The Hidden Pattern

A Patternist Philosophy of Mind

Ben Goertzel

BrownWalker Press
Boca Raton • 2006

The Hidden Pattern: A Patternist Philosophy of Mind

BrownWalker Press
Boca Raton, Florida
USA • 2006

ISBN: 1-58112-989-0 (paperback)
ISBN: 1-59942-404-5 (ebook)

BrownWalker.com

Contents

Preface v
Acknowledements ix

1. Meta-Philosophy 11
2. Kinds of Minds 17
3. Universal Mind 27
4. Intelligence 43
5. Experience 53
6. Four Levels of Mind 69
7. Complexity 81
8. Quantum Reality and Mind 91
9. Free Will 125
10. Emotion 131
11. Autopoiesis 137
12. Evolution 149
13. Science 163
14. Language 179
15. Toward Artificial Minds 195
16. Post-Embodied AI 251
17. Causation 263
18. Belief and Self Systems 275
19. Creative Intuition 295
20. Mindfulness and Evil 329
21. Immortality 339
22. Compassion and Ethics 347

Conclusion 353

Appendices

A. Toward a Mathematical Theory of Pattern 355
B. Toward a Mathematical Theory of Mind 391
C. Notes on the Formalization of Causal Inference 407

References 411

Preface

In this book I present some interim results from a quest I've been on for many years – a quest for a coherent conceptual framework making it possible to understand the various aspects of mind and intelligence in a unified way. The underlying goals of this quest have been twofold: a general desire to understand myself and the universe, and a specific interest in understanding how to create "artificial" minds (in the form of computer software or else novel engineered physical systems). So, this is a work of philosophy-of-mind, yet haunted consistently throughout by the spectre of AI. Much of what's interesting in these pages probably results from the crosspollination between these two aspects, as well as from the crosspollination between two different approaches to understanding: the informal and introspective approach, versus the nitty-gritty analytical/scientific/engineering approach, both of which are well-represented here.

The key themes of this book have been following me around at least since 1982, when I was 16 years old and first made a significant effort to synthesize everything I knew toward the goal of achieving a coherent understanding of human and digital mind. The task was well beyond my knowledge and capabilities at that point, but yet the basic concepts that occurred to me then were essentially the same ones that I present in these pages now, as I approach my (not too enthusiastically anticipated) 40th birthday.

I recall in particular a big moment about 5 months into my 16th year, when I spent a Spring Break visit back home from college writing down my thoughts on intelligence on little scraps of notebook paper, hoping to conceive something practical in the area of AI design. I failed at my attempt to make a practical AI design that Spring Break, but it was a conceptually productive time nonetheless. Various ideas from the back of my mind crystallized into a holistic understanding, and I came to the realization to which the title of this book refers – which is simply that *the mind and world are themselves nothing but pattern – patterns among patterns, patterns within patterns…* Where I got stuck was in trying to invent a general yet computationally efficient and humanly-plausible-to-program algorithm for pattern recognition (a hard problem that I'll talk a bit about in Chapter 15 below: after more than two decades of hard thinking I think I've finally made significant progress!).

When I first came to this grandiose "patternist" conclusion I had years before read Douglas Hofstadter's "Godel, Escher Bach," (1979) which had put the concept of "pattern" firmly into my mind. I hadn't yet read Charles Peirce (1935) (who modeled the universe in terms of "habits", a close-synonym for "pattern"), nor Nietzsche (1968, 1997, 2001), who spoke of "the world as will to power and morphology" (or in other words, the universe as a system of patterns struggling for power over each other). Nor had I yet read Gregory Bateson (1979), who articulated what he called "The MetaPattern: that it is pattern which connects." Nor Benjamin Lee Whorf (1964), who interpreted the universe as a web of linguistic patterns, but interpreted the notion of language so generally that he ultimately was proposing a universal pattern-theory. Each of these thinkers, when I encountered their work during the next couple years after my early Eureka experience, gave me a feeling of great familiarity, and also a stab of frustration. Each of them, within a small degree of error, seemed to be trying to get across the same basic concept as my "Hidden Pattern" insight. But none of them had developed the idea nearly as deeply or extensively as seemed necessary to me. Now, a couple decades later, I find that I haven't developed the idea nearly deeply or extensively as I would like either – but I believe I've moved it forward a fair bit, and hopefully created some tools that will accelerate future progress.

Patternist philosophy isn't something with a fixed number of axioms and conclusions. It's a fluid and shifting set of interlocking ideas – most of all, it's a *way of thinking* about the mind. Transmitting a deep, broad and idiosyncratic "way of thinking" to others isn't an easy task (it's easier just to use one's peculiar, personal "way of thinking" for oneself, to draw interesting conclusions, and then present others with the conclusions), but I've been driven to attempt it.

The word "philosophy" is carefully chosen here – though many of the ideas presented here developed in the context of my scientific work in various disciplines, I don't consider "patternist philosophy" a scientific theory per se; it is more general and abstract than that. Patternist philosophy may be used to inspire or refine scientific theories; but it may also be used for other purposes, such as non-scientific introspective self-understanding. However, as will be clear when I discuss the philosophy of science in depth in these pages, I do think the difference between science and non-science is subtler and fuzzier than generally recognized; and that, in a sense, an abstract construction like patternist philosophy may be validated or refuted in the same way as scientific research programs.

I mentioned above my twin goals of general self- and world-understanding and AI design. Another, more fine-grained goal that has motivated my ongoing work on pattern philosophy has been the desire to create a deep, abstract framework capable of unifying the multiple visions of mind I've encountered. Each of these visions seems to embody some valid insights: the introspective view I have of my own mind; the views of mind that the community of scientists has arrived at via their investigations in brain science, mathematical logic, cognitive psychology and other disciplines; and the views of mind that Eastern philosophers and "mystics" around the world have derived via their collective experiments in introspection and mind-clarification. Patternist philosophy spans all these perspectives, and I think this is one of its strengths.

The human-psychology aspects of patternism were my focus in the mid-1990s, when I was spending much of my time doing research in theoretical cognitive science. That was the stage when my study of Eastern metaphysics was also at its peak — largely as part of an (ultimately failed) attempt to build a connection between my own world-view and that of my wife at the time, who during that period was becoming a Zen Buddhist priest. From 1997 till the present, on the other hand, I've been spending much of my time designing, building and analyzing AI software systems (first a system called Webmind, and now a system called Novamente); and the recent development of patternist philosophy has thus had a strong AI bias.

Reflecting my intellectual focus in recent years, this book contains a lot of discussion of the implications of the patternist philosophy for AI, including numerous mentions of my own current AI project, the Novamente AI Engine. However, the technical details of the Novamente design aren't entered into here – many of those are described in two other books that I've been working on concurrently with this one: *Engineering General Intelligence* (coauthored with Cassio Pennachin) and *Probabilistic Logic Networks* (coauthored with Matt Ikle', Izabela Freire Goertzel and Ari Heljakka).

My schedule these last few years has been incredibly busy, far busier than I'm comfortable with – I much prefer to have more "open time" for wandering and musing. My time has been full with Novamente AI research, with several other book projects, and with a great deal of business work associated with the AI-based businesses my colleagues and I have started in the last few years…and my personal life has also been busy, what with the demands of raising three children and repeatedly moving house — not to mention getting divorced and remarried…In short, modern human life in all its variety and chaos. Given all this, it's been rather difficult for me to find time to work on this book. Every hour I've worked on this book, I've been intensely aware that I could have been spending time generating business for Biomind, or working on the details of the Novamente system, or giving my wife or kids more attention than the amount they manage to squeeze out of my over-packed schedule…But I have taken the time to write the book anyway, because I believe the ideas I present here are very important ones.

I don't think the "problem of the mind" or the "problem of the fundamental nature of the universe" are problems that can ever be completely and finally solved, and I certainly don't claim to have done so in the train of thought reported here. However, I do believe I've arrived at a number of important insights regarding the mind and world – insights which connect together in a network, and which taken together provide powerful answers for some questions that have preoccupied people for a long time.

As well as stimulating interest and further thinking and research, it is my hope that these ideas will play a part in the process of going beyond "people" altogether. Like Nietzsche I believe that "Man is something that must be overcome," both by the creation of improved humans and the creation of superhuman intelligences whose nature we can only dimly conceive. Along the path toward this self-overcoming, a deep understanding of mind and pattern is bound to be critical. I have certainly found it so in my own scientific work (both in AI and in other areas like bioinformatics), as well as in my personal quest for spiritual and mental development.

Next, I'll make a few boring comments about this prose you are reading. In writing about these ideas, I have chosen a style that is natural to me but may seem eccentric to some others – a kind of mix between informal conversational prose and more technical academic-ish prose. Whatever you think of them, my choices regarding style and level of exposition have been quite deliberate. While I admire the elegance of philosophical stylists such as Nietzsche, Goethe and Pascal (to toss out the names of a few of my favorites), and I'm sometimes tempted to try to emulate them (though I'm not sure how well I'd succeed), in this book I've opted for clarity over stylistic beauty nearly every time. Abstract philosophy ideas are hard enough to communicate when one strives for clarity; and though I love reading Nietzsche, I have less appreciation than he did for the subtle aesthetics of being misunderstood. I hope I've managed to communicate the ins and outs of my topics reasonably adequately.

As a corollary to the stylistic decision mentioned in the previous paragraph, I have chosen not to adopt a fully thorough style of referencing here. I have given a decent number of references here and there in the text, but by and large I've only referenced those sources that seemed extremely important for the subject under discussion – ones containing ideas that I directly and significantly use in the text, or else that I really think the reader should bother to look up if they want to fully understand the ideas I'm presenting. You shouldn't assume from this that I'm unaware of the vastness of the extant literature discussing the themes I consider. I've read many hundreds, probably thousands of books and papers on these topics; and by eschewing more exhaustive referencing, I'm not trying to give a false impression that everything I say and don't explicitly reference is completely original and unprecedented.

Finally, a note about my own current research direction. Part of the reason I've finally decided to write these "patternist" ideas down systematically and publish them is that they've become somewhat old and tiresome to me. I think patternist philosophy is extremely important, but it's also not something I think about that much anymore, because it's become second nature to me. I'm working on a lot of practical scientific and engineering projects these days, but in the philosophical portion of my life I'm devoting my time to a somewhat different set of ideas than the ones I discuss here: the creation of a general, qualitative theory of the development of complex pattern-systems over time. This pursuit relates closely to my current work on Novamente, which has to do with figuring out how to get a "baby Novamente" to evolve the right mind-structures through interaction with a simulated world; and also relates to other themes touched in this book, such as the possibility of deriving physical law from considerations related to subjective reality, and the possibility of general laws of emergent complex systems dynamics. So, I've been feeling I should write up these "patternist philosophy" ideas before I get so bored with them that I can't do a good job of writing them up anymore. The fact that my latest conceptual obsession (pattern-based development-theory) uses the patternist philosophy as a launching-pad is encouraging to me, and provides some additional validation within my own subjective universe for the hypothesis that the patternist philosophy is useful ("progressive" in the Lakatosian sense discussed in Chapter 13) as well as sensible and comprehensive.

Acknowledgements

Two people — my wife Izabela and my former wife Gwen — have indulged me in more lengthy, productive and creative discussions on these issues than anyone else. My primary research and business partner Cassio Pennachin has been very helpful in collaboratively working out the principles of the Novamente AI design, and many of the insights I've achieved via working with him on Novamente have found their way into this book. Other friends and colleagues over the years have also been wonderful thought-partners, a list including but by no means limited to (in no particular order): Lisa Pazer, Meg Heath, Stephan Bugaj, Mike Kalish, John Dunn, Cate Hartley, Allan Combs, Eliezer Yudkowsky, Moshe Looks, Debbie Duong, Ari Heljakka, Mike Ross and Gui Lamacie.

Ed Misch, my university philosophy professor, first got me seriously interested in Western philosophy of mind way back when I was 16 (roughly 6 months after I started obsessing on AI … Ed kindly and tactfully clued me in to how many other thinkers had already trod some of the paths I was exploring). My interest in Eastern philosophy of mind, on the other hand, was kindled by my mother's study of Oriental philosophy in my early youth, and then by my first wife Gwen's adventures in Buddhism. My interest in AI and other nonhuman types of mind grew largely out of a host of science fiction novels and stories that I read in my youth – in that sense I feel highly grateful to the genre of sci-fi for existing and urging me to think even more obsessively about "what if?" than I would have done on my own.

My mother Carol, father Ted and grandfather Leo are due copious thanks for decades ago, getting me started on the path of independent intellectual inquiry – and for encouraging me as this path has led me so many strange places over the years.

Finally, my wife Izabela and children Zarathustra, Zebulon and Scheherazade are due much gratitude for their tolerance of my seemingly endless busy-ness these days, as I've worked on this book along with a dozen or so other projects.[1]

[1] How frustrating the human body can be as an interface for the human mind, which has already, at this pre-Singularity phase, evolved so far beyond the limitations of its container – if I fail to survive into the transhuman era, only a small percentage of the interesting and potentially useful ideas in my mind are ever going to get communicated, due purely to the slowness of the physical mechanics of communication, and of the low-level cognitive processes of verbalizing and otherwise linearizing already-formed concepts and relationships.

Chapter 1
Meta-Philosophy

The main focus of this book is on the philosophy of mind – a topic that, obviously, is extremely broad with a huge variety of aspects. However, before addressing philosophy of mind, I will – in this brief chapter — give a little attention to philosophy in a more general sense, with a view toward conceptually positioning the philosophy of mind to follow.[2]

Pragmatic Relativism and the Value of Philosophy

As a not-very-dramatic prelude, I'll begin by explaining my personal meta-philosophy – my philosophy of philosophy. This is quite simply a philosophy of *pragmatic relativism*.

Relativism means I don't believe there is any one correct way of looking at the universe. I don't believe any philosophy or any science is going to find the ultimate essence of the universe, break down the universe into its essential components, give an ultimate explanation for why things are the way they are, etc. I don't believe there is any one big meaning underlying this universe we see all around us. This doesn't make me a nihilist – it's quite possible to avoid believing anything is absolutely true or objective, while still avoiding the cognitive and emotional excesses of nihilism.[3] I also note that this isn't an absolute or dogmatic belief. If someone finds a single ultimate meaning for it all, that's great, and I'm open to that possibility! But as a working assumption, I've chosen to assume that, in all probability, no such thing exists.

On the other hand, I do believe that the quest for ultimate meanings and foundational analyses and reductions of the universe is extremely valuable – even though it's ultimately bound to fail because the universe has no ultimate meaning and foundation.

The value system I've chosen for myself consists of three primary values: freedom, joy and growth. I will elaborate on this value system in the final chapter of the book, interpreting each of these concepts in a pattern theory context, and elucidating how they fit together. Of course, every one of these three things is difficult to define in a rigorous way, but nevertheless, these are the values I have chosen, fully realizing that they are defined only relative to the human

[2] In later chapters I will also have something to say about *philosophy of science* and *ethical philosophy*, but according to the patternist perspective, those topic are best addressed after the essential concepts of pattern theory and philosophy of mind have already been presented.

[3] This is one of those "existential" points that you don't *really* understand until you've experienced the opposite – i.e., until you've lived, for a while, the "cognitive and emotional excesses of nihilism." In my late teens I was highly confused on these issues, and liked to go around giving people detailed logical proofs that they didn't exist. Reading Nietzsche and taking psychedelic drugs, among other experiences such as simply growing up, nudged my mind in the direction of what I came to call "transnihilism" – or what Philip K. Dick (1991) called the recognition of "semi-reality": the attitude that the world may not be objectively real in the way some people naively think, but is incontrovertibly *there* nonetheless.

cultural and cognitive patterns existing in my human mind/brain.

And I believe that the quest for ultimate foundations of the universe is an important contributor to the values of freedom, joy and growth. As a highly relevant example of this, I've found that developing my own speculative metaphysics has been absolutely essential to developing my own philosophy of mind – and developing my own philosophy of mind has been absolutely essential to developing my Novamente AI design … which, if it works even vaguely as planned, will be a very useful thing, and even if not (egads!) will surely lead to a lot of valuable research insights and useful applications (in fact it already has done so, to an extent).

The Value of Metaphysics

For those who believe in the absolute reality of the physical world and the illusoriness of everything else, metaphysics is basically a trivial pursuit. But I'm a dyed-in-the-wool relativist[4], and I've never been able to accept *anything* as possessing absolute reality, nor to condemn subjectively apparent realities as "illusory" in any strong sense. I'm aesthetically and intuitively attracted to developing ideas with maximum explanatory power based on the minimum possible assumptions — where minimality is concerned, an absolutely real physical world seems not to fit the bill; whereas, rejecting subjective reality as totally illusory seems to fail on the grounds of weak explanatory power.

I'm strongly drawn to the "phenomenological" perspective in which physical reality is something the mind constructs, based on patterns it recognizes among its perceptions – and then based on information it gathers via communication; but communication is only recognized as a reliable information source via observation of many instances where communicated information agrees with perceived information. And yet, of course, as a would-be mind engineer, I also recognize the value of the alternate view, in which physical structures and dynamics give rise to the mental realm as an emergent phenomenon. I think both of these views – "mind creates reality" and "reality creates mind" – are important though limited.

And so I'm attracted to a more metaphysical/metamental perspective, in which one identifies some simple concepts common to both mind and reality, and develops these concepts as clearly and simply as possible, without reference to theories of either psychology or physics. One might protest that there is no basis on which to judge theories of this type – but this is just a (half-useful) half-truth. In a later chapter, I'll discuss the philosophy of science, and will argue that the quality of a scientific theory must be judged by the quantity of interestingness and surprisingness among the conclusions to which it leads. Similarly, one may judge one metaphysical theory against another by assessing the usefulness of the theories for generating more concrete, not-just-metaphysical ideas. By this standard, for example, the speculative metaphysics of Leibniz's *Monadology* (1991) has proved itself fairly useless – subsequent thinkers and doers have found little use for Leibniz's idea that the universe consists of a swarm of monads that continue forever to enact their initial programs set in place by God. And by this same standard, I believe, the patternist metaphysics given here has already begun to prove its worth, via its power at leading to interesting conclusions in domains such as psychology, artificial intelligence, and evolutionary biology.

[4] As Ben Franklin wrote, in my favorite of his maxims, "Moderation in all things – including moderation."

Patternist Metaphysics

Some metaphysical theories are concerned with identifying the most basic stuff of the universe and defining everything else in terms of this stuff.[5] One then gets involved in debates about whether entities or processes are more primal, whether time emerges from space or space emerges from time, etc. I've never found debates like this to be very productive, and I prefer to think about "identifying a foundational network of concepts for describing the universe", without worrying too much about which concepts within the network are more foundational. In a network of interrelated abstract concepts, it's often possible to create many possible "spanning subtrees" of the network (to use some terminology from graph theory), thus tracing many alternative pathways from "foundational" to "derived" concepts.

But one must begin somewhere, and – echoing Faust's "In the beginning was the Act" – in articulating my metaphysics I choose to begin with *events*. An event is simply an occurrence. Speaking phenomenologically, events may be effectively instantaneous, or events may have directionality. The "directionality" in an individual event may be thought of as a local time axis, but need not be considered as a global time axis – i.e. at the very general level of foundational metaphysics we don't need to assume all the local directionalities are aligned with each other, or even that there is any fundamental sense in which "aligned with each other" is meaningful. This kind of generality is useful when one is talking about topics like quantum gravity or quantum gravity based computation, where the very origin of the substance of "time" is a topic of analysis.

If an event is directional, it has parts, and the parts have a directional ordering to them, a before and after. These parts may be considered as events themselves. We may consider the beginning part of an event as the "input" of the event and the ending part of the event as the "output" part of the event. The middle parts of the event may be considered as "processing."

Next, suppose that events are divided into categories. This division may be done in a lot of different ways, of course. Here we may introduce the distinction between objective and subjective metaphysics. As a relativist, I don't believe in true objectivity of mind or reality or anything else, but I do think it's meaningful to talk about "quasi-universal" metaphysics, in the sense of metaphysics that holds for every subjective reality of any significant complexity. In the case of the division of events into categories, one can introduce subjectivity in a fully explicit way, by stating that the event-categories in question are defined by some specific mind's categorization of its world. Or, one can be quasi-universalist, and define "event-categories" relative to subjective reality, so that events E1 and E2 are put in the same event-category if they are judged highly similar by a particular judging mind. This is quasi-universalist because, intuitively, any significantly intelligent mind is going to have an explicit or implicit notion of similarity, and hence one may define event-categories as "similarity clusters" in any interesting subjective reality.

Now we may define a *process* as a collection of events whose inputs are all in the same category, and whose outputs are all in the same category. In logic terms, events are instances and processes are classes of these instances.[6]

[5] A friend who was involved in the creation of the SUMO formal ontology (a DARPA funded project at Teknowledge Inc.; see Niles and Pease, 2001) says that the scientists involved with SUMO spent three weeks debating whether to put "thing" or "stuff" at the top of the ontological hierarchy. (Of course, he was kidding — but only partially)

[6] As an aside, I'm aware the word "process" has a deep meaning in Whitehead's philosophy – which I've read and enjoyed but never studied carefully – and my use of the word here is definitely not intended to be

Processes set the stage for the introduction of patterns, the key player in patternist metaphysics. To obtain patterns, one needs processes, and one also needs the notion of *simplicity*. By a "simplicity measure," I mean a process whose inputs are processes, and whose outputs are entities belonging to some ordered domain. As in mathematics, by an ordered domain I mean any set of entities that is endowed with some ordering operation that is reflexive, antisymmetric, transitive and total.[7] The classic examples of ordered domains are integers and real numbers. Most of the simplicity measures used in mathematical pattern theory so far map into either the positive integers, the positive real numbers, or the interval [0,1].

Here, again, we touch the issue of subjectivism. If we are studying the subjective metaphysics of some particular mind, then we may assess simplicity as "simplicity measured relative to that mind." But this notion requires some unraveling – essentially what I mean is "X is simpler than Y relative to mind M if, when all else is equal, M prefers X as an explanation than Y." That is, I define simplicity relative to a mind as "that which, if we define it as simplicity, makes that mind work as closely as possible according to Occam's Razor."

If we are speaking in general, outside of any particular mind, then in order to define simplicity we need to adopt some kind of basic representational framework. For instance, if we define events as strings of zeros and ones, then processes become functions on bit strings, which may be represented e.g. as bit strings representing programs for some particular computer. Since we need to introduce a "reference computer" here, we have not really escaped from subjectivity, we've merely introduced a mathematical way to represent our assumption of subjectivity. Computation theory tells us that, as the bit strings we're dealing with get longer and longer, the choice of computer matters less and less – but it still matters.

Finally, with a simplicity measure in hand, we can interpret a space of processes as a space of patterns. A pattern in some X may be defined as a process P whose output is X, and so that P and its input, taken together, are simpler than X. We may then envision the space of processes as a network of patterns – patterns being processes that transform processes into processes, thus making the process network more complex by introducing a network of simplifications.

We may also define relative patterns, i.e. patterns that are to be considered relative to some particular "system" (i.e. some set of patterns). A pattern in X relative to M is a process P whose output is X, and whose input consists of some subset of M and some auxiliary input – and so that P and its auxiliary input, taken together, are simpler than X. For instance, there is an obvious pattern in the series 1, 2, 4, 8… relative to my mind (which contains a lot of mathematical background knowledge), but this pattern may not exist in the series relative to a young child's mind.

Given the notion of pattern, we can then develop the full apparatus of patternist philosophy, as will be enlarged upon in subsequent chapters. We can make definitions like:

- *Complexity*: the amount of pattern in an entity,

strictly Whiteheadian in nature. In Hegelian terms, roughly speaking, an entity is a Being and events and processs are Becomings.

[7] If we denote the ordering relation by $\leq$, then the axioms that make this relation an ordering are: $a \leq a$ (reflexivity) , if $a \leq b$ and $b \leq a$ then $a = b$ (antisymmetry), if $a \leq b$ and $b \leq c$ then $a \leq c$ (transitivity) , $a \leq b$ or $b \leq a$ (totalness) . The natural numbers are the smallest possible totally ordered domain with no upper bound.

- *Intelligence*: the ability to achieve complex goals in complex environments,
- *Mind*: the set of patterns associated with an intelligent system (i.e. an intelligent set of processes),
- *Emergence*: the existence of patterns in a set of processes that are not patterns in any of the individual processes in the set,
- *Relative complexity*: the complexity of an entity, where the patterns in the entity are defined relative to the knowledge in the system,
- *Simplicity and relative simplicity*: the inverses of complexity and relative complexity.

These definitions allow us to go full circle and return to an earlier point in our metaphysical development. We may define a simplicity measure that's dependent on a particular mind, where a mind is defined as the set of patterns associated with an intelligent system. Similarly, going back further, we may define processes in terms of categories of events, where the categorization is defined either by the categories explicitly in some mind, or by similarity as measured by the possession of similar sets of patterns relative to that mind.

Physical law, from this perspective, consists of a collection of processes that are extremely intense, powerful patterns in the event-space that is the universe. These patterns are highly "intense" in that they provide massive simplification of the universe. Quantum particles, gravitational fields and so forth are then understood as patterns defined relative to the pattern-set that is physical law. That is, when the notion "quantum particle" is used to simplify the observation of a path in a bubble chamber, this simplification is conditional on the body of knowledge that we call "physical law." Without this body of knowledge as implicit background information, the path in the bubble chamber may display patterns such as "the left half is shaped the same way as the right half" but not patterns such as "this is an intermediate vector boson."

Everyday physical reality – baseballs, galaxies, Ministers of Finance, chinchilla toenails and so forth – consists of patterns that embodied minds use to organize their sense perceptions. It's well-known that these patterns correspond only complexly and loosely with the patterns that exist conditional on physical law – for instance, humanly perceived colors are explicable in terms of electromagnetic phenomena only after a great deal of calculation and hand-wringing.

In short, patternist metaphysics portrays ultimate and everyday physical reality as well as mind in terms of sets of patterns. Patterns themselves are processes identified as patterns via the imposition of some simplicity measure; and processes are defined in terms of categories of primal events, where categories may be defined as clusters in the context of a similarity measure on sets of events. In addition to the basic concepts of directionality and similarity, the concept of simplicity is key here. Identifying similarities turns events into processes; and then, by the act of defining a way of measuring simplicity, one turns a universe of scattered and disorganized processes into a network of patterns, in which one may detect structures and processes identifiable as "mental" or "physical."

I have chosen to take pattern as the foundational concept. One could try to derive patternist philosophy from yet more basic concepts, but my feeling is that, conceptually, one obtains diminishing returns by reducing the fundamental concept-set beyond what I've done here. For instance, one can use a formalism like G. Spencer-Brown's "Laws of Form" (1994) to define events, directionality, simplicity and similarity. This is an interesting exercise but, at least in the ways I've tried so far, it doesn't really add anything to the philosophy, and has more of the ring of mathematical, computational or physics modeling.

I have here discussed the notion of pattern in a sort of semi-formal, intuitive way. One may also seek to formalize patternist metaphysics mathematically, and this is potentially very valuable, but I feel that assigning this kind of formalization a position at the heart of patternist philosophy would be a mistake. The point of a metaphysical theory is to give conceptual interpretations not precise models. In Appendices to this book, I present formalizations of many of the ideas of patternist metaphysics, in the specific context of computation theory and probability theory. Developing patternist thinking in this mathematical manner is a fascinating pursuit and I hope it will lead to deep conclusions one day, but I don't think it makes sense to put mathematics at the basis of a philosophical theory. Rather, one needs a philosophical theory to tell one what mathematics to construct. It is true that mathematics can then sometimes turn around and feed information back into philosophical theory (Godel's Theorem being the paradigm case), but this feedback occurs within the foundational philosophical theory within which the mathematics is being interpreted. For instance, to get to Godel's Theorem, one must start with a logicist philosophy, which leads one to construct formal-logic mathematics, which leads to Godel's Theorem, which then enriches logicist philosophy. Pattern-theoretic mathematics has not yet led to any major revisions or advances in patternist philosophy but this may potentially happen in future as the mathematics is further developed – this is my hope!

And so I've reached the end of my core metaphysics after just a few pages! We will revisit these basic metaphysical concepts many, many times in the rest of the book – for instance, the question "what is mind?" becomes the question "what kind of pattern network constitutes a mind?" The question "what is consciousness?" becomes "how is consciousness associated with pattern networks?" The question "how to create a thinking machine" becomes "what physical patterns cause the emergence of pattern-networks of the type characteristic of minds." And so on. I certainly don't claim that this is the only useful perspective on mind and reality and AI and so forth — but it seems to me to be significantly more useful than the other frameworks I've studied. (Time will tell if this impression is correct or not.)

Chapter 2
Kinds of Minds[8]

Now I will proceed in the direction of philosophy of mind, beginning, in this chapter, with the relatively simple step of defining the various sorts of mind that may exist. It seems to me that many (though nowhere near all) of the confusions that one finds in the literature on cognitive science, AI and philosophy of mind have to do with the failure to correctly draw distinctions between different kinds of mind. Very few theorists make the effort to distinguish properties of minds in general from properties of human minds in particular.

One may argue that this kind of conflation is reasonable — perhaps, since human minds are the only clear examples of highly generally intelligent systems that we have, we have no basis for distinguishing properties of the human mind from properties of minds in general. But this excuse doesn't hold up to scrutiny. In some cases, indeed, it's hard to tell whether some property of the human mind is a necessary aspect of mind-ness or merely a peculiarity of human cognitive nature. But even so, the distinction can be made much more often than anyone bothers to do.

Before proceeding further, I'll make a few comments on how the notion of "mind" in general may be defined in terms of patternist philosophy. Intelligence, I have defined in my prior works (Goertzel, 1993) as the achievement of complex goals in complex environments. In Chapter 4 I will enlarge upon this definition extensively, and also introduce a related notion of "efficiency-scaled intelligence", which measures intelligence per unit of processing power. A mind, then, I define as *the set of patterns associated with an intelligent system*. This means patterns in the intelligent system, and also patterns emergent between the system and its environment (including patterns emergent between the system and its tools, other intelligent systems, etc.).

Note that both intelligence and mind are fuzzy concepts[9] — intelligence comes in degrees, and a pattern belongs to a mind to a certain degree. These definitions make no commitments about the internal structure or dynamics of mind — raising the question of whether there are any

[8] This chapter was inspired largely by conversations with Meg Heath during her visit to my home in late summer 2004. In listening to her theories of distributed cognition, I became repeatedly frustrated with what seemed like confusion between properties of mind-in-general and properties of human-mind or humanlike-mind. I then realized that I had succumbed to this type of confusion a few times in my own thinking and writing, and decided it was necessary to clarify the matter explicitly by making a typology of minds, even though the effort to do so seemed a bit pedantic at first. I should add however that Meg has her own theory and typology of minds and surely doesn't agree with everything I say in this chapter. Hopefully by the time you read this she will have put her own very interesting perspective in writing, but at the moment she has not done so, so I have no pertinent references to give.

[9] "Fuzzy" is meant here not in the sense of "ill-understood" or "vague", but rather in the sense of fuzzy set theory, where a fuzzy set is defined as a set to which membership is naturally considered as having various degrees (e.g., tall, hot, etc.).

universal principles regarding what goes on inside minds, and what governs their interactions? I think that such principles do exist, but that they are few in number, and that one can give a larger number of more specific principles if one moves from analyzing "mind in general" to analyzing particular kinds of mind — still keeping things at a high level of abstraction.

But how can we meaningfully ontologize the space of possible mind-types? What are the kinds of minds? In this chapter, I will address this question, and then position with this space both the human mind and the Novamente AI system that my colleagues and I are currently developing.[10]

Huge versus Modest Resources

At the highest level, I propose, we should distinguish minds operating under severely finite resources ("modest-resources minds"), from minds operating under essentially infinite resources ("huge-resources minds").

Human minds fall into the former category, as do the minds of other animals on Earth, and the minds of any AI software programs humans will create in the foreseeable future. Most of human psychology is structured by the resource limitations of the human brain and body, and the same will be true of the psychology of our AI programs.

On the other hand, whether the latter category of minds will ever exist in reality is unclear. Marcus Hutter (2004), following up earlier work by Solomonoff (1964, 1964a) and others, has explored the mathematical nature of infinite-computational-resources-based intelligence. He has proved theorems stating, in essence, that an appropriately designed software system, if allocated arbitrarily much memory and processing power, can achieve an arbitrarily high level of intelligence.[11] He describes a particular software design called AIXI that achieves maximal intelligence, if given infinite computational resources. And then he describes an approximation to AIXI called AIXTtl that achieves close-to-maximal intelligence, if given an extremely huge — but finite — amount of computational resources.

The laws of physics, as we currently understand them, would seem to preclude the actual construction of systems like the infinite AIXI or even the very large finite AIXItl, due to the bound that special relativity places on the rate of information transmission. Since information can't spread faster than the speed of light, there is a limit to how much information processing any physical system occupying a fixed finite amount of space can do per unit time. This limit is very

[10] Note that, here and in later chapters, I will mostly discuss Novamente as it is intended to be when it's finished and fully operational, rather than in its current highly incomplete – though useful for many practical purposes — form. Whether a fully complete and operational Novamente system ever comes about depends on a lot of boring practical issues regarding funding and staffing of the project, but the design itself is interesting as a theoretical object no matter how the practical project comes out.

[11] While this may seem an obvious conclusion, to prove it rigorously using the language of mathematics is no easy trick. Hutter wrote a chapter for a book that I co-edited, and his original title was something like "A Gentle Introduction to AIXI and AIXItl" – but it was by far the most difficult and mathematical chapter in the book, so he changed the title. But his original title had a point, because his chapter was far less technical than his previous expositions of his ideas. The fact that it's so hard to mathematically formalize such a conceptually simple point is an indication that contemporary mathematics is badly suited for the analysis of intelligent systems. This is not surprising since it evolved mainly for the description of simple physical systems, and then more recently for the description of simple computational systems. Attempts to create mathematics better suited for complex living and intelligent systems – such as Robert Rosen's (2002) work – haven't yet succeeded well.

high compared to the amount of information processing done by human brains or contemporary computers, but nevertheless it's trivially small compared to what would be required to run a program like AIXI or AIXItl. Of course, our current understanding of the laws of physics is far from perfect, and it may happen that when we understand the universe better, we'll see that there are in fact ways to achieve superluminal information transmission or something with a similar impact on a computer.

One may define efficiency-scaled intelligence as the the ratio of intelligence of behavior displayed to computational resources utilized. Note that AIXI, because it uses infinite resources, doesn't have a well-defined "efficiency scaled intelligence." Whether AIXItl has a high efficiency-scaled intelligence is a tough mathematical question, but intuitively it seems clear that the answer is no. But these are basically irrelevant questions since both of the designs are most likely impossible to create in practice.

Juergen Schmidhuber, Hutter's former PhD advisor, is trying to get around AIXItl's impracticality problem with an AI design called OOPS, that embodies a vaguely AIXItl-like design, but is capable of solving some simple problems in a reasonable period of time using modest computational resources (Schmidhuber, 1997). In my view, however, OOPS has very little chance of scaling up beyond very simple problems. It seems to me that the conceptual gap between huge-resources mind-design and modest-resources mind-design is too big to be bridged by clever variations like OOPS. Huge-resources and modest-resources mind design and analysis appear to be almost entirely different problems.

The essence of a modest-resources mind consists of a set of strategies for avoiding combinatorial explosions. This is an insight achieved by AI theory during the second half of the 20th century, and it's one that's clearly applicable to human minds as well as software intelligence. A mind with limited space resources — or with space resources that are implicitly limited due to time constraints combined with information propagation limitations as described by special relativity — requires mechanisms for deciding what to retain in memory and what to forget. A mind with limited time resources requires methods to determine which possibilities to explore when deciding which action to take in a certain situation. There is then interplay between the knowledge-paring and possibility-paring processes.

Knowledge-paring and possibility-paring are difficult problems, and modest-resources minds must learn how to do them. This learning process in itself consumes resources. It is often easier to do this learning in a manner that's specialized to some particular domain of experience, rather than in a completely general way. Thus modest-resources minds will tend to contain specialized subsystems, each of which deals with a certain type of knowledge (e.g. visual knowledge, acoustic knowledge, social knowledge, self-knowledge). They must then confront problems of unification — of making their various specialized subsystems communicate with each other effectively. Furthermore, not every needed instance of knowledge and possibility paring can be handled by a specialized subsystem, so a more generalized paring subsystem is also required. A modest-resources mind hence requires some kind of "Mind OS" (OS = operating system) that is able to connect a general-purpose paring subsystem with a family of specialized paring subsystems, which may sometimes overlap with each other in functionality. Of course the form this "Mind OS" takes will depend hugely on the particular physical substrate of the mind in question; we will revisit this issue in the context of the Novamente AI design in later chapters.

These arguments show that the assumption of modest resources gives one a lot of information about a mind — so that the structures and dynamics of a modest-resources mind will necessarily be very different from that of a huge-resources mind.

Varieties of Embodiment

Within the domain of modest-resources minds, many additional abstract distinctions can be usefully drawn, including:

- Embodied versus non, and within the "embodied" category: singly versus multiply versus flexibly embodied, and tool-dependent versus non,
- Mindplexes versus fully unified minds,
- Socially-dependent versus non.

By a "singly embodied mind," I mean a mind that is particularly associated with a certain "physical" system (the "body" of the mind). Note that this "physical" system need not be physical in the narrowest sense — it could for instance be an aspect of a computer-simulation world. The important thing is that it displays the general properties of physicality — such as, involving a rich influx of "sensations" that present themselves to the mind as not being analyzable except in regard to their interrelations with one another, and that are in large part not within the mind's direct and immediate control. An embodied mind uses its body to get sensations about some portion of the physical universe within which its body is in contact; and it carries out all, or nearly all, of its direct actions on the physical world via its body. A lot of the patterns constituting the mind should be patterns in this body-system.

A multiply embodied mind is a mind that, in a similar sense, is particularly associated with a certain set of more than one physical system, which are disconnected (or nearly disconnected) from each other. "Disconnected" here means that the bandwidth of information transmission between the "disconnected" systems is vastly less than the bandwidth of transmission between the parts within the individual systems. Human and animal minds are singly embodied, but a single computer program simultaneously controlling a dozen robots might be multiply embodied.

Finally, a flexibly-embodied mind is a one that may have one or more bodies at any given point in time, but isn't specifically attached to any one of the bodies – it can flexibly switch embodiments based on whatever criteria suit it at the moment. A good example would be a human hooked into a video game environment via powerful virtual reality style sensors and actuators. The human mind would be there invariantly regardless of which character, or group of characters, was being played.

Not all minds need to be embodied at all – for instance, one could have a mind entirely devoted to proving mathematical theorems, and able to communicate only in mathematics. Such a mind would have no need for a body; its sensations and actions would consist of statements in a formal language.

One may also delineate a type of mind that is not necessarily embodied in the above sense, but possesses what I call "body-centeredness." A body-centered mind is one that is particularly associated with a certain physical system, but, most of the mind consists not of patterns in the physical system, but rather patterns emergent between the physical system and the environment. So in this case the body is not the substrate of the bulk of the mind-patterns, but only the partial substrate. Of course, a system may be singly or multiply body-centered; and the boundary between embodiment and body-centeredness is fuzzy. Humans have some embodiment and some body-centeredness to them. The development of language, it would seem, moved us further from strict embodiment toward the direction of body-centeredness. Future developments like Internet-connected neural implants might move us even further in that direction — a topic that leads us on to the next kind of mind in our ontology, mindplexes.

A mindplex is a mind that is composed (at least in part) of a set of units that are, themselves, minds. A mindplex may be a singly or multiply embodied mind, or it might not be embodied at all. The notion of a mindplex begs the question of how one distinguishes an autonomous, coherent "mind" from a part of a mind. For instance, if we conceive a mind as the set of patterns associated with intelligent system, then why isn't the set of patterns in human society considered a mind? The solution to this dilemma lies in the recognition that the notion of "mindness" is fuzzy. A human society is a mind, to an extent — and how much mindness it has relative to an individual human mind, is subject to debate. It seems clear to me, intuitively, that human society has less "efficiency-adjusted intelligence" than a typical individual human mind — but does it have less unadjusted, raw intelligence? Perhaps human society is a mindplex. But clearly there would be much more mindplexness involved if, for instance, a massive AI system were connected to the Internet and instructed to participate in online dialogues of various forms in such a way as to encourage the achievement of the essential goals of humanity. In this situation, we'd have a hybrid human/AI mindplex of the type I've called elsewhere a "global brain mindplex."

What about the population of the USA? Is this set of people a mindplex? Here mindplexness interacts with embodiment in an interesting way. Both the USA and global human society as a whole are multiply embodied minds — and mindplexes — but global human society has a much stronger degree of embodiment than the population of the USA, because the USA is not that isolated, so that very many of the patterns constituting the mind of the USA population are emergent patterns between the brains/bodies of Americans and the brains/bodies of other people. The population of the USA is more multiply body-centered, whereas the population of the Earth is more multiply embodied.

Next, the notion of body-centeredness may be decomposed. There are two cases of mind-patterns extending beyond the body centering a mind: mind-patterns spanning the body and other intelligences, and mind-patterns spanning the body and inanimate, mostly unintelligent objects. Of course this is a fuzzy distinction since intelligence is in itself a fuzzy notion. But like the other fuzzy distinctions I'm making in this ontology of minds, I think it's one worth making. A body-centered mind many of whose defining patterns span its body-center and a set of inanimate objects may be thought of as a "tool-dependent" mind, whereas a body-centered mind many of whose defining patterns span its body-center and a set of other significantly mind-bearing systems, may be thought of as a "socially-dependent" mind. Theorists of the human mind haven't reached any consensus as to the extent to which human minds are body-centered versus embodied, and the extent to which they're tool-dependent and/or socially-dependent. It seems clear that the tool-dependence and social-dependence of human minds has been drastically underestimated by the bulk of recent cognitive science theory; research during the next couple decades will likely go a long way toward telling us how much. Our tool-dependence is obvious from our dependence on various physical implements to shape our world in accordance with the concepts in our minds; as well as by our use of mnemonic and communicative tools like diagrams and writing. Our social-dependence is obvious above all from our dependence upon language, which gains much of its meaning only emergently in the interaction between individuals.

One might wonder how — in abstract, general terms – it is possible to distinguish a tool used by a mind from a part of the body about which the mind is centered. Of course, this is clear in a human context, but it might be less clear in other contexts. And it may become less clear in the human context as body-augmenting technologies become more advanced. A conventional prosthetic limb is somewhere between tool and a part of the body; but a prosthetic limb with sufficiently advanced sensors and actuators becomes definitively a part of the body. Qualities that fuzzily distinguish body parts, in an abstract sense, seem to be: one always keeps them rather than using them only occasionally and then discarding them; they grow out of other body parts rather than being found fully formed or being built by a rational process; and one can both sense and act through them. Not all human body parts meet all of these criteria though; and it's not clear that the distinction between tools and body parts will hold up in the post-human era.

Of course, varieties of embodiment lead to varieties of cognitive habits. An embodied mind will tend to use physical-world and body related metaphors for a lot of its internal thinking, and a socially dependent mind will tend to hold inner dialogues with others, even when they're not present. Modest-resources minds depend on heuristics for paring down combinatorial possibilities; embodiment and sociality push minds toward certain sorts of heuristics, based on heuristics that are useful for solving physical or social problems. Physical embodiment pushes minds toward geometric and energy-based heuristics, for example; whereas sociality pushes minds towards heuristics that are based on notions of agents and roles.

Varieties of Social Interaction

Sociality is obviously a critical aspect of human intelligence, and would seem to evolutionarily precede the development of language, since primate intelligence is also heavily focused on social interaction. Fairly strong arguments have been made that our powerful cognitive abilities arose largely out of our early ancestors' need to model one another (cf Calvin and Bickerton, 2000). It is not clear that AI's will need to be social to this extent; however, from the point of view of pragmatically teaching AI's to think, as well as from the point of view of teaching them ethical behavior, building sociality into AI's makes a lot of sense. Having other similar beings around to study is a powerful heuristic for learning self-modeling; and having other similar beings to provide feedback on one's changes through time can be a valuable guide through the process of self-modification.

An obvious question, in an AI context, is whether, given a fixed amount of computational resources, it's better to put them all into one mind, or to split them among a number of socially interacting minds. I don't think we know the answer to this – we lack a sufficiently precise theoretical cognitive science to derive an answer in advance of extensive practical experimentation. My own guess is that the optimal thing may be a community of minds that have a certain degree of mutual separation, but also have more unity than exists between human minds. Having a community of minds is good in that it allows experimentation with different ways of structuring minds – allowing "evolutionary learning" of what makes a good mind. Of course, it may be that the future will reveal learning techniques that work so well they obsolete the whole notion of evolutionary learning, but I'm guessing that explicit diversity-generation is going to have a role in learning forevermore. However, it also seems to me that, in a "society of individual minds communicating through language and physical coexistence" situation like human society, nearly everything that's learned is wasted. It would be much better, I think, if:

- We had a language that involved simply putting our ideas in a common conceptual vocabulary, without requiring a projection of thoughts into a linear sequence of symbols (or diagrams or other such strongly physical-world-limited media)

- As a sometime alternative to language, we could directly exchange thoughts with each other (understanding that in many cases the exchanged thoughts would not be comprehensible to the recipient)
- There were data-mining processes able to scan vast numbers of minds at one time and find patterns correlated with various beneficial properties of minds (in computer science terms, this would correlate to a movement from a pure evolutionary development of minds, to a mind-evolution scenario based on Estimation of Distribution Algorithms (EDA's)[12], wherein one solves a problem via maintaining an evolving population of candidate solutions, and then running a data-mining process that studies the whole population and figures out which patterns characterize good solutions and creates new candidate solutions embodying these patterns).

These kinds of communication will be easy to build into AI systems one day, and they are also in principle achievable in humans via various sorts of brain implants. A collection of telepathically-enabled mind endowed with a EDA-style central datamining process becomes a special sort of mindplex. One possible future for the human race is that it fuses with computer technology to become a "tele-EDA-mindplex" or "global brain mindplex" of this sort.

With these futuristic ideas in view, it seems important to distinguish between socially-dependent minds that live in telepathy-enabled communities, those that live in linguistics-enabled communities, and those that lack both language and telepathy. Obviously the lack of either language or telepathy greatly restricts the kinds of collective cognition that can occur (and by "language" I mean any system for abstract symbolic representation, not including the protolanguage of chimps in the wild, but including, say, systems for communication using pictures or smell, not necessarily just sounds and written words.) The need to communicate through gesture and language rather than through various forms of thought-exchange, has a huge impact on the structure of a mind, because it means that a large part of cognition has to be devoted to the careful projection of thoughts into highly simplified, relatively unambiguous form. The projection of thoughts into highly simplified, relatively unambiguous form is a useful heuristic for shaping ideas, but it is not the only possible sort of heuristic, and minds not centrally concerned with linguistic interaction would probably make far less use of this particular cognitive move. Of course, linguistically-focused minds don't need to formulate all their ideas in language – but the fact is that, in a community of minds, there is a lot of benefit obtained by getting feedback on one's ideas, so there is a strong motivation to formulate one's ideas in language whenever possible, a fact that very strongly biases the thoughts that occur.

Self-Modification

Change and self-improvement are essential to any intelligent system. However, different minds may vary in the extent to which they can modify themselves with a modest amount of effort.

In this vein, one may distinguish radically self-modifying minds from conservatively-structured minds. A radically self-modifying mind is one that has the capability to undertake arbitrarily large modifications of itself with relatively little effort – its main constraint in

[12] Examples of EDA's are the Bayesian Optimization Algorithm (BOA) developed by Martin Pelikan (2002) and to the MOSES algorithm (Looks, 2006) used in Novamente,

modifying itself is its willingness to do so, which ties in with its ability to figure out subjectively-good self-modifications. On the other hand, a conservatively-structured mind is one that cannot make large modifications in itself without undertaking huge effort – and maybe not even then, without destroying itself.

Clearly, human brains are conservatively-structured, whereas digital AI programs have the potential to be radically self-modifying. Radical self-modification is both powerful and dangerous – it brings the potential for both rapid growth and improvement, and rapid unpredictable cognitive and ethical degeneration.

Quantum versus Classical Minds

Finally, there is one more kind of distinction that is worth making, though it's basically orthogonal to the ones made above. This has to do with the sort of physics underlying a mind. Quantum computing is still in a primitive phase but it's advancing rapidly, so that it no longer seems so pie-in-the-sky to be thinking about quantum versus classical minds.

The strange properties of the quantum world have been discussed in many other books and I won't try to do full justice to them here. However, I will enlarge on this topic a bit at the end of Chapter 8. What I'll argue there is that it's almost surely possible to build quantum cognitive systems with properties fundamentally different from ordinary cognitive systems – thus creating a whole new class of mind, different from anything anyone has ever thought about in detail.

Humans versus Conjectured Future AI's

Given all these ontological categories, we may now position the human mind as being: singly-embodied, singly-body-centered, tool and socially dependent and language-enabled, conservatively-structured, and either essentially classical or mostly-classical in nature (meaning: clearly very limited in its ability to do quantum-based reasoning).

The Novamente AI system – the would-be artificial mind my colleagues and I are currently attempting to create — is intended to be a somewhat different kind of mind: flexibly embodied, flexibly body-centered, tool and socially dependent, language and telepathy enabled, and radically self-modifying, and potentially fully quantum-enabled. Also, Novamentes are explicitly designed to be formed into a community structured according to the "telepathic EDA mindplex" arrangement.

It might seem wiser, to some, to constrain one's adventures in the AI domain by sticking more closely to the nature of human intelligence. But my view is that the limitations imposed by the nature of humans' physical embodiment pose significant impediments to the development of intelligence, as well as to the development of positive ethics. Single embodiment and the lack of any form of telepathy are profound shortcomings, and there seems no need to build these shortcomings into our AI systems. Rather, the path to creating highly intelligent software will be shorter and simpler if we make use of the capability digital technology presents for overcoming these limitations of human-style embodiment. And the minds created in this way will lack some of the self-centeredness and parochialism displayed by humans – much of which is rooted precisely in our single-bodiedness and our lack of telepathic interaction.

I would argue, tentatively, that flexible embodiment and telepathic enablement and EDA mindplexing are not only cognitively but ethically superior to the human arrangement. Being able to exchange bodies and minds with other beings will lead to a psychology of one-ness and sharing quite different from human psychology. And, being able to choose the best mind-forms from an evolving population and perpetuate them will avoid the manifold idiocies of the purely evolutionary process, in which minds are stuck with habits and structures that were appropriate only for their ancient evolutionary predecessors, but are still around because evolution is slow at getting rid of things.

Of course, there could be hidden problems with the flexibly-embodied, telepathic EDA-mindplex approach. However, we lack a mathematical framework adequate to uncover such problems in advance of doing extensive experimentation. So our plan with the Novamente project is to do the experiments. If these experiments lead to the conclusion that singly-embodied, no-telepathy minds are cognitively and/or ethically superior, then we may end up building something more humanlike. But this seems an unlikely outcome.