uploading, self, category theory

1. Introduction

The comparison of one mind – a certain cognitive system, existing over a certain interval of time – with another is a tricky business. In everyday life, we conventionally assume that the various minds associated with the same physical body during that body's lifetime are "the same" – i.e. are different versions of the same mind. But consideration of advanced technologies like mind uploading, brain computer interfacing and artificial general intelligence (AGI) forces us to go beyond this sort of conventional understanding, and craft a more fundamental conceptualization of what it means for two minds to be sensibly considerable as "different versions of the same mind."

If a person's mind is uploaded via, say, creating an atomically precise digital computer simulation of the person's brain, then there's not much of an issue here. Pretty clearly, the uploaded mind is a different version of the original biologically-based mind. There are philosophical issues here regarding consciousness and identity, which have been much discussed in

the literature (for recent discussions see e.g. Agar, 2001; Shores, 2011; Walker, 2011), but in this paper we will purposefully sidestep these, focusing on questions regarding the substantive contents and the dynamics of minds. Consciousness and uploading is a worthy topic but has been treated very thoroughly elsewhere.

On the other hand, approximate uploads raise thornier questions. What if, as Martine Rothblatt (2011) and Bill Bainbridge (2011) have suggested, an "upload" of a person were created based on data about them such as questionnaires they answered, texts they wrote, phone calls they made, and videos of them going about their lives? What if such a "weak upload" were made by a very intelligent AI with the goal of creating a system operating within the constraints of human brain structure and dynamics, that would give rise to a mind reasonably likely to produce the behaviors constituted by the data provides? Would this just be a simulacrum of the "uploaded" person? Or would it be the real thing? How could the difference between these possibilities be rationally assessed?

In the context of ordinary human life, we take for granted that when we wake up in the morning, we are the same person as – a new instance of the same mind as – the person who occupied our body when we went to sleep the previous night. We even take for granted that we are the same person as we were 40 years ago – although the similarities between my current 45 year old self and my previous 5 year old self are arguably not that large. We assume that continuity of body implies continuity of mind, and this assumption generally works OK in practical life situations. But, once mind uploading has become feasible, and once radical intentional mind and body modification become feasible, such simplistic assumptions won't work so well anymore. Suppose one could increase their intelligence by a factor of 1000 overnight – would the morning version of "oneself" sensibly be describable as a new version of the late-night version of "oneself", as opposed to a basically different mind? Suppose you took the left half of your brain and paired it with the right half of George W. or Barbara Bush's brain – to what extent would the result still be "you"?

In this paper I suggest a "mind-mind correspondence principle" that I believe resolves these issues – at least in theory! A formal statement of the principle requires some mathematics, which I'll give in a later section, but for right now I'll state an informal version.

A key aspect of the treatment given here is its *abstract* nature. In the perspective presented, mind is not about the particular entities of which a system is composed, but rather about the patterns by which these entities are arranged. Thus, the correspondence between two minds is treated using abstract mappings between state-transition sequences associated with different minds – without worrying about the specific contents of these mind-states, let alone the physical or digital substrate via which these contents and transitions are realized. Category theory is used, because it is the branch of mathematics that deals most elegantly with mappings and their properties, without making any commitments regarding the underlying entities being mapped.

For the present purposes, I consider a "mind" to be a series of states of an intelligent system, where intelligence is conceived as the manifested ability to achieve complex goals in complex environments. So, for a state-series to be a mind, it must be possible for a suitably intelligent observer to infer from the state-series, the action of a system working to achieve some complex goals in a complex environment. In this framework, a crude, informal version of the Mind-Mind Correspondence Principle would be:

MIND-MIND CORRESPONDENCE PRINCIPLE (quasi-formal version):

- ***For two minds M_1 and M_2 to be considered close instances of one another, there should be a "nice" mapping from sequences of the first mind's states into sequences of the second mind's states -- where "nice" means that a mind-state-sequence S_1 in M_1 composed by sequencing together two subsequences S_{11} and S_{12}, gets mapped (within a close degree of approximation) into a mind-state-sequence S_2 in M_2 composed of sequencing together two corresponding subsequences S_{21} and S_{22}***
- ***For two minds M_1 and M_2 to be considered distant instances of one another, there should exist in reality (at various time points) intervening minds $M_1 = M_{i(1)}$, … , $M_{i(n)} = M_2$ so that: for all $k=1,\ldots,n$, it holds that $M_{i(k)}$ and $M_{i(k+1)}$ are close instances of one another, and $M_{i(k)}$ immediately temporally precedes $M_{i(k+1)}$***

As an example of the subsequences mentioned above, suppose that for mind M_1 , the process of remembering a person's name (S_1) often involves a two-stage process: first summoning up that person's face to memory

(S_{11}); then searching the memory for cases where that face was associated with some name (S_{12}). Then a nice mapping from M_1 into M_2 might involve mapping M_2's process of remembering a person's name (S_2) into two subprocesses: one (S_{21}) corresponding to S_{11} and involving summarizing a person's face to memory; and the next (S_{22}) corresponding to S_{12} and involving trying to remember a name associated with that face. The specific dynamics and structures associated with the corresponding subprocesses might or might not be similar between the two minds. To the extent that M_1's mind-state sequences can be decomposed into subsequences that naturally map onto subsequences of M_2's corresponding mind-states, in the manner of this example, we may say there is M_1 and M_2 are close instances.

Note that, for the "distant instance" relation to hold, the intervening minds must actually exist in reality. Otherwise all minds would be distant instances of each other, because one can morph any mind into any other mind via a series of small steps. The idea is that one mind is a distant instance of another if it has actually, in reality, been morphed into or out of the other via a series of small steps.

These notions are related, I suspect, to the notion of "continuity of self". If one mind progresses to another continuously, so that the latter is a distance instance of the former in the above sense, then it will likely be the case that the mind at each step during the progression is able to incorporate its predecessors and successors into its self-modeling process in an subjectively meaningful way – thus, in a sense, "owning its own growth process."

The treatment here is fairly abstract, but the ultimate goal of the theory presented is practical application. To practically apply the ideas presented, one would need a tool mapping minds into abstract structures like state-transition graphs. This is infeasible for human brains at present, due to limitations of brain imaging technology; and also infeasible (though less so) for complex AI systems at present, due to limitations of scalable real-time pattern recognition technology such as one would need to abstract such structures from a large, complex, rapidly-changing system. But it's not implausible to imagine that appropriately constructed AI systems may in future be able to carry out useful approximate modeling of complex minds as state-transition graphs, and thus enable the pragmatic application of the ideas described here. AGI and mind uploading may result in minds allowing us to study the nature of AGIs and mind uploads

using the conceptual tools presented here, and others (including, almost surely, better tools of their own invention).

Exactly how interesting these ideas are, depends on one's value system. If one values both radical mental growth, and the quasi-comprehensibility of this growth as it proceeds, then it's interesting to explore the ways and senses in which this might be achievable. If exact uploading proves infeasible or very difficult, whereas approximate uploading proves easier, then some personal value will attach to the question of what an approximate upload means in terms of cognitive theory.

2. A Category-Theoretic Model of Mind

Now I will introduce some formalism, aimed at representing minds in a manner that renders the ideas discussed above formally addressable. At this stage of development of the theory proposed here, mathematics is used mainly as a device to ensure clarity of expression. However, once the theory is further developed, it may possibly become useful for purposes of calculation as well.

Suppose one has any system S (which could be an AI system, or a human, or an environment that a human or AI is interacting with, or the combination of an environment and a human or AI's body, etc.). One may then construct an uncertain **transition graph** associated with that system S, in the following way:

- The **nodes** of the graph represent **fuzzy sets of states** of system S (I'll call these "state-sets" from here on, leaving the fuzziness implicit)
- The (directed) **links** of the graph represent **probabilistically weighted transitions** between state-sets

Specifically, the weight of the link from A to B should be defined as

$$\text{Prob}(o(S,A,t(T)) \mid o(S,B,T))$$

where

$$o(S,A,T)$$

denotes the presence of the system S in the state-set A during time-distribution T, and t() is a temporal succession function defined so that

t(T) refers to a time-distribution conceived as "after" T. A time-distribution is a probability distribution over time-points.*

An intelligent system may, as a working definition, be considered as a system that achieves complex goals in relation to complex environments (Goertzel, 2006, 2010). It's not hard to ties this understanding of intelligence into the transition-graph framework described above. Suppose one has a transition graph corresponding to an environment; then a goal relative to that environment may be defined as a particular node in the transition graph. The goals of a particular system acting in that environment may then be conceived as one or more nodes in the transition graph. The system's situation in the environment at any point in time may also be associated with one or more nodes in the transition graph; then, the system's movement toward goal-achievement may be associated with paths through the environment's transition graph leading from its current state to goal states.

Note, it may be useful for some purposes to filter the uncertain transition graph into a **crisp transition graph** by placing a threshold on the link weights, and removing links with weights below the threshold.

Now one may look at the space of **mind-paths** associated with a given mind. A mind-path is a path through the transition graph associated with a given intelligent system. Given two mind-paths P and Q, it's obvious how to define the composition P*Q – one follows P and then, after that, follows Q, thus obtaining a longer path. In category theory terms, we are constructing the free category associated with the graph: the objects of the category are the nodes, and the morphisms of the category are the paths.

3. Mappings Between Minds

Now I will bring this mathematics to bear on the problem of mind uploading and the continuity of mental evolution, via explaining how to use the above formalism to describe mappings between different minds.

* The interaction of fuzziness and probability here is fairly straightforward, and I'm suggesting to handle it here the way it's done in PLN (Goertzel et al, 2008). Note that the definition of link weights is dependent on the specific implementation of the temporal succession function, which includes an implicit time-scale.

Suppose one has two different minds, represented by two mathematical categories as described above. Then, a functor F between one mind-category and another is a mapping that preserves object identities and so that

$$F(P * Q) = F(P) * F(Q)$$

We may also introduce the notion of an **approximate functor**, meaning a mapping F so that the average of

$$d(F(P * Q) , F(P) * F(Q))$$

is small.

One can introduce a prior distribution into the average here. This could be the Levin universal distribution (Hutter, 2005) or some variant (the Levin distribution assigns higher probability to computationally simpler entities). Or it could be something more purpose specific: for example, one can give a higher weight to paths leading toward a certain set of nodes (e.g. goal nodes). Or one can use a distribution that weights based on a combination of simplicity and directedness toward a certain set of nodes. The latter seems particularly interesting, and I will define a **goal-weighted approximate functor** as an approximate functor, defined with averaging relative to a distribution that balances simplicity with directedness toward a certain set of goal nodes.

The move to approximate functors is simple conceptually, but mathematically it's a fairly big step, because it requires us to introduce a geometric structure on our categories. But there are plenty of natural metrics defined on paths in graphs (weighted or not), so there's no real problem here. There are also some interesting links with topos theory, which I have not yet carefully elaborated.

4. The Mind-Mind Correspondence Principle

Now we finally have the formalism set up to make a non-trivial statement about the relationship between different minds. Namely, the hypothesis that:

MIND-MIND CORRESPONDENCE PRINCIPLE:

- ***For two minds M_1 and M_2 to be considered close instances of one another, there should be an approximate functor (with a high degree of approximation) mapping M_1 into M_2***
- ***For two minds M_1 and M_2 to be considered distant instances of one another, there should exist in reality (at various time points) intervening minds $M_1 = M_{i(1)}$, … , $M_{i(n)} = M_2$ so that: for all k=1,…,n, it holds that $M_{i(k)}$ and $M_{i(k+1)}$ are close instances of one another, and $M_{i(k)}$ immediately temporally precedes $M_{i(k+1)}$***

Comparing with the informal statement of the same principle given in the Introduction, one sees that the formal statement is more concise, and differs only via introducing the concept of a functor, in place of the ill-elaborated phrase "nice mapping" used in the informal statement.

That is, a little more loosely: the hypothesis is that,

- *for two minds to be close instances, there has to be a natural correspondence between the transition-sequences of the mind-states of one mind and the corresponding transition-sequences of the mind-states of the other. If one wishes one can restrict or bias this toward -sequences leading to relevant goals.*
- *For two minds to be distant instances, there must have been an actually realized sequence of minds leading from the former to the latter, where each mind in the sequence is a close instance of the minds immediate before and after it in the sequence.*

5. The Self-Continuity Principle

I suspect that there is a close connection between this notion of a "distance instance" and the notion of "continuity of self", presented in (Goertzel, 2006). In short, "continuity of self" means that, as a mind grows and changes, at each step it maintains a self-model that includes a model of its current state, and a model of its recent past and near future state. The model of the near-future state need not be wholly accurate (in fact this would be an odd case), but the growth process must allow the mind to compare its current self to its prior conjectural model of what its current self was expected to look like, and integrate the results of this comparison into its self-model, in a way that would have had a reasonable degree of meaning to its prior self, in the context of its prior self-model. Put less formally, this means that the mind is able to "own" its growth and change

process – to model the process as it occurs, and model its own relationship to the process along the way. This modeling may be approximate and in some regard inaccurate, but it must be linked meaningfully (in the system's view) to the system's intelligence (its achievement of goals in its environment).

What kind of growth process would not display continuity of self? Death, for example! Or, suddenly and incredibly rapidly becoming 1000 times more intelligent than one was previously. In the latter case, after the intelligence increase, the only way to integrate the changes into one's self-model, would be using concepts and methods that would have been completely alien to one's prior self. The mind would not be able to feel itself becoming more intelligent, and integrate each step of the process into its self-model in a subjectively meaningful way; rather, it would suddenly become incredibly more intelligent, and as a consequence suddenly become a totally different mind.

Self-continuity is a thornier notion to define than the category-theoretic inter-mind relations described above. However, given a more advanced mathematics of self-modeling, it might be possible to formally prove a connection between these various notions – say, something along the lines of a

SELF-CONTINUITY PRINCIPLE:

Suppose one has a sequence of minds that are all actively engaged in self-modeling. This sequence will realize a "distant instance" relationship between the beginning and end of the sequence (with a reasonable degree of approximation), iff all the minds along the path display continuity of self.

In other words, if one is dealing with minds that have selves, then the only way to achieve a path of minds that are smoothly morph-able into each other, is to create a path where the self at each point in time can effectively model the selves at closely previous and subsequent points in time. This is not quite starkly obvious but seems plausible, and I suspect that careful precisiation of the terms involved could produce a rigorous truth along these lines.

6. How Might These Ideas Be Useful?

Suppose one believes the Mind-Mind Correspondence and Self-Continuity Principles as laid out above – so what?

The answer to this depends on one's values. One may adopt many different value systems, in relation to the possibilities of mind uploading and radical self-modification. At two extremes, one might:

- Value strict self-preservation, in the sense of preserving the existence of a mind quite similar to one's current mind (regardless of how the environment may change in future, or the novel neural-upgrade possibilities the future may provide)
- Dismiss the value of self-preservation, valuing rather the creation of amazing or intelligent new forms, whether or not they have any particular continuity with one's current self or preoccupations

In either of these cases, the concerns of the present paper are largely irrelevant. However, there is an alternate attitude one may adopt, something like

Value growth wildly beyond one's current self, but also value having a growth process that is gradual enough so that, at each stage, one's mind can appreciate and somewhat understand the nature of the growth process, and the new mind it is about to become

If one adopts a value system that prizes "semi-comprehensible growth" of this nature, then the notions discussed here become quite relevant. For they constitute an articulation of what "gradual enough" means. Gradual enough is, I suggest, gradual enough that one's current self and one's future selves are "distant instances" in the sense described here. Gradual enough is gradual enough that there is some reasonable degree of continuity of self between each of one's future minds, and the new mind that it spawns.

Of course, it's highly possible that, if I do succeed in gradually but massively increasing my intelligence, I will eventually reach a point where the notions in this paper seem absurd – and, for reasons I am now incapable of understanding, the notions of continuity of self and distance instances will seem ridiculously irrelevant and childlike or worse! There's no way to confidently avoid this sort of eventuality. The best one can do is to chart a path forward that seems sensible and valuable according to one's current capability to understand.

Acknowledgments

Thanks are due to Nil Geisweiller for comments on an earlier version of this paper.

References

Agar, Nicholas (2011). Ray Kurzweil and Uploading: Just Say No!, Journal of Evolution and Technology 22(1) , pp. 23-36

Bainbridge, Bill (2011). This volume.

Goertzel, Ben (2010). Toward a Formal Characterization of Real-World General Intelligence. Proceedings of AGI-10. http://agi-conf.org/2010/wp-content/uploads/2009/06/paper_14.pdf

Goertzel, Ben, Matthew Ikle', Izabela Freire Goertzel and Ari Heljakka (2008). Probabilistic Logic Networks. Springer.

Hutter, Marcus (2005). Universal AI. Springer.

Rothblatt, Martine (2011). This Volume.

Shores, Corry (2011). Misbehaving Machines: The Emulated Brains of Transhumanist Dreams, Journal of Evolution and Technology 22(1), pp. 10-22

Walker, Mark (2011). Personal Identity and Uploading, Journal of Evolution and Technology 22(1), pp. 37-52