Levels of Organization in General Intelligence

Research Fellow, Singularity Institute for Artificial Intelligence, Inc.

To appear in:
Real AI: New Approaches to Artificial General Intelligence
Ben Goertzel and Cassio Pennachin, eds.

Abstract

Part I discusses the conceptual foundations of general intelligence as a discipline, orienting it within the Integrated Causal Model of Tooby and Cosmides. Part II constitutes the bulk of the paper and discusses the functional decomposition of general intelligence into a complex supersystem of interdependent internally specialized processes, and structures the description using five successive levels of functional organization: Code, sensory modalities, concepts, thoughts, and deliberation. Part III discusses probable differences between humans and AIs and points out several fundamental advantages that minds-in-general potentially possess relative to current evolved intelligences, especially with respect to recursive self-improvement.

1: Part I: Foundations of general intelligence

What is intelligence? In humans, intelligence *is* a brain with a hundred billion neurons and a hundred trillion synapses; a brain in which the cerebral cortex alone is organized into 52 cytoarchitecturally distinct areas per hemisphere. Intelligence is not the complex expression of a simple principle; intelligence is the complex expression of a complex set of principles. Intelligence is a supersystem composed of many mutually interdependent subsystems - subsystems specialized not only for particular environmental skills but for particular *internal* functions. The heart is not a specialized organ that enables us to run down prey; the heart is a specialized organ that supplies oxygen to the body. Remove the heart and the result is not a less efficient human, or a less specialized human; the result is a system that ceases to function.

Why is intelligence? The cause of human intelligence is evolution - the operation of natural selection on a genetic population in which organisms reproduce differentially depending on heritable variation in traits. Intelligence is an evolutionary advantage because it enables us to model, predict, and manipulate reality. Evolutionary problems are not limited to stereotypical ancestral contexts such as fleeing lions or chipping spears; our intelligence includes the ability to model social realities consisting of other humans, and the ability to predict and manipulate the *internal* reality of the mind. Philosophers of the mind sometimes define "knowledge" as cognitive patterns that map to external reality [Newell80], but a surface mapping has no inherent evolutionary utility. Intelligence requires more than passive correspondence between internal representations and sensory data, or between sensory data and reality. Cognition goes beyond passive denotation; it can predict future sensory data from past experience. Intelligence requires correspondences strong enough for the organism to choose between futures by choosing actions on the basis of their future results. Intelligence in the fully human sense requires the ability to manipulate the world by reasoning backward from a mental image of the desired outcome to create a mental image of the

necessary actions. (In Part II, these ascending tests of ability are formalized as *sensory, predictive, decisive,* and *manipulative* bindings between a model and a referent.)

Understanding the evolution of the human mind requires more than classical Darwinism; it requires the modern "neo-Darwinian" or "population genetics" understanding of evolution - the Integrated Causal Model set forth by [Tooby92]. One of the most important concepts in the ICM is that of "complex functional adaptation". Evolutionary adaptations are driven by selection pressures acting on genes. A given gene's contribution to fitness is determined by regularities of the *total* environment, including both the *external* environment and the *genetic* environment. Adaptation occurs in response to statistically present genetic complexity, not just statistically present environmental contexts. A new adaptation that requires the presence of a previous adaptation cannot spread unless the prerequisite adaptation is present in the genetic environment with sufficient statistical regularity to make the new adaptation a recurring evolutionary advantage. Evolution uses existing genetic complexity to build new genetic complexity, but evolution exhibits no foresight. Evolution does not construct genetic complexity unless it is an *immediate* advantage, and this is a fundamental constraint on accounts of the evolution of complex systems.

Complex functional adaptations - adaptations that require multiple genetic features to build a complex interdependent system in the phenotype - are usually, and necessarily, universal within a species. Independent variance in each of the genes making up a complex interdependent system would quickly reduce to insignificance the probability of any phenotype possessing a full functioning system. To give an example in a simplified world, if independent genes for "retina", "lens", "cornea", "iris", and "optic nerve" each had an independent 20% frequency in the genetic population, the random-chance probability of any individual being born with a complete eyeball would be 1 in 3125.

Natural selection, while feeding on variation, uses it up [Sober84]. The bulk of genetic complexity in any single organism consists of a deep pool of panspecies complex functional adaptations, with selection pressures operating on a surface froth of individual variations. *The target matter of Artificial Intelligence is not the surface variation that makes one human slightly smarter than another human, but rather the vast store of complexity that separates a human from an amoeba.* We must avoid distraction by the surface variations that occupy the whole of our day-to-day *social* universe. The differences between humans are the points on which we compete and the features we use to recognize our fellows, and thus it is easy to slip into paying them too much attention.

A still greater problem for would-be analysts of panhuman complexity is that the foundations of the mind are not open to introspection. We perceive only the highest levels of organization of the mind. You can remember a birthday party, but you cannot remember your hippocampus encoding the memory.

Is either introspection or evolutionary argument relevant to AI? To what extent can truths about humans be used to predict truths about AIs, and to what extent does knowledge about humans enable us to create AI designs? If the sole purpose of AI as a research field is to test theories about human cognition, then only truths about human cognition are relevant. But while human cognitive science constitutes a legitimate purpose, it is not the sole reason to pursue AI; one may also pursue AI as a goal in its own right, in the belief that AI will be useful and beneficial. From this perspective, what matters is the quality of the resulting intelligence, and not the means through which it is achieved. However, proper use of this egalitarian viewpoint should be distinguished from historical

uses of the "bait-and-switch technique" in which "intelligent AI" is redefined away from its intuitive meaning of "AI as recognizable person", simultaneously with the presentation of a AI design which leaves out most of the functional elements of human intelligence and offers no replacement for them. There is a difference between relaxing constraints on the means by which "intelligence" can permissibly be achieved, and lowering the standards by which we judge the results as "intelligence". It is thus permitted to depart from the methods adopted by evolution, but is it wise?

Evolution often finds good ways, but rarely the best ways. Evolution is a useful inspiration but a dangerous template. Evolution is a good teacher, but it's up to us to apply the lessons wisely. Humans are not good examples of minds-in-general; humans are an evolved species with a cognitive and emotional architecture adapted to hunter-gatherer contexts and cognitive processes tuned to run on a substrate of massively parallel 200Hz biological neurons. Humans were created by evolution, an unintelligent process; AI will be created by the intelligent processes that are humans.

Because evolution lacks foresight, complex functions cannot evolve unless their prerequisites are evolutionary advantages for other reasons. The human evolutionary line did not evolve *toward* general intelligence; rather, the hominid line evolved smarter and more complex systems that *lacked* general intelligence, until finally the cumulative store of existing complexity contained all the tools and subsystems needed for evolution to stumble across general intelligence. Even this is too anthropocentric; we should say rather that primate evolution stumbled across a fitness gradient whose path includes the subspecies *Homo sapiens sapiens,* which subspecies exhibits one particular kind of general intelligence.

The human designers of an AI, unlike evolution, will possess the ability to plan ahead for general intelligence. Furthermore, unlike evolution, a human planner can jump sharp fitness gradients by executing multiple simultaneous actions; a human designer can use foresight to plan multiple new system components as part of a coordinated upgrade. A human can take present actions based on anticipated forward compatibility with future plans.

Thus, the ontogeny of an AI need not recapitulate human phylogeny. Because evolution cannot stumble across grand supersystem designs until the subsystems have evolved for other reasons, the phylogeny of the human line is characterized by development from very complex non-general intelligence to very complex general intelligence through the layered accretion of adaptive complexity lying within successive levels of organization. In contrast, a deliberately designed AI is likely to begin as a set of subsystems in a relatively primitive and undeveloped state, but nonetheless already designed to form a functioning supersystem[1]. Because human intelligence is evolutionarily recent, the vast bulk of the complexity making up a human evolved in the *absence* of general intelligence; the rest of the system has not yet had time to adapt. Once an AI supersystem possesses any degree of intelligence at all, no matter how primitive, that intelligence becomes a tool which can be used in the construction of further complexity.

Where the human line developed from very complex non-general intelligence into very complex general intelligence, a successful AI project is more likely to develop from a primitive general intelligence into a complex general intelligence. Note that *primitive* does not mean *architecturally simple.* The right set of subsystems, even in a primitive and simplified state, may be able to function together as a complete but imbecilic mind which then provides a framework for further development. This does *not* imply that AI can be reduced to a single algorithm containing the

"essence of intelligence". A cognitive supersystem may be "primitive" relative to a human and still require a tremendous amount of functional complexity.

I am admittedly biased against the search for a single essence of intelligence; I believe that the search for a single essence of intelligence lies at the center of AI's previous failures. Simplicity is the grail of physics, not AI. Physicists win Nobel Prizes when they discover a previously unknown underlying layer and explain its behaviors. We already know what the ultimate bottom layer of an Artificial Intelligence looks like; it looks like ones and zeroes. Our job is to build something interesting out of those ones and zeroes. The Turing formalism does not solve this problem any more than quantum electrodynamics tells us how to build a bicycle; knowing the abstract fact that a bicycle is built from atoms doesn't tell you *how* to build a bicycle out of atoms - which atoms to use and where to put them. Similarly, the abstract knowledge that biological neurons implement human intelligence does not explain human intelligence. The classical hype of early neural networks, that they used "the same parallel architecture as the human brain", should, at most, have been a claim of using the same parallel architecture as an earthworm's brain. (And given the complexity of biological neurons, the claim would still have been wrong.)

> "The science of understanding living organization is very different from physics or chemistry, where parsimony makes sense as a theoretical criterion. The study of organisms is more like reverse engineering, where one may be dealing with a large array of very different components whose heterogeneous organization is explained by the way in which they interact to produce a functional outcome. Evolution, the constructor of living organisms, has no privileged tendency to build into designs principles of operation that are simple and general."
>
> -- Leda Cosmides and John Tooby, "The Psychological Foundations of Culture" [Tooby92]

The field of Artificial Intelligence suffers from a heavy, lingering dose of genericity and black-box, blank-slate, tabula-rasa concepts seeping in from the Standard Social Sciences Model (SSSM) identified by [Tooby92]. The general project of liberating AI from the clutches of the SSSM is more work than I wish to undertake in this paper, but one problem that must be dealt with immediately is *physics envy*. The development of physics over the last few centuries has been characterized by the discovery of unifying equations which neatly underlie many complex phenomena. Most of the past fifty years in AI might be described as the search for a similar unifying principle believed to underlie the complex phenomenon of intelligence.

Physics envy in AI is the search for a *single, simple* underlying process, with the expectation that this one discovery will lay bare all the secrets of intelligence. The tendency to treat new approaches to AI as if they were new theories of physics may at least partially explain AI's past history of *overpromise* and *oversimplification*. Attributing all the vast functionality of human intelligence to some single descriptive facet - that brains are "parallel", or "distributed", or "stochastic"; that minds use "deduction" or "induction" - results in a failure (an *overhyped* failure) as the project promises that all the functionality of human intelligence will slide out from some simple principle.

The effects of physics envy can be more subtle; they also appear in the lack of interaction between AI projects. Physics envy has given rise to a series of AI projects that could only use *one* idea, as each new hypothesis for the *one true essence of intelligence* was tested and discarded. Douglas Lenat's

AM and EURISKO programs [Douglas83] - though the results were controversial and may have been mildly exaggerated [Ritchie84] - nonetheless used very intriguing and fundamental design patterns to deliver significant and unprecedented results. Despite this, the design patterns of EURISKO, such as self-modifying decomposable heuristics, have seen almost no reuse in later AIs. Even Lenat's subsequent Cyc project [Lenat86] apparently does not reuse the ideas developed in EURISKO. From the perspective of a modern-day programmer, accustomed to hoarding design patterns and code libraries, the lack of crossfertilization is a surprising anomaly. One would think that self-optimizing heuristics would be useful as an external tool, e.g. for parameter tuning, even if the overall cognitive architecture did not allow for the internal use of such heuristics. The AI field seems to have treated EURISKO as a *failed hypothesis,* or even a *competing hypothesis,* rather than an *incremental success* or a *reusable tool.*

The most common paradigms of traditional AI - search trees, neural networks, genetic algorithms, evolutionary computation, semantic nets - have in common the property that they can be implemented without requiring a store of preexisting complexity. The processes that have become traditional, that *have* been reused, are the tools that stand alone and are immediately useful. A semantic network is a "knowledge" representation so simple that it is literally writable on paper. An AI project adding a semantic network need not design a hippocampus-equivalent to form memories, nor build a sensory modality to represent mental imagery. The traditional AI processes accompanying semantic nets - such as theorem proving, case-based reasoning, production systems, and expert systems - are again standalone algorithms. Neural networks and evolutionary computations are not generally intelligent but they are generically intelligent; they can be trained on any problem that has a sufficiently shallow fitness gradient relative to available computing power. (Though EURISKO's self-modifying heuristics probably had generality equaling or exceeding these more typical tools, the source code was not open and the system design was far too complex to build over an afternoon, so the design pattern was not reused - or so I would guess.)

The standalone nature of the traditional processes may make them useful tools for shoring up the initial stages of a general AI supersystem - with the exception of the semantic network; I regard semantic nets as poisonous to AI research for reasons which should shortly become clear. But standalone algorithms are not substitutes for intelligence and they are not complete systems. Genericity is not the same as generality.

"Physics envy" (trying to replace the human cognitive supersystem with a single process or method) should be distinguished from the less ambitious attempt to *clean up* the human mind design while leaving the essential architecture intact. Cleanup is probably inevitable while human programmers are involved, but it is nonetheless a problem to be approached with extreme caution. Although the population genetics model of evolution admits of many theoretical reasons why the presence of a feature may not imply adaptiveness (much less optimality), in practice the adaptationists usually win. The spandrels of San Marco may not have been built for decorative elegance [Gould79], but they are still holding the roof up. Cleanup should be undertaken, not with pride in the greater simplicity of human design relative to evolutionary design, but with a healthy dose of anxiety that we will leave out something important.

An example: Humans are currently believed to have a modular adaptation for visual face recognition, generally identified with a portion of inferotemporal cortex, though this is a simplification [Rodman99]. At first glance this brainware appears to be an archetypal example of

human-specific functionality, an adaptation to an evolutionary context with no obvious analogue for an early-stage AI. However, [Carey92] has suggested from neuropathological evidence (associated deficits) that face recognition brainware is also responsible for the generalized task of *acquiring very fine expertise in the visual domain*; thus, the dynamics of face recognition may be of general significance for builders of sensory modalities.

Another example is the sensory modalities themselves. As described in greater detail in Part II, the human cognitive supersystem is built to require the use of the sensory modalities which we originally evolved for other purposes. One good reason why the human supersystem uses sensory modalities is that the sensory modalities are *there*. Sensory modalities are evolutionarily ancient; they would have existed, in primitive or complex form, during the evolution of all higher levels of organization. Neural tissue was already dedicated to sensory modalities, and would go on consuming ATP[2] even if inactive, albeit at a lesser rate. Consider the incremental nature of adaptation, so that in the very beginnings of hominid intelligence only a very small amount of *de novo* complexity would have been involved; consider that evolution has no inherent drive toward design elegance; consider that adaptation is in response to the total environment, which includes both the external environment and the genetic environment - these are all plausible reasons to suspect evolution of offloading the computational burden onto pre-existing neural circuitry, even where a human designer would have chosen to employ a separate subsystem. Thus, it was not inherently absurd for AI's first devotees to try for general intelligence that employed no sensory modalities.

Today we have at least one reason to believe that nonsensory intelligence is a bad approach; we tried it and it didn't work. Of course this is far too general an argument - it applies equally to "we tried non-face-recognizing intelligence and it didn't work" or even "we tried non-bipedal intelligence and it didn't work". The argument's real force derives from specific hypotheses about the functional role of sensory modalities in general intelligence (discussed in Part II). But in retrospect we can identify at least one *methodological* problem: Rather than identifying the role played by modalities in intelligence, and then attempting to "clean up" the design by *substituting* a simpler process into the functional role played by modalities[3], the first explorers of AI simply assumed that sensory modalities were irrelevant to general intelligence.

Leaving out key design elements, without replacement, on the basis of the mistaken belief that they are not relevant to general intelligence, is an error that displays a terrifying synergy with "physics envy". In extreme cases - and most historical cases *have* been extreme - the design ignores *everything* about the human mind except one characteristic (logic, distributed parallelism, fuzziness, etc.), which is held to be "the key to intelligence".

I argue strongly for "supersystems", but I do not believe that "supersystems" are the necessary and sufficient Key to AI. Human general intelligence requires the *right* supersystem, with the right cognitive subsystems, doing the right things in the right way. Humans are not intelligent by virtue of being "supersystems", but by virtue of being a *particular* supersystem which implements human intelligence. I emphasize supersystem design because I believe that the field of AI has been crippled by the *wrong kind of simplicity* - a simplicity which, as a design constraint, rules out workable designs for intelligence; a simplicity which, as a methodology, rules out incremental progress toward an understanding of general intelligence; a simplicity which, as a viewpoint, renders most of the mind invisible except for whichever single aspect is currently promoted as the Key to AI.

If the quest for design simplicity is to be "considered harmful"[4], what should replace it? I believe that rather than simplicity, we should pursue *sufficiently complex explanations* and *usefully deep designs*. In ordinary programming, there is no reason to assume *a priori* that the task is enormously large. In AI the rule should be that the problem is always harder and deeper than it looks, even after you take this rule into account. Knowing that the task is large does not enable us to meet the challenge just by making our designs larger or more complicated; certain *specific* complexity is required, and complexity for the sake of complexity is worse than useless. Nonetheless, the presumption that we are more likely to underdesign than overdesign implies a different attitude towards design, in which victory is never declared, and even after a problem appears to be solved, we go on trying to solve it. If this creed were to be summed up in a single phrase, it would be: "Necessary but not sufficient." In accordance with this creed, it should be emphasized that supersystems thinking is only one part of a larger paradigm, and that an open-ended design process is itself "necessary but not sufficient". These are first steps toward AI, but not the only first steps, and certainly not the last steps.

2: Part II: Levels of organization in deliberative general intelligence

Intelligence in the human cognitive supersystem is the result of the many cognitive processes taking place on multiple levels of organization. However, this statement is vague without hypotheses about specific levels of organization and specific cognitive phenomena. The concrete theory presented in Part II goes under the name of "deliberative general intelligence" (DGI).

The human mind, owing to its accretive evolutionary origin, has several major distinct candidates for the mind's "center of gravity". For example, the limbic system is an evolutionarily ancient part of the brain that now coordinates activities in many of the other systems that later grew up around it. However, in (cautiously) considering what a more foresightful and less accretive design for intelligence might look like, I find that a single center of gravity stands out as having the most complexity and doing most of the substantive work of intelligence, such that in an AI, to an even greater degree than in humans, this center of gravity would probably become the central supersystem of the mind. This center of gravity is the cognitive superprocess which is introspectively observed by humans through the internal narrative - the process whose workings are reflected in the mental sentences that we internally "speak" and internally "hear" when thinking about a problem. To avoid the awkward phrase "stream of consciousness" and the loaded word "consciousness", this cognitive superprocess will hereafter be referred to as *deliberation*.

2.1: An illustration of principles

My chosen entry point into deliberation is words - that is, the words we mentally speak and mentally hear in our internal narrative. Let us take the word "lightbulb" (or the wordlike phrase "light bulb") as an example[5]. When you see the letters spelling "light bulb", the phonemes for *light bulb* flow through your auditory cortex. If a mental task requires it, a visual exemplar for the "light bulb" category may be retrieved as mental imagery in your visual cortex (and associated visual areas). Some of your past memories and experiences, such as accidentally breaking a light bulb and carefully sweeping up the sharp pieces, may be associated with or stored under the "light bulb" concept. "Light bulb" is associated to other concepts; in cognitive priming experiments, it has been shown that hearing a phrase such as "light bulb"[6] will prime associated words such as "fluorescent" or "fragile", increasing the recognition speed or reaction speed when associated words are presented

[Meyer71]. The "light bulb" concept can act as a mental category; it describes some referents in perceived sensory experiences or internal mental imagery, but not other referents; and, among the referents it describes, it describes some strongly and others only weakly.

To further expose the internal complexity of the "light bulb" concept, I would like to offer an introspective illustration. I apologize to any readers who possess strong philosophical prejudices against introspection; I emphasize that the exercise is not intended as *evidence* for a theory, but rather as a means of introducing and grounding concepts that will be argued in more detail later. That said:

Close your eyes, and try to immediately (without conscious reasoning) visualize a *triangular light bulb* - now. Did you do so? What did you see? On personally performing this test for the first time, I saw a pyramidal light bulb, with smoothed edges, with a bulb on the square base. Perhaps you saw a tetrahedral light bulb instead of a pyramidal one, or a light bulb with sharp edges instead of smooth edges, or even a fluorescent tube bent into an equilateral triangle. The specific result varies; what matters is the process you used to arrive at the mental imagery.

Our mental image for "triangular light bulb" would intuitively appear to be the result of imposing "triangular", the adjectival form of "triangle", on the "light bulb" concept. That is, the novel mental image of a triangular light bulb is apparently the result of combining the sensory content of two pre-existing concepts. (DGI[7] does not hold otherwise, but the assumption deserves to be pointed out explicitly.) Similarly, the combination of the two concepts is not a collision, but a structured imposition; "triangular" is imposed on "light bulb", and not "light-bulb-like" on "triangle".

The structured combination of two concepts is a major cognitive process. I emphasize that I am not talking about interesting complexity which is supposedly to be found in the overall pattern of relations between concepts; I am talking about complexity which is directly visible in the *specific* example of imposing "triangular" on "light bulb". I am not "zooming out" to look at the overall terrain of concepts, but "zooming in" to look at the cognitive processes needed to handle this *single* case. The specific example of imposing "triangular" on "light bulb" is a nontrivial feat of mind; "triangular light bulb" is a trickier concept combination than "green light bulb" or "triangular parking lot".

The mental process of visualizing a "triangular light bulb" flashes through the mind very quickly; it may be possible to glimpse subjective flashes of the concept combination, but the process is not really open to human introspection. For example, when first imposing "triangular" on "light bulb", I would report a brief subjective flash of a *conflict* arising from trying to impose the planar 2-D shape of "triangular" on the 3-D "light bulb" concept. However, before this conflict could take place, it would seem necessary that some cognitive process have already selected the *shape* facet of "triangular" for imposition - as opposed to, say, the color or line width of the "triangle" exemplar that appears when I try to visualize a "triangle" as such. However, this initial selection of *shape* as the key facet did not rise to the level of conscious attention. I can guess at the underlying selection process - in this case, that past experience with the usage had already "cached" *shape* as the salient facet for the concept *triangular,* and that the concept was abstracted from an experiential base in which shape, but not color, was the perceived similarity within the group of experiences. However, I cannot actually introspect on this selection process.

Likewise, I may have glimpsed the existence of a conflict, and that it was a conflict resulting from the 2D nature of "triangular" versus the 3D nature of "light bulb", but *how* the conflict was detected is not apparent in the subjective glimpse. And the resolution of the conflict, the transformation of the 2D *triangle* shape into a 3D *pyramid* shape, was apparently instantaneous from my introspective vantage point. Again, I can guess at the underlying process - in this case, that several already-associated conceptual neighbors of "triangle" were imposed on "light bulb" in parallel, and the best fit selected. But even if this explanation is correct, the process occurred too fast to be visible to direct introspection. I cannot rule out the possibility that a more complex, more deeply creative process was involved in the transition from *triangle* to *pyramid*, although basic constraints on human information-processing (the 200 spike/second speed limit of the underlying neurons) still apply. Nor can I rule out the possibility that there was a unique serial route from *triangle* to *pyramid*.

The creation of an actual visuospatial image of a pyramidal light bulb is, presumably, a complex visual process - one that implies the ability of the visuospatial modality to reverse the usual flow of information and send commands from high-level features to low-level features, instead of detecting high-level features from low-level features. DGI hypothesizes that visualization occurs through a flow from high-level *feature controllers* to low-level feature controllers, creating an articulated mental image within a sensory modality through a multistage process that allows the detection of conflicts at higher levels before proceeding to lower levels. The final mental imagery is introspectively visible, but the process that creates it is mostly opaque.

Some theorists defy introspection to assert that our mental imagery is purely abstract [Pylyshyn81]. Yet there exists evidence from neuroanatomy, functional neuroimaging, pathology of neurological disorders, and cognitive psychology to support the contention that mental imagery is directly represented in sensory modalities [Kosslyn94]. [Finke77] show that mental imagery can create visual afterimages[8] similar to, though weaker than, the afterimages resulting from real visual experience. [Sherman86] estimate that while the cat has roughly 10^6 fibers from the lateral geniculate nucleus[9] to the visual cortex, there are approximately 10^7 fibers running in the opposite direction. No explanatory consensus currently exists for the existence of the massive corticothalamic feedback projections, though there are many competing theories; the puzzle is of obvious interest to an AI researcher positing a theory in which inventing novel mental imagery is more computationally intensive than sensory perception.

To return to the "triangular lightbulb" example: Once the visuospatial image of a pyramidal light bulb was fully articulated, the next introspective glimpse was of a conflict in visualizing a *glass* pyramid - a pyramid has sharp edges, and sharp glass can cut the user. This implies the mental imagery had semantic content (knowledge about the material composition of the pyramidal light bulb), imported from the original "light bulb" concept, and well-integrated with the visual representation. Like most modern-day humans, I know from early parental warnings and later real-life confirmation that sharp glass is dangerous. Thus the rapid visual detection of sharp glass is important when dealing with real-life sensory experience. I say this to emphasize that no extended line of intelligent reasoning (which would exceed the 200Hz speed limit of biological neurons) is required to react negatively to a fleeting mental image of sharp glass. This reaction could reasonably happen in a single perceptual step, so long as the same perceptual system which detects the visual signature of sharp glass in real-world sensory experience also reacts to mental imagery.

The conflict detected was resolved by the imposition of smooth edges on the glass pyramid making up the pyramidal light bulb. Again, this apparently occurred instantly; again, nontrivial hidden complexity is implied. To frame the problem in the terms suggested by [Hofstadter85], the imaginative process needed to possess or create a "knob" governing the image's transition from sharp edges to rounded edges, and the possession or creation of this knob is the most interesting part of the process, not the selection of one knob from many. If the "knob" was created on the fly, it implies a much higher degree of systemic creativity than selecting from among pre-existing options.

Once the final conflict was resolved by the perceptual imposition of smoothed edges, the final mental image took on a stable form. Again, in this example, all of the mental events appeared introspectively to happen automatically and without conscious decisions on my part; I would estimate that the whole process took less than one second.

In concept combination, a few flashes of the intermediate stages of processing may be visible as introspective glimpses - especially those conflicts that arise to the level of conscious attention before being resolved automatically. But the extreme rapidity of the process means the glimpses are even more unreliable than ordinary introspection - where introspection is traditionally considered unreliable to begin with. To some extent, this is the *point* of the illustration narrated above; almost all of the internal complexity of concepts is hidden away from human introspection, and many theories of AI (even in the modern era) thus attempt to implement concepts on the token level, e.g., "lightbulb" as a raw LISP atom.

This traditional problem is why I have carefully avoided using the word *symbol* in the exposition above. In AI, the term "symbol" carries implicit connotations about representation - that the symbol is a naked LISP atom (Prolog variable, etc.) whose supposed meaning derives from its relation to the surrounding atoms in a semantic net; or at most a LISP atom whose content is a "frame-based" LISP structure (that is, whose content is another semantic net). Even attempts to argue against the design assumptions of Good Old-Fashioned AI (GOFAI) are often phrased in GOFAI's terms; for example, the "symbol grounding problem". Much discussion of the symbol grounding problem has approached the problem as if the design *starts out* with symbols and "grounding" is then *added*. In some cases this viewpoint has directly translated to AI architectures; e.g., a traditional semantic net is loosely coupled to a connectionist sensorimotor system [Hexmoor93].

DGI belongs to the existing tradition that asks, not "How do we ground our semantic nets?", but rather "What is the underlying stuff making up these rich high-level objects we call 'symbols'?" - an approach presented most beautifully in [Hofstadter79]; see also [Chalmers92]. From this viewpoint, without the right underlying "symbolstuff", there *are* no symbols; merely LISP tokens carved in mockery of real concepts and brought to unholy life by the naming-makes-it-so fallacy.

Imagine sensory modalities as solid objects with a metaphorical surface composed of the layered feature detectors and their inverse functions as feature controllers. The metaphorical "symbolstuff" is a pattern that interacts with the feature detectors to test for the presence of complex patterns in sensory data, or inversely, interacts with the feature controllers to produce complex mental imagery. Symbols combine through the faceted combination of their symbolstuffs, using a process that might be called "holonic conflict resolution", where information flows from high-level feature

controllers to low-level feature controllers, and conflicts are detected at each layer as the flow proceeds. ("Holonic" is a useful word to describe the simultaneous application of reductionism and holism, in which a single quality is simultaneously a combination of parts and a part of a greater whole [Koestler67]. Note that "holonic" does not imply strict hierarchy, only a general flow from high-level to low-level and vice versa. For example, a single feature detector may make use of the output of lower-level feature detectors, and act in turn as an input to higher-level feature detectors. The information contained in a mid-level feature is then the holistic sum of many lower-level features, and also an element in the sums produced by higher-level features. If you pick one vantage point in a holonic structure and "look down" (reductionism) you find parts composing the local whole, with simpler behaviors that contribute to local complexity; if you "look up" (holism) you find a greater whole to which local parts contribute, and more complex processes which local behaviors support. See also [Hofstadter79].)

I apologize for adding yet another term, "holonic conflict resolution", to a namespace already crowded with terms such as "computational temperature" [Mitchell93], "Prägnanz" [Koffka35], "Hopfield networks" [Hopfield85], "constraint propagation" [Kumar92], and many others. Holonic conflict resolution is certainly not a wholly new idea, and may even be wholly unoriginal on a feature-by-feature basis, but the combination of features I wish to describe does not exactly match the existing common usage of any of the terms above. "Holonic conflict resolution" is intended to convey the image of a process that flows serially through the layered, holonic structure of perception, with detected conflicts resolved locally or propagated to the level above, with a final solution that satisfices. Many of the terms above, in their common usage, refer to an iterated annealing process which seeks a global minimum. Holonic conflict resolution is intended to be biologically plausible; i.e., to involve a smooth flow of visualization which is computationally tractable for parallel but speed-limited neurons.

Holonic conflict resolution is not proposed as a complete solution to perceptual problems, but rather as the active canvas for the interaction of concepts with mental imagery. In theoretical terms, holonic conflict resolution is a structural framework within which to posit specific conflict-detection and conflict-resolution methods. Holonic imagery is the artist's medium within which symbolstuff paints mental pictures such as "triangular light bulb".

A constructive account of concepts and symbolstuff would need to supply:

- (a) A description of how a concept is *satisfied by* and *imposed on* referents in a sensory modality.
- (b) A symbolstuff representation satisfying (a) that can contain the internal complexity needed for faceted concept combination.
- (c) A representation satisfying (a) and (b), such that it is computationally tractable to abstract new concepts using sensory experience as raw material.

This is *not* an exhaustive list of concept functionality; these are just the three most "interesting" challenges[10]. These challenges are interesting because the difficulty of solving them simultaneously seems to be the multiplicative (rather than additive) product of the difficulties of solving them individually. Other design requirements for a constructive account of concepts would include: association to nearby concepts; supercategories and subcategories; exemplars stored in memory; prototype and typicality effects [Rosch78]; and many others (see, e.g., [Lakoff87]).

The interaction of concepts with modalities, and the interaction of concepts with each other, illustrate what I believe to be several important rules about how to approach AI.

The first principle is that of *multiple levels of organization.* The human phenotype is composed of atoms[11], molecules, proteins, cells, tissues, organs, organ systems, and finally the complete body - eight distinguishable layers of organization, each successive layer built above the preceding one, each successive layer incorporating evolved adaptive complexity. Some useful properties of the higher level may emerge naturally from lower-level behaviors, but not all of them; higher-level properties are also subject to selection pressures on heritable variation and the elaboration of complex functional adaptations. In postulating multiple levels of organization, I am not positing that the behaviors of all higher layers emerge automatically from the lowest layer.

If I had to pick one *single* mistake that has been the *most* debilitating in AI, it would be *implementing a process too close to the token level* - trying to implement a high-level process without implementing the underlying layers of organization. Many proverbial AI pathologies result at least partially from omitting lower levels of organization from the design.

Take, for example, that version of the "frame problem" - sometimes also considered a form of the "commonsense problem" - in which intelligent reasoning appears to require knowledge of an infinite number of special cases. Consider a CPU which adds two 32-bit numbers. The higher level consists of two integers which are added to produce a third integer. On a lower level, the computational objects are not regarded as opaque "integers", but as ordered structures of 32 bits. When the CPU performs an arithmetic operation, two structures of 32 bits collide, under certain rules which govern the local interactions between bits, and the result is a new structure of 32 bits. Now consider the woes of a research team, with no knowledge of the CPU's underlying implementation, that tries to create an arithmetic "expert system" by encoding a vast semantic network containing the "knowledge" that two and two make four, twenty-one and sixteen make thirty-seven, and so on. This giant lookup table requires eighteen billion billion entries for completion.

In this hypothetical world where the lower-level process of addition is not understood, we can imagine the "common-sense" problem for addition; the launching of distributed Internet projects to "encode all the detailed knowledge necessary for addition"; the frame problem for addition; the philosophies of formal semantics under which the LISP token thirty-seven is meaningful because it refers to thirty-seven objects in the external world; the design principle that the token thirty-seven has no internal complexity and is rather given meaning by its network of relations to other tokens; the "number grounding problem"; the hopeful futurists arguing that past projects to create Artificial Addition failed because of inadequate computing power; and so on.

To some extent this is an unfair analogy. Even if the thought experiment is basically correct, and the woes described would result from an attempt to capture a high-level description of arithmetic without implementing the underlying lower level, this does not prove the analogous mistake is the source of these woes in the real field of AI. And to some extent the above description is unfair even as a thought experiment; an arithmetical expert system would not be as bankrupt as semantic nets. The regularities in an "expert system for arithmetic" would be real, noticeable by simple and computationally feasible means, and could be used to deduce that arithmetic was the underlying process being represented, even by a Martian reading the program code with no hint as to the

intended purpose of the system. The gap between the higher level and the lower level is not absolute and uncrossable, as it is in semantic nets.

An arithmetic expert system that leaves out one level of organization may be recoverable. Semantic nets leave out *multiple* levels of organization. Omitting all the experiential and sensory grounding of human symbols leaves no raw material to work with. If all the LISP tokens in a semantic net were given random new names, there would be no way to deduce whether G0025 formerly meant hamburger or chair. [Harnad90] describes the symbol grounding problem arising out of semantic nets as similar to learning Chinese as a first language using only a Chinese-to-Chinese dictionary.

I believe that many (though not all) cases of the "commonsense problem" or "frame problem" arise from trying to store all possible descriptions of high-level behaviors that, in the human mind, are modeled by visualizing the lower level of organization from which those behaviors emerge. For example, [Lakoff99] give a sample list of "built-in inferences" emerging from what they identify as the Source-Path-Goal metaphor:

- If you have traversed a route to a current location, you have been at all previous locations on that route.
- If you travel from A to B and from B to C, then you have traveled from A to C.
- If X and Y are traveling along a direct route from A to B and X passes Y, then X is farther from A and closer to B than Y is.
- (et cetera)

A general intelligence with a visual modality has no need to explicitly store an infinite number of such statements in a theorem-proving production system. The above statements can be perceived on the fly by inspecting depictive mental imagery. Rather than storing *knowledge about* trajectories, a visual modality actually *simulates the behavior of* trajectories. A visual modality uses low-level elements, metaphorical "pixels" and their holonic feature structure, whose behaviors locally correspond to the real-world behaviors of the referent. There is a mapping from representation to referent, but it is a mapping on a lower level of organization than traditional semantic nets attempt to capture. The correspondence happens on the level where 13 is the structure 00001101, not on the level where it is the number thirteen.

I occasionally encounter some confusion about the difference between a *visual modality* and a *microtheory of vision*. Admittedly, microtheories in theorem-proving systems are well known in AI, so some confusion is understandable. But layered feature extraction in the visual modality - which is an established fact of neuroscience - is also very well known even in the pure computer science tradition of AI, and has been well-known ever since David Marr's tremendously influential 1982 book *Vision* [Marr82] and earlier papers. To make the difference explicit, the human visual cortex "knows" about edge detection, shading, textures of curved surfaces, binocular disparities, color constancy under natural lighting, motion relative to the plane of fixation, and so on. The visual cortex does not know about butterflies. In fact, a visual cortex "knows" nothing; a sensory modality contains behaviors which correspond to environmental invariants, not knowledge about environmental regularities.

This illustrates the second-worst error in AI, the failure to distinguish between *things that can be hardwired* and *things that must be learned*. We are not preprogrammed to know about

butterflies. Evolution wired us with visual circuitry that makes sense of the sensory image of the butterfly, and with object-recognition systems that form visual categories. When we see a butterfly, we are then able to recognize future butterflies as belonging to the same kind. Sometimes evolution bypasses this system to give us visual instincts, but this constitutes a tiny fraction of visual knowledge. A modern human recognizes a vast number of visual categories with no analogues in the ancestral environment.

What problems result from failing to distinguish between things that can be hardwired and things that must be learned? "Hardwiring what should be learned" is so universally combined with "collapsing the levels of organization" that it is difficult to sort out the resulting pathologies. An expert systems engineer, in addition to acting on the assumption that knowledge of butterflies can be preprogrammed, is also likely to act on the assumption that knowledge about butterflies consists of a butterfly LISP token which derives meaning from relations to other LISP tokens - rather than *butterfly* being a stored pattern that interacts with the visual modality and recognizes a butterfly. A semantic net not only lacks richness, it lacks the capacity to represent richness. Thus, I would attribute the symbol grounding problem to "collapsing the levels of organization", rather than "hardwiring what should be learned".

But even if a programmer who understood the levels of organization tried to create butterfly-recognizing symbolstuff by hand, I would still expect the resulting butterfly pattern to lack the richness of the learned butterfly pattern in a human mind. When the human visual system creates a *butterfly* visual category, it does not write an opaque, procedural butterfly-recognition codelet using abstract knowledge about butterflies and then tag the codelet onto a butterfly frame. Human visual categorization abstracts the butterfly category from a store of visual experiences of butterflies.

Furthermore, visual categorization - the general concept-formation process, not just the temporal visual processing stream - leaves behind an association between the butterfly concept and the stored memories from which "butterfly" was abstracted; it associates one or more exemplars with the butterfly category; it associates the butterfly category through overlapping territory to other visual categories such as *fluttering;* it creates butterfly symbolstuff that can combine with other symbolstuffs to produce mental imagery of a *blue butterfly;* and so on. To the extent that a human lacks the patience to do these things, or to the extent that a human does them in fragile and hand-coded ways rather than using robust abstraction from a messy experiential base, *lack of richness* will result. Even if an AI needs programmer-created concepts to bootstrap further concept formation, bootstrap concepts should be created using programmer-directed tool versions of the corresponding AI subsystems, and the bootstrap concepts should be replaced with AI-formed concepts as early as possible.

Two other potential problems emerging from the use of programmer-created content are *opacity* and *isolation.*

Opacity refers to the potential inability of an AI's subsystems to modify content that originated outside the AI. If a programmer is creating cognitive content, it should at least be the kind of content that the AI *could have* created on its own; it should be content in a form that the AI's cognitive subsystems can manipulate. The best way to ensure that the AI can modify and use internal content is to have the AI create the content. If an AI's cognitive subsystems are powerful enough to create content independently, then hopefully those same subsystems will be capable of

adding to that content, manipulating it, bending it in response to pressures exerted by a problem, and so on. What the AI creates, the AI can use and improve. Whatever the AI accomplishes on its own is a part of the AI's mind; the AI "owns" it and is not simply borrowing it from the programmers. This is a principle that extends far beyond abstracting concepts!

Isolation means that if a concept, or a piece of knowledge, is handed to the AI on a silver platter, the AI may be isolated from the things that the AI would have needed to learn first in order to acquire that knowledge naturally, in the course of building up successive layers of understanding to handle problems of increasing complexity. The concept may also be isolated from similar concepts and related concepts that the AI would otherwise have learned at around the same time, denying the AI useful associations and slippages. Conceivably programmers could try to second-guess isolation by hardwiring many similar "knowledges", but this is no substitute for a natural ecology of cognition.

2.2: Levels of organization in deliberation

The model of intelligence presented in this chapter - "Deliberative General Intelligence" or "DGI" - requires five distinct layers of organization, each layer built on top of the underlying layer.

- The bottom layer is *source code* and *data structures* - complexity that is manipulated directly by the programmer. The equivalent layer for humans is *neurons* and *neural circuitry*.

- The next layer is *sensory modalities*. In humans, the archetypal examples of sensory modalities are sight, sound, touch, taste, smell, and so on[12]; implemented by the visual areas, auditory areas, et cetera. In biological brains, sensory modalities come the closest to being "hardwired"; they generally involve clearly defined stages of information-processing and feature-extraction, sometimes with individual neurons playing clearly defined roles. Thus, sensory modalities are some of the best candidates for processes that can be directly coded by programmers without rendering the system crystalline and fragile.

- The next layer is *concepts*. Concepts (also sometimes known as "categories", or "symbols") are abstracted from our experiences. Abstraction reifies a perceived similarity within a group of experiences. Once reified, the common quality can then be used to determine whether new mental imagery *satisfies* the quality, and the quality can be *imposed* on a mental image, altering it. Having abstracted the concept "red", we can take a mental image of a non-red object (for example, grass) and imagine "red grass". Concepts are patterns that mesh with sensory imagery; concepts are *complex, flexible, reusable* patterns that have been reified and placed in long-term storage.

- The next layer is *thoughts,* built from structures of concepts. By imposing concepts in targeted series, it becomes possible to build up complex mental images within the workspace provided by one or more sensory modalities. The archetypal example of a thought is a human "sentence" - an arrangement of concepts, invoked by their symbolic tags, with internal structure and targeting information that can be reconstructed from a linear series of words using the constraints of syntax, constructing a complex mental image that can be used in reasoning. Thoughts (and their corresponding mental imagery) are the disposable one-time structures, built from reusable concepts, that implement a non-recurrent mind in a non-

recurrent world.

- Finally, it is sequences of thoughts that implement *deliberation* - explanation, prediction, planning, design, discovery, and the other activities used to solve knowledge problems in the pursuit of real-world goals.

Although the five-layer model is central to the DGI theory of intelligence, the rule of *Necessary But Not Sufficient* still holds. An AI project will not succeed by virtue of "implementing a five-layer model of intelligence, just like the human brain". It must be the right five layers. It must be the *right* modalities, used in the *right* concepts, coming together to create the *right* thoughts seeking out the *right* goals. (An AI might use different modalities, but will still need *a* right set of modalities.)

The five-layer model of deliberation is not inclusive of everything in the DGI theory of mind, but it covers substantial territory, and can be extended beyond the deliberation superprocess to provide a loose sense of which level of organization any cognitive process lies upon. Observing that the human body is composed of molecules, proteins, cells, tissues, and organs is not a complete design for a human body, but it is nonetheless important to know whether something is an organ or a protein. Blood, for example, is not a prototypical tissue, but it is composed of cells, and is generally said to occupy the tissue level of organization of the human body. Similarly, the hippocampus, in its role as a memory-formation subsystem, is not a sensory modality, but it can be said to occupy the "modality level": It is brainware (a discrete, modular chunk of neural circuitry); it lies above the neuron/code level; it has a characteristic tiling/wiring pattern as the result of genetic complexity; it interacts as an equal with the subsystems comprising sensory modalities.

Generalized definitions of the five levels of organization might be as follows:

- *Code-level, hardware-level:* No generalized definition is needed, except that the biological equivalent is the *neural level* or *wetware level.*

- *Modality-level:* Subsystems which, in humans, derive their adaptive complexity from genetic specification - or rather from the genetic specification of an initial tiling pattern and a self-wiring algorithm, and from exposure to invariant environmental complexity[13]. The AI equivalent is complexity which is known in advance to the programmer and which is directly specified through programmer efforts. Full systems on this level are modular parts of the cognitive supersystem - one of a large but limited number of major parts making up the mind. Where the system in question is a sensory modality or a system which clearly interrelates to the sensory modalities and performs modality-related tasks, the system can be referred to as *modality-level.* Similarly, a subsystem or subprocess of a major modality-level system, or a minor function of such a subsystem, may also be referred to as modality-level. Where this term is inappropriate, because a subsystem has little or no relation to sensory modalities, the subsystem may be referred to as *brainware*[14].

- *Concept-level:* Concepts are cognitive objects which are placed in long-term storage, and reused as the building blocks of thoughts. The generalization for this level of organization is *learned complexity:* cognitive content which is derived from the environment and placed in long-term storage, and which thereby becomes part of the permanent reservoir of complexity with which the AI challenges future problems. The term *concept-level* might

optionally be applied to any learned complexity that resembles categories; i.e., learned complexity that interacts with sensory modalities and acts on sensory modalities. Regardless of whether they are conceptlike (an issue considered later), other examples of learned complexity include declarative beliefs and episodic memories.

- *Thought-level:* A thought is a specific structure of combinatorial symbols which builds or alters mental imagery. The generalizable property of thoughts is their *immediacy*. Thoughts are not evolved/programmed brainware, or a long-term reservoir of learned complexity; thoughts are constructed on a moment-by-moment basis. Thoughts make up the life history of a non-recurrent mind in a non-recurrent universe. The generalized thought level extends beyond the mentally spoken sentences in our stream of consciousness; it includes all the major cognitive events occurring within the world of active mental imagery, especially events that involve structuring the combinatorial complexity of the concept level.

- *Deliberation,* like the code level, needs no generalization. Deliberation describes the activities carried out by patterns of thoughts. The patterns in deliberation are not just epiphenomenal properties of thought sequences; the deliberation level is a complete layer of organization, with complexity specific to that layer. In a deliberative AI, it is patterns of thoughts that plan and design, transforming abstract high-level goal patterns into specific low-level goal patterns; it is patterns of thoughts that reason from current knowledge to predictions about unknown variables or future sensory data; it is patterns of thoughts that reason about unexplained observations to invent hypotheses about possible causes. In general, deliberation uses organized sequences of thoughts to solve knowledge problems in the pursuit of real-world goals.

Even for the generalized levels of organization, not everything fits cleanly into one level or another. While the hardwired-learned-invented trichotomy usually matches the modality-concept-thought trichotomy, the two are conceptually distinct, and sometimes the correspondence is broken. But the levels of organization are almost always useful - even exceptions to the rule are more easily seen as partial departures than as complete special cases.

2.3: The code level

The code level is composed of functions, classes, modules, packages; data types, data structures, data repositories; all the purely programmatic challenges of creating AI. Artificial Intelligence has traditionally been much more intertwined with computer programming than it should be, mostly because of attempts to overcompress the levels of organization and implement thought sequences directly as programmatic procedures, or implement concepts directly as LISP atoms or LISP frames. The code level lies directly beneath the modality level or brainware level; bleedover from modality-level challenges may show up as legitimate programmatic problems, but little else - not thoughts, cognitive content, or high-level problem-solving methods.

Any good programmer - a programmer with a feeling for aesthetics - knows the tedium of solving the same special case, over and over, in slightly different ways; and also the triumph of thinking through the metaproblem and creating a general solution that solves all the special cases simultaneously. As the hacker Jargon File observes, "Real hackers generalize uninteresting problems enough to make them interesting and solve them -- thus solving the original problem as a special

case (and, it must be admitted, occasionally turning a molehill into a mountain, or a mountain into a tectonic plate)." [Raymond01a]. This idiom *does not work* for general AI! A real AI would be the ultimate general solution because it would encapsulate the cognitive processes that human programmers use to write *any* specific piece of code, but this ultimate solution cannot be obtained through the technique of successively generalizing uninteresting problems into interesting ones.

Programming is the art of translating a human's mental model of a problem-solution into a computer program; that is, the art of translating thoughts into code. Programming *inherently* violates the levels of organization; it leads directly into the pitfalls of classical AI. The underlying low-level processes that implement intelligence are of a fundamentally different character than high-level intelligence itself. When we translate our thoughts about a problem into code, we are establishing a correspondence between code and the high-level *content* of our minds, not a correspondence between code and the dynamic process of a human mind. In ordinary programming, the task is to get a computer to solve a specific problem; it may be an "interesting" problem, with a very large domain, but it will still be a specific problem. In ordinary programming the problem is solved by taking the human thought process that would be used to solve an instance of the problem, and translating that thought process into code that can also solve instances of the problem. Programmers are humans who have learned the art of inventing thought processes, called "algorithms", that rely only on capabilities an ordinary computer possesses.

The reflexes learned by a good, artistic programmer represent a fundamental danger when embarking on a general AI project. Programmers are trained to solve problems, and trying to create general AI means solving the programming problem of creating a mind that solves problems. There is the danger of a short-circuit, of misinterpreting the problem task as writing code that directly solves some specific challenge posed to the mind, instead of building a mind that can solve the challenge with general intelligence. Code, when abused, is an excellent tool for creating long-term problems in the guise of short-term solutions.

Having described what we are *forbidden* to do with code, what *legitimate* challenges lie on this level of organization?

Some programming challenges are universal. Any modern programmer should be familiar with the world of compilers, interpreters, debuggers, Integrated Development Environments, multithreaded programming, object orientation, code reuse, code maintenance, and the other tools and traditions of modern-day programming. It is difficult to imagine anyone successfully coding the brainware level of general intelligence in assembly language - at least if the code is being developed for the first time. In that sense object orientation and other features of modern-day languages are "required" for AI development; but they are necessary as productivity tools, not because of any deep similarity between the structure of the programming language and the structure of general intelligence. Good programming tools help with AI *development* but do not help with *AI*.

Some programming challenges, although universal, are likely to be unusually severe in AI development. AI development is *exploratory, parallelized,* and *large*. Writing a great deal of exploratory code means that IDEs with refactoring support and version control are important, and that modular code is even more important than it is usually - or at least, code that is as modular as possible given the highly interconnected nature of the cognitive supersystem.

Parallelism on the hardware level is currently supported by symmetric multiprocessing chip architectures [Hwang98], NOW (network-of-workstations) clustering [Anderson95] and Beowulf clustering [Becker95], and message-passing APIs such as PVM [Geist93] and MPI [Gropp94]. However, software-level parallelism is not handled well by present-day languages and is therefore likely to present one of the greatest challenges. Even if software parallelism *were* well-supported, AI developers will still need to spend time explicitly thinking on how to parallelize cognitive processes - human cognition may be massively parallel on the lower levels, but the overall flow of cognition is still serial.

Finally, there are some programming challenges that are likely to be unique to AI.

We know it is possible to evolve a general intelligence that runs on a hundred trillion synapses with characteristic limiting speeds of approximately 200 spikes per second. An interesting property of human neurobiology is that, at a limiting speed of 150 meters per second for myelinated axons, each neuron is potentially within roughly a single "clock tick" of any other neuron in the brain[15]. [Sandberg99] describes a quantity S that translates to the wait time, in clock cycles, between different parts of a cognitive system - the minimum time it could take for a signal to travel between the most distant parts of the system, measured in the system's clock ticks. For the human brain, S is on the rough order of 1 - in theory, at least. In practice, axons take up space and myelinated axons take up even more space, so the brain uses a highly modular architecture, but there are still long-distance pipes such as the corpus callosum. Currently, S is much greater than 1 for clustered computing systems. S is greater than 1 even within a single-processor computer system; Moore's Law for intrasystem communications bandwidth describes a substantially slower doubling time than processor speeds. Increasingly the limiting resource of modern computing systems is not processor speed but memory bandwidth [Wulf95] (and this problem has gotten worse, rather than better, since 1995).

One class of purely programmatic problems that are unique to AI arise from the need to "port" intelligence from massively parallel neurons to clustered computing systems (or other human-programmable substrate). It is conceivable, for example, that the human mind handles the cognitive process of *memory association* by comparing current working imagery to all stored memories, in parallel. We have no particular evidence that the human mind uses a brute force comparison, but it *could* be brute-forced. The human brain acknowledges no distinction between CPU and RAM. If there are enough neurons to store a memory, then the same neurons may presumably be called upon to compare that memory to current experience. (This holds true even if the correspondence between neural groups and stored memories is many-to-many instead of one-to-one.)

Memory association may or may not use a "compare" operation (brute force or otherwise) of current imagery against all stored memories, but it seems likely that the brain uses a massively parallel algorithm at one point or another of its operation; memory association is simply a plausible candidate. Suppose that memory association is a brute-force task, performed by asking all neurons engaged in memory storage to perform a "compare" against patterns broadcast from current working imagery. Faced with the design requirement of matching the brute force of 10^{14} massively parallel synapses with a mere clustered system, a programmer may be tempted to despair. There is no *a priori* reason why such a task should be possible.

Faced with a problem of this class, there are two courses the programmer can take. The first is to implement an analogous "massive compare" as efficiently as possible on the available hardware - an algorithmic challenge worthy of Hercules, but past programmers have overcome massive computational barriers through heroic efforts and the relentless grinding of Moore's Law. The second road - much scarier, with even less of a guarantee that success is possible - is to *redesign the cognitive process* for different hardware.

The human brain's most fundamental limit is its speed. Anything that happens in less than a second perforce must use less than 200 *sequential* operations, however massively parallelized. If the human brain really does use a massively parallel brute-force compare against all stored memories to handle the problem of association, it's probably because there isn't time to do anything else! The human brain is massively parallel because massive parallelism is the only way to do *anything* in 200 clock ticks. If modern computers ran at 200Hz instead of 2GHz, PCs would also need 10^{14} processors to do anything interesting in realtime.

A sufficiently bold general AI developer, instead of trying to reimplement the cognitive process of association as it developed in humans, might instead ask: *What would this cognitive subsystem look like, if it had evolved on hardware instead of wetware?* If we remove the old constraint of needing to complete in a handful of clock ticks, and add the new constraint of not being able to offhandedly "parallelize against all stored memories", what is the *new* best algorithm for memory association? For example, suppose that you find a method of "fuzzy hashing" a memory, such that mostly similar memories automatically collide within a container space, but where the fuzzy hash inherently requires an extended linear series of sequential operations that would have placed "fuzzy hashing" out of reach for realtime neural operations. "Fuzzy hashing" would then be a strong candidate for an alternative implementation of memory association.

A computationally cheaper association subsystem that exploits serial speed instead of parallel speed, whether based around "fuzzy hashing" or something else entirely, might still be qualitatively less intelligent than the corresponding association system within the human brain. For example, memory recognition might be limited to clustered contexts rather than being fully general across all past experience, with the AI often missing "obvious" associations (where "obvious" has the anthropocentric meaning of "computationally easy for a human observer"). In this case, the question would be whether the overall general intelligence could function well enough to get by, perhaps compensating for lack of associational breadth by using longer linear chains of reasoning. The difference between serialism and parallelism, on a low level, would propagate upward to create cognitive differences that compensate for the loss of human advantages or exploit new advantages not shared by humans.

Another class of problem stems from "porting" across the extremely different *programming styles* of evolution versus human coding. Human-written programs typically involve a long series of chained dependencies that intersect at single points of failure - "crystalline" is a good term to describe most human code. Computation in neurons has a different character. Over time our pictures of biological neurons have evolved from simple integrators of synaptic inputs that fire when a threshold input level is reached, to sophisticated biological processors with mixed analog-digital logics, adaptive plasticity, dendritic computing, and functionally relevant dendritic and synaptic morphologies [Koch00]. What remains true is that, from an algorithmic perspective, neural computing uses roughly arithmetical operations[16] that proceed along multiple intertwining channels

in which information is represented redundantly and processed stochastically. Hence, it is easier to "train" neural networks - even nonbiological connectionist networks - than to train a piece of human-written code. Flipping a random bit inside the state of a running program, or flipping a random bit in an assembly-language instruction, has a much greater effect than a similar perturbation of a neural network. For neural networks the *fitness landscapes* are smoother. Why is this? Biological neural networks need to tolerate greater environmental noise (data error) and processor noise (computational error), but this is only the beginning of the explanation.

Smooth fitness landscapes are a useful, necessary, and fundamental outcome of evolution. Every evolutionary success starts as a mutation - an error - or as a novel genetic combination. A modern organism, powerfully adaptive with a large reservoir of genetic complexity, necessarily possesses a very long evolutionary history; that is, the genotype has necessarily passed through a very large number of successful mutations and recombinations along the road to its current form. The "evolution of evolvability" is most commonly justified by reference to this historical constraint [Dawkins96], but there have also been attempts to demonstrate local selection pressures for the characteristics that give rise to evolvability [Wagner96], thus averting the need to invoke the controversial agency of species selection. Either way, smooth fitness landscapes are part of the design signature of evolution.

"Smooth fitness landscapes" imply, among other things, that a small perturbation in the program code (genetic noise), in the input (environmental noise), or in the state of the executing program (processor noise), is likely to produce at most a small degradation in output quality. In most human-written code, a small perturbation of any kind usually causes a crash. Genomes are built by a cumulative series of point mutations and random recombinations. Human-written programs start out as high-level goals which are translated, by an extended serial thought process, into code. A perturbation to human-written code perturbs the code's final form, rather than its first cause, and the code's final form has no history of successful mutation. The thoughts that *gave rise* to the code probably have a smooth fitness metric, in the sense that a slight perturbation to the programmer's state of mind will probably produce code that is at most a little worse, and possibly a little better. Human thoughts, which are the original source of human-written code, are resilient; the code itself is fragile.

The dream solution would be a programming language in which human-written, top-down code somehow had the smooth fitness landscapes that are characteristic of accreted evolved complexity, but this is probably far too much to ask of a programming language. The difference between evolution and design runs deeper than the difference between stochastic neural circuitry and fragile chip architectures. On the other hand, using fragile building blocks can't possibly *help,* so a language-level solution might solve at least some of the problem.

The importance of smooth fitness landscapes holds true for all levels of organization. Concepts and thoughts should not break as the result of small changes. The code level is being singled out because smoothness on the code level represents a different kind of problem than smoothness on the higher levels. On the higher levels, smoothness is a product of correctly designed cognitive processes; a learned concept will apply to messy new data because it was abstracted from a messy experiential base. Given that AI complexity lying within the concept level requires smooth fitness landscapes, the correct strategy is to duplicate the smoothness on that level - to accept as a high-

level design requirement that the AI produce error-tolerant concepts abstracted from messy experiential bases.

On the code level, neural circuitry is smooth and stochastic by the nature of neurons and by the nature of evolutionary design. Human-written programs are sharp and fragile ("crystalline") by the nature of modern chip architectures and by the nature of human programming. The distinction is not likely to be erased by programmer effort or new programming languages. The long-term solution might be an AI with a sensory modality for code (see Part III), but that is not likely to be attainable in the early stages. The basic code-level "stuff" of the human brain has built-in support for smooth fitness landscapes, and the basic code-level "stuff" of human-written computer programs does not. Where human processes rely on neural circuitry being *automatically* error-tolerant and trainable, it will take additional programmatic work to "port" that cognitive process to a new substrate where the built-in support is absent. The final compromise solution may have error tolerance as one explicit design feature among many, rather than error-tolerance naturally emerging from the code level.

There are other important features that are also supported by biological neural networks - that are "natural" to neural substrate. These features probably include:

- Optimization for recurring problems;
- Completion of partial patterns;
- Similarity recognition (detection of static pattern repetition);
- Recurrence recognition (detection of temporal repetition);
- Clustering detection, cluster identification, and sorting into identified clusters;
- Training for pattern recognition and pattern completion;
- Massive parallelism.

Again, this does not imply an unbeatable advantage for biological neural networks. In some cases wetware has very *poor* feature support, relative to contemporary hardware. Contemporary hardware has better support for:

- Reflectivity and execution traces;
- Lossless serialization (storage and retrieval) and lossless pattern transformations;
- Very-high-precision quantitative calculations;
- Low-level algorithms which involve extended iteration, deep recursion, and complex branching;
- "Massive serialism"; the ability to execute hundreds of millions of *sequential* steps per second.

The challenge is using new advantages to compensate for the loss of old advantages, and replacing substrate-level support with design-level support.

This concludes the account of *exceptional* issues that arise at the code level. An enumeration of *all* issues that arise at the code level - for example, serializing the current contents of a sensory modality for efficient transmission to a duplicate modality on a different node of a distributed network - would constitute at least a third of a complete constructive account of a general AI. But programming is not all the work of AI, perhaps not even most of the work of AI; much of the effort needed to construct an intelligence will go into prodding the AI into forming certain concepts,

undergoing certain experiences, discovering certain beliefs, and learning various high-level skills. These tasks cannot be accomplished with an IDE. Coding the wrong thing successfully can mess up an AI project worse than any number of programming failures. I believe that the most important skill an AI developer can have is knowing what *not* to program.

2.4: The modality level

2.4.1: The evolutionary design of modalities in humans

Most students of AI are familiar with the high-level computational processes of at least one human sensory modality, vision, at least to the extent of being acquainted with David Marr's "2 1/2D world" and the concept of layered feature extraction [Marr82]. Further investigations in computational neuroscience have both confirmed Marr's theory and rendered it enormously more complex. Although many writers, including myself, have been known to use the phrase "visual cortex" when talking about the entire visual modality, this is like talking about the United States by referring to New York. About 50% of the neocortex of nonhuman primates is devoted exclusively to visual processing, with over 30 distinct visual areas identified in the macaque monkey [Felleman91].

The major visual stream is the retinal-geniculate-cortical stream, which goes from the retina to the lateral geniculate nucleus to the striate cortex[17] to the higher visual areas. Beyond the visual cortex, processing splits into two major secondary streams; the ventral stream heading toward the temporal lobe for object recognition, and the dorsal stream heading toward the parietal lobe for spatial processing. The visual stream begins in the retina, which contains around 100 million rods and 5 million cones, but feeds into an optic cable containing only around 1 million axons. Visual preprocessing begins in the first layer of the retina, which converts the raw intensities into center-surround gradients, a representation that forms the basis of all further visual processing. After several further layers of retinal processing, the final retinal layer is composed of a wide variety of ganglion types that include directionally selective motion detectors, slow-moving edge detectors, fast movement detectors, uniformity detectors, and subtractive color channels. The axons of these ganglions form the optic nerve and project to the magnocellular, parvocellular, and koniocellular layers of the lateral geniculate nucleus; currently it appears that each class of ganglion projects to only one of these layers. It is widely assumed that further feature detection takes place in the lateral geniculate nucleus, but the specifics are not currently clear. From the lateral geniculate nucleus, the visual information stream continues to area V1, the primary visual cortex, which begins feature extraction for information about motion, orientation, color and depth. From primary visual cortex the information stream continues, making its way to the higher visual areas, V2 through V6. Beyond the visual cortex, the information stream continues to temporal areas (object recognition) and parietal areas (spatial processing).

As mentioned earlier, primary visual cortex sends massive corticothalamic feedback projections to the lateral geniculate nucleus [Sherman86]. Corticocortical connections are also typically accompanied by feedback projections of equal strength [Felleman91]. There is currently no standard explanation for these feedback connections. DGI[18] requires sensory modalities with *feature controllers* that are the inverse complements of the *feature detectors;* this fits with the existence of the feedback projections. However, it should be noted that this assertion is not part of contemporary neuroscience. The existence of feature controllers is allowed for, but not asserted, by current theory;

their existence is asserted, and required, by DGI. (The hypothesis that feedback projections play a role in mental imagery is not limited to DGI; for example, [Kosslyn94] cites the existence of corticocortical feedback projections as providing an underlying mechanism for higher-level cognitive functions to control depictive mental imagery.)

The general lesson learned from the human visual modality is that modalities are not microtheories, that modalities are not flat representations of the pixel level, and that modalities are functionally characterized by successive layers of successively more elaborate feature structure. Modalities are one of the best exhibitions of this evolutionary design pattern - ascending layers of adaptive complexity - which also appears, albeit in very different form, in the ascending code-modality-concept-thought-deliberation model of the human mind. Each ascending layer is more elaborate, more complex, more flexible, and more computationally expensive. Each layer requires the complexity of the layer underneath - both functionally within a single organism, and evolutionarily within a genetic population.

The concept layer is evolvable in a series of short steps if, and only if, there already exists substantial complexity within the modality layer. The same design pattern - ascending layers of adaptive complexity - also appears *within* an evolved sensory modality. The first features detected are simple, and can evolve in a single step or a small series of adaptive short steps. The ability to detect these first features can be adaptive even in the absence of a complete sensory modality. The eye, which is currently believed to have independently evolved in many different species, may have begun, each time, as a single light-sensitive spot on the organism's skin.

In modalities, each additional layer of feature detectors makes use of the information provided by the first layer of feature detectors. In the absence of the first layer of feature detectors, the "code" for the second layer of feature detectors would be too complex to evolve in one chunk. With the first layer of feature detectors already present, feature detectors in the second layer can evolve in a single step, or in a short series of locally adaptive steps. The successive layers of organization in a sensory modality are a beautiful illustration of evolution's design signature, the functional ontogeny of the information recapitulating the evolutionary phylogeny.

Evolution is a good teacher but a poor role model; is this design a bug or a feature? I would argue that it is generally a feature. There is a deep correspondence between *evolutionarily* smooth fitness landscapes and *computationally* smooth fitness landscapes. There is a deep correspondence between each successive layer of feature detectors being evolvable, and each successive layer of feature detectors being computable in a way that is "smooth" rather than "fragile", as described in the earlier discussion of the code layer. Smooth computations are more evolvable, so evolution, in constructing a system incrementally, tends to construct linear sequences or ascending layers of smooth operations.

An AI designer may conceivably discard the requirement that each ascending layer of feature detection be incrementally useful/adaptive - although this may make the subsystem harder to incrementally develop and test! It is cognitively important, however, that successive layers of feature detectors be computationally "smooth" in one specific sense. DGI concepts interact with inverse feature detectors, "feature controllers", in order to construct mental imagery. For the task of *imposing a concept* and the still more difficult task of *abstracting a concept* to be *simultaneously* tractable, it is necessary that sensory modalities be a continuum of locally smooth layers, rather than consisting of

enormous, intractable, opaque chunks. There is a deep correspondence between the smooth design that renders concepts tractable and the smooth architecture emergent from incremental evolution.

The feature controllers used to create mental imagery are evolvable and preadaptive in the absence of mental imagery; feature controllers could begin as top-down constraints in perceptual processing, or even more simply as a perceptual step which happens to be best computed by a recurrent network. In both cases, the easiest (most evolvable) architecture is generally one in which the feedback connection reciprocates the feedforward connection. Thus, the feature controller layers are not a separate system independent from the feature detector layers; rather, I expect that what is locally a feature detector is also locally a feature controller. Again, this smooth reversibility helps render it possible to learn a single concept which can act as a category detector *or* a category imposer. It is the *simultaneous* solution of *concept imposition, concept satisfaction, concept faceting,* and *concept abstraction* that requires reversible features - feature controllers which are the local inverses of the feature detectors. I doubt that feature controllers reach all the way down to the first layers of the retina (I have not heard of any feedback connections reaching this far), but direct evidence from neuroimaging shows that mental imagery activates primary visual cortex [Kosslyn93]; I am not sure whether analogous tests have been performed for the lateral geniculate nucleus, but the feedback connections are there.

2.4.2: The human design of modalities in AI

An AI needs sensory modalities - but which modalities? How do those modalities contribute materially to general intelligence outside the immediate modality?

Does an AI need a visuospatial system modeled after the grand complexity of the visuospatial system in primates and humans? We know more about the human visual modality than about any other aspect of human neurology, but that doesn't mean we know enough to build a visual modality from scratch. Furthermore, the human visual modality is enormously complex, computationally intensive, and fitted to an environment which an AI does not necessarily have an immediate need to comprehend. Should humanlike 3D vision[19] be one of the *first* modalities attempted?

I believe it will prove best to discard the human modalities or to use them as inspiration only - to use a completely different set of sensory modalities during the AI's early stages. An AI occupies a different environment than a human and direct imitation of human modalities would not be appropriate. For an AI's initial learning experiences, I would advocate placing the AI in complex virtual environments, where the virtual environments are *internal to the computer* but *external to the AI*. The programmers would then attempt to develop sensory modalities corresponding to the virtual environments. Henceforth I may use the term "microenvironment" to indicate a complex virtual environment. The term "microworld" is less unwieldy, but should not be taken as having the Good Old-Fashioned AI connotation of "microworlds" in which all features are directly represented by predicate logic, e.g., SHRDLU's simplified world of blocks and tables [Winograd72].

Abandoning the human modalities appears to introduce an additional fragile dependency on the correctness of the AI theory, in that substituting novel sensory modalities for the human ones would appear to require a correct understanding of the nature of sensory modalities and how they contribute to intelligence. This is true, but I would argue that the existence of an *additional* dependency is illusory. An attempt to blindly imitate the human visual modality, without

understanding the role of modalities in intelligence, would be unlikely to contribute to general intelligence except by accident. Our modern understanding of the human visual modality is not so perfect that we could rely on the functional completeness of a neurologically inspired design; for example, a design based only on consensus contemporary theory might omit feature controllers! However, shifting to microworlds does require that experience in the microworlds reproduce functionally relevant aspects of experience in real life, including unpredictability, uncertainty, real-time process control, holonic (part-whole) organization, et cetera. I do not believe that this introduces an *additional* dependency on theoretic understanding, over and above the theoretic understanding that would be required to build an AI that absorbed complexity from these aspects of real-world environments, but it nonetheless represents a strong dependency on theoretic understanding.

Suppose that we are designing, *de novo,* a sensory modality and virtual environment. Three possible modalities that come to mind as reasonable for a *very primitive and early-stage AI,* in ascending order of implementational difficulty, would be:

1. A modality for Newtonian billiard balls;
2. A modality for a 100x100 "Go" board;
3. A modality for some type of interpreted code (a metaphorical "codic cortex").

In human vision, the very first visual neurons are the "rods and cones" which transduce impinging environmental photons to a neural representation as sensory information. For each of the three modalities above, the "rods and cones" level would probably use essentially the same representation as the data structures used to create the microworld, or virtual environment, in which the AI is embodied. This is a major departure from the design of naturally evolved modalities, in which the basic level - the quark level, as far as we know - is many layers removed from the high-level objects that give rise to the indirect information that reaches the senses. Evolved sensory modalities devote most of their complexity to *reconstructing* the world that gives rise to the incoming sensory impressions - to reconstructing the 3D moving objects that give rise to the photons impinging on the rods-and-cones layer of the retina. Of course, choosing vision as an example is arguably a biased selection; sound is not as complex as vision, and smell and taste are not as complex as sound. Nonetheless, eliminating the uncertainty and intervening layers between the true environment and the organism's sensory data is a major step. It should significantly reduce the challenges of early AI development, but is a dangerous step nonetheless because of its distance from the biological paradigm and its elimination of a significant complexity source.

I recommend eliminating environmental reconstruction as a complexity source in *early* AI development. Visualizing the prospect of deliberately degrading the quality of the AI's environmental information on one end, and elaborating the AI's sensory modality on the other end, I find it likely that the entire operation will cancel out, contributing nothing. An AI that had to learn to reconstruct the environment, in the same way that evolution learned to construct sensory modalities, might produce interesting complexity as a result; but if the same programmer is creating environmental complexity and modality complexity, I would expect the two operations to cancel out. While environmental reconstruction is a nontrivial complexity source within the human brain, I consider the *ratio* between the difficulty of programmer development of the complexity, and the contribution of that complexity to general intelligence, to be relatively small. Adding complexity for environmental reconstruction, by introducing additional layers of complexity in the microworld and

deliberately introducing information losses between the topmost layer of the microworld and the AI's sensory receptors, and then attempting to create an AI modality which could reconstruct the original microworld content from the final sensory signal, would require a relatively great investment of effort in return for what I suspect would be a relatively small boost to general intelligence.

Suppose that for each of the three modalities - billiards, Go, code - the "pre-retinal" level consists of true and accurate information about the quark level of the virtual microworld, although perhaps not complete information, and that the essential complexity which renders the model a "sensory modality" rests in the feature structure, the ascending layers of feature detectors and descending layers of feature controllers. Which features, then, are appropriate? And how do they contribute materially to general intelligence?

The usual statement is that the complexity in a sensory modality reflects regularities of the environment, but I wish to offer a slightly different viewpoint. To illustrate this view, I must borrow and severely simplify the punchline of a truly elegant paper, "The Perceptual Organization of Colors" by Roger Shepard [Shepard92]. Among other questions, this paper seeks to answer the question of trichromancy: Why are there three kinds of cones in the human retina, and not two, or four? Why is human visual perception organized into a three-dimensional color space? Historically, it was often theorized that trichromancy represented an arbitrary compromise between chromatic resolution and spatial resolution; that is, between the number of colors perceived and the grain size of visual resolution. As it turns out, there is a more fundamental reason why three color channels are needed.

To clarify the question, consider that surfaces possess a potentially infinite number of spectral reflectance distributions. We will focus on spectral reflectance distributions, rather than spectral power distributions, because adaptively relevant objects that emit their own light are environmentally rare. Hence the physically constant property of most objects is the spectral reflectance distribution, which combines with the spectral power distribution of light impinging on the object to give rise to the spectral power distribution received by the human eye. The spectral reflectance distribution is defined over the wavelengths from 400nm to 700nm (the visible range), and since wavelength is a continuum, the spectral reflectance distribution can theoretically require an unlimited number of quantities to specify. Hence, it is not possible to exactly constrain a spectral reflectance distribution using only three quantities, which is the amount of information transduced by human cones.

The human eye is not capable of discriminating among all physically possible reflecting surfaces. However, it is possible that for "natural" surfaces - surfaces of the kind commonly encountered in the ancestral environment - reflectance for each pure frequency does not vary independently of reflectance for all other frequencies. For example, there might exist some set of *basis reflectance functions,* such that the reflectance distributions of almost all natural surfaces could be expressed as a weighted sum of the basis vectors. If so, one possible explanation for the trichromancy of human vision would be that three color channels are just enough to perform adequate discrimination in a "natural" color space of limited dimensionality.

The ability to discriminate between all natural surfaces would be the design recommended by the "environmental regularity" philosophy of sensory modalities. The dimensionality of the internal model would mirror the dimensionality of the environment.

As it turns out, natural surfaces have spectral reflectance distributions that vary along roughly five to seven dimensions [Maloney86]. There thus exist natural surfaces that, although appearing to trichromatic viewers as "the same color", nonetheless possess different spectral reflectance distributions.

[Shepard92] instead asks how many color channels are needed to ensure that the color we perceive is the *same* color each time the surface is viewed under different lighting conditions. The amount of ambient light can also potentially vary along an unlimited number of dimensions, and the actual light reaching the eye is the product of the spectral power distribution and the spectral reflectance distribution. A reddish object in bluish light may reflect the same number of photons of each wavelength as a bluish object in reddish light. Similarly, a white object in reddish light may reflect mostly red photons, while the same white object in bluish light may reflect mostly blue photons. And yet the human visual system manages to maintain the property of *color constancy;* the same object will appear to be the same color under different lighting conditions.

[Judd64] measured 622 spectral power distributions for natural lighting, under 622 widely varying natural conditions of weather and times of day, and found that variations in natural lighting reduce to three degrees of freedom. Furthermore, these three degrees of freedom bear a close correspondence to the three dimensions of color opponency that were proposed for the human visual system based on experimental examination [Hurvich57]. The three degrees of freedom are:

- The *light-dark* variation, which depends on the total light reaching the object.
- The *yellow-blue* variation, which depends on whether a surface is illuminated by direct sunlight or is in shade. In shade the surface is illuminated by the Raleigh-scattered blue light of the sky, but is not directly illuminated by the sun. The corresponding yellow extreme occurs when an object is illuminated only by direct sunlight; e.g., if sunlight enters through a small channel and skylight is cut off.
- The *red-green* variation, which depends on both the elevation of the sun (how much atmosphere the sun travels through), and the amount of atmospheric water vapor. E.g., illumination by a red sunset versus illumination at midday. Red wavelengths are the wavelengths least scattered by dust and most absorbed by water.

The three color channels of the human visual system are precisely the number of channels needed in order to maintain color constancy under natural lighting conditions[20]. Three color channels are not enough to discriminate between all natural surface reflectances, but three color channels are the exact number required to compensate for ambient natural lighting and thereby ensure that the same surface is perceptually the "same color" on any two occasions. This simplifies the adaptively important task of recognizing a previously experienced object on future encounters.

The lesson I would learn from this tale of color constancy is that sensory modalities are about *invariants* and not just *regularities*. Consider the task of designing a sensory modality for some form of interpreted code. (This is a very challenging task because human programming languages tend toward non-smooth fitness landscapes, as previously discussed.) When considering which features to extract, the question I would ask is not "What regularities are found in code?" but rather "What feature structure is needed for the AI to perceive two identical algorithms with slightly different implementations as 'the same piece of code'?" Or more concretely: "What features does this

modality need to extract to perceive the recursive algorithm for the Fibonacci sequence and the iterative algorithm for the Fibonacci sequence as 'the same piece of code'?"

Tip your head slightly to the left, then slightly to the right. Every retinal receptor may receive a different signal, but the experienced visual field remains almost exactly the "same". Hold up a chess pawn, and tip it slightly to the left or slightly to the right. Despite the changes in retinal reception, we see the "same" pawn with a slightly different orientation. Could a sensory modality for code look at two sets of interpreted bytecodes (or other program listing), completely different on a byte-by-byte basis, and see these two listings as the "same" algorithm in two slightly different "orientations"?

The modality level of organization, like the code level, has a characteristic kind of work that it performs. Formulating a *butterfly* concept and seeing two butterflies as members of the same category is the work of the concept level, but seeing a chess pawn in two orientations as the same pawn is the work of the modality level. There is overlap between the modality level and the concept level, just as there is overlap between the code level and the modality level. But on the whole, the modality level is about *invariants* rather than *regularities* and *identities* rather than *categories*.

Similarly, the understanding conferred by the modality level should not be confused with the analytic understanding characteristic of thoughts and deliberation. Returning to the example of a codic modality, one possible indication of a serious design error would be constructing a modality that could analyze *any possible* piece of code equally well. The very first layer of the retina - rods and cones - is the *only* part of the human visual system that will work on all possible pixel fields. The rest of the visual system will only work for the low-entropy pixel fields experienced by a low-entropy organism in a low-entropy environment. The very next layer, after rods and cones, already relies on center-surround organization being a useful way to compress visual information; this only holds true in a low-entropy visual environment.

Designing a modality that worked equally well for any possible computer program would probably be an indication that the modality was extracting the wrong kind of information. Thus, one should be wary of an alleged "feature structure" that looks as if it would work equally well for all possible pieces of code. It may be a valid analytical method but it probably belongs on the deliberation level, not the modality level. (Admittedly not *every* local step of a modality *must* be dependent on low-entropy input; some local stages of processing may have the mathematical nature of a lossless transform that works equally well on any possible input. Also, hardware is probably better suited than wetware to lossless transforms.)

The human brain is constrained by a characteristic serial speed of 200 sequential steps per second, and by the ubiquitous internal use of the synchronous arrival of associated information, to arrange processing stages that flow smoothly forward. High-level "if-then" or "switch-case" logic is harder to arrive at neurally, and extended complex "if-then" or "switch-case" logic is probably almost impossible unless implemented through branching *parallel* circuitry that remains synchronized. Probably an exceptional condition must be ignored, averaged out, or otherwise handled using the same algorithms that would apply to any other modality content. Can an AI modality use an architecture that applies different algorithms to different pieces of modality content? Can an AI modality handle exceptional conditions through special-case code? I would advise caution, for several reasons. First, major "if-then" branches are characteristic of deliberative

processes, and being tempted to use such a branch may indicate a level confusion. Second, making exceptions to the smooth flow of processing will probably complicate the meshing of concepts and modalities. Third, modalities are imperfect but fault-tolerant processes, and the fault tolerance plays a role in smoothing out the fitness landscapes and letting the higher levels of organization be built on top; thus, trying to handle *all* the data by detecting exceptional conditions and correcting them, a standard pattern in human programming, may indicate that the modality is insufficiently fault-tolerant. Fourth, handling all exceptions is characteristic of trying to handle all inputs and not just low-entropy inputs. Hence, on the whole, sensory modalities are characterized by the smooth flow of information through ascending layers of feature detectors. Of course, detecting an exceptional condition *as a feature* may turn out to be entirely appropriate!

Another issue which may arise in artificial sensory modalities is that unsophisticated artificial modalities may turn out to be significantly more expensive, computationally, for the effective intelligence they deliver. Sophisticated evolved modalities conserve computing power in ways that might be very difficult for a human programmer to duplicate. An example would be the use of partial imagery, modeling only the features that are needed for a high-level task [Hayhoe98]; a simplified modality that does not support partial imagery may consume more computing power. Another example would be the human visual system's selective concentration on the center of the visual field - the "foveal architecture", in which areas of the visual field closer to the center are allocated a greater number of neurons. The cortical magnification factor for primates is inverse-linear [Tootell85]; the complex logarithm is the only two-dimensional map function that has this property [Schwartz77], as confirmed experimentally by [Schwartz89]. A constant-resolution version of the visual cortex, with the maximum human visual resolution across the full human visual field, would require 10,000 times as many cells as our actual cortex [Rojer90].

But consider the programmatic problems introduced by the use of a logarithmic map. Depending on where an object lies in the visual field, its internal representation on a retinotopic map will be completely different; no direct comparison of the data structures would show the identity or even hint at the identity. That an off-center object in our visual field can rotate without *perceptually* distorting, as its image distorts wildly within the physical retinotopic map, presents a nontrivial computational problem[21].

Evolution conserves computing power by complicating the algorithm. Evolution, considered as a *design pressure,* exerts a steady equipotential design pressure across all existing complexity; a human programmer wields general intelligence like a scalpel. It is not much harder for evolution to "design" and "debug" a logarithmic visual map because of this steady "design pressure"; further adaptations can build on top of a logarithmic visual map almost as easily as a constant-resolution map. A human programmer's general intelligence would run into difficulty keeping track of all the simultaneous design complications created by a logarithmic map. It might be possible, but it would be difficult, especially in the context of exploratory research; the logarithmic map transforms simple design problems into complex design problems and hence transforms complex design problems into nightmares.

I would suggest using constant-resolution sensory modalities during the early stages of an AI - as implied above by suggesting a sensory modality modeled around a 100x100 Go board - but the implication is that these early modalities will be lower-resolution, will have a smaller field, and will be less efficient computationally. An opposing theoretic view would be that complex but efficient

modalities introduce necessary issues for intelligence. An opposing pragmatic view would be that complex but efficient modalities are easier to accommodate in a mature AI if they have been included in the architecture from the beginning, so as to avoid metaphorical "Y2K" issues (ubiquitous dependencies on a simplifying assumption which is later invalidated).

2.5: The concept level

DGI uses the term *concept* to refer to the mental stuffs underlying the words that we combine into sentences; concepts are the combinatorial building blocks of thoughts and mental imagery. These building blocks are learned complexity, rather than innate complexity; they are abstracted from experience. Concept structure is absorbed from recurring regularities in perceived reality.

A concept is *abstracted* from experiences that exist as sensory patterns in one or more modalities. Once abstracted, a concept can be compared to a new sensory experience to determine whether the new experience *satisfies* the concept, or equivalently, whether the concept *describes* a facet of the experience. Concepts can describe both environmental sensory experience and internally generated mental imagery. Concepts can also be *imposed* on current working imagery. In the simplest case, an exemplar associated with the concept can be loaded into the working imagery, but constructing complex mental imagery requires that a concept target a piece of existing mental imagery, which the concept then transforms. Concepts are faceted; they have internal structure and associational structure which comes into play when imposition or description encounters a bump in the road. Faceting can also be invoked purposefully; for example, "tastes like chocolate" versus "looks like chocolate".

A "concept kernel" is the pseudo-sensory pattern produced by abstracting from sensory experience. During concept satisfaction, this kernel interacts with the layered feature detectors to determine whether the reported imagery matches the kernel; during concept imposition, the kernel interacts with the layered feature controllers to produce new imagery or alter existing imagery. A programmer seeking a good representation for concept kernels must find a representation that *simultaneously* fulfills these requirements:

- (a) The kernel representation can be *satisfied by* and *imposed on* referents in a sensory modality.
- (b) The kernel representation or concept representation contains the internal structure needed for faceted concept combination, as in "triangular lightbulb" previously given as an example.
- (c) It is computationally tractable to *abstract* new kernel representations using sensory experience as raw material.

It would be a serious challenge to solve any one of these problems individually, with sufficient generality and using a computationally tractable method; solving all three problems simultaneously is the fundamental challenge of building a system that learns complexity in combinatorial chunks.

Concepts have other properties besides their complex kernels. Kernels relate concepts to sensory imagery and hence to the modality level. Concepts also have complexity that relates to the concept level; i.e., concepts have complexity that derives from their relation to other concepts. In Good Old-Fashioned AI this aspect of concepts has been emphasized at the expense of all others[22], but this is no excuse for ignoring concept-concept relations in a new theory. For example, concepts are

supercategories and subcategories of each other; there are concepts that describe concepts; there are concepts that describe relations between concepts; there are mutually exclusive concepts which cannot simultaneously describe the same referent. (Further examples of concept relations are given later.)

In formal logic, the traditional idea of concepts is that concepts are categories defined by a set of individually necessary and together sufficient requisites; that a category's extensional referent is the set of events or objects that are members of the category; and that the combination of two categories is the sum of their requisites and hence the intersection of their sets of referents. This formulation is inadequate to the complex, messy, overlapping category structure of reality and is incompatible with a wide range of established cognitive effects [Lakoff87]. Properties such as *usually* necessary and *usually* sufficient requisites, and concept combinations that are *sometimes* the sum of their requisites or the intersection of their extensional classes, are emergent from the underlying representation of concepts - along with other important properties, such as prototype effects in which different category members are assigned different degrees of typicality [Rosch78].

Concepts relate to the thought level primarily in that they are the building blocks of thoughts, but there are other level-crossings as well. Introspective concepts can describe beliefs and thoughts and even deliberation; the concept "thought" is an example. Inductive generalizations are often "about" concepts in the sense that they apply to the referents of a concept; for example, "Triangular lightbulbs are red." Deliberation may focus on a concept in order to arrive at conclusions about the extensional category, and introspective deliberation may focus on a concept in its role as a cognitive object. Concept structure is ubiquitously invoked within perceptual and cognitive processes because category structure is ubiquitous in the low-entropy processes of our low-entropy universe.

2.5.1: The substance of concepts

One of the meanings of "abstraction" is "removal"; in chemistry, to *abstract* an atom means subtracting it from a molecular group. Using the term "abstraction" to describe the process of creating concepts could be taken as implying two views: First, that to create a concept is to generalize; second, that to generalize is to lose information. Abstraction as information loss is implicit in the classical view of concepts (that is, the view of concepts under GOFAI and formal logic). Forming the concept "red" is taken to consist of focusing only on color, at the expense of other features such as size and shape; all concept usage is held to consist of purposeful information-loss.

The problem with the classical view is that it allows only a limited repertoire of concepts. True, some concepts apparently work out to straightforward information-loss. The task of arriving at a concept kernel for the concept "red" - a kernel capable of interacting with visual imagery to distinguish between red objects and non-red objects - is relatively trivial. Even simultaneously satisfying the abstraction and satisfaction problems for "red" is relatively trivial. Well-known, fully general tools such as neural nets or evolutionary computation would suffice. To learn to solve the satisfaction problem, a neural net need only to learn to fire when the modality-level feature detectors for "color" report a certain color - a point falling within a specific volume of color space - across a broad area, and not to fire otherwise. A piece of code need only evolve to test for the same characteristic. (The neural net would probably train faster for this task.)

A sufficiently sophisticated modality would simplify the task even further, doing most of the work of grouping visual imagery into objects and detecting solid-color or same-hue or mostly-the-same-hue surfaces. The human visual modality goes still farther and precategorizes colors, dividing them up into a complex color space [Boynton87], said color space having eleven culturally universal focal volumes [Berlin69], said focal volumes having comparatively sharp internal boundaries relative to physically continuous variations in wavelength (see [Shepard92], or just look at the *bands* in a rainbow). Distinguishing across innate color boundaries is easy; distinguishing within color boundaries is hard [Mervis75]. Thus, the human visual modality provides very strong suggestions as to where the boundaries lie in color space, although the final step of categorization is still required [Dedrick98].

Given a visual modality, the concept of *red* lies very close to the metaphorical "surface" of the modality. In humans *red* is probably at the surface, a direct output of the modality's feature-detectors. In AIs with less sophisticated visual modalities, "redness" as a category would need to be abstracted as a fuzzy volume within a smooth color space lacking the human boundaries. The *red* concept kernel (in humans and AIs) needs to be more complex than a simple binary test or fuzzy color clustering test, since "redness" as we understand it describes visual areas and not single pixels (although *red* can describe a "visual area" consisting of a small point). Even so, the complexity involved in the redness concept lies almost entirely within the sensory modality, rather than the concept kernel. We might call such concepts *surface concepts*.

Even for surface concepts, simultaneously solving abstraction, satisfaction, and imposition would probably be far more tractable with a special representation for concept kernels, rather than generically trained neural nets or evolutionary programs. Imposition requires a concept kernel which can be selectively applied to imagery within a visual modality, transforming that imagery such that the final result satisfies the concept. In the case of the concept "red", the concept kernel would interact with the feature controllers for color, and the targeted mental imagery would become red. This cannot be done by painting each individual pixel the same shade of red; such a transformation would obliterate edges, surfaces, textures, and many other high-level features that intuitively ought to be preserved. Visualizing a "red lemon" does not cause the mind to picture a bright red patch with the outline of a lemon. The concept kernel does not send separate color commands to the low-level feature controller of each individual visual element; rather the concept kernel imposes *red* in combination with other currently activated features, to depict a *red lemon* that retains the edge, shape, surface curvature, texture, and other visualized features of the starting *lemon* image. Probably this occurs because perceived coloration is a property of surfaces and visual objects rather than, or as well as, individual visual elements, and our redness concept kernel interacts with this high-level feature, which then ripples down in coherent combination with other features.

Abstracting an impose-able concept kernel for "red" is a problem of different scope than abstracting a satisfy-able kernel for "red". There is an immediately obvious way to train a neural net to detect satisfaction of "red", given a training set of known "red" and non-"red" experiences, but there is no equally obvious teaching procedure for the problem of *imposing* "red". The most straightforward success metric is the degree to which the transformed imagery satisfies a neural network already trained to detect "red", but a bright red lemon-shaped patch is likely to be more "red" than a visualized red lemon. How does the kernel arrive at a transformation which makes a coherent change in object coloration, rather than a transformation which paints all visual elements an

indiscriminate shade of red, or a transformation which loads a random red object into memory? Any of these transformations would satisfy the "red" concept.

Conceivably fully general neural nets could be trained to impose *minimal* transformations, although I am not sure that "minimal transformation" is the rule which should govern concept imposition. Regardless of the real tractability of this problem, I strongly doubt that human cognitive systems create concepts by training generic neural nets on satisfaction and imposition. I suspect that concepts do not have independent procedures for satisfaction and imposition; I also suspect that neither satisfaction nor imposition are the product of reinforcement learning on a fully general procedure. Rather, I suspect that a concept kernel consists of a pattern in a representation related to (but not identical with) the representation of sensory imagery, that this pattern is produced by transforming the experiences from which the concept is abstracted, and that this pattern interacts with the modality to implement both concept satisfaction and concept imposition.

A very simple example of a non-procedural, pattern-based concept kernel would be "clustering on a single feature". *Red* might be abstracted from an experiential base by observing an unusual clustering of point values for the *color* feature. Suppose that the AI is challenged with a virtual game in which the goal is to find the "keys" to a "lock" by selecting objects from a large sample set. When the AI successfully passes five trials by selecting the correct object on the first try, the AI is assumed to have learned the rule. Let us suppose that the game rule is that "red" objects open the lock, and that the AI has already accumulated an experiential base from its past failures and successes on individual trials.

Assuming the use of a three-dimensional color space, the *color* values of the correct keys would represent a tight cluster relative to the distribution among all potential keys. Hence the abstracted concept kernel might take the form of a feature-cluster pair, where the feature is *color* and the cluster is a central point plus some measure of standard deviation. This creates a concept kernel with a prototype and quantitative satisfiability; the concept has a central point and fuzzy but real boundaries. The same concept kernel can also be imposed on a selected piece of mental imagery by loading the central color point into the *color* feature controller - that is, loading the clustered value into the feature controller corresponding to the feature detector clustered upon.

Clustering of this type also has indirect implications for concept-concept relations: The *red* concept's "color volume" might overlap a nearby concept such as *burgundy,* or might turn out to enclose that concept; a modality-level fact which over time might naturally give rise to an association relationship, or a supercategory relationship, on the concept level. This would not humanly occur through direct comparison of the representations of the concept kernels, but through the observation of overlap or inclusion within the categories of extensional referents. A more strongly introspective AI might occasionally benefit from inspecting kernel representations, but this should be an adjunct to experiential detection of category relationships, not a substitute for it.

Clustering on a single feature is definitely not a complete conceptual system. Single-feature clustering cannot notice a correlation between two features where neither feature is clustered alone; single-feature clustering cannot cross-correlate two features in any way at all. Concepts which are limited to clustering on a single feature will always be limited to concepts at the immediate surface of a given sensory modality.

At the same time, a concept system is not a general intelligence and need not be capable of representing *every possible* relation. Suppose a human were challenged with a game in which the "correct key" always had a color that lay on the exact surface of a sphere in color space; could the human concept-formation system *directly* abstract this property? I would guess not; I would guess that, at most, a human might notice that the key tended to belong to a certain group of colors; i.e., might slice up the surface of this color sphere into separate regions, and postulate that solution keys belong to one of several color regions. Thus, even though in this case the underlying "rule" is *computationally* very simple, it is unlikely that a human will create a concept that directly incorporates the rule; it may even be impossible for a human to abstract a kernel that performs this simple computation. A concept-formation system need not be generally intelligent in itself; need not represent all possible perceptual regularities; just enough for the overall mind to work.

I suspect that the system design used by humans, and a good design for AIs, will turn out to be a repertoire of different concept-formation methods. ("Clustering on a single feature" could be one such method, or could be a special case of a more general method.) Concept faceting could then result either from concepts with multiple kernels, so that a concept employs more than one categorization method against its perceptual referents, or from internal structure in a single kernel, or both. If some aspects of perceptual referents are more salient, then kernels which match those aspects are likely to have greater weight within the concept. Faceting within a concept, arising out of multiple unequal kernels or faceting within a single complex kernel, seems like the most probable source of prototype effects within a category.

2.5.2: Stages in concept processes

Concept formation is a multi-stage process. For an AI to form a new concept, the AI must have the relevant experiences, perceptually group the experiences, notice possible underlying similarities within members of a group (this may be the same perceived similarity that led to the original experiential grouping), verify the generalization, initiate the new concept as distinguished cognitive content, create the concept kernel(s) by abstraction from the experiential base, and integrate the new concept into the system. (This checklist is intended as an interim approximation; actual mind designs may differ, but presumably a temporal sequence will still be involved.)

In the example given earlier, an AI abstracts *redness* starting with a bottom-up, experience-driven event: noticing the possible clustering of the *color* feature within the preexisting category *keys*. Conceivably the act of checking for color clustering could have been suggested top-down, for example by some heuristic belief, but in this example we will assume the seminal perception of similar coloration was an unexpected, bottom-up event; the product of continuous and automatic checks for clustering on a single feature across all high-level features in currently salient experiential categories. Rather than being part of an existing train of thought, the detection of clustering creates an "Aha!" event, a new cognitive event with high salience that becomes the focus of attention, temporarily shunting aside the previous train of thought. (See the discussion of the thought level.)

If the scan for clustering and other categorizable similarities is a continuous background task, it may imply a major expenditure of computational resources - perhaps a major percentage of the computing power used by the AI. This is probably the price of having a cognitive process that can be driven by bottom-up interrupts as well as top-down sequences, and the price of having a cognitive process that can occasionally notice the unexpected. Hence, the efficiency, optimization,

and scalability of algorithms for such continuous background tasks may play a major role in determining the AI's performance. If imagery stays in place long enough, I would speculate that it may be possible to farm out the task of *noticing* a possible clustering to distant parts of a distributed network, while keeping the task of *verifying* the clustering, and all subsequent cognitive actions, within the local process. Most of the computing power is required to find the hint, not to verify the match, and a false hint does no damage (assuming the false hints are not malicious attacks from untrusted nodes).

Once the suspicion of similarity is triggered by a cue picked up by a continuous background process, and the actual degree of similarity is verified, the AI would be able to create the concept as cognitive content. Within the above example, the process that notices the possible clustering is essentially the same process that would verify the clustering and compute the degree of clustering, center of clustering, and variance within the cluster. Thus, clustering on a single feature may compress into a single stage the cueing, description, and abstraction of the underlying similarity. Given the expense of a continuous background process, however, I suspect it will usually be best to separate out a less expensive cueing mechanism as the background process, and use this cueing mechanism to suggest more detailed and expensive scans. (Note that this is a "parallel terraced scan"; see [Rehling97] and [Hofstadter95].)

After the creation of the concept and the concept kernel(s), it would then be possible for the AI to notice concept-concept relations, such as supercategory and subcategory relations. I do not believe that concept-concept relations are computed by directly comparing kernel representations; I think that concept-concept relations are learned by generalizing across the concept's usage. It may be a good heuristic to look for concept-concept relations immediately after forming a new concept, but that would be a separate track within deliberation, not an automatic part of concept formation.

After a concept has been formed, the new concept must be integrated into the system. For us to concede that a concept has really been "integrated into the system" and is now contributing to intelligence, the concept must be used. Scanning across the stored base of concepts, in order to find which concepts are satisfied by current mental imagery, promises to be an even more computationally expensive process than continuous background checks for clustering. An individual satisfaction check is probably less computationally intensive than carrying out a concept imposition - but satisfaction checks seem likely to be a continuous background operation, at least in humans.

As discussed earlier, humans and AIs have different computational substrates: Humans are slow but hugely parallel; AIs are fast, but resource-poor. If humans turn out to routinely parallelize against all learned concepts, an AI may simply be unable to afford it. The AI optimum may involve comparing working imagery against a smaller subset of learned complexity - only a few concepts, beliefs, or memories would be scanned against working imagery at any given point. Alternatively, an AI may be able to use terraced scanning[23], fuzzy hashing[24], or branched sorting[25] to render the problem tractable. One hopeful sign is the phenomenon of cognitive priming on related concepts [Meyer71], which suggests that humans, despite their parallelism, are not using pure brute force. Regardless, I conjecture that matching imagery against large concept sets will be one of the most computationally intensive subprocesses in AI, perhaps *the* most expensive subprocess. Concept matching is hence another good candidate for distribution under "notice distantly, verify locally"; note also that the concept base could be sliced up among distributed processors, although this might prevent matching algorithms from exploiting regularities within the concept base and matching process.

2.5.3: Complex concepts and the structure of "five"

Under the classical philosophy of category abstraction, abstraction consists solely of selective focus on information which is already known; focusing on the "color" or "redness" of an object as opposed to its shape, position, or velocity. In DGI's "concept kernels", the internal representation of a concept has complexity extending beyond information loss - even for the case of "redness" and other concepts which lie almost directly on the surface of a sensory modality. The only concept that is pure information-loss is a concept that lies *entirely* on the surface of a modality; a concept whose satisfaction exactly equals the satisfaction of some single feature detector.

The concept for "red", described earlier, is actually a fuzzy percept for degrees of redness. Given that the AI has a flat color space, rather than a human color space with innate focal volumes and color boundaries, the "redness" percept would contain at least as much additional complexity - over and above the modality-level complexity - as is used to describe the clustering. For example, "clustering on a single feature" might take the form of describing a Gaussian distribution around a central point. The specific use of a Gaussian distribution does not contribute to useful intelligence unless the environment also exhibits Gaussian clustering, but a Gaussian distribution is probably useful for allowing an AI to notice a wide class of clusterings around a central point, even clusterings that do not actually follow a Gaussian distribution.

Even in the absence of an immediate environmental regularity, a concept can contribute to effective intelligence by enabling the perception of more complex regularities. For example, an alternating sequence of "red" and "green" key objects may fail the modality-level tests for clustering because no Gaussian cluster contains (almost) all successes and excludes (almost) all failures. However, if the AI has already previously developed concepts for "red" and "green", the alternating repetition of the satisfaction of the "red" and "green" concepts is potentially detectable by higher-level repetition detectors. Slicing up the color space with surface-level concepts renders computationally tractable the detection of higher-order alternation. Even the formation of simple concepts - concepts lying on the surface of a modality - expands the perceptual capabilities of the AI and the range of problems the AI can solve.

Concepts can also embody regularities which are not directly represented in any sensory modality, and which are not any covariance or clustering of feature detectors already in a sensory modality.

Melanie Mitchell and Douglas Hofstadter's "Copycat" program works in the domain of letter-strings, such as "abc", "xyz", "onml", "ddd", "cwj", etc. The function of Copycat is to complete analogy problems such as "abc:abd::ace:?" [Hofstadter88]. Since Copycat is a model of perceptual analogy-making, rather than a model of category formation, Copycat has a limited store of preprogrammed concepts and does not learn further concepts through experience. (This should *not* be taken as criticism of the Copycat project; the researchers explicitly noted that concept formation was not being studied.)

Suppose that a general AI (not Copycat), working in the toy domain of letter strings, encounters a problem that can only be solved by discovering what makes the letter-strings "hcfrb", "yhumd", "exbvb", and "gxqrc" similar to each other but dissimilar to the strings "ndaxfw", "qiqa", "r", "rvm", and "zinw". Copycat has the built-in ability to count the letters in a string or group; in DGI's terms Copycat might be said to extract *number* as a modality-level feature. There is extensive evidence that

humans also have brainware support for subitizing (directly perceiving) small numbers, and brainware support for perceiving the approximate quantities of large numbers (see [Dehaene97] for a review). Suppose, however, that a general AI does *not* possess a modality-level counting ability. How would the AI go about forming the category of "five", or even "groups-of-five-letters"?

This challenge points up the inherent deficit of the "information loss" viewpoint of abstraction. For an AI with no subitization support - or for a human challenged with a number like "nine", which is out-of-range for human subitization - the distinguishing feature, cardinality, is not represented by the modality (or in humans, represented only approximately). For both humans and AIs, the ability to form concepts for non-subitizable exact numbers requires more than the ability to selectively focus on the facet of "number" rather than the facet of "location" or "letter" (or "color", "shape", or "pitch"). The fundamental challenge is not *focusing* on the numerical facet but rather *perceiving* a "numerical facet" in the first place. For the purposes of this discussion, we are not speaking of the ability to understand numbers, arithmetic, or mathematics, only an AI's ability to form the category "five". Possession of the category "five" does not even imply the possession of the categories "four" or "six", much less the formulation of the abstract supercategory "number".

Similarly, the "discovery" of fiveness is not being alleged as mathematically significant. In mathematical terms almost any set of cognitive building blocks will suffice to discover numbers; numbers are fundamental and can be constructed through a wide variety of different surface procedures. The significant accomplishment is not "squeezing" numbers out of a system so sparse that it apparently lacks the usual precursors of number. Rather, the challenge is to give an account of the discovery of "fiveness" in a way that generalizes to the discovery of other complex concepts as well. The hypothesized building blocks of the concept should be general (useful in building other, non-numerical concepts), and the hypothesized relations between building blocks should be general. It is acceptable for the discovery of "fiveness" to be straightforward, but the discovery method must be general.

A working but primitive procedure for satisfying the "five" concept, *after* the discovery of fiveness, might look something like this: Focus on a target group (the group which may or may not satisfy "five"). Retrieve from memory an exemplar for "five" (that is, some specific past experience that has become an exemplar for the "five" concept). Picture the "five" exemplar in a separate mental workspace. Draw a correspondence from an object within the group that is the five exemplar to an object within the group that is the target. Repeat this procedure until there are no objects remaining in the exemplar imagery or there are no objects remaining in the target imagery. Do not draw a correspondence from one object to another if a correspondence already exists. If, when this procedure completes, there are no dangling objects in the exemplar or in the target group, label the target group as satisfying the "five" concept.

In this example, the "five" property translates to the property: "I can construct a complete mapping, with no dangling elements, using unique correspondences, between this target group of objects, and a certain group of objects whose mental image I retrieved from memory."

This is mathematically straightforward, but cognitively general. In support of the proposition that "correspondence", "unique correspondence", and "complete mapping with no dangling elements" are all general conceptual primitives, rather than constructs useful solely for discovering numbers, please note that Copycat incorporates correspondences, unique correspondences, and a perceptual

drive toward complete mappings [Mitchell93]. Copycat has a direct procedural implementation of number sense and does not use these mapping constructs to build numerical concepts. The mapping constructs I have invoked for number are *independently* necessary for Copycat's theory of analogy-making as perception.

Once the procedure ends by labeling imagery with the "five" concept, that imagery becomes an experiential instance of the "five" concept. If the examples associated with a procedurally defined concept have any universal features or frequent features that are perceptually noticeable, the concept can acquire kernels after the fact, although the kernel may express itself as a hint or as an expectation, rather than being a necessary and sufficient condition for concept satisfaction. Concepts with procedural definitions are regular concepts and may possess kernels, exemplars, associated memories, and so on.

What is the benefit of decomposing "fiveness" into a complex procedure, rather than simply writing a codelet, or a modality-level feature detector, which directly counts (subitizes) the members of a group? The fundamental reason for preferring a non-modality solution in this example is to demonstrate that an AI must be capable of solving problems that were not anticipated during design. From this perspective "fiveness" is a bad example to use, since it would be very unlikely for an AI developer to not anticipate numericity during the design phase.

However, a decomposable concept for "five", and a modality-level feature detector which subitizes all numbers up to $(2^{32} - 1)$, can also be compared in terms of how well they support general intelligence. Despite its far greater computational overhead, I would argue that the decomposable concept is superior to a modality-level feature detector.

A billiards modality with a feature detector that subitizes all the billiard balls in a perceptual grouping and outputs a perceptually distinct label - a "numeron detector" - will suffice to solve many immediate problems that require a number sense. However, an AI that uses this feature detector to form a *surface* concept for "five" will not be able to subitize "five" groups of billiards within a supergroup, unless the programmer also had the foresight to extend the subitizing feature detector to count groups as well as specific objects[26]. Similarly, this universal subitizing ability will not extend across multiple modalities, unless the programmer had the foresight to extend the feature detector there as well[27]. Brainware is limited to what the programmer was thinking about at the time. Does an AI understand "fiveness" when it becomes able to count five apples? Or when the AI can also count five events in two different modalities? Or when the AI can count five of its own thoughts? It is programmatically trivial to extend the feature detector to handle any of these as a special case, but that is a path which ends in requiring an infinite amount of tinkering to implement routine thought processes (i.e., non-decomposability causes a "commonsense problem").

The *most* important reason for decomposability is that concepts with organized internal structures are more mutable. A human-programmed numeron detector, mutated on the code level, would probably simply break. A concept with internal structure or procedural structure, created by the AI's own thought processes in response to experience, is mutable by the AI's thought processes in response to further experience. For example, Douglas Lenat attests (see [Lenat83] and [Lenat84]) that the most difficult part of building EURISKO[28] was inventing a decomposable representation for heuristics, so that the class of transformations accessible to EURISKO would occasionally result in improvements rather than broken code fragments and LISP errors. To describe this as *smooth*

fitness landscapes is probably stretching the metaphor too much, but "smoothing" in some form is definitely involved. Raw code has only a single level of organization, and changing a random instruction on this level usually simply breaks the overall function. A EURISKO heuristic was broken up into chunks, and could be manipulated (by EURISKO's heuristics) on the chunk level.

Local shifts in the chunks of the "five"-ness procedure yield many useful offspring. By selectively relaxing the requirement of "no dangling objects" in the target image, we get the concept "less than or equal to five"-ness. By relaxing the requirement of "no dangling objects" in the exemplar image, we get the concept "greater than or equal to five"-ness. By requiring one or more dangling objects in the target image, we get the concept "more than five"-ness. By comparing two target images, instead of an exemplar and an image, we get the concept "one-to-one correspondence between group members" (what we would call "same-number-as" under a different procedure), and from there "less than" or "less than or equal to", and so on.

One of these concepts, the one-to-one correspondence between two mental images, is not just a useful offspring of the "fiveness" concept, but a *simpler* offspring. Thus it is probably not an "offspring" at all, but a *prerequisite* concept that suggests a real-world path to the apprehension of fiveness. Many physical tasks in our world require equal numbers (corresponding sets) for some group; four pegs for four holes, two shoes for two feet.

2.5.4: Experiential pathways to complex concepts.

Consider the real-world task of placing four pegs in four holes. A peg cannot fill two holes; two pegs will not fit in one hole. Solid objects cannot occupy the same location, cannot appear in multiple locations simultaneously, and do not appear or disappear spontaneously. These rules of the physical environment are reflected in the default behaviors of our own visuospatial modality; even early infants represent objects as continuous and will look longer at scenes which imply continuity violations [Spelke90].

From real-world problems such as pegs and holes, or their microworld analogues, an AI can develop concepts such as *unique correspondence:* a peg cannot fill multiple holes, multiple pegs will not fit in one hole. The AI can learn rules for drawing a *unique correspondence*, and test the rules against experience, before encountering the need to form the more complex concept for "fiveness". The presence of an immediate, local test of utility means that observed failures and successes can contribute unambiguously to forming a concept that is "simple" relative to the already-trained base of concepts. If a new concept contains many new untested parts, and a mistake occurs, then it may be unclear to the AI which local error caused the global failure. If the AI tries to chunk "fiveness" all in a single step, and the current procedure for "fiveness" satisfaction fails - is positively satisfied by a non-five-group, or unsatisfied by a five-group - it may be unclear to the AI that the global failure resulted from the local error of a nonunique correspondence.

The full path to fiveness would probably involve:

1. Learning *physical continuity;* acquiring expectations in which objects do not spontaneously disappear or reappear. In humans, this viewpoint is likely very strongly supported by modality-level visuospatial intuitions in which continuity is the default, and the same should hold true of AIs.

2. Learning *unique correspondence.* Unique correspondence, as a mental skill, tends to be reinforced by any goal-oriented challenge in which a useful object cannot be in two places at once.
3. Learning *complete mapping.* Completeness, along with symmetry, is one of the chief cognitive pressures implemented by Copycat in its model of analogy-making as a *perceptual* operation [Mitchell93]. A drive toward completeness implies that dangling, unmapped objects detract from the perceived "goodness" of a perceptual mapping. Thus, there may be modality-level support for noticing dangling, unmapped objects within an image.
4. With these three underlying concepts present, it is possible to abstract the concept of *complete mapping using the unique-correspondence relation,* also known as *one-to-one mapping.* We, using an entirely different procedure, would call this relation *same-number-as* ("identity of numeron produced by counting").
5. With *one-to-one mapping,* it is possible for an AI to notice that all the answers on a challenge task are related to a common prototype by the *one-to-one mapping* relation. The AI could then abstract the "five" concept using the prototype as the exemplar and the relation as a test.
6. Where do we go from here? Carl Feynman (personal communication) observes at this point that the *one-to-one mapping* relation is commutative and transitive, and therefore defines a set of equivalence classes; these equivalence classes turn out to be the natural numbers. At first, using "equivalence class detection" as a cognitive method sounded like cheating, but on reflection it's hard to see why a general intelligence should not notice when objects with a common relation to a prototype are similarly related to each other. "Equivalence class" may be a mathematical concept that happens to roughly (or even exactly) correspond to a perceptual property.
7. Forming the superclass concept of *number* is not dealt with in this paper, due to space constraints.

A deliberative intelligence must build up complex concepts from simple concepts, in the same way that evolution builds high-level feature detectors above low-level feature detectors, or builds organs using tissues, or builds thoughts over concepts or modalities. There are holonic[29] ecologies within the learned complexity of concepts, in the same way and for roughly the same reason that there is genetically specified holonic structure in modality-level feature detection. Categories describe regularities in perception, and in doing so, become part of the perceptual structure in which further regularities are detected.

If the programmer hardwires a subitizer that outputs numerons (unique number tags) as detected features, the AI may be able to chunk "five" very rapidly, but the resulting concept will suffer from *opacity* and *isolation.* The concept will not have the lower levels of organization that would enable the AI's native cognitive abilities to disassemble and reassemble the concept in useful new shapes; the inability of the AI to decompose the concept is *opacity.* The concept will not have a surrounding ecology of similar concepts and prerequisite concepts, such as would result from natural knowledge acquisition by the AI. Cognitive processes that require well-populated concept ecologies will be unable to operate; an AI that has "triangle" but not "pyramid" is less likely to successfully visualize "triangular lightbulb". This is *isolation.*

2.5.5: Microtasks

In the DGI model of AI development, concepts are abstracted from an experiential base; experiences are cognitive content within sensory modalities; and sensory modalities are targeted on a complex virtual microenvironment. Learning a concept requires (necessary, but not sufficient) having experiences from which to abstract the concept. How does an AI obtain these experiences? It would be possible to teach the AI about "fiveness" simply by presenting the AI with a series of sensory images (programmatically manipulating the AI's microenvironment) and prompting the AI's perceptual processes to generalize them, but this severs the task of concept formation from its ecological validity (metaphorically speaking). Knowledge goals (discussed in later sections) are not arbitrary; they derive from real-world goals or higher-level knowledge goals. Knowledge goals exist in a holonic goal ecology; the goal ecology shapes our knowledge goals and thereby often shapes the knowledge itself.

A first approximation to ecological validity is presenting the AI with a "challenge" in one of the virtual microenvironments previously advocated - for example, the billiards microenvironment. Henceforth, I will shorten "microenvironmental challenge" to "microtask". Microtasks can tutor concepts by presenting the AI with a challenge that must be solved using the concept the programmer wishes to tutor. For scrupulous ecological validity the key concept should be part of a larger problem, but even playing "one of these things is not like the others" would still be better than manipulating the AI's perceptual processes directly.

Tutoring a concept as the key to a microtask ensures that the concept's basic "shape", and associated experiences, are those required to solve problems, and that the AI has an experience of the concept being necessary, the experience of discovering the concept, and the experience of using the concept successfully. Effective intelligence is produced not by *having* concepts but by *using* concepts; one learns to use concepts by using them. The AI needs to possess the experiences of discovering and using the concept, just as the AI needs to possess the actual experiential referents that the concept generalizes; the AI needs experience of the contexts in which the concept is useful.

Forming a complex concept requires an incremental path to that complex concept - a series of building-block concepts and precursor concepts so that the final step is a leap of manageable size. Under the microtask developmental model, this would be implemented by a series of microtasks of ascending difficulty and complexity, in order to coax the AI into forming the precursor concepts leading up to the formation of complex concepts and abstract concepts. This is a major expense in programmer effort, but I would argue that it is a necessary expense for the creation of rich concepts with goal-oriented experiential bases.

The experiential path to "fiveness" would culminate with a microtask that could only be solved by abstracting and using the fiveness concept, and would lead up to that challenge through microtasks that could only be solved by abstracting and using concepts such as "object continuity", "unique correspondence", "mapping", "dangling group members", and the penultimate concept of "one-to-one mapping".

With respect to the specific microtask protocol for presenting a "challenge" to the AI, there are many possible strategies. Personally, I visualize a simple microtask protocol (on the level of "one of these things is not like the others") as consisting of a number of "gates", each of which must be "passed" by taking one of a set of possible actions, depending on what the AI believes to be the rule indicating the correct action. Passing ten successive gates *on the first try* is the indicator of

success. (For a binary choice, the chance of this happening accidentally is 1024:1. If the AI thinks fast enough that this may happen randomly (which seems rather unlikely), the number of successive gates required can be raised to twenty or higher.) This way, the AI can succeed or fail on individual gates, gathering data about individual examples of the common rule, but will not be able to win through the entire microtask until the common rule is successfully formulated. This requires a microenvironment programmed to provide an infinite (or merely "relatively large") number of variations on the underlying challenge - enough variations to prevent the AI from solving the problem through simple memory.

The sensory appearance of a microtask would vary depending on the modality. For a Newtonian billiards modality, an individual "gate" (subtask) might consist of four "option systems", each option system grouped into an "option" and a "button". Spatial separations in the Newtonian modality would be used to signal grouping; the distance between option systems would be large relative to the distance within option systems, and the distance between an option and a button would be large relative to the distance between subelements of an option. Each option would have a different configuration; the AI would choose one of the four options based on its current hypothesis about the governing rule. For example, the AI might select an option that consists of *four* billiards, or an option with *two large* billiards and *one small* billiard, or an option with *moving* billiards. Having chosen an option, the AI would manipulate a motor effector billiard - the AI's embodiment in that environment - into contact with the button *belonging to* (grouped with) the selected option. The AI would then receive a signal - perhaps a movement on the part of some billiard acting as a "flag" - which symbolized success or failure. The environment would then shift to the next "gate", causing a corresponding shift in the sensory input to the AI's billiards modality.

(Since the format of the microtask is complex and requires the AI to *start out* with an understanding of notions like "button" or "the button which belongs to the chosen option", there is an obvious chicken-and-egg problem with teaching the AI the format of the microtask before microtasks can be used to tutor other concepts. For the moment we will assume the bootstrapping of a small concept base, perhaps by "cheating" and using programmer-created cognitive content as *temporary* scaffolding.)

Given this challenge format, a simple microtask for "fiveness" seems straightforward: The option containing five billiards, regardless of their size or relative positions or movement patterns, is the key to the gate. In practice, setting up the fiveness microtask may prove more difficult because of the need to eliminate various false ways of arriving at a solution. In particular, if the AI has a sufficiently wide variety of quantitative feature detectors, then the AI will almost certainly possess an emergent Accumulator Model (see [Meck83]) of numeracy. If the AI takes a relatively fixed amount of time to mentally process each object, then single-feature clustering on the subjectively perceived time to mentally process a group could yield the microtask solution without a complex concept of fiveness. Rather than fiveness, the AI would have formed the concept "things-it-takes-about-20-milliseconds-to-understand". The real-world analogue of this situation has already occurred when an experiment formerly thought to show evidence for infant numeracy on small visual sets was demonstrated to show sensitivity to the contour length (perimeter) of the visual set, but not to the cardinality of the visual set [Clearfield99]. Even with all precursor concepts already present, a complex microtask might be necessary to make fiveness the *simplest* correct answer.

Also, the microtasks for the earlier concepts leading up to fiveness might inherently require greater complexity than the "option set" protocol described above. The concept of unique correspondence derives its behavior from physical properties. Choosing the right option set is a perceptual *decision* task rather than a physical *manipulation* task; in a decision microtask, the only manipulative subtask is maneuvering an effector billiard to touch a selected button. Concepts such as "dangling objects" or "one-to-one mapping" might require manipulation subtasks rather than decision subtasks, in order to incorporate feedback about physical (microenvironmental) outcomes into the concept.

For example, the microtask for teaching "one-to-one mapping" might incorporate the microworlds equivalent of a peg-and-hole problem. The microtask might be to divide up 9 "pegs" among 9 "holes" - where the 9 "holes" are divided into three subgroups of 4, 3, and 2, and the AI must allocate the peg supply among these subgroups in advance. For example, in the first stage of the microtask, the AI might be permitted to move pegs between three "rooms", but not permitted to place pegs in holes. In the second stage of the microtask the AI would attempt to place pegs in holes, and would then succeed or fail depending on whether the initial allocation between rooms was correct. Because of the complexity of this microtask, it might require other microtasks simply to explain the problem format - to teach the AI about pegs and holes and rooms. ("Pegs and holes" are universal and translate easily to a billiards modality; "holes", for example, might be immobile billiards, and "pegs" moveable billiards to be placed in contact with the "holes".)

Placing virtual pegs in virtual holes is admittedly not an inherently impressive result. In this case the AI is being taught to solve a simple problem so that the learned complexity will carry over into solving complex problems. If the learned complexity does carry over, and the AI later goes on to solve more difficult challenges, then, *in retrospect,* getting the AI to think coherently enough to navigate a microtask will "have been" an impressive result.

2.5.6: Interactions on the concept level

Concept-concept interactions are more readily accessible to introspection and to experimental techniques, and are relatively well-known in AI and in cognitive psychology. To summarize some of the complexity bound up in concept-concept interactions:

- Concepts are associated with other concepts. Activating a concept can "prime" a nearby concept, where "priming" is usually experimentally measured in terms of decreased reaction times [Meyer71]. This suggests that more computational resources should be devoted to scanning for primed concepts, or that primed concepts should be scanned first. (This viewpoint is too mechanomorphic to be considered as an explanation of priming in humans. Preactivation or advance binding of a neural network would be more realistic.)
- Nearby concepts may sometimes "slip" under cognitive pressures; for example, "triangle" to "pyramid". Such slippages play a major role in analogies under the Copycat system [Mitchell93]. Slippages occurring in complex design and planning problems probably incorporate context sensitivity and even goal orientation; see the later discussion of conflict and resonance in mental imagery.
- Concepts, in their role as categories, share territory. An individual sparrow, as an object, is described by the concepts "sparrow" and "bird". All objects that can be described as "sparrow" will also be described by "bird". Thus, information arriving through "bird" will usually, though not always, affect the entire territory of "sparrow". This form of inheritance

can take place without an explicit "is-a" rule connecting "sparrow" to "bird"; it is enough that "bird" happens to describe all referents of "sparrow".

- Concepts, in their role as categories, have supercategory and subcategory relationships. Declarative beliefs targeted on concepts can sometimes be inherited through such links. For example, "At least one X is an A" is inherited by the supercategory Y of X: If all referents of X are referents of Y, then "At least one referent of X is an A" implies that "At least one referent of Y is an A". Conversely, rules such as "All X are A" are inherited by subcategories of X but not supercategories of X. Inheritance that occurs on the concept level, through an "is-a" rule, should be distinguished from pseudo-inheritance that occurs through shared territory in specific mental imagery. Mental quantifiers such as "all X are Y" usually translate to "most X are Y" or "X, by default, are Y"; all beliefs are subject to controlled exception. It is possible to reason about category hierarchies deliberatively rather than perceptually, but our speed in doing so suggests a perceptual shortcut.
- Concepts possess transformation relations, which are again illustrated in Copycat. For example, in Copycat, "a" is the "predecessor" of "b", and "1" is the "predecessor" of "2". In a general intelligence these concept-concept relations would refer to, and would be generalized from, observation of transformational processes acting on experiential referents which causes the same continuous object to move from one category to another. Often categories related by transformational processes are subcategories of the same supercategory.
- Concepts act as verbs, adjectives, and adverbs as well as nouns. In humans, concepts act as one-place, two-place, and three-place predicates, as illustrated by the "subject", "direct object", and "indirect object" in the human parts of speech; "X gives Y to Z". For humans, four-place and higher predicates are probably represented through procedural rules rather than perceptually; spontaneously noticing a four-place predicate could be very computationally expensive. Discovering a predicate relation is assisted by categorizing the predicate's subjects, factoring out the complexity not germane to the predicate.
- Concepts, in their role as symbols with auditory, visual, or gestural tags, play a fundamental role in both human communication and internal human conceptualization. The short, snappy auditory tag "five" can stand in for the complexity bound up in the fiveness concept. Two humans that share a common lexical base can communicate a complex mental image by interpreting the image using concepts, describing the image with a concept structure, translating the concepts within the structure into socially shared auditory tags, transforming the concept structure into a linear sequence using shared syntax, and emitting the auditory tags in that linear sequence. (To translate the previous sentence into English: We communicate with sentences that use words and syntax from a shared language.) The same base of complexity is apparently also used to summarize and compactly manipulate thoughts internally; see the next section.

I also recommend George Lakoff's *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* [Lakoff87] for descriptions of many concept-level phenomena.

2.6: The thought level

Concepts are *combinatorial* learned complexity. Concepts represent regularities that recur, not in isolation, but in combination and interaction with other such regularities. Regularities are not isolated and independent, but are similar to other regularities, and there are simpler regularities and more complex regularities, forming a metaphorical "ecology" of regularities. This essential fact

about the structure of our low-entropy universe is what makes intelligence possible, computationally tractable, evolvable within a genotype, and learnable within a phenotype.

The thought level lies above the learned complexity of the concept level. Thoughts are structures of combinatorial concepts that alter imagery within the workspace of sensory modalities. Thoughts are the disposable one-time structures implementing a non-recurrent mind in a non-recurrent world. Modalities are wired; concepts are learned; thoughts are invented.

Where concepts are building blocks, thoughts are immediate. Sometimes the distance between a concept and a thought is very short; *bird* is a concept, but with little effort it can become a thought that retrieves a bird exemplar as specific mental imagery. Nonetheless, there is still a conceptual difference between a brick and a house that happens to be built from one brick. Concepts, considered as concepts, are building blocks with ready-to-use concept kernels. A thought fills in all the blanks and translates combinatorial concepts into specific mental imagery, even if the thought is built from a single concept. Concepts reside in long-term storage; thoughts affect specific imagery.

The spectra for "learned vs. invented", "combinatorial vs. specific", "stored vs. instantiated", and "recurrent vs. nonrecurrent" are conceptually separate, although deeply interrelated and usually correlated. Some cognitive content straddles the concept and thought levels. "Beliefs" (declarative knowledge) are learned, specific, stored, and recurrent. An episodic memory in storage is learned, specific, stored, and nonrecurrent. Even finer gradations are possible: A retrieved episodic memory is learned, specific, and immediate; the memory may recur as mental content, but its external referent is nonrecurrent. Similarly, a concept which refers to a specific external object is learned, specific, stored, and "semi-recurrent" in the sense that it may apply to more than one sensory image, since the object may be encountered more than once, but still referring to only one object and not a general category.

	Modalities:	Concepts:	Thoughts:
Source:	Wired	Learned	Invented
Degrees of freedom:	Representing	Combinatorial	Specific
Cognitive immediacy:	(Not applicable.)	Stored	Instantiated
Regularity:	Invariant	Recurrent	Nonrecurrent
Amount of complexity:	Bounded	Open-ended	Open-ended

2.6.1: Thoughts and language

The archetypal examples of "thoughts" (invented, specific, instantiated, nonrecurrent) are the sentences mentally "spoken" and mentally "heard" within the human stream of consciousness. We use the same kind of sentences, spoken aloud, to communicate thoughts between humans.

Words are the phonemic tags (speech), visual tags (writing), gestural tags (sign language), or haptic tags (Braile) used to invoke concepts. Henceforth, I will use speech to stand for all language

modalities; "auditory tag" or "phonemic tag" should be understood as standing for a tag in any modality.

When roughly the same concept shares roughly the same phonemic tag within a group of humans, words can be used to communicate concepts between humans, and sentences can be used to communicate complex imagery. The phonemes of a word can evoke all the functionality of the real concept associated with the auditory tag. A spoken sentence is a linear sequence of words; the human brain uses grammatical and syntactical rules to assemble the linear sequence into a structure of concepts, complete with internal and external targeting information. "Triangular lightbulb", an adjective followed by a noun, becomes "triangular" targeting "light bulb". "That is a telephone", anaphor-verb-article-noun, becomes a statement about the telephone-ness of a previously referred-to object. "That" is a backreference to a previously invoked mental target, so the accompanying cognitive description ("is a telephone") is imposed on the cognitive imagery representing the referent of "that".

The cognitive process that builds a concept structure from a word sequence combines syntactic constraints and semantic constraints; pure syntax is faster and races ahead of semantics, but semantic disharmonies can break up syntactically produced cognitive structures. Semantic guides to interpretation also reach to the word level, affecting the interpretation of homophones and ambiguous phonemes.

For the moment I will leave open the question of why we hear "mental sentences" *internally* - that is, the reason why the transformation of concept structures into linear word sequences, obviously necessary for spoken communication, also occurs internally within the stream of consciousness. I later attempt to explain this as arising from the coevolution of thoughts and language. For the moment, let it stand that the combinatorial structure of words and sentences in our internal narrative reflects the combinatorial structure of concepts and thoughts.

2.6.2: Mental imagery

The complexity of the thought level of organization arises from the cyclic interaction of thoughts and mental imagery. Thoughts modify mental imagery, and in turn, mental imagery gives rise to thoughts.

Mental imagery exists within the representational workspace of sensory modalities. Sensory imagery arises from environmental information (whether the environment is "real" or "virtual"); imaginative imagery arises from the manipulation of modality workspace through concept imposition and memory retrieval.

Mental imagery, whether sensory or imaginative, exhibits holonic organization: from the "pixel" level into objects and chunks; from objects and chunks into groups and superobjects; from groups and superobjects into mental scenes. In human vision, examples of specific principles governing grouping are proximity, similarity of color, similarity of size, common fate, and closure [Wertheimer23]; continuation [Moore98]; common region and connectedness [Palmer94]; and collinearity [Lavie96]. Some of the paradigms that have been proposed for resolving the positive inputs from grouping principles, and the negative inputs from detected conflicts, into a consistent global organization, include: Holonic conflict resolution (described earlier), computational

temperature [Mitchell93], Prägnanz [Koffka35], Hopfield networks [Hopfield85], the likelihood principle [Helmholtz67]; [Lowe85], minimum description length [Hochberg57], and constraint propagation [Kumar92].

Mental imagery provides a workspace for specific perceptions of concepts and concept structures. A chunk of sensory imagery may be mentally labeled with the concept structure "yellow box", and that description will remain bound to the object - a part of the perception of the object - even beyond the scope of the immediate thought. Learned categories and learned expectations also affect the gestalt organization of mental imagery [Zemel02].

Mental imagery is the active canvas on which deliberative thought is painted - "active canvas" implying a dynamic process and not just a static representation. The gestalt of mental imagery is the product of many local relations between elements. Because automatic cognitive processes maintain the gestalt, a local change in imagery can have consequences for connected elements in working imagery, without those changes needing to be specified within the proximate thought that caused the modification. The gestalt coherence of imagery also provides feedback on which possible changes will cohere well, and is therefore one of the verifying factors affecting which potential thoughts rise to the status of actuality (see below).

Imagery supports *abstract* percepts. It is possible for a human to reason about an object which is known to cost $1000, but for which no other mental information is available. Abstract reasoning about this object requires a means of representing mental objects that occupy no *a priori* modality; however, this does not mean that abstract reasoning operates independently of all modalities. Abstract reasoning might operate through a modality-level "object tracker" which can operate independently of the modalities it tracks; or by borrowing an existing modality using metaphor (see below); or the first option could be used routinely, and the second option when necessary. Given an abstract "object which costs $1000", it is then possible to attach concept structures that describe the object without having any specific sensory imagery to describe. If I impose the concept "red" on the existing abstract imagery for "an object which costs $1000", to yield "a red object which costs $1000", the "red" concept hangs there, ready to activate when it can, but not yielding specific visual imagery as yet.

Similarly, knowledge generalized from experience with concept-concept relations can be used to detect abstract conflicts. If I know that all penguins are green, I can deduce that "a red object which costs $1000" is not a penguin. It is possible to detect the conflict between "red" and "green" by a concept-level comparison of the two abstract descriptions, even in the absence of visualized mental imagery. However, this does *not* mean that it is possible for AI development to implement only "abstract reasoning" and leave out the sensory modalities. First, a real mind uses the rich concept-level complexity acquired from sensory experience, and from experience with reasoning that uses fully visualized imaginative imagery, to support abstract reasoning; we know that "red" conflicts with "green" because of prior sensory experience with *red* and *green*. Second, merely because some steps in reasoning appear as if they could theoretically be carried out purely on the concept level does not mean that a complete deliberative process can be carried out purely on the concept level. Third, abstract reasoning often employs metaphor to contribute modality behaviors to an abstract reasoning process.

The idea of "pure" abstract reasoning has historically given rise to AI pathologies and should be considered harmful. With that caution in mind, it is nonetheless possible that human minds visualize concepts only to the extent required by the current train of thought, thus conserving mental resources. An early-stage AI is likely to be less adept at this trick, meaning that early AIs may need to use full visualizations where a human could use abstract reasoning.

Abstract reasoning is a means by which inductively acquired generalizations can be used in deductive reasoning. If empirical induction from an experiential base in which all observed penguins are green leads to the formation of the belief "penguins are green", then this belief may apply abstractly to "a red object which costs $1000" to conclude that this object is probably not a penguin. In this example, an abstract belief is combined with abstract imagery about a specific object to lead to a further abstract conclusion about that specific object. Humans go beyond this, employing the very powerful technique of "deductive reasoning". We use abstract beliefs to reason about abstract mental imagery that describes classes and not just specific objects, and arrive at conclusions which then become new abstract beliefs; we can use deductive reasoning, as well as inductive reasoning, to acquire new beliefs. "Pure" deductive reasoning, like "pure" abstract reasoning, should be considered harmful; deductive reasoning is usually grounded in our ability to visualize specific test cases and by the intersection of inductive confirmation with the deductive conclusions.

Imagery supports tracking of reliances, a cognitive function which is conceptually separate from the perception of event causation. Another way of thinking about this is that perceived *cognitive* causation should not be confused with perceived causation in real-world referents. I may believe that the sun will rise soon; the cause of this belief may be that I heard a rooster crow; I may know that my confidence in sunrise's nearness relies on my confidence in the rooster's accuracy; but I do not believe that the rooster crowing causes the sun to rise.

Imagery supports complex percepts for "confidence" by tracking reliances on uncertainty sources. Given an assertion A with 50% confidence that "object X is blue", and a belief B with 50% confidence that "blue objects are large", the classical deduction would be the assertion "object X is large" with 25% confidence. However, this simple arithmetical method omits the possibility, important even under classical logic, that A and B are both mutually dependent on a third uncertainty C - in which case the combined confidence is greater than 25%. For example, in the case where "object X is blue" and "blue objects are large" are both straightforward deductions from a third assertion C with 50% confidence, and neither A nor B have any inherent uncertainty of their own, then "object X is large" is also a straightforward deduction from C, and has confidence 50% rather than 25%.

Confidence should not be thought of as a single quantitative probability; *confidence* is a percept that sums up a network of reliances on uncertainty sources. Straightforward links - that is, links whose local uncertainty is so low as to be unsalient - may be eliminated from the perceived reliances of forward deductions: "object X is large" is seen as a deduction from assertion C, not a deduction from C plus "object X is blue" plus "blue objects are large". If, however, the assertion "object X is blue" is contradicted by independent evidence supporting the inconsistent assertion "object X is red", then the reliance on "object X is blue" is an independent source of uncertainty, over and above the derived reliance on C. That is, the confidence of an assertion may be evaluated by weighing it against the support for the negation of the assertion [Tversky94]. Although the global structure of reliances is that of a network, the local percept of confidence is more likely derived from a set of

reliances on supporting and contradicting assertions whose uncertainty is salient. That the *local* percept of confidence is a set, and not a bag or a directed network, accounts for the elimination of common reliances in further derived propositions and the preservation of the global network structure. In humans, the percept of confidence happens to exhibit a roughly quantitative strength, and this quantity behaves in some ways like the mathematical formalism we call "probability".

Confidence and probability are not identical; for humans, this is both an advantage and a disadvantage. Seeing an assertion relying on four independent assertions of 80% confidence as psychologically different from an assertion relying on a single assertion of 40% confidence may contribute to useful intelligence. On the other hand, the human inability to use an arithmetically precise handling of probabilities may contribute to known cases of non-normative reasoning, such as not taking into account Bayesian priors, overestimating conjunctive probabilities and underestimating disjunctive probabilities, and the other classical errors described in [Tversky74]. See however [Cosmides96] for some cautions against underestimating the ecological validity of human reasoning; an AI might best begin with separate percepts for "humanlike" confidence and "arithmetical" confidence.

Imagery interacts with sensory information about its referent. Expectational imagery is confirmed or violated by the actual event. Abstract imagery created and left hanging binds to the sensory percept of its referent when and if a sensory percept becomes available. Imagery interacts with Bayesian information about its referent: assertions that make predictions about future sensory information are confirmed or disconfirmed when sensory information arrives to satisfy or contradict the prediction. Confirmation or disconfirmation of a belief may backpropagate to act as Bayesian confirmation or disconfirmation on its sources of support. (Normative reasoning in these cases is generally said to be governed by the Bayesian Probability Theorem.) The ability of imagery to bind to its referent is determined by the "matching" ability of the imagery - its ability to distinguish a sensory percept as belonging to itself - which in turn is a property of the way that abstract imagery interacts with incoming sensory imagery on the active canvas of working memory. A classical AI with a symbol for "hamburger" may be able to distinguish correctly spelled keystrokes typing out "hamburger", but lacks the matching ability to bind to hamburgers in any other way, such as visually or olfactorily. In humans, the abstract imagery for "a red object" may not involve a specific red image, but the "red" concept is still bound to the abstract imagery, and the abstract imagery can use the "red" kernel to match a referent in sensory imagery.

Imagery may bind to its referent in different ways. A mental image may be an immediate, environmental sensory experience; it may be a recalled memory; it may be a prediction of future events; it may refer to the world's present or past; it may be a subjunctive or counterfactual scenario. We can fork off a subjunctive scenario from a descriptive scene by thinking "What if?" and extrapolating, and we can fork off a separate subjunctive scenario from the first by thinking "What if?" again. Humans cannot continue the process indefinitely, because we run out of short-term memory to track all the reliances, but we have the native tracking ability. Note that mental imagery does not have an opaque tag selected from the finite set "subjunctive", "counterfactual", and so on. This would constitute code abuse: directly programming, as a special case, that which should result from general behaviors or emerge from a lower level of organization. An assertion within counterfactual imagery is not necessarily marked with the special tag "counterfactual"; rather, "counterfactual" may be the name we give to a set of internally consistent assertions with a common dependency on an assertion that is strongly disconfirmed. Similarly, a prediction is not necessarily

an assertion tagged with the opaque marker "prediction"; a prediction is better regarded as an assertion with deductive support whose referent is a future event or other referent for which no sensory information has yet arrived; the prediction imagery then binds to sensory information when it arrives, permitting the detection of confirmation or disconfirmation. The distinction between "prediction", "counterfactual", and "subjunctive scenario" can arise out of more general behaviors for confidence, reliance, and reference.

Mental imagery supports the perception of similarity and other comparative relations, organized into complex mappings, correspondences, and analogies (with Copycat being the best existing example of an AI implementation; see [Mitchell93]). Mental imagery supports expectations and the detection of violated expectations (where "prediction", above, refers to a product of deliberation, "expectations" are created by concept applications, modality behaviors, or gestalt interactions). Mental imagery supports temporal imagery and the active imagination of temporal processes. Mental imagery supports the description of causal relations between events and between assertions, forming complex causal networks which distinguish between implication and direct causation [Pearl00]. Mental imagery supports the binding relation of "metaphor" to allow extended reasoning by analogy, so that, e.g., the visuospatial percept of a forking path can be used to represent and reason about the behavior of if-then-else branches, with conclusions drawn from the metaphor (tentatively) applied to the referent [Lakoff99]. Imagery supports annotation of arbitrary objects with arbitrary percepts; if I wish to mentally label my watch as "X", then "X" it shall be, and if I also label my headphones and remote control as "X", then "X" will form a new (though arbitrary) category.

2.6.3: The origin of thoughts

Thoughts are the cognitive events that change mental imagery. In turn, thoughts are created by processes that relate to mental imagery, so that deliberation is implemented by the cyclic interaction of thoughts modifying mental imagery which gives rise to further thoughts. This does not mean that the deliberation level is "naturally emergent" from thought. The thought level has specific features allowing thought in paragraphs and not just sentences - "trains of thought" with internal momentum, although not so much momentum that interruption is impossible.

At any one moment, out of the vast space of possible thoughts, a single thought ends up being "spoken" within deliberation. Actually, "one thought at a time" is just the human way of doing things, and a sufficiently advanced AI might multiplex or multithread deliberation, but this doesn't change the basic question: Where do thoughts come from? I suggest that it is best to split our conceptual view of this process into two parts; first, the production of suggested thoughts, and second, the selection of thoughts that appear "useful" or "possibly useful" or "important" or otherwise interesting. In some cases, the process that invents or suggests thoughts may do most of the work, with winnowing relatively unimportant; when you accidentally rest your hand on a hot stove, the resulting bottom-up event immediately hijacks deliberation. In other cases, the selection process may comprise most of the useful intelligence, with a large number of possible thoughts being tested in parallel. In addition to being conceptually useful, distinguishing between *suggestion* and *verification* is useful on a design level if "verifiers" and "suggesters" can take advantage of modular organization. Multiple suggesters can be judged by one verifier and multiple verifiers can summate the goodness of a suggestion. This does not necessarily imply hard-bounded processing

stages in which "suggestion" runs, terminates and is strictly followed by "verification", but it implies a common ground in which repertoires of suggestion processes and verification processes interact.

I use the term *sequitur* to refer to a cognitive process which suggests thoughts. "Sequitur" refers, not to the way that two thoughts follow each other - that is the realm of deliberation - but rather to the source from which a *single* thought arises, following from mental imagery. Even before a suggested thought rises to the surface, the suggestion may interact with mental imagery to determine whether the thought is interesting and possibly to influence the thought's final form. I refer to specific interactions as *resonances;* a suggested thought resonates with mental imagery during verification. Both positive resonances and negative resonances (conflicts) can make a thought more interesting, but a thought with no resonances at all is unlikely to be interesting.

An example of a sequitur might be noticing that a piece of mental imagery satisfies a concept; for a human, this would translate to the thought "X is a Y!" In this example, the concept is cued and satisfied by a continuous background process, rather than being suggested by top-down deliberation; thus, noticing that X is a Y comes as a surprise which may shift the current train of thought. How *much* of a surprise - how salient the discovery becomes - will depend on an array of surrounding factors, most of which are probably the same resonances that promoted the candidate suggestion "concept Y matches X" to the real thought "X is a Y!". (The difference between the suggestion and the thought is that the real thought persistently changes current mental imagery by binding the Y concept to X, and shifts the focus of attention.)

What are the factors that determine the resonance of the suggestion "concept Y matches X" or "concept Y may match X" and the salience of the thought "X is a Y"? Some of these factors will be inherent properties of the concept Y, such as Y's past value, the rarity of Y, the complexity of Y, et cetera; in AI, these are already-known methods for ranking the relative value of heuristics and the relative salience of categories. Other factors are inherent in X, such as the degree to which X is the focus of attention.

Trickier factors emerge from the interaction of X (the targeted imagery), Y (the stored concept that potentially matches X), the suggested mental imagery for Y describing X, the surrounding imagery, and the task context. A human programmer examining this design problem naturally sees an unlimited range of potential correlations. To avoid panic, it should be remembered that evolution did not begin by contemplating the entire search space and attempting to constrain it; evolution would have incrementally developed a repertoire of correlations in which adequate thoughts resonated some of the time. Just as concept kernels are not AI-complete, sequiturs and resonances are not AI-complete. Sequiturs and resonances also may not need to be human-equivalent to minimally support deliberation; it is acceptable for an early AI to miss out on many humanly obvious thoughts, so long as those thoughts which are successfully generated sum to fully general deliberation.

Specific sequiturs and resonances often seem reminiscent of general heuristics in Lenat's EURISKO [Lenat83] or other AI programs intended to search for interesting concepts and conjectures [Colton00]. The resemblance is further heightened by the idea of adding learned associations to the mix; for example, correlating which concepts Y are frequently useful when dealing with imagery described by concepts X, or correlating concepts found useful against categorizations of the current task domain, bears some resemblance to EURISKO trying to learn specific heuristics about when

specific concepts are useful. Similarly, the general sequitur that searches among associated concepts to match them against working imagery bears some resemblance to EURISKO applying a heuristic. Despite the structural resemblance, sequiturs are not heuristics. Sequiturs are general cognitive subprocesses lying on the brainware level of organization. The subprocess is the sequitur that handles thoughts of the general form "X is a Y"; any cognitive content relating to specific Xs and Ys is learned complexity, whether it takes the form of heuristic beliefs or correlative associations. Since our internal narrative is open to introspection, it is not surprising if sequiturs produce some thoughts resembling the application of heuristics; the mental sentences produced by sequiturs are open to introspection, and AI researchers were looking at these mental sentences when heuristics were invented.

Some thoughts that might follow from "X is a Y!" (unexpected concept satisfaction) are: "Why is X a Y?" (searching for explanation); or "Z means X can't be a Y!" (detection of belief violation); or "X is not a Y" (rechecking of a tentative conclusion). Any sequence of two or more thoughts is technically the realm of deliberation, but connected deliberation is supported by properties of the thought level such as focus of attention. The reason that "Why is X a Y?" is likely to follow from "X is a Y!" is that the thought "X is a Y" shifts the focus of attention to the Y-ness of X (the mental imagery for the Y concept binding to X), so that sequitur processes tend to focus selectively on this piece of mental imagery and try to discover thoughts that involve it.

The interplay of thoughts and imagery has further properties that support deliberation. "Why is X a Y?" is a thought that creates, or focuses attention on, a question - a thought magnet that attracts possible answers. *Question imagery* is both like and unlike *goal imagery*. (More about goals later; currently what matters is how the thought level interacts with goals, and the intuitive definition of goals should suffice for that.) A goal in the classic sense might be defined as abstract imagery that "wants to be true", which affects cognition by affecting the AI's decisions and actions; the AI makes decisions and takes actions based on whether the AI predicts those decisions and actions will lead to the goal referent. *Questions* primarily affect which thoughts arise, rather than which decisions are made. Questions are thought-level complexity, a property of mental imagery, and should not be confused with *reflective goals* asserting that a piece of knowledge is desirable; the two interrelate very strongly but are conceptually distinct. A question is a thought magnet and a goal is an action magnet. Since stray thoughts are (hopefully!) less dangerous than stray actions, question-ness (*inquiry*) can spread in much more unstructured ways than goal-ness (*desirability*).

Goal imagery is abstract imagery whose referent is brought into correspondence with the goal description by the AI's actions. Question imagery is also abstract imagery, since the answer is not yet known, but question imagery has a more open-ended satisfaction criterion. Goal imagery tends to want its referent to take on a *specific* value; question imagery tends to want its referent to take on *any* value. Question imagery for "the outcome of event E" attracts any thoughts about the outcome of event E; it is the agnostic question "What, if anything, is the predicted outcome of E?" Goal imagery for "the outcome of event E" tends to require some *specific* outcome for E.

The creation of question imagery is one of the major contributing factors to the continuity of thought sequences, and therefore necessary for deliberation. However, just as goal imagery must affect actual decisions and actual actions before we concede that the AI has something which deserves to be called a "goal", question imagery must affect actual thoughts - actual sequiturs and actual verifiers - to be considered a cognitively real question. If there is salient question imagery for

"the outcome of event E", it becomes the target of sequiturs that search for beliefs about implication or causation whose antecedents are satisfied by aspects of E; in other words, sequiturs searching for beliefs of the form "E usually leads to F" or "E causes F". If there is open question imagery for "the cause of the Y-ness of X", and a thought suggested for some other reason happens to intersect with "the cause of the Y-ness of X", the thought resonates strongly and will rise to the surface of cognition.

A similar and especially famous sequitur is the search for a causal belief whose consequent matches goal imagery, and whose antecedent is then visualized as imagery describing an event which is predicted to lead to the goal. The event imagery created may become new goal imagery - a subgoal - if the predictive link is confirmed and no obnoxious side effects are separately predicted (see the discussion of the deliberation level for more about goals and subgoals). Many classical theories of AI, in particular "theorem proving" and "planning" [Newell63], hold up a simplified form of the "subgoal seeker" sequitur as the core algorithm of human thought. However, this sequitur does not in itself implement planning. The process of seeking subgoals is more than the one cognitive process of searching for belief consequents that match existing goals. There are other roads to finding subgoal candidates aside from backward chaining on existing goals; for example, forward reasoning from available actions. There may be several different real sequiturs (cognitive processes) that search for relevant beliefs; evolution's design approach would have been "find cognitive processes that make useful suggestions", not "constrain an exhaustive search through all beliefs to make it computationally efficient", and this means there may be several sequiturs in the repertoire that selectively search on different kinds of causal beliefs. Finding a belief whose consequent matches goal imagery is not the same as finding an event which is predicted to lead to the goal event; and even finding an action predicted to lead to at least one goal event is not the same as verifying the net desirability of that action.

The sequitur that seeks beliefs whose consequents match goal imagery is only one component of the thought level of organization. But it is a component that looks like the "exclamation mark of thought" from the perspective of many traditional theories, so it is worthwhile to review how the other levels of organization contribute to the effective intelligence of the "subgoal seeker" sequitur.

A goal is descriptive mental imagery, probably taking the form of a concept or concept structure describing an event; goal-oriented thinking uses the combinatorial regularities of the concept layer to describe regularities in the structure of goal-relevant events. The search for a belief whose consequent matches a goal description is organized using the category structure of the concept layer; concepts match against concepts, rather than unparsed sensory imagery matching against unparsed sensory imagery. Searching through beliefs is computationally tractable because of learned resonances and learned associations which are "learned complexity" in themselves, and moreover represent regularities in a conceptually described model rather than a raw sensory imagery. Goal-oriented thinking as used by humans is often abstract, which requires support from properties of mental imagery; it requires that the mind maintain descriptive imagery which is not fully visualized or completely satisfied by a sensory referent, but which binds to specific referents when these become available. Sensory modalities provide a space in which all this imagery can exist and interprets the environment from which learned complexity is learned. The feature structure of modalities renders learning computationally tractable. Without feature structure, concepts are computationally intractable; without category structure, thoughts are computationally intractable. Without modalities there are no experiences and no mental imagery; without learned

complexity there are no concepts to structure experience and no beliefs generalized from experience. In addition to supporting basic requirements, modalities contribute directly to intelligence in any case where referent behaviors coincide with modality behaviors, and indirectly in cases where there are valid metaphors between modality behaviors and referent behaviors.

Even if inventing a new subgoal is the "exclamation mark of thought" from the perspective of many traditional theories, it is an exclamation mark at the end of a very long sentence. The rise of a single thought is an event that occurs within a whole mind - an intact reasoning process with a past history.

2.6.4: Beliefs

Beliefs - declarative knowledge - straddle the division between the concept level and the thought level. In terms of the level characteristics noted earlier, beliefs are learned, specific, stored, and recurrent. From this perspective beliefs should be classified as learned complexity and therefore a part of the generalized concept level. However, beliefs bear a greater surface resemblance to mental sentences than to individual words. Their internal structure appears to resemble concept structures more than concepts; and beliefs possess characteristics, such as structured antecedents and consequents, which are difficult to describe except in the context of the thought level of organization. I have thus chosen to discuss beliefs within the thought level[30].

Beliefs are acquired through cognitive processes that fall into two major classes, *inductive* and *deductive,* respectively referring to generalization over experience, and reasoning from previous beliefs. The strongest beliefs have both inductive and deductive support: deductive conclusions with experiential confirmation, or inductive generalizations with causal explanations.

Induction and deduction can intersect because both involve abstraction. Inductive generalization produces a description containing categories that act as variables - abstract imagery that varies over the experiential base and describes it. Abstract deduction takes several inductively or deductively acquired generalizations, and chains together their abstract antecedents and abstract consequents to produce an abstract conclusion, as illustrated in the earlier discussion of abstract mental imagery. Even completely specific beliefs confirmed by a single experience, such as "New Year's Eve of Y2K took place on a Friday night", are still "abstract" in that they have a concept-based, category-structure description existing above the immediate sensory memory, and this conceptual description can be more easily chained with abstract beliefs that reference the same concepts.

Beliefs can be suggested by generalization across an experiential base, and supported by generalization across an experiential base, but there are limits to how much support pure induction can generate (a common complaint of philosophers); there could always be a disconfirming instance you don't know about. Inductive generalization probably resembles concept generalization, more or less; there is the process of initially noticing a regularity across an experiential base, the process of verifying it, and possibly even a process producing something akin to concept kernels for cueing frequently relevant beliefs. Beliefs have a different structure than concepts; concepts are either *useful* or *not useful,* but beliefs are either *true* or *false.* Concepts apply to referents, while beliefs describe relations between antecedents and consequents. While this implies a different repertoire of generalizations that produce inductive beliefs, and a different verification procedure, the computational task of noticing a generalization across antecedents and consequents seems strongly reminiscent of generalizing a two-place predicate.

Beliefs are well-known in traditional AI, and are often dangerously misused; while any process whatever can be *described with* beliefs, this does not mean that a cognitive process is *implemented by* beliefs. I possess a visual modality that implements edge detection, and I possess beliefs about my visual modality, but the latter aspect of mind does not affect the former. I could possess no beliefs about edge detection, or wildly wrong beliefs about edge detection, and my visual modality would continue working without a hiccup. An AI may be able to introspect on lower levels of organization (see Part III), and an AI's cognitive subsystems may interact with an AI's beliefs more than the equivalent subsystems in humans (again, see Part III), but beliefs and brainware remain distinct - not only distinct, but occupying different levels of organization. When we seek the functional consequences of beliefs - their material effects on the AI's intelligence - we should look for the effect on the AI's reasoning and its subsequent decisions and actions. Anything can be described by a belief, including every event that happens within a mind, but not all events within a mind are implemented by the possession of a belief which describes the rules governing that event.

In formal, classical terms, the cognitive effect of possessing a belief is sometimes defined to mean that when the antecedent of a belief is satisfied, its consequent is concluded. I would regard this as one sequitur out of many, but it is nonetheless a good example of a sequitur - searching for beliefs whose antecedents are satisfied by current imagery, and concluding the consequent (with reliances on the belief itself and on the imagery matched by the antecedent). However, this sequitur, if applied in the blind sense evoked by classical logic, will produce a multitude of useless conclusions; the sequitur needs to be considered in the context of verifiers such as "How rare is it for this belief to be found applicable?", "How often is this belief useful when it is applicable?", or "Does the consequent produced intersect with any other imagery, such as open question imagery?"

Some other sequiturs involving beliefs: Associating backward from question imagery to find a belief whose consequent touches the question imagery, and then seeing if the belief's antecedent can be satisfied by current imagery, or possibly turning the belief's antecedent into question imagery. Finding a causal belief whose consequent corresponds to a goal; the antecedent may then become a subgoal. Detecting a case where a belief is *violated* - this will usually be highly salient.

Suppose an AI with a billiards modality has inductively formed the belief "all billiards which are 'red' are 'gigantic'". Suppose further that 'red' and 'gigantic' are concepts formed by single-feature clustering, so that a clustered size range indicates 'gigantic', and a clustered volume of color space indicates 'red'. If this belief is salient enough, relative to the current task, to be routinely checked against all mental imagery, then several cognitive properties should hold if AI really possesses a belief about the size of red billiards. In subjunctive imagery, used to imagine non-sensory billiards, any billiard imagined to be red (within the clustered color volume of the 'red' concept) would need to be imagined as being gigantic (within the clustered size range of the 'gigantic' concept). If the belief "all red billiards are gigantic" has salient uncertainty, then the conclusion of gigantism would have a reliance on this uncertainty source and would share the perceived doubt. Given external sensory imagery, if a billiard is seen which is red and small, this must be perceived as violating the belief. Given sensory imagery, if a billiard is somehow seen as "red" in advance of its size being perceived (it's hard to imagine how this would happen in a human), then the belief must create the prediction or expectation that the billiard will be gigantic, binding a hanging abstract concept for 'gigantic' to the sensory imagery for the red billiard. If the sensory image is completed later and the concept kernel for 'gigantic' is not satisfied by the completed sensory image for the red billiard, then

the result should be a violated expectation, and this conflict should propagate back to the source of the expectation to be perceived as a violated belief.

Generally, beliefs used within subjunctive imagery control the imagery directly, while beliefs used to interpret sensory information govern expectations and determine when an expectation has been violated. However, "sensory" and "subjunctive" are relative; subjunctive imagery governed by one belief may intersect and violate another belief - any imagery is "sensory" relative to a belief if that imagery is not directly controlled by the belief. Thus, abstract reasoning can detect inconsistencies in beliefs. (An inconsistency should not cause a real mind to shriek in horror and collapse, but it should be a salient event that shifts the train of thought to hunting down the source of the inconsistency, looking at the beliefs and assertions relied upon and checking their confidences. Inconsistency detections, expressed as thoughts, tend to create question imagery and knowledge goals which direct deliberation toward resolving the inconsistency.)

2.6.5: Coevolution of thoughts and language: Origins of the internal narrative

Why is the transformation of concept structures into linear word sequences, obviously necessary for spoken communication, also carried out within the internal stream of consciousness? Why not use only the concept structures? Why do we transform concept structures into grammatical sentences if nobody is listening? Is this a necessary part of intelligence? Must an AI do the same in order to function?

The dispute over which came first, thought or language, is ancient in philosophy. Modern students of the evolution of language try to break down the evolution of language into incrementally adaptive stages, describe multiple functions that are together required for language, and account for how preadaptations for those functions could have arisen [Hurford99]. Functional decompositions avoid some of the chicken-and-egg paradoxes that result from viewing language as a monolithic function. Unfortunately, there are further paradoxes that result from viewing language independently from thought, or from viewing thought as a monolithic function.

From the perspective of a cognitive theorist, language is only one function of a modern-day human's cognitive supersystem, but from the perspective of an evolutionary theorist, linguistic features determine which *social* selection pressures apply to the evolution of cognition at any given point. Hence "coevolution of thought and language" rather than "evolution of language as one part of thought". An evolutionary account of language alone will become "stuck" the first time it reaches a feature which is adaptive for cognition and preadaptive for language, but for which no independent linguistic selection pressure exists in the absence of an already-existent language. Since there is currently no consensus on the functional decomposition of intelligence, contemporary language evolution theorists are sometimes unable to avoid such sticking points.

On a first look DGI might appear to explain the evolvability of language merely by virtue of distinguishing between the concept level and the thought level; as long as there are simple reflexes that make use of learned category structure, elaboration of the concept level will be independently adaptive, even in the absence of a humanlike thought level. The elaboration of the concept level to support cross-modality associations would appear to enable crossing the gap between a signal and a concept, and the elaboration of the concept level to support the blending or combination of concepts (adaptive because it enables the organism to perceive simple combinatorial regularities)

would appear to enable primitive, nonsyntactical word sequences. Overall this resembles Bickerton's [Bickerton90] picture of *protolanguage* as an evolutionary intermediate, in which learned signals convey learned concepts and multiple concepts blend, but without syntax to convey targeting information. Once protolanguage existed, linguistic selection pressures proper could take over.

However, as [Deacon97] points out, this picture does not explain why other species have not developed protolanguage. Cross-modal association is not limited to humans or even primates. Deacon suggests that some necessary mental steps in language are not only *unintuitive* but actually *counterintuitive* for nonhuman species, in the same way that the Wason Selection Test is counterintuitive for humans. Deacon's account of this "awkward step" uses a different theory of intelligence as background, and I would hence take a different view of the nature of the awkward step: my guess is that chimpanzees find it extraordinarily hard to learn symbols as we understand them because language, even protolanguage, requires creating abstract mental imagery which can hang unsupported and then bind to a sensory referent later encountered. The key difficulty in language - the step that is awkward for other species - is not the ability to associate signals; primates (and rats, for that matter) can readily associate a perceptual signal with a required action or a state of the world. The awkward step is for a signal to evoke a category as abstract imagery, apart from immediate sensory referents, which can bind to a referent later encountered. This step is completely routine for us, but could easily be almost impossible in the absence of design support for "hanging concepts in midair". In the absence of thought, there are few reasons why a species would find it useful to hang concepts in midair. In the absence of language, there are even fewer reasons to associate a perceptual signal with the evocation of a concept as abstract imagery. Language is hard for other species, not because of a gap between the signal and the concept, but because language uses a feature of mental imagery for which there is insufficient design support in other species. I suspect it may have been an adaptive context for abstract imagery, rather than linguistic selection pressures, which resulted in the adaptation which turned out to be preadaptive for symbolization and hence started some primate species sliding down a fitness gradient that included coevolution of thought and language.

If, as this picture suggests, pre-hominid evolution primarily elaborated the concept layer (in the sense of elaborating brainware processes that support categories, not in the sense of adding learned concepts as such), it implies that the concept layer may contain the bulk of supporting functional complexity for human cognition. This does not follow necessarily, since evolution may have spent much time but gotten little in return, but it is at least suggestive. (This paper's section on the concept level is, in fact, the longest section.) The above picture also suggests that the hominid family may have coevolved combinatorial concept structures that modify mental imagery internally (thoughts) and combinatorial concept structures that evoke mental imagery in conspecifics (language). It is obvious that language makes use of many functions originally developed to support internal cognition, but *coevolution* of thought and language implies a corresponding opportunity for evolutionary elaboration of hominid thought to coopt functions originally evolved to support hominid language.

The apparent necessity of the internal narrative for human deliberation could turn out to be an introspective illusion, but if real, it strongly suggests that linguistic functionality has been coopted for cognitive functionality during human evolution. Linguistic features such as special processing of the tags that invoke concepts, or the use of syntax to organize complex internal targeting information for structures of combinatorial concepts, could also be adaptive or preadaptive for

efficient thought. Only a few such linguistic features would need to be coopted as necessary parts of thought before the "stream of consciousness" became an entrenched part of human intelligence. This is probably a sufficient explanation for the existence of an internal narrative, possibly making the internal narrative a pure spandrel (emergent but nonadaptive feature). However, caution in AI, rather than caution in evolutionary psychology, should impel us to wonder if our internal narrative serves an adaptive function. For example, our internal narrative could express deliberation in a form that we can more readily process as (internal) sensory experience for purposes of introspection and memory; or the cognitive process of imposing internal thoughts on mental imagery could coopt a linguistic mechanism that also translates external communications into mental imagery; or the internal narrative may coopt social intelligence that models other humans by relating to their communications, in order to model the self. But even if hominid evolution has coopted the internal narrative, the overall model still suggests that - while we cannot disentangle language from intelligence or disentangle the evolution of thought from the evolution of language - a *de novo* mind design could disentangle intelligence from language.

This in turn suggests that an AI could use concept structures without serializing them as grammatical sentences forming a natural-language internal narrative, as long as all linguistic functionality coopted for human intelligence were reproduced in non-linguistic terms - including the expression of thoughts in an introspectively accessible form, and the use of complex internal targeting in concept structures. Observing the AI may require recording the AI's thoughts and translating those thoughts into humanly understandable forms, and the programmers may need to communicate concept structures to the AI, but this need not imply an AI capable of understanding or producing human language. True linguistic communication between humans and AIs might come much later in development, perhaps as an ordinary domain competency rather than a brainware-supported talent. Of course, human-language understanding and natural human conversation is an *extremely* attractive goal, and would undoubtedly be attempted as early as possible; however, it appears that language need not be implemented immediately or as a necessary prerequisite of deliberation.

2.7: The deliberation level

2.7.1: From thoughts to deliberation

In humans, higher levels of organization are generally more accessible to introspection. It is not surprising if the internal cognitive events called "thoughts", as described in the last section, seem strangely familiar; we listen to thoughts all day. The danger for AI developers is that cognitive content which is open to introspection is sometimes temptingly easy to translate directly into code. But if humans have evolved a cyclic interaction of thought and imagery, this fact alone does not prove (or even argue) that the design is a good one. What is the material benefit to intelligence of using blackboard mental imagery and sequiturs, instead of the simpler fixed algorithms of "reasoning" under classical AI?

Evolution is characterized by ascending levels of organization of increasing elaboration, complexity, flexibility, richness, and computational costliness; the complexity of the higher layers is not automatically emergent solely from the bottom layer, but is instead subject to selection pressures and the evolution of complex functional adaptation - adaptation which is relevant at that level, and, as it turns out, sometimes preadaptive for the emergence of higher levels of organization. This design

signature emerges at least in part from the characteristic blindness of evolution, and may not be a necessary idiom of minds-in-general. Nonetheless, past attempts to directly program cognitive phenomena which arise on post-modality levels of organization have failed profoundly. There are specific AI pathologies that emerge from the attempt, such as the symbol grounding problem and the commonsense problem. In humans concepts are smoothly flexible and expressive because they arise from modalities; thoughts are smoothly flexible and expressive because they arise from concepts. Even considering the value of blackboard imagery and sequiturs in isolation - for example, by considering an AI architecture that used fixed algorithms of deliberation but used those algorithms to create and invoke DGI thoughts - there are still necessary reasons why deliberative patterns must be built on behaviors of the thought level, rather than being implemented as independent code; there are AI pathologies that would result from the attempt to implement deliberation in a purely top-down way. There is top-down complexity in deliberation - adaptive functionality that is best viewed as applying to the deliberation level and not the thought level - but this complexity is mostly incarnated as behaviors of the thought level that support deliberative patterns.

Because the deliberation level is flexibly emergent out of the sequiturs of the thought level, a train of thought can be diverted without being destroyed. To use the example given earlier, if a deliberative mind wonders "Why is X a Y?" but no explanation is found, this local failure is not a disaster for deliberation as a whole. The mind can mentally note the question as an unsolved puzzle and continue with other sequiturs. A belief violation does not destroy a mind; it becomes a focus of attention and one more thing to ponder. Discovering inconsistent beliefs does not cause a meltdown, as it would in a system of monotonic logic, but instead shifts the focus of attention to checking and revising the deductive logic. Deliberation weaves multiple, intersecting threads of reasoning through intersecting imagery, with the waystations and even the final destination not always known in advance.

In the universe of bad TV shows, speaking the Epimenides Paradox[31] "This sentence is false" to an artificial mind causes that mind to scream in horror and collapse into a heap of smoldering parts. This is based on a stereotype of thought processes that cannot divert, cannot halt, and possess no bottom-up ability to notice regularities across an extended thought sequence. Given how deliberation emerges from the thought level, it is possible to imagine a sufficiently sophisticated, sufficiently reflective AI that could naturally surmount the Epimenides Paradox. Encountering the paradox "This sentence is false" would probably indeed lead to a looping thought sequence at first, but this would not cause the AI to become permanently stuck; it would instead lead to categorization across repeated thoughts (like a human noticing the paradox after a few cycles), which categorization would then become salient and could be pondered in its own right by other sequiturs. If the AI is sufficiently competent at deductive reasoning and introspective generalization, it could generalize across the specific instances of "If the statement is true, it must be false" and "If the statement is false, it must be true" as two general classes of thoughts produced by the paradox, and show that reasoning from a thought of one class leads to a thought of the other class; if so the AI could deduce - not just inductively notice, but deductively confirm - that the thought process is an eternal loop. Of course, we won't know whether it really works this way until we try it.

The use of a blackboard sequitur model is not automatically sufficient for deep reflectivity; an AI that possessed a limited repertoire of sequiturs, no reflectivity, no ability to employ reflective categorization, and no ability to notice when a train of thought hasn't yielded anything useful for a

while, might still loop eternally through the paradox as the emergent but useless product of the sequitur repertoire. Transcending the Epimenides Paradox requires the ability to perform inductive generalization and deductive reasoning on introspective experiences. But it also requires bottom-up organization in deliberation, so that a spontaneous introspective generalization can capture the focus of attention. Deliberation must *emerge* from thoughts, not just *use* thoughts to implement rigid algorithms.

Having reached the deliberation level, we finally turn from our long description of what a mind *is,* and focus at last on what a mind *does* - the useful operations implemented by sequences of thoughts that are structures of concepts that are abstracted from sensory experience in sensory modalities.

2.7.2: The dimensions of intelligence

Philosophers frequently define "truth" as an agreement between belief and reality; formally, this is known as the "correspondence theory" of truth [James11]. Under the correspondence theory of truth, philosophers of Artificial Intelligence have often defined "knowledge" as a mapping between internal data structures and external physical reality [Newell80]. Considered in isolation, the correspondence theory of knowledge is easily abused; it can be used to argue on the basis of mappings which turn out to exist entirely in the mind of the programmer.

Intelligence is an evolutionary advantage because it enables us to model *and* predict *and* manipulate reality. In saying this, I am not advocating the philosophical position that only useful knowledge can be true. There is enough regularity in the activity of acquiring knowledge, over a broad spectrum of problems that require knowledge, that evolution has tended to create independent cognitive forces for truthseeking. Individual organisms are best thought of as adaptation-executers rather than fitness-maximizers [Tooby92]. "Seeking truth", even when viewed as a mere local subtask of a larger problem, has sufficient functional autonomy that many human adaptations are better thought of as "truthseeking" than "useful-belief-seeking". Furthermore, under my own philosophy, I would say that beliefs are useful because they are true, not "true" because they are useful.

But usefulness is a *stronger* and *more reliable* test of truth; it is harder to cheat. The social process of science applies prediction as a test of models, and the same models that yield successful predictions are often good enough approximations to construct technology (manipulation).

I would distinguish four successively stronger grades of *binding* between a model and reality:

- A *sensory* binding occurs when there is a mapping between cognitive content in the model and characteristics of external reality. Without tests of usefulness, there is no formal way to prevent abuse of claimed sensory bindings; the supposed mapping may lie mostly in the mind of the observer. However, if the system as a whole undergoes tests of usefulness, much of the task of extending and improving the model will still locally consist of discovering good sensory bindings - finding beliefs that are true under the intuitive "correspondence theory" of truth.
- A *predictive* binding occurs when a model can be used to correctly predict future events. From the AI's internal perspective, a predictive binding occurs when the model can be used to correctly predict future sensory inputs. The AI may be called upon to make successful predictions about external reality (outside the computer), virtual

microenvironments (inside the computer but outside the AI), or the outcome of cognitive processes (inside the AI, but proceeding distinct from the prediction). A "sensory input" can derive not only from a sensory device targeted on external reality, but also from sensory cognition targeted on *any* process whose outcome, on the level predicted, is not subject to direct control. (Of course, from our perspective, prediction of the "real world" remains the strongest test.)

- A *decisive* binding occurs when the model can predict the effects of several possible actions on reality, and choose whichever action yields the best result under some goal system (see below). By predicting outcomes under several possible world-states, consisting of the present world-state plus each of several possible actions, it becomes possible to choose between futures.
- A *manipulative* binding occurs when the AI can describe a desirable future with subjunctive imagery, and invent a sequence of actions which leads to that future. Where *decision* involves selecting one action from a predetermined and bounded set, *manipulation* involves inventing new actions, perhaps actions previously unperceived because the set of possible actions is unbounded or computationally large. The simplest form of manipulation is backward chaining from parent goals to child goals using causal beliefs; this is not the only form of manipulation, but it is superior to exhaustive forward search from all possible actions.

I also distinguish three successive grades of *variable complexity:*

- A *discrete* variable has referents selected from a bounded set which is computationally small - for example, a set of twenty possible actions, or a set of twenty-six possible lowercase letters. The binary presence or absence of a feature is also a discrete variable.
- A *quantitative* variable is selected from the set of real numbers, or from a computationally large set which approximates a smoothly varying scalar quantity (such as the set of floating-point numbers).
- A *patterned* variable is composed of a finite number of quantitative or discrete elements. Examples: A finite string of lowercase letters, e.g. "mkrznye". A real point in 3D space (three quantitative elements). A 2D black-and-white image (2D array of binary pixels).

The dimension of variable complexity is orthogonal to the SPDM (sensory-predictive-decisive-manipulative) dimension, but like SPDM it describes successively tougher tests of intelligence. A decisive binding from desired result to desirable action is computationally feasible only when the "action" is a discrete variable chosen from a small set - small enough that each possible action can be modeled. When the action is a quantitative variable, selected from computationally large sets such as the floating-point numbers in the interval [0, 1], some form of manipulative binding, such as backward chaining, is necessary to arrive at the specific action required. (Note that adding a continuous time parameter to a discrete action renders it quantitative.) Binding precise quantitative goal imagery to a precise quantitative action cannot be done by exhaustive testing of the alternatives; it requires a way to transform the goal imagery so as to arrive at subgoal imagery or action imagery. The simplest transformation is the identity relation - but even the identity transformation is not possible to a *purely* forward-search mechanism. The next most straightforward method would be to employ a causal belief that specifies a reversible relation between the antecedent and the consequent. In real-time control tasks, motor modalities (in humans, the entire sensorimotor system) may automatically produce action symphonies in order to achieve quantitative or patterned goals.

A string of several discrete or quantitative variables creates a patterned variable, which is also likely to be computationally intractable for exhaustive forward search. Binding a patterned goal to a patterned action, if the relation is not one of direct identity, requires (again) a causal belief that specifies a reversible relation between the antecedent and the consequent, or (if no such belief is forthcoming) deliberative analysis of complex regularities in the relation between the action and the outcome, or exploratory tweaking followed by induction on which tweaks increase the apparent similarity between the outcome and the desired outcome.

There are levels of organization within bindings; a loose binding at one level can give rise to a tighter binding at a higher level. The rods and cones of the retina correspond to incoming photons that correspond to points on the surface of an object. The binding between a metaphorical pixel in the retina and a point in a real-world surface is very weak, very breakable; a stray ray of light can wildly change the detected optical intensity. But the actual sensory experience occupies one level of organization *above* individual pixels. The fragile sensory binding between retinal pixels and surface points, on a lower level of organization, gives rise to a solid sensory binding between our perception of the entire object and the object itself. A match between two discrete variables or two rough quantitative variables can arise by chance; a match between two patterned variables on a higher holonic level of organization is far less likely to arise from complete coincidence, though it may arise from a cause other than the obvious. The concept kernels in human visual recognition likewise bind to the entire perceptual experience of an object, not to individual pixels of the object. On an even higher level of organization, the *manipulative* binding between human intelligence and the real world is nailed down by many individually tight *sensory* bindings between conceptual imagery and real-world referents. Under the human implementation, there are at least three levels of organization *within* the correspondence theory of truth! The AI pathology that we perceive as "weak semantics" - which is very hard to define, but is an intuitive impression shared by many AI philosophers - may arise from omitting levels of organization in the binding between a model and its referent.

2.7.3: Actions

The series of motor actions I use to strike a key on my keyboard have enough degrees of freedom that "which key I strike", as a discrete variable, or "the sequence of keys struck", as a patterned variable, are both subject to direct specification. I do not need to engage in complex planning to strike the key sequence "hello world" or "labm4"; I can specify the words or letters directly and without need for complex planning. My motor areas and cerebellum do an enormous amount of work behind the scenes, but it is work that has been optimized to the point of subjective invisibility. A keystroke is thus an *action* for pragmatic purposes, although for a novice typist it might be a goal. As a first approximation, goal imagery has been reduced to action imagery when the imagery can direct a realtime skill in the relevant modality. This does not necessarily mean that actions are handed off to skills with no further interaction; realtime manipulations sometimes go wrong, in which case the interrelation between goals and actions and skills becomes more intricate, sometimes with multiple changing goals interacting with realtime skills. Imagery approaches the action level as it becomes able to interact with realtime skills.

Sometimes a goal does not directly reduce to actions because the goal referent is physically distant or physically separated from the "effectors" - the motor appendages or their virtual equivalents - so that manipulating the goal referent depends on first overcoming the physical separation as a subproblem. However, in the routine activity of modern-day humans, another very common reason

why goal imagery does not translate directly into action imagery is that the goal imagery is a high-level abstract characteristic, *cognitively* separated from the realm of direct actions. I can control every keystroke of my typing, but the quantitative percept of *writing quality*[32] referred to by the goal imagery of *high writing quality* is not subject to direct manipulation. I cannot directly set my writing quality to equal that of Shakespeare, in the way that I can directly set a keystroke to equal "H", because *writing quality* is a derived, abstract quantity. A better word than "abstract" is "holonic", the term used earlier from [Koestler67] and used to describe the way in which a single quality may simultaneously be a whole composed of parts, and a part in a greater whole. *Writing quality* is a quantitative *holon* which is eventually bound to the series of discrete keystrokes. I can directly choose keystrokes, but cannot directly choose the writing-quality holon. To increase the writing quality of a paragraph I must link the writing-quality holon to lower-level holons such as *correct spelling* and *omitting needless words,* which are qualities of the *sentences* holons, which are created through keystroke actions. Action imagery is typically, though not always, the level on which variables are completely free (directly specifiable with many degrees of freedom); higher levels involve interacting constraints which must be resolved through deliberation.

2.7.4: Goals

The very-high-level abstract goal imagery for *writing quality* is bound to directly specifiable action imagery for *words* and *keystrokes* through an intermediate series of child goals which inherit desirability from parent goals. But what are goals? What is desirability? So far I have been using an intuitive definition of these terms, which often suffices for describing how the goal system interacts with other systems, but is not a description of the goal system itself.

Unfortunately, the human goal system is somewhat... *confused*... as you know if you're a human. Most of the human goal system originally evolved in the absence of deliberative intelligence, and as a result, behaviors that contribute to survival and reproduction tend to be evolved as independent drives. Taking the intentionalist stance toward evolution, we would say that the sex drive is a child goal of reproduction. Over evolutionary time this might be a valid stance. But individual organisms are best regarded as adaptation-executers rather than fitness-maximizers, and the sex drive is not *cognitively* a child goal of reproduction; hence the modern use of contraception. Further complications are introduced at the primate level by the existence of complex social groups; consequently primates have "moral" adaptations, such as reciprocal altruism, third-party intervention to resolve conflicts ("community concern"), and moralistic aggression against community offenders [Flack00]. Still further complications are introduced by the existence of deliberative reasoning and linguistic communication in humans; humans are imperfectly deceptive social organisms that argue about each other's motives in adaptive contexts. This has produced what I can only call "philosophical" adaptations, such as the ways we reason about causation in moral arguments - ultimately giving us the ability to pass (negative!) judgement on the moral worth of our evolved goal systems and evolution itself.

It is not my intent to untangle that vast web of causality in this paper, although I have written (informally but at length) about the problem elsewhere [Yudkowsky01], including a description of the cognitive and motivational architectures required for a mind to engage in such apparently paradoxical behaviors as passing coherent judgement on its own top-level goals. (For example, a mind may regard the current representation of morals as a probabilistic approximation to a moral referent that can be reasoned about.) The architecture of morality is a pursuit that goes along with

the pursuit of general intelligence, and the two should not be parted, for reasons that should be obvious and will become even more obvious in Part III; but unfortunately there is simply not enough room to deal with the issues here. I will note, however, that the human goal system sometimes does the Wrong Thing[33] and I do not believe AI should follow in those footsteps; a mind may share our moral frame of reference without being a functional duplicate of the human goal supersystem.

Within this paper I will set aside the question of moral reasoning and take for granted that the system supports moral content. The question then becomes how moral content binds to goal imagery and ultimately to actions.

The imagery that describes the supergoal is the moral content and describes the events or world-states that the mind regards as having intrinsic value. In classical terms, the supergoal description is analogous to the intrinsic utility function. Classically, the total utility of an event or world-state is its intrinsic utility, plus the sum of the intrinsic utilities (positive or negative) of the future events to which that event is predicted to lead, multiplied in each case by the predicted probability of the future event as a consequence. (Note that predicted consequences include both direct and indirect consequences, i.e., consequences of consequences are included in the sum.) This may appear at first glance to be yet another oversimplified Good Old-Fashioned AI definition, but for once I shall argue in favor; the classical definition is more fruitful of complex behaviors than first apparent. The property *desirability* should be coextensive with, and should behave identically to, the property *is-predicted-to-lead-to-intrinsic-utility*.

Determining which actions are predicted to lead to the greatest total intrinsic utility, and inventing actions which lead to greater intrinsic utility, has subjective regularities when considered as a cognitive problem and external regularities when considered as an event structure. These regularities are called *subgoals*. Subgoals define areas where the problem can be efficiently viewed from a local perspective. Rather than the mind needing to rethink the entire chain of reasoning "Action A leads to B, which leads to C, which leads to D, [...], which leads to actual intrinsic utility Z", there is a useful regularity that actions which lead to B are mostly predicted to lead through the chain to Z. Similarly, the mind can consider which of subgoals B1, B2, B3 are most likely to lead to C, or consider which subgoals C1, C2, C3 are together sufficient for D, without rethinking the rest of the logic to Z.

This network (not hierarchical) event structure is an *imperfect* regularity; desirability is heritable only to the extent, and exactly to the extent, that predicted-to-lead-to-Z-ness is heritable. Our low-entropy universe has category structure, but not perfect category structure. Using imagery to describe an event E which is predicted to lead to event F is never perfect; perhaps *most* real-world states that fit description E lead to events that fit description F, but it would be very rare, outside of pure mathematics, to find a case where the prediction is perfect. There will always be some states in the volume carved out by the description E that lead to states outside the volume carved out by description F. If C is predicted to lead to D, and B is predicted to lead to C, then usually B will inherit C's predicted-to-lead-to-D-ness. However, it may be that B leads to a special case of C which does not lead to D; in this case, B would not inherit C's predicted-to-lead-to-D-ness. Therefore, if C had inherited desirability from D, B would not inherit C's desirability either.

To deal with a world of imperfect regularities, goal systems model the regularities in the irregularities, using descriptive constraints, distant entanglements, and global heuristics. If events fitting description E usually but not always lead to events fitting description F, then the mental imagery describing E, or even the concepts making up the description of E, may be refined to narrow the extensional class to eliminate events that seem to fit E but that don't turn out to lead to F. These "descriptive constraints" drive the AI to focus on concepts and categories that expose predictive, causal, and manipulable regularities in reality, rather than just surface regularities.

A further refinement is "distant entanglements"; for example, an action A that leads to B which leads to C, but which also simultaneously has side effects that block D, which is C's source of desirability. Another kind of entanglement is when action A leads to unrelated side effect S, which has negative utility outweighing the desirability inherited from B.

"Global heuristics" describe goal regularities that are general across many problem contexts, and which can therefore be used to rapidly recognize positive and negative characteristics; the concept "margin for error" is a category that describes an important feature of many plans, and the belief "margin for error supports the local goal" is a global heuristic that positively links members of the perceptual category *margin for error* to the local goal context, without requiring separate recapitulation of the inductive and deductive support for the general heuristic. Similarly, in self-modifying or at least self-regulating AIs, "minimize memory usage" is a subgoal that many other subgoals and actions may impact, so the perceptual recognition of events in the "memory usage" category or "leads to memory usage" categories implies entanglement with a particular distant goal.

Descriptive constraints, distant entanglements, and global heuristics do not violate the desirability-as-prediction model; descriptive constraints, distant entanglements, and global heuristics are also useful for modeling complex predictions, in the same way and for the same reasons as they are useful in modeling goals. However, there are at least three reasons for the activity of *planning* to differ from the activity of *prediction*. First, prediction typically proceeds forward from a definite state of the universe to determine what comes after, while planning often (though not always) reasons backward from goal imagery to pick out one point in a space of possible universes, with the space's dimensions determined by degrees of freedom in available actions. Second, desirabilities are differential, unlike predictions; if A and ~A both lead to the same endpoint E, then from a predictive standpoint this may increase the confidence in E, but from a planning standpoint it means that neither A nor ~A will inherit *net* desirability from E. The final effect of desirability is that an AI chooses the *most* desirable action, an operation which is comparative rather than absolute; if both A and ~A lead to E, neither A nor ~A transmit differential desirability to actions.

Third, while both *implication* and *causation* are useful for reasoning about predictions, only causal links are useful in reasoning about goals. If the observation of A is usually followed by the observation of B, then this makes A a good predictor of B - regardless of whether A is the direct cause of B, or whether there is a hidden third cause C which is the direct cause of both A and B. I would regard *implication* as an emergent property of a directed network of events whose underlying behavior is that of causation; if C causes A, and then causes B, then A will imply B. Both "A causes B" (direct causal link) and "A implies B" (mutual causal link from C) are useful in prediction. However, in planning, the distinction between "A directly causes B" and "A and B are both effects of C" leads to a distinction between "Actions that lead to A, as such, are likely to lead to B" and "Actions that lead directly to A, without first leading through C, are unlikely to have any effect on B". This distinction

also means that experiments in manipulation tend to single out real causal links in a way that predictive tests do not. If A implies B then it is often the case that C causes both A and B, but it is rarer in most real-world problems for an action intended to affect A to separately and invisibly affect the hidden third cause C, giving rise to false confirmation of direct causality[34]. (Although it happens, especially in economic and psychological experiments.)

2.7.5: Activities of intelligence: Explanation, prediction, discovery, planning, design

So far, this section has introduced the distinction between sensory, predictive, decisive, and manipulative models; discrete, quantitative, and patterned variables; the holonic model of high-level and low-level patterns; and supergoal referents, goal imagery, and actions. These ideas provide a framework for understanding the immediate subtasks of intelligence - the moment-to-moment activities of deliberation. In carrying out a high-level cognitive task such as *design a bicycle,* the subtasks consist of crossing gaps from very high-level holons such as *good transport* to the holon *fast propulsion* to the holon *pushing on the ground* to the holon *wheel* to the holons for *spokes* and *tires,* until finally the holons become directly specifiable in terms of design components and design materials directly available to the AI.

The activities of intelligence can be described as *knowledge completion* in the service of *goal completion.* To complete a bicycle, one must first complete a design for a bicycle. To carry out a plan, one must complete a mental picture of a plan. Because both planning and design make heavy use of knowledge, they often spawn purely knowledge-directed activities such as explanation, prediction, and discovery. These activities are messy, non-inclusive categories, but they illustrate the general sorts of things that general minds do.

Knowledge activities are carried out both on a large scale, as major strategic goals, and on a small scale, in routine subtasks. For example, "explanation" seeks to extend current knowledge, through deduction or induction or experiment, to fill the gap left by the unknown cause of a known effect. The unknown cause will at least be the referent of question imagery, which will bring into play sequiturs and verifiers which react to open questions. If the problem becomes salient enough, and difficult enough, finding the unknown cause may be promoted from question imagery to an internal goal, allowing the AI to reason deliberatively about which problem-solving strategies to deploy. The knowledge goal for "building a plan" inherits desirability from the objective of the plan, since creating a plan is required for (is a subgoal of) achieving the objective of the plan. The knowledge goal for explaining an observed failure might inherit desirability from the goal achievable when the failure is fixed. Since knowledge goals can govern actual actions and not just the flow of sequiturs, they should be distinguished from question imagery. Knowledge goals also permit reflective reasoning about what kind of internal actions are likely to lead to solving the problem; knowledge goals may invoke sequiturs that search for beliefs about *solving knowledge problems,* not just beliefs about the specific problem at hand.

Explanation fills holes in knowledge about the past. Prediction fills holes in knowledge about the future. Discovery fills holes in knowledge about the present. Design fills gaps in the mental model of a tool. Planning fills gaps in a model of future strategies and actions. Explanation, prediction, discovery, and design may be employed in the pursuit of a specific real-world goal, or as an independent pursuit in the anticipation of the resulting knowledge being useful in future goals - "curiosity". Curiosity fills completely general gaps (rather than being targeted on specific, already-

known gaps), and involves the use of forward-looking reasoning and experimentation, rather than backward chaining from specific desired knowledge goals; curiosity might be thought of as filling the very abstract goal of "finding out X, where X refers to anything that will turn out to be a good thing to know later on, even though I don't know specifically what X is." (Curiosity involves a very abstract link to intrinsic utility, but one which is nonetheless completely true - curiosity *is* useful.)

What all the activities have in common is that they involve reasoning about a complex, holonic model of causes and effects. "Explanation" fills in holes about the past, which is a complex system of cause and effect. "Prediction" fills in holes in the future, which is a complex system of cause and effect. "Design" reasons about tools, which are complex holonic systems of cause and effect. "Planning" reasons about strategies, which are complex holonic systems of cause and effect. Intelligent reasoning completes knowledge goals and answers questions in a complex holonic causal model, in order to achieve goal referents in a complex holonic causal system.

This gives us the three elements of DGI:

- The *what* of intelligence: Intelligence consists*, in humans,* of a highly modular brain with dozens of areas, which implements a deliberative process (built on thoughts built of concepts built on sensory modalities built on neurons); plus contributing subsystems (e.g. memory); plus surrounding subsystems (e.g. autonomic regulation); plus leftover subsystems implementing pre-deliberative approximations of deliberative processes; plus emotions, instincts, intuitions and other systems that influence the deliberative process in ways that were adaptive in the ancestral environment; plus everything else. A similar system is contemplated for AIs, of roughly the same order of complexity, but inevitably less messy. Both supersystems are characterized by levels of organization: Code / neurons, modalities, concepts, thoughts, and deliberation.
- The *why* of intelligence: The cause of human intelligence is evolution. Intelligence is an evolutionary advantage because it enables us to model reality, including external reality, social reality and internal reality, which in turn enables us to predict, decide, and manipulate reality. AIs will have intelligence because we, the human programmers, wish to accomplish a goal that can best be reached through smart AI, or because we regard the act of creating AI as having intrinsic utility; in either case, building AI requires building a deliberative supersystem that manipulates reality.
- The *how* of intelligence: Intelligence (deliberate reasoning) completes knowledge goals and answers questions in a complex holonic causal model, in order to achieve goal referents in a complex holonic causal system.

2.7.6: General intelligence

The evolutionary context of intelligence has historically included environmental adaptive contexts, social adaptive contexts (modeling of other minds), and reflective adaptive contexts (modeling of internal reality). In evolving to fit a wide variety of adaptive contexts, we have acquired much cognitive functionality that is visibly specialized for particular adaptive problems, but we have also acquired cognitive functionality that is adaptive across many contexts, and adaptive functionality that coopts previously specialized functionality for wider use. Humans can acquire substantial competence in modeling, predicting, and manipulating fully general regularities of our low-entropy universe. We call this ability "general intelligence". In some ways our ability is very weak; we often

solve general problems abstractly instead of perceptually, so we can't deliberatively solve problems on the order of realtime visual interpretation of a 3D scene. But we can often say something which is true enough to be useful and simple enough to be tractable. We can deliberate on how vision works, even though we can't deliberate fast enough to perform realtime visual processing.

There is currently a broad trend toward one-to-one mappings of cognitive subsystems to domain competencies. While in popular psychology this often degenerates into phrenology, such abuses are of course irrelevant to genuine hypotheses about mappings between specialized domain competencies and specialized computational subsystems, or decisions to pursue specialized AI. In DGI, human intelligence is held to consist of a supersystem with complex interdependent subsystems that exhibit *internal* functional specialization, but this does not rule out the existence of other subsystems that contribute solely or primarily to specific cognitive talents and domain competencies, or subsystems that contribute more heavily to some cognitive talents than others. The mapping from computational subsystems to cognitive talents is many-to-many, and the mapping from cognitive talents plus acquired expertise to domain competencies is also many-to-many, but this does not rule out specific correspondences between human variances in the "computing power" (generalized cognitive resources) allocated to computational subsystems and observed variances in cognitive talents or domain competencies.

However, the subject matter of AI is not the variance between humans, but the base of adaptive complexity common to all humans (or at least all neurologically intact humans). If increasing the resources allocated to a cognitive subsystem yields an increase in a cognitive talent or domain competency, it does not follow that the talent or competency can be implemented by that subsystem alone. It should also be noted that under the traditional paradigm of programming, programmers' thoughts about solving specific problems are translated into code, and this is the idiom underlying most branches of classical AI; for example, expert systems engineers supposedly translate the beliefs in specific domains directly into the cognitive content of the AI. This would naturally tend to yield a view of intelligence in which there is a one-to-one mapping between subsystems and competencies. I believe this is the underlying cause of the atmosphere in which the quest for intelligent AI is greeted with the reply: "AI that is intelligent in what domain?"

This does not mean that exploration in specialized AI is entirely worthless; in fact, DGI's levels of organization suggest a specific class of cases where specialized AI may prove fruitful. Sensory modalities lie directly above the code level; sensory modalities were some of the first specialized cognitive subsystems to evolve and hence are not as reliant on a supporting supersystem framework, although other parts of the supersystem depend heavily on modalities. This suggests a specialized approach, with programmers directly writing code, may prove fruitful if the project is constructing a sensory modality. And indeed, AI research that focuses on creating sensory systems and sensorimotor systems continues to yield real progress. Such researchers are following evolution's incremental path, often knowingly so, and thereby avoiding the pitfalls that result from violating the levels of organization.

However, I still do not believe it is possible to match the deliberative supersystem's inherently broad applicability by implementing a separate computational subsystem for each problem context. Not only is it impossible to duplicate general intelligence through the sum of such subsystems, I suspect it is impossible to achieve humanlike performance in most *single* contexts using specialized AI. Occasionally we use abstract deliberation to solve modality-level problems for which we lack

sensory modalities, and in this case it is possible for AI projects to solve the problem on the modality level, but the resulting problem-solving method will be very different from the human one, and will not generalize outside the specific domain. Hence Deep Blue.

Even on the level of individual domain competencies, not all competencies are unrelated to each other. Different minds may have different abilities in different domains; a mind may have an "ability surface", with hills and spikes in areas of high ability; but a spike in an area such as *learning* or *self-improvement* tends to raise the rest of the ability surface [Voss01]. The talents and subsystems that are general in the sense of contributing to many domain competencies - and the domain competencies of self-improvement; see Part III - occupy a strategic position in AI analogous to the central squares in chess.

2.7.7: Self

When can an AI legitimately use the word "I"?

(For the sake of this discussion, I must give the AI a temporary proper name; I will use "Aisa" during this discussion.)

A classical AI that contains a LISP token for "hamburger" knows nothing about hamburgers; at most the AI can recognize recurring instances of a letter-sequence typed by programmers. Giving an AI a suggestively named data structure or function does not make that component the functional analogue of the similarly named human feature [McDermott76]. At what point can Aisa talk about something called "Aisa" without Drew McDermott popping up and accusing us of using a term that might as well translate to "G0025"?

Suppose that Aisa, in addition to modeling virtual environments and/or the outside world, also models certain aspects of internal reality, such as the effectiveness of heuristic beliefs used on various occasions. The degrees of binding between a model and reality are sensory, predictive, decisive, and manipulative. Suppose that Aisa can sense when a heuristic is employed, notice that heuristics tend to be employed in certain contexts and that they tend to have certain results, and use this inductive evidence to formulate expectations about when a heuristic will be employed and predict the results on its employment. Aisa now predictively models Aisa; it forms beliefs about its operation by observing the introspectively visible effects of its underlying mechanisms. Tightening the binding from predictive to manipulative requires that Aisa link introspective observations to internal actions; for example, Aisa may observe that devoting discretionary computational power to a certain subprocess yields thoughts of a certain kind, and that thoughts of this kind are useful in certain contexts, and subsequently devote discretionary power to that subprocess in those contexts.

A manipulative binding between Aisa and Aisa's model of Aisa is enough to let Aisa legitimately say "Aisa is using heuristic X", such that using the term "Aisa" is materially different from using "hamburger" or "G0025". But can Aisa legitimately say, "*I* am using heuristic X"?

My favorite quote on this subject comes from Douglas Lenat, although I cannot find the reference and am thus quoting from memory: "While Cyc knows that there is a thing called Cyc, and that Cyc is a computer, it does not know that *it* is *Cyc*." Personally, I would question whether Cyc knows that

Cyc is a computer - but regardless, Lenat has made a legitimate and fundamental distinction. Aisa modeling a thing called Aisa is not the same as Aisa modeling itself.

In an odd sense, assuming that the problem exists is enough to solve the problem. If another step is required before Aisa can say "I am using heuristic X", then there must be a material difference between saying "Aisa is using heuristic X" and "I am using heuristic X". And that is one possible answer: Aisa can say "I" when the behavior of modeling itself is materially different, because of the self-reference, from the behavior of modeling another AI that happens to look like Aisa.

One specific case where self-modeling is materially different than other-modeling is in planning. Employing a complex plan in which a linear sequence of actions A, B, C are individually necessary and together sufficient to accomplish goal G requires an implicit assumption that the AI will follow through on its own plans; action A is useless unless it is followed by actions B and C, and action A is therefore not desirable unless actions B and C are predicted to follow. Making complex plans does not actually *require* self-modeling, since many classical AIs engage in planning-like behaviors using programmatic assumptions in place of reflective reasoning, and in humans the assumption is usually automatic rather than being the subject of deliberation. However, deliberate reflective reasoning about complex plans requires an understanding that the future actions of the AI are determined by the decisions of the AI's future self, that there is some degree of continuity (although not perfect continuity) between present and future selves, and that there is thus some degree of continuity between present decisions and future actions.

An intelligent mind navigates a universe with four major classes of variables: Random factors, variables with hidden values, the actions of other agents, and the actions of the self. The space of possible actions differs from the spaces carved out by other variables because the space of possible actions is under the AI's control. One difference between "Aisa will use heuristic X" and "*I* will use heuristic X" is the degree to which heuristic usage is under Aisa's deliberate control - the degree to which Aisa has goals relating to heuristic usage, and hence the degree to which the observation "I predict that I will use heuristic X" affects Aisa's subsequent actions. Aisa, if sufficiently competent at modeling other minds, might predict that a similar AI named Aileen would also use heuristic X, but beliefs about Aileen's behaviors would be derived from predictive modeling of Aileen, and not decisive planning of internal actions based on goal-oriented selection from the space of possibilities. There is a cognitive difference between Aisa saying "I predict Aileen will use heuristic X" and "I plan to use heuristic X". On a systemic level, the global specialness of "I" would be nailed down by those heuristics, beliefs, and expectations that individually relate specially to "I" because of introspective reflectivity or the space of undecided but decidable actions. It is my opinion that such an AI would be able to legitimately use the word "I", although in humans the specialness of "I" may be nailed down by additional cognitive forces as well. (Legitimate use of "I" is explicitly *not* offered as a necessary and sufficient condition for the "hard problem of conscious experience" [Chalmers95] or social, legal, and moral personhood.)

3: Part III: Seed AI

In the space between the theory of human intelligence and the theory of general AI is the ghostly outline of a theory of *minds in general,* specialized for humans and AIs. I have not tried to lay out such a theory explicitly, confining myself to discussing those specific similarities and differences of humans and AIs that I feel are worth guessing in advance. The Copernican revolution for cognitive

science - humans as a noncentral special case - is not yet ready; two points are not enough to draw a curve, and currently we only have one. Nonetheless, humans *are in fact* a noncentral special case, and this abstract fact is knowable even if our current theories are anthropocentric.

There is a fundamental rift between evolutionary design and deliberative design. From the perspective of a deliberative intelligence - a human, for instance - evolution is the degenerate case of design-and-test where intelligence equals zero. Mutations are atomic; recombinations are random; changes are made on the genotype's lowest level of organization (flipping genetic bits); the grain size of the component tested is the whole organism; and the goodness metric operates solely through induction on historically encountered cases, without deductive reasoning about which contextual factors may later change[35]. The evolution of evolvability [Wagner96] improves this picture somewhat. There is a tendency for low-level genetic bits to exert control over high-level complexity, so that changes to those genes can create high-level changes. Blind selection pressures can create self-wiring and self-repairing systems that turn out to be highly evolvable because of their ability to phenotypically adapt to genotypical changes. Nonetheless, the evolution of evolvability is not a substitute for intelligent design. Evolution works, despite local inefficiencies, because evolution exerts vast cumulative design pressure over time.

However, the total amount of design pressure exerted over a given time is limited; there is only a limited amount of selection pressure to be divided up among all the genetic variances selected on in any given generation [Worden95]. One obvious consequence is that evolutionarily recent adaptations will probably be less optimized than those which are evolutionarily ancient. In DGI, the evolutionary phylogeny of intelligence roughly recapitulates its functional ontogeny; it follows that higher levels of organization may contain less total complexity than lower levels, although sometimes higher levels of organization are also more evolvable. Therefore, a subtler consequence is that the lower levels of organization are likely to be less well adapted to evolutionarily recent innovations (such as deliberation) than those higher levels to the lower levels - an effect enhanced by evolution's structure-preserving properties, including the preservation of structure that evolved in the absence of deliberation. Any design possibilities that first opened up with the appearance of *Homo sapiens sapiens* remain unexploited because *Homo sapiens sapiens* has only existed for 50,000-100,000 years; this is enough time to select among variances in quantitative tendencies, but not really enough time to construct complex functional adaptation. Since only *Homo sapiens sapiens* in its most modern form is known to engage in computer programming, this may explain why we do not yet have the capacity to reprogram our own neurons (said with tongue firmly in cheek, but there's still a grain of truth). And evolution is *extremely* conservative when it comes to wholesale revision of architectures; the homeotic genes controlling the embryonic differentiation of the forebrain, midbrain, and hindbrain have identifiable homologues in the developing head of the *Drosophila* fly(!) [Holland92].

Evolution never refactors its code. It is far easier for evolution to stumble over a thousand individual optimizations than for evolution to stumble over two simultaneous changes which are together beneficial and separately harmful. The genetic code that specifies the mapping between codons (a codon is three DNA bases) and the 20 amino acids is inefficient; it maps 64 possible codons to 20 amino acids plus the stop code. Why hasn't evolution shifted one of the currently redundant codons to a new amino acid, thus expanding the range of possible proteins? Because for any complex organism, the smallest change to the behavior of DNA - the lowest level of genetic organization - would destroy virtually all higher levels of adaptive complexity, unless the change

were accompanied by millions of other simultaneous changes throughout the genome to shift every suddenly-nonstandard codon to one of its former equivalents. Evolution simply cannot handle simultaneous dependencies, unless individual changes can be deployed incrementally, or multiple phenotypical effects occur as the consequence of a single genetic change. For humans, planning coordinated changes is routine; for evolution, impossible. Evolution is hit with an enormous discount rate when exchanging the paper currency of incremental optimization for the hard coin of complex design.

We should expect the human design to incorporate an intimidatingly huge number of simple functional optimizations. But it is also understandable if there are deficits in the higher design. While the higher levels of organization (including deliberation) have emerged from the lower levels and hence are fairly well adapted to them, the lower levels of organization are not as adapted to the existence of deliberate intelligence. Humans were constructed by accretive evolutionary processes, moving from very complex nongeneral intelligence to very complex general intelligence, with deliberation the last layer of icing on the cake.

Can we exchange the hard coin of complex design for the paper currency of low-level optimization? "Optimizing compilers" are an obvious step but a tiny one; program optimization makes programs faster but exerts no design pressure for better functional organization, even for simple functions of the sort easily optimized by evolution. Directed evolution, used on modular subtasks with clearly defined performance metrics, would be a somewhat larger step. But even directed evolution is still the degenerate case of design-and-test where individual steps are unintelligent. We are, by assumption, building an AI. Why use *unintelligent* design-and-test?

Admittedly, there is a chicken-and-egg limit on relying on an AI's intelligence to help build an AI. Until a stably functioning cognitive supersystem is achieved, only the nondeliberative intelligence exhibited by pieces of the system will be available. Even after the achievement of a functioning supersystem - a heroic feat in itself - the intelligence exhibited by this supersystem will initially be very weak. The weaker an AI's intelligence, the less ability the AI will show in understanding complex holonic systems. The weaker an AI's abilities at holonic design, the smaller the parts of itself that the AI will be able to understand. At whatever time the AI finally becomes smart enough to participate in its own creation, the AI will initially need to concentrate on improving small parts of itself with simple and clear-cut performance metrics supplied by the programmers. This is not a special case of a stupid AI trying to understand itself, but a special case of a stupid AI trying to understand any complex holonic system; when the AI is "young" it is likely to be limited to understanding simple elements of a system, or small organizations of elements, and only where clear-cut goal contexts exist (probably programmer-explained). But even a primitive holonic design capability could cover a human gap; we don't like fiddling around with little things because we get bored, and we lack the ability to trade our massive parallelized power on complex problems for greater serial speed on simple problems. Similarly, it would be unhealthy (would result in AI pathologies) for human programming abilities to play a permanent role in learning or optimizing concept kernels - but at the points where interference seems tempting, it is perfectly acceptable for the *AI's* deliberative processes to play a role, if the AI has advanced that far.

Human intelligence, created by evolution, is characterized by evolution's design signature. The vast majority of our genetic history took place in the *absence* of deliberative intelligence; our older cognitive systems are poorly adapted to the possibilities inherent in deliberation. Evolution has

applied vast design pressures to us but has done so very unevenly; evolution's design pressures are filtered through an unusual methodology that works far better for hand-massaging code than for refactoring program architectures.

Now imagine a mind built in its own presence by intelligent designers, beginning from primitive and awkward subsystems that nonetheless form a complete supersystem. Imagine a development process in which the elaboration and occasional refactoring of the subsystems can coopt any degree of intelligence, however small, exhibited by the supersystem. The result would be a fundamentally different design signature, and a new approach to Artificial Intelligence which I call *seed AI*.

A seed AI is an AI designed for self-understanding, self-modification, and recursive self-improvement. This has implications both for the functional architectures needed to achieve primitive intelligence, and for the later development of the AI if and when its holonic self-understanding begins to improve. Seed AI is not a *workaround* that avoids the challenge of general intelligence by bootstrapping from an unintelligent core; seed AI only begins to yield benefits once there is some degree of available intelligence to be utilized. The later consequences of seed AI (such as true recursive self-improvement) only show up after the AI has achieved significant holonic understanding and general intelligence. The bulk of this paper, Part II, describes the general intelligence that is prerequisite to seed AI; Part III assumes some degree of success in constructing general intelligence and asks what may happen afterward. This may seem like hubris, but there are interesting things to be learned thereby, some of which imply design considerations for earlier architecture.

3.1: Advantages of minds-in-general

From the standpoint of computer science it may seem like breathtaking audacity if I dare to predict any advantages for AIs in advance of their construction, given past failures. But from the standpoint of evolutionary psychology, the human mind has surprising flaws to match its surprising strengths. If discussing the potential advantages of "AIs" strikes you as too audacious, then consider what follows, not as discussing the potential advantages of "AIs", but as discussing the potential advantages of *minds in general* relative to humans. One may then consider separately the audacity involved in claiming that a given AI approach can achieve one of these advantages, or that it can be done in less than fifty years.

Humans definitely possess the following advantages, relative to *current* AIs:

- We are smart, flexible, generally intelligent organisms with an enormous base of evolved complexity, years of real-world experience, and 10^{14} parallelized synapses, and current AIs are not.

Humans probably possess the following advantages, relative to intelligences developed by humans on foreseeable extensions of current hardware:

- Considering each synaptic signal as roughly equivalent to a floating-point operation, the raw computational power of a human is enormously in excess of any current supercomputer or clustered computing system, although Moore's Law continues to eat up this ground [Moravec98].

- Human neural hardware - the wetware layer - offers built-in support for operations such as pattern recognition, pattern completion, optimization for recurring problems, et cetera; this support was added from below, taking advantage of microbiological features of neurons, and could be enormously expensive to simulate computationally to the same degree of ubiquity.
- With respect to the holonically simpler levels of the system, the total amount of "design pressure" exerted by evolution over time is probably considerably in excess of the design pressure that a reasonably-sized programming team could expect to personally exert.
- Humans have an extended history as intelligences; we are proven software.

Current computer programs definitely possess these mutually synergetic advantages relative to humans:

- Computer programs can perform highly repetitive tasks without boredom.
- Computer programs can execute complex extended tasks without making that class of human errors caused by distraction or short-term memory overflow in abstract deliberation.
- Computer hardware can perform extended sequences of simple steps at much greater *serial* speeds than human abstract deliberation or even human 200Hz neurons.
- Computer programs are fully configurable by the general intelligences called humans. (Evolution, the designer of humans, cannot invoke general intelligence.)

These advantages will not necessarily carry over to real AI. A real AI is not a computer program any more than a human is a cell. The relevant complexity exists at a much higher layer of organization, and it would be inappropriate to generalize stereotypical characteristics of computers to real AIs, just as it would be inappropriate to generalize the stereotypical characteristics of amoebas to modern-day humans. One might say that a real AI *consumes* computing power but is not a computer. This basic distinction has been confused by many cases in which the label "AI" has been applied to constructs that turn out to be only computer programs; but we should still expect the distinction to hold true of real AI, when and if achieved.

The potential cognitive advantages of *minds-in-general,* relative to human minds, probably include:

- *New sensory modalities.* Human programmers, lacking a sensory modality for assembly language, are stuck with abstract reasoning plus compilers. We are not entirely helpless, even this far outside our ancestral environment - but the traditional fragility of computer programs bears witness to our awkwardness. Minds-in-general may be able to exceed human programming ability with relatively primitive *general* intelligence, given a sensory modality for code.
- *Blending-over of deliberative and automatic processes.* Human wetware has very poor support for the realtime diversion of processing power from one subsystem to another. Furthermore, a computer can burn serial speed to generate parallel power but neurons cannot do the reverse. Minds-in-general may be able to carry out an uncomplicated, relatively uncreative track of deliberate thought using simplified mental processes that run at higher speeds - an idiom that blurs the line between "deliberate" and "algorithmic" cognition. Another instance of the blurring line is coopting deliberation into processes that are algorithmic in humans; for example, minds-in-general may choose to make use of top-level intelligence in forming and encoding the concept kernels of categories. Finally, a sufficiently intelligent AI might be able to incorporate *de novo* programmatic functions into deliberative processes - as if Gary

Kasparov[36] could interface his brain to a computer and write search trees to contribute to his intuitive perception of a chessboard.

- *Better support for introspective perception and manipulation.* The comparatively poor support of the human architecture for low-level introspection is most apparent in the extreme case of modifying code; we can think thoughts about thoughts, but not thoughts about individual neurons. However, other cross-level introspections are also closed to us. We lack the ability to introspect on concept kernels, focus-of-attention allocation, sequiturs in the thought process, memory formation, skill reinforcement, et cetera; we lack the ability to introspectively notice, induce beliefs about, or take deliberate actions in these domains.
- *The ability to add and absorb new hardware.* The human brain is instantiated with a species-typical upper limit on computing power and loses neurons as it ages. In the computer industry, computing power continually becomes exponentially cheaper, and serial speeds exponentially faster, with sufficient regularity that "Moore's Law" [Moore97] is said to govern its progress. Nor is an AI project limited to waiting for Moore's Law; an AI project that displays an important result may conceivably receive new funding which enables the project to buy a much larger clustered system (or rent a larger computing grid), perhaps allowing the AI to absorb hundreds of times as much computing power. By comparison, the 5-million-year transition from *Australopithecus* to *Homo sapiens sapiens* involved a tripling of cranial capacity relative to body size, and a further doubling of prefrontal volume relative to the expected prefrontal volume for a primate with a brain our size, for a total sixfold increase in prefrontal capacity relative to primates [Deacon90]. At 18 months per doubling, it requires 3.9 years for Moore's Law to cover this much ground. Even granted that intelligence is more software than hardware, this is still impressive.
- *Agglomerativity.* An advanced AI is likely to be able to communicate with other AIs at much higher bandwidth than humans communicate with other humans - including sharing of thoughts, memories, and skills, in their underlying cognitive representations. An advanced AI may also choose to internally employ multithreaded thought processes to simulate different points of view. The traditional hard distinction between "groups" and "individuals" may be a special case of human cognition rather than a property of minds-in-general. It is even possible that no one project would ever choose to split up available hardware among more than one AI. Much is said about the benefits of cooperation between humans, but this is because there is a species limit on individual brainpower. We solve difficult problems using many humans because we cannot solve difficult problems using *one big* human. Six humans have a fair advantage relative to one human, but one human has a tremendous advantage relative to six chimpanzees.
- *Hardware that has different, but still powerful, advantages.* Current computing systems lack good built-in support for biological neural functions such as automatic optimization, pattern completion, massive parallelism, etc. However, the bottom layer of a computer system is well-suited to operations such as reflectivity, execution traces, lossless serialization, lossless pattern transformations, very-high-precision quantitative calculations, and algorithms which involve iteration, recursion, and extended complex branching. Also in this category, but important enough to deserve its own section, is:
- *Massive serialism: Different 'limiting speed' for simple cognitive processes.* No matter how simple or computationally inexpensive, the speed of a human cognitive process is bounded by the 200Hz limiting speed of spike trains in the underlying neurons. Modern computer chips can execute billions of *sequential* steps per second. Even if an AI must "burn" this serial speed to imitate parallelism, simple (routine, noncreative, nonparallel) deliberation might be carried out substantially (orders of magnitude) faster than more computationally intensive thought

processes. If enough hardware is available to an AI, or if an AI is sufficiently optimized, it is possible that even the AI's full intelligence may run substantially faster than human deliberation.

- *Freedom from evolutionary misoptimizations.* The term "misoptimization" here indicates an evolved feature that was adaptive for inclusive reproductive fitness in the ancestral environment, but which today conflicts with the goals professed by modern-day humans. If we could modify our own source code, we would eat Hershey's lettuce bars, enjoy our stays on the treadmill, and use a volume control on "boredom" at tax time.
- *Everything evolution just didn't think of.* This catchall category is the flip side of the human advantage of "tested software" - humans aren't necessarily *good* software, just *old* software. Evolution cannot create design improvements which surmount simultaneous dependencies unless there exists an incremental path, and even then will not execute those design improvements unless that particular incremental path happens to be adaptive for other reasons. Evolution exhibits no predictive foresight and is strongly constrained by the need to preserve existing complexity. Human programmers are free to be creative.
- *Recursive self-enhancement.* If a seed AI can improve itself, each local improvement to a design feature means that the AI is now partially the *source* of that feature, in partnership with the original programmers. Improvements to the AI are now improvements to the *source* of the feature, and may thus trigger further improvement in that feature. Similarly, where the seed AI idiom means that a cognitive talent coopts a domain competency in internal manipulations, improvements to intelligence may improve the domain competency and thereby improve the cognitive talent. From a broad perspective, a mind-in-general's self-improvements may result in a higher level of intelligence and thus an increased ability to originate new self-improvements.

3.2: Recursive self-enhancement

Fully recursive self-enhancement is a potential advantage of minds-in-general that has no analogue in nature - not just no analogue in human intelligence, but no analogue in *any* known process. Since the divergence of the hominid family within the primate order, further developments have occurred at an accelerating pace - but this is not because the character of the evolutionary process changed or became "smarter"; successive adaptations for intelligence and language opened up new design possibilities and also tended to increase the selection pressures for intelligence and language. Similarly, the exponentially accelerating increase of cultural knowledge in *Homo sapiens sapiens* was triggered by an underlying change in the human brain, but has not itself had time to create any significant changes in the human brain. Once *Homo sapiens sapiens* arose, the subsequent runaway acceleration of cultural knowledge took place with essentially constant brainware. The exponential increase of culture occurs because acquiring new knowledge makes it easier to acquire more knowledge.

The accelerating development of the hominid family and the exponential increase in human culture are both instances of *weakly self-improving processes,* characterized by an externally constant process (evolution, modern human brains) acting on a complexity pool (hominid genes, cultural knowledge) whose elements interact synergetically. If we divide the process into an improver and a content base, then weakly self-improving processes are characterized by an external improving process with roughly constant characteristic intelligence, and a content base within which positive feedback takes place under the dynamics imposed by the external process.

If a seed AI begins to improve itself, this will mark the beginning of the AI's *self-encapsulation.* Whatever component the AI improves will no longer be *caused* entirely by humans; the cause of that component will become, at least in part, the AI. Any improvement to the AI will be an improvement to the *cause* of a component of the AI. If the AI is improved further - either by the external programmers, or by internal self-enhancement - the AI may have a chance to re-improve that component. That is, any improvement to the AI's global intelligence may indirectly result in the AI improving local components. This secondary enhancement does not necessarily enable the AI to make a further, tertiary round of improvements. If only a few small components have been self-encapsulated, then secondary self-enhancement effects are likely to be small, not on the same order as improvements made by the human programmers.

If computational subsystems give rise to cognitive talents, and cognitive talents plus acquired expertise give rise to domain competencies, then self-improvement is a means by which domain competencies can wrap around and improve computational subsystems, just as the seed AI idiom of coopting deliberative functions into cognition enables improvements in domain competencies to wrap around and improve cognitive talents, and the ordinary idiom of intelligent learning enables domain competencies to wrap around and improve acquired expertise[37]. The degree to which domain competencies improve underlying processes will depend on the AI's degree of advancement; successively more advanced intelligence is required to improve expertise, cognitive talents, and computational subsystems. The degree to which an improvement in intelligence cascades into further improvements will be determined by how much self-encapsulation has already taken place on different levels of the system.

A seed AI is a *strongly self-improving process,* characterized by improvements to the content base that exert direct positive feedback on the intelligence of the underlying improving process. The exponential surge of human cultural knowledge was driven by the action of an already-powerful but constant force, human intelligence, upon a synergetic content base of cultural knowledge. Since strong self-improvement in seed AI involves an initially very weak but improving intelligence, it is not possible to conclude from analogies with human cultural progress that strongly recursive self-improvement will obey an exponential lower bound during early stages, nor that it will obey an exponential upper bound during later stages. Strong self-improvement is a mixed blessing in development. During earlier epochs of seed AI, the dual process of programmer improvement and self-improvement probably sums to a process entirely dominated by the human programmers. We cannot rely on exponential bootstrapping from an unintelligent core. However, we may be able to achieve powerful results by bootstrapping from an *intelligent* core, if and when such a core is achieved. *Recursive* self-improvement is a consequence of seed AI, not a cheap way to achieve AI.

It is possible that self-improvement will become cognitively significant relatively early in development, but the wraparound of domain competencies to improve expertise, cognition, and subsystems does not imply strong effects from *recursive* self-improvement. Precision in discussing seed AI trajectories requires distinguishing between epochs for holonic understanding, epochs for programmer-dominated and AI-dominated development, epochs for recursive and nonrecursive self-improvement, and epochs for overall intelligence.

(Readers averse to advance discussion of sophisticated AI may consider these epochs as referring to minds-in-general that possess physical access to their own code and some degree of general

intelligence with which to manipulate it; the rationale for distinguishing between epochs may be considered separately from the audacity of suggesting that AI can progress to any given epoch.)

- Epochs for holonic understanding and holonic programming:
    - First epoch: The AI can transform code in ways that do not affect the algorithm implemented. ("Understanding" on the order of an optimizing compiler; i.e., not "understanding" in any real sense.)
    - Second epoch: The AI can transform algorithms in ways that fit simple abstract beliefs about the design purposes of code. That is, the AI would understand what a stack implemented as a linked list and a stack implemented as an array have in common. (Note that this is already out of range of current AI, at least if you want the AI to figure it out on its own.)
    - Third epoch: The AI can draw a holonic line from simple internal metrics of *cognitive* usefulness (how fast a concept is cued, the usefulness of the concept returned) to specific algorithms. Consequently the AI would have the theoretical capability to invent and test new algorithms. This does not mean the AI would have the ability to invent *good* algorithms or *better* algorithms, just that invention in this domain would be theoretically possible. (An AI's theoretical capacity for invention does not imply capacity for improvement over and above the programmers' efforts. This is determined by relative domain competencies and by relative effort expended at a given focal point.)
    - Fourth epoch: The AI has a concept of "intelligence" as the final product of a continuous holonic supersystem. The AI can draw a continuous line from (a) its abstract understanding of intelligence to (b) its introspective understanding of cognition to (c) its understanding of source code and stored data. The AI would be able to invent an algorithm or cognitive process that contributes to intelligence in a novel way and integrate that process into the system. (Again, this does not automatically imply that the AI's inventions are improvements relative to existing processes.)
- Epochs for sparse, continuous, and recursive self-improvement:
    - First epoch: The AI has a limited set of rigid routines which it applies uniformly. Once all visible opportunities are exhausted, the routines are used up. This is essentially analogous to the externally driven improvement of an optimizing compiler. An optimizing compiler may make a large number of improvements, but they are not self-improvements, and they are not design improvements. An optimizing compiler tweaks assembly language but leaves the program constant.
    - Second epoch: The cognitive processes which create improvements have characteristic complexity on the order of a classical search tree, rather than on the order of an optimizing compiler. Sufficient investments of computing power can sometimes yield extra improvements, but it is essentially an exponential investment for a linear improvement, and no matter how much computing power is invested, the total kind of improvements conceivable are limited.
    - Third epoch: Cognitive complexity in the AI's domain competency for programming is high enough that at any given point there is a large number of visible possibilities for complex improvements, albeit perhaps minor improvements. The AI usually does not exhaust all visible opportunities before the programmers

improve the AI enough to make new improvements visible. However, it is only programmer-driven improvements in intelligence which are powerful enough to open up new volumes of the design space.
    - Fourth epoch: Self-improvements sometimes result in genuine improvements to "smartness", "creativity", or "holonic understanding", enough to open up a new volume of the design space and make new possible improvements visible.
- Epochs for relative human-driven and AI-driven improvement:
    - First epoch: The AI can make optimizations at most on the order of an optimizing compiler, and cannot make design improvements or increase functional complexity. The combination of AI and programmer is not noticeably more effective than a programmer armed with an ordinary optimizing compiler.
    - Second epoch: The AI can understand a small handful of components and make improvements to them, but the total amount of AI-driven improvement is small by comparison with programmer-driven development. Sufficiently major programmer improvements do very occasionally trigger secondary improvements. The total amount of work done by the AI on its own subsystems serves only as a measurement of progress and does not significantly accelerate work on AI programming.
    - Third epoch: AI-driven improvement is significant, but development is "strongly" programmer-dominated in the sense that overall systemic progress is driven almost entirely by the creativity of the programmers. The AI may have taken over some significant portion of the work from the programmers. The AI's domain competencies for programming may play a critical role in the AI's continued functioning.
    - Fourth epoch: AI-driven improvement is significant, but development is "weakly" programmer-dominated. AI-driven improvements and programmer-driven improvements are of roughly the same kind, but the programmers are better at it. Alternatively, the programmers have more subjective time in which to make improvements, due to the number of programmers or the slowness of the AI.
- Epochs for overall intelligence:
    - Tool-level AI: The AI's behaviors are immediately and directly specified by the programmers, or the AI "learns" in a single domain using prespecified learning algorithms. (In my opinion, tool-level AI as an alleged step on the path to more complex AI is highly overrated.)
    - Prehuman AI: The AI's intelligence is not a significant subset of human intelligence. Nonetheless, the AI is a cognitive supersystem, with some subsystems we would recognize, and at least some mind-like behaviors. A toaster oven does not qualify as a "prehuman chef", but a general kitchen robot might do so.
    - Infrahuman AI: The AI's intelligence is, overall, of the same basic character as human intelligence, but substantially inferior. The AI may excel in a few domains where it possesses new sensory modalities or other brainware advantages not available to humans. I believe that a worthwhile test of infrahumanity is whether humans talking to the AI recognize a mind on the other end. (An AI that lacks even a primitive ability to communicate with and model external minds, and cannot be taught to do so, does not qualify as infrahuman.)

It should again be emphasized that this entire discussion *assumes* that the problem of building a general intelligence is solvable. Without significant *existing* intelligence an alleged "AI" will remain

permanently stuck in the first epoch of holonic programming - it will remain nothing more than an optimizing compiler. It is true that so far attempts at computer-based intelligence have failed, and perhaps there is a barrier which states that while 750 megabytes of DNA can specify physical systems which learn, reason, and display general intelligence, no amount of human design can do the same.

But if no such barrier exists - if it is possible for an artificial system to match DNA and display human-equivalent general intelligence - then it seems very probable that seed AI is achievable as well. It would be the height of biological chauvinism to assert that, while it *is* possible for humans to build an AI and improve this AI to the point of roughly human-equivalent general intelligence, this same human-equivalent AI can never master the (humanly solved) programming problem of making improvements to the AI's source code.

Furthermore, the above statement misstates the likely interrelation of the epochs. An AI does not need to wait for full human-equivalence to begin improving on the programmer's work. An optimizing compiler can "improve" over human work by expending greater *relative* effort on the assembly-language level. That is, an optimizing compiler uses the programmatic advantages of *greater serial speed* and *immunity to boredom* to apply much greater design pressures to the assembly-language level than a human could exert *in equal time*. Even an optimizing compiler might fail to match a human at hand-massaging a small chunk of time-critical assembly language. But, at least in today's programming environments, humans no longer hand-massage most code - in part because the task is best left to optimizing compilers, and in part because it's extremely boring and wouldn't yield much benefit relative to making further high-level improvements. A sufficiently advanced AI that takes advantage of *massive serialism* and *freedom from evolutionary misoptimizations* may be able to apply massive design pressures to higher holonic levels of the system.

Even at our best, humans are not very good programmers; programming is not a task commonly encountered in the ancestral environment. A human programmer is metaphorically a blind painter - not just a blind painter, but a painter entirely lacking a visual cortex. We create our programs like an artist drawing one pixel at a time, and our programs are fragile as a consequence. If the AI's human programmers can master the essential design pattern of sensory modalities, they can gift the AI with a sensory modality for codelike structures. Such a modality might perceptually interpret: a simplified interpreted language used to tutor basic concepts; any internal procedural languages used by cognitive processes; the programming language in which the AI's code level is written; and finally the native machine code of the AI's hardware. An AI that takes advantage of a codic modality may not need to wait for human-equivalent *general* intelligence to beat a human in the *specific domain competency* of programming. Informally, an AI is native to the world of programming, and a human is not.

This leads inevitably to the question of how much programming ability would be exhibited by a seed AI with human-equivalent general intelligence *plus* a codic modality. Unfortunately, this leads into territory that is generally considered taboo within the field of AI. Some readers may have noted a visible incompleteness in the above list of seed AI epochs; for example, the last stage listed for human-driven and AI-driven improvement is "weak domination" of the improvement process by human programmers (the AI and the programmers make the same kind of improvements, but the programmers make more improvements than the AI). The obvious succeeding epoch is one in which AI-driven development roughly equals human development, and the epoch after that one in

which AI-driven development exceeds human-driven development. Similarly, the discussion of epochs for recursive self-improvement stops at the point where AI-driven improvement sometimes opens up new portions of the opportunity landscape, but does not discuss the possibility of open-ended self-improvement: a point beyond which progress can continue in the absence of human programmers, so that by the time the AI uses up all the improvements visible at a given level, that improvement is enough to "climb the next step of the intelligence ladder" and make a new set of improvements visible. The epochs for overall intelligence define tool-level, prehuman, and infrahuman AI, but do not define human-equivalence or transhumanity.

3.3: Infrahumanity and transhumanity: "Human-equivalence" as anthropocentrism

It is interesting to contrast the separate perspectives of modern-day Artificial Intelligence researchers and modern-day evolutionary psychologists with respect to the particular level of intelligence exhibited by *Homo sapiens sapiens*. Modern-day AI researchers are strongly reluctant to discuss human equivalence, let alone what might lie beyond it, as a result of past claims for "human equivalence" that fell short. Even among those rare AI researchers who are still willing to discuss general cognition, the attitude appears to be: "First we'll achieve general cognition, then we'll talk human-equivalence. As for transhumanity, forget it."

In contrast, modern-day evolutionary theorists are strongly trained against Panglossian or anthropocentric views of evolution, i.e., those in which humanity occupies any special or best place in evolution. Here it is socially unacceptable to suggest that *Homo sapiens sapiens* represents cognition in an optimal or maximally developed form; in the field of evolutionary psychology, the overhanging past is one of Panglossian optimism. Rather than modeling the primate order and hominid family as evolving *toward* modern-day humanity, evolutionary psychologists try to model the hominid family as evolving somewhere, which then decided to call itself "humanity". (This view is beautifully explicated in Terrence Deacon's "The Symbolic Species" [Deacon97].) Looking back on the history of the hominid family and the human line, there is no reason to believe that evolution has hit a hard upper limit. *Homo sapiens* has existed for a short time by comparison with the immediately preceding species, *Homo erectus*. We look back on our evolutionary history from this vantage point, not because evolution stopped at this point, but because the subspecies *Homo sapiens sapiens* is the very first elaboration of primate cognition to cross over the minimum line that supports rapid cultural growth and the development of evolutionary psychologists. We observe human-level intelligence in our vicinity, not because human intelligence is optimal or because it represents a developmental limit, but because of the Anthropic Principle; we are the first intelligences smart enough to look around. Should basic design limits on intelligence exist, it would be an astonishing coincidence if they centered on the human level.

Strictly speaking, the attitudes of AI and evolutionary psychology are not irreconcilable. One could hold that achieving general cognition will be extremely hard and that this constitutes the immediate research challenge, while simultaneously holding that once AI is achieved, only ungrounded anthropocentrism would predict that AIs will develop to a human level and then stop. This hybrid position is the actual stance I have tried to maintain throughout this paper - for example, by decoupling discussion of developmental epochs and advantages of minds-in-general from the audacious question of whether AI can achieve a given epoch or advantage.

But it would be silly to pretend that the tremendous difficulty of achieving general cognition licenses us to sweep its enormous consequences under the rug. Despite AI's glacial slowness by comparison with more tractable research areas, Artificial Intelligence is still improving at an *enormously* faster rate than human intelligence. A human may contain millions or hundreds of millions of times as much processing power as a personal computer circa 2002, but computing power per dollar is (still) doubling every eighteen months, and human brainpower is not.

Many have speculated whether the development of human-equivalent AI, however and whenever it occurs, will be shortly followed by the development of transhuman AI [Moravec88]; [Vinge93]; [Minsky94]; [Kurzweil99]; [Hofstadter00]; [Hawking01]. Once AI exists it can develop in a number of different ways; for an AI to develop to the point of human-equivalence and then remain at the point of human-equivalence for an extended period would require that all liberties be simultaneously blocked[38] at exactly the level which happens to be occupied by *Homo sapiens sapiens*. This is too much coincidence. Again, we observe *Homo sapiens sapiens* intelligence in our vicinity, not because *Homo sapiens sapiens* represents a basic limit, but because *Homo sapiens sapiens* is the very first hominid subspecies to cross the minimum line that permits the development of evolutionary psychologists.

Even if this were not the case - if, for example, we were now looking back on an unusually long period of stagnation for *Homo sapiens* - it would still be an unlicensed conclusion that the fundamental design bounds which hold for *evolution* acting on *neurons* would hold for *programmers* acting on *transistors*. Given the different design methods and different hardware, it would again be too much of a coincidence.

This holds doubly true for seed AI. The behavior of a strongly self-improving process (a mind with access to its own source code) is not the same as the behavior of a weakly self-improving process (evolution improving humans, humans improving knowledge). The ladder question for recursive self-improvement - whether climbing one rung yields a vantage point from which enough opportunities are visible that they suffice to reach the next rung - means that effects need not be proportional to causes. The question is not how much of an effect any *given* improvement has, but rather how much of an effect the improvement plus further triggered improvements and *their* triggered improvements have. It is literally a domino effect - the universal metaphor for small causes with disproportionate results. Our instincts for system behaviors may be enough to give us an intuitive feel for the results of any single improvement, but in this case we are asking not about the fall of a single domino, but rather about how the dominos are arranged. We are asking whether the tipping of one domino is likely to result in an isolated fall, two isolated falls, a small handful of toppled dominos, or whether it will knock over the entire chain.

If I may be permitted to adopt the antipolarity of "conservatism" - i.e., asking how *soon* things could conceivably happen, rather than how late - then I must observe that we have *no idea* where the point of open-ended self-improvement is located, and furthermore, *no idea* how fast progress will occur after this point is reached. Lest we overestimate the total amount of intelligence required, it should be noted that nondeliberate evolution did eventually stumble across general intelligence; it just took a very long time. We do not know how much improvement over evolution's incremental steps is required for a strongly self-improving system to knock over dominos of sufficient size that each one triggers the next domino. Currently, I believe the best strategy for AI development is to try for general cognition as a necessary prerequisite of achieving the domino effect. But in theory, general cognition might not be required. Evolution managed without it. (In a sense this is disturbing, since,

while I can see how it would be theoretically possible to bootstrap from a nondeliberative core, I cannot think of a way to place such a nondeliberative system within the human moral frame of reference.)

It is conceptually possible that a basic bound rules out all improvement of effective intelligence past our current level, but we have no evidence supporting such a bound. I find it difficult to credit that a bound holding for minds in general on all physical substrates coincidentally limits intelligence to the exact level of the very first hominid subspecies to evolve to the point of developing computer scientists. I find it equally hard to credit bounds that limit strongly self-improving processes to the characteristic speed and behavior of weakly self-improving processes. "Human equivalence", commonly held up as the great unattainable challenge of AI, is a chimera - in the sense of being both a "mythical creature" and an "awkward hybrid". Infrahuman AI and transhuman AI are both plausible as self-consistent durable entities. Human-equivalent AI is not.

Given the tremendous architectural and substrate differences between humans and AIs, and the different expected cognitive advantages, there are no current grounds for depicting an AI that strikes an anthropomorphic balance of domain competencies. Given the difference between weakly recursive self-improvement and strongly recursive self-improvement; given the ladder effect and domino effect in self-enhancement; given the different limiting subjective rates of neurons and transistors; given the potential of minds-in-general to expand hardware; and given that evolutionary history provides no grounds for theorizing that the *Homo sapiens sapiens* intelligence range represents a special slow zone or limiting point with respect to the development of cognitive systems; therefore, there are no current grounds for expecting AI to spend an extended period in the *Homo sapiens sapiens* range of general intelligence. *Homo sapiens sapiens* is not the center of the cognitive universe; we are a noncentral special case.

Under standard folk psychology, whether a task is easy or hard or extremely hard does not change the default assumption that people undertaking a task do so because they expect positive consequences for success. AI researchers continue to try and move humanity closer to achieving AI. However near or distant that goal, AI's critics are licensed under folk psychology to conclude that these researchers believe AI to be desirable. AI's critics may legitimately ask for an immediate defense of this belief, whether AI is held to be five years away or fifty. Although the topic is not covered in this paper, I personally pursue general cognition as a means to seed AI, and seed AI as a means to transhuman AI, because I believe human civilization will benefit greatly from breaching the upper bounds on intelligence that have held for the last fifty thousand years, and furthermore, that we are rapidly heading toward the point where we *must* breach the current upper bounds on intelligence for human civilization to survive. I would not have written a paper on recursively self-improving minds if I believed that recursively self-improving minds were inherently a bad thing, whether I expected construction to take fifty years or fifty thousand.

Conclusion

> "People are curious about how things began, and especially about the origins of things they deem important. Besides satisfying such curiosity, accounts of origin may acquire broader theoretical or practical interest when they go beyond narrating historical accident, to impart insight into more enduring forces, tendencies, or sources from which the phenomena of interest more generally proceed. Accounts of evolutionary adaptation do this when they

> explain how and why a complex adaptation first arose over time, or how and why it has been conserved since then, in terms of selection on heritable variation. [...] In such cases, evolutionary accounts of origin may provide much of what early Greek thinkers sought in an *arche,* or origin - a unified understanding of something's original formation, source of continuing existence, and underlying principle."
>
> -- Leonard D. Katz, ed., "Evolutionary Origins of Morality" [Katz00]

On the cover of Douglas Hofstadter's *Gödel, Escher, Bach: An Eternal Golden Braid* are two trip-lets - wooden blocks carved so that three orthogonal spotlights shining through the 3D block cast three different 2D shadows - the letters "G", "E", "B". The trip-let is a metaphor for the way in which a deep underlying phenomenon can give rise to a number of different surface phenomena. It is a metaphor about intersecting constraints that give rise to a whole that is *deeper* than the sum of the requirements, the multiplicative and not additive sum. It is a metaphor for arriving at a solid core by asking what casts the shadows, and how the core can be stronger than the shadows by reason of its solidity. (In fact, the trip-let itself could stand as a metaphor for the different metaphors cast by the trip-let concept.)

In seeking the *arche* of intelligence, I have striven to neither overstate nor understate its elegance. The central shape of cognition is a messy 4D object that casts the thousand subfields of cognitive science as 3D shadows. Using the relative handful of fields with which I have some small acquaintance, I have tried to arrive at a central shape which is no more and no less coherent than we would expect of evolution as a designer.

I have used the levels of organization as structural support for the theory, but have tried to avoid turning the levels of organization into Aristotelian straitjackets - permitting discussion of "beliefs", cognitive content that combines the nature of concept structures and learned complexity; or discussion of "sequiturs", brainware adaptations whose function is best understood on the thought level. The levels of organization are visibly pregnant with evolvability and plead to be fit into specific accounts of human evolution - but this does not mean that our evolutionary history enacted a formal progress through Modalities, Concepts, and Thoughts, with each level finished and complete before moving on to the next. The levels of organization structure the functional decomposition of intelligence; they are not in themselves such a decomposition. Similarly, the levels of organization structure accounts of human evolution without being in themselves an account of evolution. We should not say that Thoughts evolved from Concepts; rather, we should consider a specific thought-level function and ask which specific concept-level functions are necessary and preadaptive for its evolution.

In building this theory, I have tried to avoid those psychological sources of error that I believe have given rise to past failures in AI; physics envy, Aristotelian straitjackets, magical analogies with human intelligence, and others too numerous to list. I have tried to give some explanation of past failures of AI, not just in terms of "*This* is the magic key we were missing all along (take two)", but in terms of "This is what the past researchers were looking at when they made the oversimplification, these are the psychological forces underlying the initial oversimplification and its subsequent social propagation, and this explains the functional consequences of the oversimplification in terms of the specific subsequent results as they appeared to a human observer." Or so I would *like* to say, but alas, I had no room in this paper for such a complete account. Nonetheless I have tried, not only to give an account of some of AI's past failures, but also to give an account of how successive failures

tried and failed to account for past failures. I have only discussed a few of the best-known and most-studied AI pathologies, such as the "symbol grounding problem" and "common-sense problem", but in doing so, I have tried to give accounts of their specific effects and specific origins.

Despite AI's repeated failures, and despite even AI's repeated failed attempts to dig itself out from under past failures, AI still has not dug itself in so deep that no possible new theory could dig itself out. If you show that a new theory does not contain a set of causes of failure in past theories - where the causes of failure include both surface scientific errors and underlying psychological errors, and these causes are together sufficient to account for observed pathologies - then this does not prove you have identified *all* the old causes of failure, or prove that the new theory will succeed, but it is sufficient to set the new approach aside from aversive reinforcement on past attempts. I can't promise that DGI will succeed - but I believe that even if DGI is slain, it won't be the AI dragon that slays it, but a new and different dragon. At the least I hope I have shown that, as a new approach, DGI-based seed AI is different enough to be worth trying.

As presented here, the theory of DGI has a great deal of potential for expansion. To put it less kindly, the present paper is far too short. The paper gives a descriptive rather than a constructive account of a functional decomposition of intelligence; the paper tries to show evolvability, but does not give a specific account of hominid evolution; the paper analyzes a few examples of past failures but does not fully reframe the history of AI. I particularly regret that the paper fails to give the amount of background explanation that is usually considered standard for interdisciplinary explanations. In assembling the pieces of the puzzle, I have not been able to explain each of the pieces for those unfamiliar with it. I have been forced to the opposite extreme. On more than one occasion I have compressed someone else's entire lifework into one sentence and a bibliographic reference, treating it as a jigsaw piece to be snapped in without further explanation.

At this point in the ritual progress of a general theory of cognition, there are two possible paths forward. One can embrace the test of fire in evolutionary psychology, cognitive psychology, and neuroscience, and try to show that the proposed new explanation is the most probable explanation for previously known evidence, and that it makes useful new predictions. Or, one can embrace the test of fire in Artificial Intelligence and try to build a mind. I intend to take the latter path as soon as my host organization finds funding, but this may subtract from the time available to mend the gaps in the present paper. Hopefully my efforts in this paper will serve to argue that DGI is promising enough to be worth the significant funding needed for the acid test of building AI.

In today's world it is commonly acknowledged that we have a responsibility to discuss the moral and ethical questions raised by our work. I would take this a step farther and say that we not only have a responsibility to discuss those questions, but also to arrive at interim answers and guide our actions based on those answers - still expecting future improvements to the ethical model, but also willing to take action based on the best current answers. Artificial Intelligence is too profound a matter for us to have no better reply to such pointed questions as "Why?" than "Because we can!" or "I've got to make a living somehow." If *Homo sapiens sapiens* is a noncentral and nonoptimal special case of intelligence, then a world full of nothing but *Homo sapiens sapiens* is not necessarily the happiest world we could live in. For the last fifty thousand years, we've been trying to solve the problems of the world with *Homo sapiens sapiens* intelligence. We've made a lot of progress, but there are also problems that we've hit and bounced. Maybe it's time to use a bigger hammer.

Acknowledgements

This paper would not have been written without the support and assistance of a large number of people whose names I unfortunately failed to accumulate in a single location. At the least I would like to thank Peter Voss, Ben Goertzel, and Carl Feynman for discussing some of the ideas found in this paper. Any minor blemishes remaining in this document are, of course, my fault. (Any major hideous errors or gaping logical flaws were probably smuggled in while I wasn't looking.) Without the Singularity Institute for Artificial Intelligence, this paper would not exist. To all the donors, supporters, and volunteers of the Singularity Institute, my deepest thanks, but we're not finished with you yet. We still need to build an AI, and for that to happen, we need a lot more of you.

I apologize to the horde of authors whom I have inevitably slighted by failing to credit them for originating an idea or argument inadvertently duplicated in this paper; the body of literature in cognitive science is too large for any one person to be personally familiar with more than an infinitesimal fraction. In particular, as I was editing a draft of this paper, I discovered the paper "Perceptual Symbol Systems" by Lawrence Barsalou [Barsalou99], presenting a model in which concepts reify perceptual experiences and bind to perceptual imagery, and in which combinatorial concept structures create complex depictive mental imagery. Barsalou should receive full credit for first publication of this idea, which is one of the major theoretical foundations of DGI.

Bibliography

Anderson,T.E., D.E. Culler, D.A. Patterson, and the NOW team. (1995.) A Case for NOW (networks of Workstations). IEEE Micro, 15(1), pp. 54-64.

Barsalou, L.W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577-609.

Becker, D. J., T. Sterling, D. Savarese, J. E. Dorband, U. A. Ranawak, and C. V. Packer. (1995.) Beowulf: A parallel workstation for scientific computation. Proceedings of the International Conference on Parallel Processing, pp. 11-14, 1995.

Berlin, B., and P. Kay. (1969). Basic Color Terms: Their Universality and Evolution. Berkeley: University of California.

Bickerton, Derek, (1990) Language and Species, Chicago: University of Chicago Press.

Bonmassar, G., and E. Schwartz. (1997). Space-variant Fourier analysis: The exponential chirp transform. IEEE Pattern Analysis and Machine Vision, vol. 19, pp. 1080-1089.

Boynton, Robert M. and Olson, Conrad X. (1987), "Locating basic colors in the OSA space", Color Research and Application, 12(2):94--105.

Brown, Roger. 1958. How Shall a Thing Be Called? *Psychological Review* 65:14-21.

Carey, S. (1992). Becoming a face expert. Philosophical Transactions of the Royal Society of London B 335:95-103.

Chalmers, D., French, R., & Hofstadter, D. (1992). High Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology. Journal of Experimental and Theoretical AI, 4 (3), 185-211.

Chalmers, David. (1995.) "Facing Up to the Problem of Consciousness." Journal of Consciousness Studies. 2, pp. 200--219.

Church, R. M., & Meck, W. H. (1984). The numerical attribute of stimuli. In H. L. Roitblat, T. G. Bever, & H. S. Terrace (Eds.), Animal cognition (pp. 445-464). Hillsdale, NJ: Erlbaum.

Clearfield, M. W., & Mix, K. S. (1999). Number versus contour length in infants' discrimination of small visual sets. Psychological Science, 10(5), 408-411.

Colton, S., A. Bundy, and T. Walsh. (2000.) On the notion of interestingness in automated mathematical discovery. International Journal of Human Computer Studies, 53(3):351-375.

Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty. *Cognition,* 58, 1-73.

Dawkins, R. (1996). Climbing Mount Improbable. New York: W.W. Norton and Co.

Deacon, T. (1990). Rethinking mammalian brain evolution. American Zoologist, 30, 629-705.

Deacon, Terrence. (1997). The Symbolic Species. 1997. Allen Lane: The Penguin Press: London.

Dedrick, D. (1998) Naming the rainbow: Colour language, colour science, and culture. Kluwer Academic Publishers.

Dehaene S. (1997), The Number Sense : How the Mind Creates Mathematics, Oxford University Press.

Dijkstra, E. W. 1968. Go To Statement Considered Harmful. Communications of the ACM, Vol. 11, No. 3, pp. 147-148.

Douglas B. Lenat. Eurisko: A program which learns new heuristics and domain concepts. Artificial Intelligence, 21, 1983.

Felleman, D. J. and D. C. van Essen. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex 1: 1-47.

Finke, R. A. & Schmidt, M. J. (1977). Orientation-specific color after-effects following imagination. Journal of Experimental Psychology: Human Perception and Performance, 3:599-606.

Flack, J. & F. de Wall. 2000. 'Any Animal Whatever': Darwinian Building Blocks of Morality in Monkeys and Apes. In Evolutionary Origins of Morality: Cross-Disciplinary Perspectives. Katz, L. (Ed.) Thorverton, UK: Imprint Academic.

Geist, A., A. Beguelin, J. J. Dongarra, W. Jiang, R. Manchek, and V. S. Sunderam. (1993.) PVM 3 User's Guide and Reference Manual. Technical Report ORNL/TM-12187, Oak Ridge National Laboratory, May 1993.

Gould, S.J. and R.C. Lewontin (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. In: Proceedings of the Royal Society of London 205. pp. 281-288

Gropp, W., E. Lusk, and A. Skjellum. (1994.) Using MPI: Portable Parallel Programming with the Message-Passing Interface. MIT Press.

Harnad, S. (1990). The Symbol Grounding Problem. Physica D 42: 335-346.

Hawking, Stephen. (2001.) Interview in *Focus Magazine*. Corrected English translation provided to KurzweilAI.net. URL: http://www.kurzweilai.net/news/frame.html?main=news_single.html?id=495

Hayhoe, M. M., Bensigner, D. G., and Ballard, D. H. (1998). Task constraints in visual working memory. Vision Research, 38(1)125-137.

Helmholtz, H. (1867). Treatise on Physiological Optics. Dover, New York 1962. First published 1867.

Hexmoor, H.; Lammens, J.; and Shapiro, S. C. (1993.) "Embodiment in GLAIR: a grounded layered architecture with integrated reasoning for autonomous agents." In Dankel II, D. D., and Stewman, J., eds., Proceedings of The Sixth Florida AI Research Symposium (FLAIRS 93). The Florida AI Research Society. 325-329.

Hochberg, J. E. (1957). Effects of the Gestalt revolution: the Cornell symposium on perception. Psychological Review, 64(2), 73-84.

Hofstadter, D. R., Gödel, Escher, Bach: an Eternal Golden Braid, NY: Basic Books, 1979.

Hofstadter, D.R., & Mitchell, M. (1988.) Conceptual slippage and analogy-making: A report on the copy-cat project. In Proceedings of the Tenth Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Hofstadter, D.R., with the Fluid Analogies Research Group. (1995.) Fluid Concepts and Creative Analogies. Basic Books.

Hofstadter, Douglas, R. (1985). Variations on a theme as the crux of creativity. In Metamagical Themas, pp. 232-259. New York: Basic Books.

Hofstadter, Douglas. (2000.) Moderator, Stanford Symposium on Spiritual Robots. April 2000, Stanford University, Palo Alto, CA. URL: http://www.stanford.edu/dept/symbol/Hofstadter-event.html

Holland, P.W.H., Ingham, P.W. and Krauss, S. (1992) Development and Evolution: Mice and flies head to head. Nature 358, 627-628.

Hopfield, J. J. and Tank, D. (1985). "Neural" computation of decisions in optimization problems. Biological Cybernetics, 52:141-152.

Hurford, J. (1999.) The Evolution of Language and Languages. In Robin Dunbar, Chris Knight and Camilla Power, editors, The Evolution of Culture. Edinburgh University Press.

Hurvich, L. M. and D. Jameson. (1957.) An opponent-process theory of color vision. Psychological Review, 64, 384-390.

Hwang, K. and Z. Xu. (1998.) Scalable Parallel Computing. McGraw-Hill.

James, W. (1911.) "The Meaning of Truth." New York: Longman Green and Co. 217-220.

Judd, D.B., D.L. MacAdam, and G. Wyszecki. (1964). Spectral Distribution of Typical Daylight as a Function of Correlated Color Temperature. Journal of the Optical Society of America, 54:1031-1040, August 1964.

Katz, L. D. (2000.) "Toward Good and Evil: Evolutionary Approaches to Aspects of Human Morality." In Evolutionary Origins of Morality: Cross-Disciplinary Perspectives. Katz, L. (Ed.) Thorverton, UK: Imprint Academic.

Koch, C. (1999.) Biophysics of Computation. Oxford Univ. Press, New York.

Koch, C. and Segev, I. (2000). The role of single neurons in information processing. Nature Neuroscience, 3(Supp):1160-1211.

Koch, C., Poggio, T. (1992) Multiplying with synapses and neurons. In T. McKenna, J.L. Davis and S.F. Zornetzer (Eds.), Single Neuron Computation, Academic Press, Cambridge Masachussetts, pp. 315-345.

Koestler, A., "The Ghost in the Machine". Hutchinson & Co, London, 1967.

Koffka, K. (1935). Principles of Gestalt Psychology. New York: Harcourt Brace.

Kosslyn, Alpert, Thompson, Maljkovic, Weise, Chabris, Hamilton, Rauch, & Buonanno (1993). Visual mental imagery activates topographically organized visual cortex: PET investigations. Journal of Cognitive Neuroscience, 5, 263-287.

Kosslyn, S. M. (1994). Image and Brain: The Resolution of the Imagery Debate. Cambridge, MA: MIT Press.

Kumar, V. (1992.) Algorithms for constraint-satisfaction problems: a survey. AI Magazine, 13:32-44.

Kurzweil, Ray. (1999.) The Age of Spiritual Machines: When Computers Exceed Human Intelligence. Viking Press, New York

Lakoff, G. & M. Johnson. (1999.) Philosophy In The Flesh: The Embodied Mind and its Challenge to Western Thought. New York: Basic Books.

Lakoff, G. (1987.) Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. Chicago: University of Chicago Press.

Lavie, N. and Driver, J. (1996). On the spatial extent of attention in object-based visual selection. Perception and Psychophysics, 58:1238–1251.

Lenat, D. B. (1983.) EURISKO: A program that learns new heuristics and domain concepts. Artificial Intelligence 21, pp. 61-98.

Lenat, D. B., Prakash, M., & Shepherd, M. (1986). CYC: Using Commonsense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks. AI Magazine, 6, 65-85.

Lenat, D. B., and Brown, J. S. 1984. Why AM and EURISKO appear to work. Artificial Intelligence 23(3):269--294.

Lowe, D.G. (1985). Perceptual Organization and Visual Recognition. Kluwer Academic Publishers, Massachusetts.

Maloney, L. T. (1986), Evaluation of linear models of surface spectral reflectance with small numbers of parameters. Journal of the Optical Society of America A, 3, 1673-1683.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco: W.H. Freeman and Company.

McDermott, D. 1976. Artificial Intelligence Meets Natural Stupidity. SIGART Newsletter 57.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. Journal of Experimental Psychology: Animal Behavior Processes, 9, 320-334.

Mervis, Carolyn B.; Catlin, Jack; Rosch, Eleanor. Development of the Structure of Color Categories. Developmental Psychology 11. 1 (1975 Jan): 54-60.

Meyer, D. E. & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words. Journal of Experimental Psychology, 90, 227-234.

Minsky, M. (1994.) Will robots inherit the earth? Scientific American, 271(4):109-113.

Mitchell, M. (1993.) Analogy-making as perception: A computer model. Cambridge, MASS: MIT press.

Moore, C. M., Yantis, S., and Vaughan, B. (1998). Object-based visual selection: Evidence from perceptual completion. Psychological Science, 9:104–110.

Moore, G.E. (1997.) "An Update on Moore's Law." Intel Developer Forum Keynote: San Francisco.

Moravec, H. (1998.) "When will computer hardware match the human brain?" Journal of Evolution and Technology, Vol 1.

Moravec, Hans. (1988.) Mind Children: The Future of Robot and Human Intellegence. Harvard University Press, 1988.

Newell, A. (1980.) Physical Symbol Systems. Cognitive Science, 4, 135-183.

Newell, A., and H. A. Simon. (1963). GPS, a program that simulates human thought. In E. A. Fiegenbaum and J. Feldman, Eds., Computers and Thought. New York: McGraw Hill, pp. 279-293.

Palmer, S. and Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. Psychonomic Bulletin and Review, 1:29–55.

Pearl, J. (1996). Causation, action, and counterfactuals. In Y. Shoham, Ed., Theoretical Aspects of Rationality and Knowledge: Proceedings of the Sixth Conference. San Francisco: Morgan Kaufmann, pp. 51-73.

Pearl, J. (2000.) Causality: Models, Reasoning, and Inference. Cambridge Univ. Press.

Pylyshyn, Z.W. (1981). The imagery debate: Analogue media versus tacit knowledge. Psychological Review, 88, 16-45.

Raymond, E. S. (ed). 2001b: "Wrong Thing." The on-line hacker Jargon File, version 4.3.1, 29 Jun 2001. http://www.tuxedo.org/~esr/jargon/html/entry/Wrong-Thing.html

Raymond, E. S. (ed.) 2001a: "Uninteresting". The on-line hacker Jargon File, version 4.3.1, 29 Jun 2001. URL: http://www.tuxedo.org/~esr/jargon/html/entry/uninteresting.html

Rehling, J. and Hofstadter, D.R. (1997.) The Parallel Terraced Scan: An Optimization for an Agent-Oriented Architecture. Proceedings of the IEEE International Conference on Intelligent Processing Systems, 1997. Beijing, China.

Ritchie, G.D. and F.K. Hanna (1984). AM: A Case Study in AI Methodology. Artificial Intelligence, 23.

Rodman, H. R. (1999). Face Recognition. In Wilson, R. and Keil, F. (eds.), The MIT Encyclopedia of the Cognitive Sciences. Cambridge, MA: MIT Press.

Rojer, A., and E. L. Schwartz. (1990). Design considerations for a space-variant visual sensor with complex-logarithmic geometry. In 10th International Conference on Pattern Recognition, vol. 2. pp. 278-285.

Rosch, E. (1978). Principles of Categorization, in Rosch, E. and Lloyd, B. B. (Eds.), Cognition and Categorization, Hillsdale NJ: Lawrence Erlbaum.

Rosch, E., Mervis, C., Gray, W., Johnson, D. and Boyes-Braem, P., "Basic Objects in Natural Categories," Cognitive Psychology, Vol. 8, 1976, pp. 382-439.

Sandberg, Anders. (1999). The Physics of Information Processing Superobjects. Journal of Evolution and Technology, 5.

Schwartz, E. L. (1977). Spatial mapping in primate sensory projection: Analytic structure and relevance to perception. Biological Cybernetics 25:181-194.

Schwartz, E. L., A. Munsif, and T. D. Albright. (1989). The topographic map of macaque V1 measured via 3D computer reconstruction of serial sections, numerical flattening of cortex, and conformal image modeling. Investigative Opthalmol. Supplement, p. 298.

Shepard, R. (1992). The perceptual organization of colors. In J. Barkow, L. Cosmides, and J. Tooby, Eds., The Adapted Mind. Oxford: Oxford University Press.

Sherman, S. M., & Koch, C. (1986). The control of retinogeniculate transmission in the mammalian lateral geniculate nucleus. Experimental Brain Research, 63, 1-20.

Shipley, E.F. and B. Shepperson 1990: Countable entities: Developmental changes. Cognition 34: 109-136.

Sober, E. (1984) The nature of selection. MIT Press.

Spelke, E.S. (1990.) Principles of object perception. Cognitive Science, 14(1):29-56.

Tooby, J. and Cosmides, L. (1992). *The psychological foundations of culture.* In J. Barkow, L. Cosmides, and J. Tooby (Eds.), *The Adapted Mind: Evolutionary psychology and the generation of culture.* New York: Oxford University Press.

Tootell, R. B., M. S. Silverman, E. Switkes, and R. deValois. (1985). Deoxyglucose, retinotopic mapping and the complex log model in striate cortex. *Science* 227: 1066.

Tversky, A. and D. K. Koehler, "Support Theory: A Nonexistential Representation of Subjective Probability," Psychological Review, 101 (1994), 547-567.

Tversky, A. and D. Kahneman. (1974.) Judgement under uncertainty: Heuristics and biases. Science, 185:1124-1131.

Vinge, Vernor. (1993.) "Technological Singularity." VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, March, 1993. URL: http://www.frc.ri.cmu.edu/~hpm/book98/com.ch1/vinge.singularity.html

Voss, Peter. (2001.) Presentation at Extro 5 (fifth convention of the Extropy Institute). San Jose, California: June 2001.

Wagner, G.P. and Altenberg, L. (1996). Complex adaptations and the evolution of evolvability. Evolution, 50, 967-976.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt, II. Psychologische Forschung 4:301-350. Condensed translation published as Laws of organization in perceptual forms, in W. D. Ellis (1938), A Sourcebook of Gestalt Psychology. New York: Harcourt, Brace, pp. 71 - 88.

Winograd, T. (1972) Understanding Natural Language. Edinburgh, Edinburgh University Press.

Worden, R. (1995). A speed limit for evolution. Journal of Theoretical Biology, 176, pp. 137-152.

Wulf, W. and S. McKee. Hitting the memory wall: Implications of the obvious. Computer Architecture News, 23(1), 1995.

Yudkowsky, E. (2001). Creating Friendly AI. Publication of the Singularity Institute: http://singinst.org/CFAI/

Zemel, R. S., Behrmann, M., Mozer, M. C., & Bavelier, D. (2002). Experience-dependent perceptual grouping and object-based attention. Journal of Experimental Psychology: Human Perception and Performance, Vol. 28, No. 1, 202–217.

Endnotes

- 1: This does not rule out the possibility of discoveries in cognitive science occurring through less intentional and more evolutionary means. For example, a commercial AI project with a wide range of customers might begin with a shallow central architecture loosely integrating domain-specific functionality across a wide variety of tasks, but later find that their research tends to produce specialized internal functionality hinting at a deeper, more integrated supersystem architecture.
- 2: Adenine triphosphate, the standard unit of currency in the economy of the human metabolism.
- 3: I cannot think of any plausible way to do this, and do not advocate such an approach.
- 4: A phrase due to [Dijkstra68] in "Go To Statement Considered Harmful"; today it indicates that a prevalent practice has more penalties than benefits and should be discarded.
- 5: Note that "lightbulb" is a *basic-level category* [Brown58]. "Basic-level" categories tend to lie on the highest level at which category members have similar shapes, the highest level at which a single mental image can reflect the entire category, the highest level at which a person uses similar motor actions for interacting with category members, et cetera [Rosch76]. "Chair" is a basic-level category but "furniture" is not; "red" is a basic-level category but "scarlet" is not. Basic-level categories generally have short, compact names, are among the first terms learned within a language, and are the easiest to process cognitively. [Lakoff87] cautions against inadvertant generalization from basic-level categories to categories in general, noting that most researchers, in trying to think of examples of categories, almost always select examples of basic-level categories.

- 6: I don't know of a specific case of priming tests conducted on the specific word-pair "lightbulb" and "fluorescent", but this is a typical example.
- 7: "DGI" stands for "Deliberative General Intelligence", the theory of mind presented in this paper.
- 8: Finke and Schmidt showed that afterimages from mental imagery can recreate the McCullough effect. The McCullough effect is a striking illustration of the selective fatiguing of higher-level feature detectors, in which, following the presentation of alternating green horizontal and red vertical bars, differently colored afterimages are perceived in the white space of a background image depending on whether the background image has horizontal black-and-white bars (red afterimage) or vertical black-and-white bars (green afterimage). This is an unusual and counterintuitive visual effect, and not one that a typical study volunteer would know about and subconsciously "fake" (as Pylyshyn contends).
- 9: The lateral geniculate nucleus is a thalamic body which implements an intermediate stage in visual processing between the retina and the visual cortex.
- 10: "Interesting" is here used in its idiomatic sense of "extremely hard".
- 11: The levels begin with "atoms" rather than "quarks" or "molecules" because the atomic level is the highest layer selected from a bounded set of possible elements (ions and isotopes notwithstanding). "Quarks" are omitted from the list of layers because no adaptive complexity is involved; evolution exercises no control over how quarks come together to form atoms.
- 12: Other human modalities include, e.g., proprioception and vestibular coordination.
- 13: Environmental complexity of this type is reliably present and is thus "known in advance" to the genetic specification, and in some sense can be said to be a constant and reliable part of the genetic design.
- 14: The term "brainware" is not necessarily anthropomorphic, since the term "brain" can be extended to refer to nonbiological minds. The biology-only equivalent is often half-jokingly referred to as *wetware,* but the term "wetware" should denote the human equivalent of the code level, since only neurons and synapses are actually wet.
- 15: The statement that each neuron is "potentially" within one clock tick of any other neuron is meant as a statement about the genome, not a statement about developmental neurology - that is, it would probably require a genetic change to produce a previously forbidden connection.
- 16: Note that biological neurons can easily implement multiplication as well as addition and subtraction [Koch92], plus low- and band-pass filtering, normalization, gain control, saturation, amplification, thresholding, and coincidence detection [Koch99].
- 17: The striate cortex is also known as "primary visual cortex", "area 17", and "V1".
- 18: Deliberative General Intelligence, the theory of mind presented in this paper.
- 19: I say "human-like" and not "primate-like" or "mammal-like" because of the possibility that the human visual modality has further adaptations that support the use of mental imagery in deliberation.
- 20: Artificial lighting, which has an "unnatural" spectral power distribution (one that is not the weighted sum of the natural basis vectors), can cause objects to appear as a different color to the human visual system. Hence the manufacture and sale of "natural lighting" or "full spectrum" light sources.
- 21: For one suggested solution, see [Bonmassar97].
- 22: This does not imply that GOFAI handles concept-concept relations *correctly.* The links in a classical "semantic net" are as oversimplified as the nodes.

- 23: The use of computationally inexpensive cues to determine when more expensive checks should be performed.
- 24: An algorithm which reduces complex representations to a form that can be more easily compared or scanned.
- 25: Rather than comparing against each potential match in turn, an algorithm would be used which eliminates half the potential matches by asking a question, then eliminates half the remaining potential matches by asking a new question pre-optimized against that set, and so on until the remaining potential matches are computationally tractable. Branched sorting of this kind could conceivably be implemented by spatial properties of a parallel neural network as well.
- 26: There is some indication that young humans possess a tendency to count discrete physical objects and that this indeed interferes with the ability of human children to count groups of groups or count abstract properties [Shipley90].
- 27: In animals, experiments with cross-modality numeracy sometimes exhibit surprisingly positive results. For example, rats trained to press lever A on hearing two tones *or* seeing two flashes, and to press lever B on hearing four tones *or* seeing four flashes, spontaneously press lever B on hearing two tones *and* seeing two flashes [Church84]. This may indicate that rats categorize on (approximate) quantities by categorizing on an internal accumulator which is cross-modality. Evolution, however, tends to write much smoother code than human programmers; I am speaking now of the likely consequence of a "naive" AI programmer setting out to create a numeron-detector feature.
- 28: EURISKO was a self-modifying AI that used heuristics to modify heuristics, including modification of the heuristics modifying the heuristics.
- 29: As described earlier, "holonic" describes the simultaneous application of reductionism and holism, in which a single quality is simultaneously a combination of parts and a part of a greater whole.
- 30: Whether a belief is *really* more like a concept or more like a thought is a "wrong question". The specific similarities and differences say all there is to say. The levels of organization are aids to understanding, not Aristotelian straitjackets.
- 31: "This sentence is false" is properly known as the Eubulides Paradox rather than the Epimenides Paradox, but "Epimenides Paradox" seems to have become the standard term.
- 32: Of course, writing quality is made up of a number of components and is not a true scalar variable. A more accurate description would be that "writing quality" is the summation of a number of other percepts, and that we conceive of this summated quality as increasing or decreasing. Some writing qualities may be definitely less than or greater than others, but this does not imply that the complete set of percepts is well-ordered or that the percept itself is cognitively implemented by a simple scalar magnitude.
- 33: As opposed to the Right Thing. See the Jargon File entry for "Wrong Thing", [Raymond01b].
- 34: I believe this is the underlying distinction which [Pearl96] is attempting to model when he suggests that agent actions be represented as surgery on a causal graph.
- 35: Viewing *evolution itself* through the lens provided by DGI is just *barely* possible. There are so many differences as to render the comparison one of "loose analogy" rather than "special case". This is as expected; evolution is not intelligent, although it may sometimes appear so.
- 36: Former world champion in chess, beaten by the computer Deep Blue.
- 37: It is sometimes objected that an intelligence modifying itself is "circular" and therefore impossible. This strikes me as a complete *non sequitur,* but even if it were not, the objection is still based on the idea of intelligence as an opaque monolithic function. The character of

the computational subsystems making up intelligence is fundamentally different from the character of the high-level intelligence that exists atop the subsystems. High-level intelligence can wrap around to make improvements to the subsystems in their role as computational processes without ever *directly* confronting the allegedly sterile problem of "improving itself" - though as said, I see nothing sterile about this.

- 38: This is a metaphor from the game Go, where you capture an opponent's group of stones by eliminating all adjoining clear spaces, which are known as "liberties".