

The Mind-World Correspondence Principle
(Toward a General Theory of General Intelligence)

Ben Goertzel
September 17, 2011

Abstract. *A novel “Mind-World Correspondence Principle” is proposed – which, given an environment and goal-set, heavily constrains the structure of any intelligent system capable of efficiently achieving those goals in that environment. This is seen as a first step toward a “general theory of general intelligence.”*

An approximate gloss of the proposed principle is: “For a mind to work intelligently toward certain goals in a certain world, there should be a nice mapping from goal-directed sequences of world-states into sequences of mind-states, where “nice” means that a world-state-sequence W composed of two parts $W1$ and $W2$, gets mapped into a mind-state-sequence M composed of two corresponding parts $M1$ and $M2$.”

The principle is formulated using the mathematical language of category theory, but refinement of the principle into a precise theorem is left for later works. Discussion is given regarding the use of the principle to explain common properties of real-world intelligences such as the presence of hierarchical structure. The hope is articulated that the principle will eventually be useful for deriving and refining practical designs for Artificial General Intelligence.

Introduction

The human mind is not the only possible system with a high level of general intelligence. Various other animal minds are quite different, and artificially intelligent minds may be yet more different. The space of possible minds seems quite vast, and there is no obvious way to categorize the space.

A clue, however, is found in the fact that real-world minds are always adapted to certain classes of environments and goals. Even a system of vast general intelligence, subject to real-world space and time constraints, will necessarily be more efficient at some kinds of learning than others. Thus, one approach to formulating a general theory of general intelligence is to look at the relationship between minds and worlds – where a “world” is conceived as an environment and a set of goals defined in terms of that environment.

In this paper we move toward a general theory of general intelligence by formulating a specific principle binding together worlds and the minds that are

intelligent in these worlds. A careful statement of the principle requires introduction of a number of technical concepts, and will be given later on in the paper. A crude, informal version of the principle would be:

MIND-WORLD CORRESPONDENCE-PRINCIPLE: For a mind to work intelligently toward certain goals in a certain world, there should be a nice mapping from goal-directed sequences of world-states into sequences of mind-states, where “nice” means that a world-state-sequence W composed of two parts $W1$ and $W2$, gets mapped into a mind-state-sequence M composed of two corresponding parts $M1$ and $M2$.

What’s nice about this principle is that it relates the decomposition of the world into parts, to the decomposition of the mind into parts.

The level of presentation here is “semi-formal.” The paper doesn’t contain any mathematical theorems – but it does indicate what kind of theorem could be proved to make the theory fully precise.

The treatment here is fairly abstract, but the ultimate goal of the theory presented is practical application. My point in pursuing the general theory of general intelligence is not pure theoretical interest, but rather an intuition that it can help in terms of creating new Artificial General Intelligence (AGI; Goertzel and Pennachin, 2001) systems, and sculpting and shaping existing ones.

What Might a General Theory of General Intelligence Look Like?

In a recent conference talk (Goertzel, 2011), I suggested the need for a “general theory of general intelligence” confronting the two questions

- How does one design a world to foster the development of a certain sort of mind
- How does one design a mind to match the particular challenges posed by a certain sort of world?

Toward this end, I suggested the creation of a theory that, given a description of an environment and some associated goals, would output a description of the structure and dynamics that a system should possess to be intelligent in that environment relative to those goals, using limited computational resources.

There is already a mathematical theory of general intelligence (Hutter, 2011), but currently it gives useful conclusions only about general intelligences with infinite or infeasibly massive computational resources. On the other hand, my suggestion regards the creation of a theory of real-world general intelligences utilizing realistic amounts of computational power, but still possessing general intelligence comparable to human beings or greater.

This reflects a vision of intelligence as largely concerned with adaptation to particular classes of environments and goals. This may seem contradictory to the notion of “general” intelligence, but I think it actually embodies a realistic understanding of general intelligence. Maximally general intelligence is not pragmatically feasible; it could only be achieved using infinite computational resources (Hutter, 2005). Real-world systems are inevitably limited in the intelligence they can display in any real situation, because real situations involve finite resources, including finite amounts of time. One may say that, in principle, a certain system could solve *any* problem given enough resources and time – but, even when this is true, it’s not necessarily the most interesting way to look at the system’s intelligence. I think it’s more important to look at what a system can do given the resources at its disposal in reality. And this perspective leads one to ask questions like the ones I posed above: which bounded-resources systems are well-disposed to display intelligence in which classes of situations?

As I’ve noted in a prior paper (Goertzel, 2010), one can assess the generality of a system’s intelligence via looking at the entropy of the class of situations across which it displays a high level of intelligence (where “high” is measured relative to its total level of intelligence across all situations). A system with a high generality of intelligence will tend to be roughly equally intelligent across a wide variety of situations; whereas a system with lower generality of intelligence will tend to be much more intelligent in a small subclass of situations, than in any other. The definitions given in (Goertzel, 2010) embody this notion in a formal and quantitative way.

If one wishes to create a general theory of general intelligence according to this sort of perspective, the main question then becomes how to represent goals/environments and systems in such a way as to render transparent the natural correspondence between the specifics of the former and the latter, in the context of resource-bounded intelligence. This is the business of the next section.

Steps Toward A (Formal) General Theory of General Intelligence

Now begins the formalism. 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

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

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.

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 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.

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.

The next concept to introduce is the **world-mind transfer function**, which maps world (environment) state-sets into organism (e.g. AI system) state-sets in a specific way. Given a world state-set W, the world-mind transfer function M maps W into various organism state-sets with various probabilities, so that we may say: **M(W) is the probability distribution of state-sets the organism tends to be in, when its environment is in state-set W.** (Recall also that state-sets are fuzzy.)

Now one may look at the spaces of **world-paths** and **mind-paths**. A world-path is a path through the world’s transition graph, and a mind-path is a path through the organism’s transition graph. Given two world-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. Similarly for mind-paths.

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. And category theory is the right way to be thinking here – we want to be thinking about the relationship between the world category and the mind category.

The world-mind transfer function can be interpreted as a mapping from paths to subgraphs: Given a world-path, it produces a set of mind state-sets, which have a number of links between them. One can then define a **world-mind path transfer function** $M(P)$ via taking the mind-graph $M(\text{nodes}(P))$, and looking at the highest-weight path spanning $M(\text{nodes}(P))$. (Here $\text{nodes}(P)$ obviously means the set of nodes of the path P .)

A functor F between the world category and the mind category 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 most 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 haven't thought about much.

The Mind-World Correspondence Principle

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

MIND-WORLD CORRESPONDENCE PRINCIPLE: For an organism with a reasonably high level of intelligence in a certain world, relative to a certain set of goals, the mind-world path transfer function is a goal-weighted approximate functor

That is, a little more loosely: the hypothesis is that, *for intelligence to occur, there has to be a natural correspondence between the transition-sequences of world-states and the corresponding transition-sequences of mind-states, at least in the cases of transition-sequences leading to relevant goals.*

I suspect that a variant of the above proposition can be formally proved, using the definition of general intelligence given in (Goertzel, 2010) and mentioned above. The proof of a theorem corresponding to the above would, I think, be a great start toward a general formal theory of general intelligence. Note that proving anything of this nature would require some attention to the time-scale-dependence of the link weights in the transition graphs involved.

A formally proved variant of the above proposition would be in short, a ***MIND-WORLD CORRESPONDENCE THEOREM.***

Recall that at the start of the paper, I expressed the same idea as:

MIND-WORLD CORRESPONDENCE-PRINCIPLE: For a mind to work intelligently toward certain goals in a certain world, there should be a nice mapping from goal-directed sequences of world-states into sequences of mind-states, where "nice" means that a world-state-sequence W composed of two parts $W1$ and $W2$, gets mapped into a mind-state-sequence M composed of two corresponding parts $M1$ and $M2$.

That is a reasonable gloss of the principle, but it's clunkier and less accurate, than the statement in terms of functors and path transfer functions, because it tries to use only common-language vocabulary, which doesn't really contain all the needed concepts.

How Might the Mind-World Correspondence Principle Be Useful?

Suppose one believes the Mind-World Correspondence Principle as laid out above – so what?

My hope is that the principle can be useful in actually figuring out how to architect intelligent systems biased toward particular sorts of environment. And of course,

this is said with the understanding that any finite intelligence must be biased toward some sorts of environment.

One next step, aside from (and potentially building on) full formalization of the principle, would be an exploration of real-world environments in terms of transition graphs. What properties do the transition graphs induced from the real world have?

One such property, I suggest, is successive refinement. Often the path toward a goal involves first gaining an approximate understanding of a situation, then a slightly more accurate understanding, and so forth – until finally one has achieved a detailed enough understanding to actually achieve the goal. This would be represented by a world-path whose nodes are state-sets involving the gathering of progressively more detailed information.

Via pursuing to the mind-world correspondence property in this context, I believe we will find that world-paths reflecting successive refinement correspond to mind-paths embodying successive refinement. This will be found to relate to the hierarchical structures found so frequently in both the physical world and the human mind-brain. Hierarchical structures allow many relevant goals to be approached via successive refinement, which I believe is the ultimate reason why hierarchical structures are so common in the human mind-brain.

Another next step would be exploring what mind-world correspondence means for the structure and dynamics of a limited-resources intelligence. If an organism O has limited resources and, to be intelligent, needs to make

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

high for particular world state-sets A and B, then what's the organism's best approach? Arguably, it should represent M(A) and M(B) internally in such a way that very little computational effort is required for it to transition between M(A) and M(B). For instance, this could be done by coding its knowledge in such a way that M(A) and M(B) share many common bits; or it could be done in other more complicated ways.

If, for instance, A is a subset of B, then it may prove beneficial for the organism to represent M(A) physically as a subset of its representation of M(B).

Pursuing this line of thinking, one could likely derive specific properties of an intelligent organism's internal information-flow, from properties of the environment and goals with respect to which it's supposed to be intelligent.

This would allow us to achieve the holy grail of intelligence theory as I understand it: given a description of an environment and goals, to be able to derive an

architectural description for an organism that will display a high level of intelligence relative to those goals, given limited computational resources.

While this “holy grail” is obviously a far way off, what I’ve tried to do here is outline a clear mathematical and conceptual direction for moving toward it. In my own mind, at least, the idea is a lot less vague and wishy-washy than it was yesterday.

Conclusion

The Mind-World Correspondence Principle presented here goes a fairly long way today toward fleshing out the concept of a general theory of general intelligence. But obviously the theory is still rather abstract, and also not completely rigorous. So there’s a lot more work to be done.

The Mind-World Correspondence Principle as articulated above is not quite a formal mathematical statement. It would take a little work to put in all the needed quantifiers to formulate it as one, and it’s not clear the best way to do so – the details would perhaps become clear in the course of trying to prove a version of it rigorously.

On the other hand, as *philosophy of mind*, I believe the Principle is fairly articulately formulated, compared to most philosophical theories. One could interpret the ideas presented here as a philosophical theory that hopes to be turned into a mathematical theory and to play a key role in a scientific theory.

Most importantly, my hope is that a few years from now, a more refined version of these ideas will be useful to me and others in refining practical aspects of real-world AGI systems.

References

- Goertzel, Ben and Cassio Pennachin (2001). Artificial General intelligence. Springer.
- Goertzel, Ben (2011). Designing Worlds and Minds. Talk at "Humanity+ @ Parsons: Transhumanism Meets Design", June 2011.
- 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.