

Faster Than You Think

Artificial General Intelligence Will Soon Change Everything

PRELIMINARY DRAFT—
NOT FOR DISTRIBUTION

Ben Goertzel
2013

All contents © Ben Goertzel 2013

Foreword.....	5
Who I Am and Why I Wrote This Book	6
Who I am.....	6
Why I wrote this book.....	6
Some Basic Questions	11
A Rough Definition of Artificial General Intelligence.....	15
PART ONE.....	16
The Coming Technological Singularity	17
The Logic of Exponential Growth	21
Critical Technologies are Advancing Exponentially	24
The Exponential Advancement of AI	29
The Possibility of “Hard Takeoff”	30
The Transformative Power of Intelligence	33
AGI Will Transform Every Aspect of Human Endeavor	33
Aerospace.....	34
Agriculture, Food & Water	35
Automobiles	36
Chemicals.....	39
Computers & Software	40
Construction.....	45
Defense & Intelligence	47
Energy	50
Entertainment & Arts	51
Finance & Insurance	54
Hospitality.....	57
Manufacturing.....	58
News media.....	59
Pharmaceuticals & Health Care	60
Space	66
Telecommunications	68
AGI Will Vastly Transcend Humanity	71
The Psychological Singularity	72
The Illusion of Free Will.....	77
Now is the Time for AGI.....	81
Advances in Computer Hardware, Software, and Cognitive Science.....	81
The Absurd Underfunding of Singularity-Enabling Technologies.....	84
AGI Sputnik	88
PART TWO.....	91
What is Artificial General Intelligence, Really?.....	92
AI vs. AGI vs. SCADS	93
The Meaning of “AI” has Drifted	94
The Origins of the Term “Artificial General Intelligence”	95
Advanced AGIs won’t really be “Artificial”	98
Real-world Intelligence can’t truly be “General”	99

AGI and Narrow AI Deliver Different Kinds of Value	100
Intelligence Itself is a Somewhat Limiting Concept.....	102
How Minds Work	105
How I Think the Brain Works	105
Why Neuroscience is Not the Best Guide to AGI	116
The Insights of Cognitive Science	129
Dividing the Mind Into Parts	131
Why Is It All So Damn Complicated?	140
Motivation and Action Selection	145
Emotion.....	148
Deliberative Processing	154
Perception, Action & Language Hierarchies	160
Language.....	163
Cognitive Synergy	166
Mind as a Complex System	167
A Strategy for Building Minds	170
AGI and Human Childhood Development	171
Embodiment.....	179
Embodiment & Environment.....	193
What Would a General Theory of General Intelligence Look Like?.....	194
How to Proceed?	195
The OpenCog Project	197
OpenCog's AtomSpace.....	204
Truth and Attention Values.....	206
Mixing Neural and Symbolic.....	209
A Big Scary Diagram.....	214
Interfacing Mind and World	220
OpenCog's Cognitive Processes	221
Probabilistic Reasoning	221
The Consistent Pursuit of Goals.....	223
The Limitations of Logic – And Everything Else.....	226
A Fiendishly Common Conceptual Mistake.....	227
Concept Blending.....	228
Evolving Procedural Knowledge	231
Procedure Learning in OpenCog	233
The Mind's Eye.....	238
Deciding What to Pay Attention to	239
Attention in the Brain.....	240
OpenCog's "Economic" Attention Allocation.....	241
Embodying OpenCog.....	241
Making the Robot Talk	244
From Here to AGI.....	247
Toward the AGI Robot Sputnik.....	251
PART THREE	253
AGI Can Help Achieve Radical Human Longevity	254
Abolishing the Plague of Involuntary Death	258
Biology Has Become an Information Science.....	260

AGI and Longevity	261
Quick Review of Basic Genetics	263
Biomind's bioinformatic AI	270
Aubrey de Grey's SENS Approach to Increasing Human Lifespan.....	281
Cryonics as a Backup Plan.....	289
Why So Little Focus on Longevity?	290
PART FOUR	292
The Risks and Rewards of AGI.....	293
Will There Be Cyborgs?	297
The Possibility of Femtotech Superintelligences.....	301
Would Femtotech Superminds Bother to Exterminate Humans?.....	305
The Global Brain.....	306
The Risks and Rewards of Advanced AGI	310
Does Humanity Need an AGI Nanny?.....	311
Why AGI?.....	314
Transcending the Discontents of Civilization.....	317
The Cosmist Perspective.....	320
Shaping the Singularity.....	325
FURTHER READING	327
A Few Relevant Websites.....	328
A Few Relevant Books	329
Some Online Background Reading	329
Acknowledgements	330

Foreword

by To Be Determined

(Insert foreword here)

Who I Am and Why I Wrote This Book

Who I am

I am a human being, male, born on Planet Earth in 1966.

- I am a scientist and entrepreneur, currently occupied with various applications of artificial intelligence technology, and most deeply with the quest to create thinking machines with greater than human intelligence...
- I am an avid dreamer about wild future possibilities that science and technology have the potential to make real.
- I am interested – passionately interested -- in traveling far beyond the Earth; and living far beyond the date of 2050 or so when I would be expected to die according to today's prevailing views; and exploring modes of life and experience far beyond those possible in my current human form.

Why I wrote this book

I want to tell you why I think somebody (maybe my colleagues and me, maybe someone else) is going to create very powerful artificial intelligence, sometime this century, maybe even this decade, and what implications I think this is going to have for life, the universe and everything.

...

When I was a young child in the early 1970s, Artificial Intelligence was basically the stuff of science fiction novels. Yeah, there were some professors doing AI research in universities, but it wasn't something you heard about much, and the achievements of their AI software programs and robots weren't very interesting. There certainly weren't any major AI companies occupying news headlines.

Today, in 2013 as I write these words, things look pretty different. Google, one of the most successful companies around, is basically an AI company; relying on AI text-processing tricks to place ads on web pages in such a way that people will be more likely to click on them. Search engines like Google and Bing are not branded as AI, but that's what they are. Google is also, just for fun, using AI technology to create self-driving cars that go around San Francisco. Apple's smartphones now include an AI chatbot, Siri, which admittedly falls far short of human conversational ability, but does resolve

some useful queries correctly. AI has helped Apple sell a lot of phones. Google's smartphone operating system Android has some pretty good "AI" voice control too.

Computer programs now do most of the trading on the US stock exchange, and among these are many programs using techniques taught in university AI courses, such as machine learning and natural language processing. The US military relies on AI not only for flashy applications like unmanned automated vehicles -- but also for dull, necessary stuff like planning and logistics, and supply chain management. PayPal uses AI for automated fraud detection, enabling them to provide affordable credit card processing across the Internet. Video games regularly feature AIs with intelligence customized to their game worlds; and AI players can beat the best human players at chess, checkers and many other games (though not yet poker or Go)!



Figure 1: According to Arthur C. Clarke's novel 2001 and the influential Stanley Kubrick film of the same name, we were supposed to have human-level AI's like HAL 9000 by now. Not quite. (If you didn't read the book or movie, you should. HAL was a mentally unbalanced AI controlling a spaceship, not always in accordance with human instructions.) Though, given HAL 9000's defective ethics, it's just as well he didn't come about. What we do have now, in 2013, is a host of highly useful narrow AI applications – and the technology and science base to let us finally seriously approach the AGI problem, and with a bit more work create AGI systems that are both smarter than HAL 9000 and more ethical.

We don't yet have R2D2, C3PO or Hal 9000, let alone the massively superhuman AIs that some of the more philosophical science fiction writers, Stanislaw Lem among others, envisioned. However, given the progress made in specialized AI technologies like I've mentioned above, along with the simultaneous progress that's been made in understanding the human brain and mind, and the massive ongoing improvement in computer hardware and software, the advent of AI software and robots with real human-like general intelligence (and more) seems a lot less like total pie-in-the-sky science fiction these days. Some fairly mainstream figures, like Justin Rattner, the CTO of Intel, are joining an increasing chorus of futurists predicting the likely advent of superhumanly intelligent AI by the middle of the century.

As a child, I was very interested in AI, but just as one of a host of really cool, speculative, futuristic technologies: AI, robots, time travel, immortality pills, psychic powers, spaceships, teleporters, and the like. I didn't see much prospect of any of these getting created on Earth during my lifetime, so I mused about creating a starship that would travel away from Earth at near light speed, exploiting the wonders of relativistic physics to enable me to return to Earth just a couple years older, but finding the Earth a million years in the future. After a million years or so, I figured, a lot of cool technologies would have been developed, and a lot of amazing new life possibilities would be open to me.

During the mid-1990s, when I was in my late 20s, I took a hard look at all the fantastic sciences and technologies that fascinated me, and decided that AI was the one I should spend the most time pursuing. I had received a PhD in math in 1989, at age 22, but even as I was working on that degree, I knew I was more passionate about applying math to understand the world and make amazing things happen, than about pure math for its own sake. I calculated that one could possibly build a time machine by spinning a football-shaped star fast enough, but this seemed to involve some significant engineering difficulties. Similarly, immortality pills and human brain modification or genetic engineering involved cumbersome human experimentation; nanotechnology or spaceships required very expensive hardware to set up. Dramatically world-changing AI, it seemed to me, could plausibly be achieved by just typing lines of computer code into plain ordinary computers. The trick was just to know which lines of code to type, to enable one's program to be intelligent! It seemed fairly likely to me that someone would figure this out during my lifetime, and I figured that someone might as well be me.

Since that decision to focus on AI nearly 20 years ago, I've been spending a lot of my time on the quest

for advanced, autonomous AI systems: or in a phrase, AGI, Artificial General Intelligence, as I've come to call it. I'm now helping to lead an open source software project, called OpenCog, which is aimed at creating an AGI with the rough intelligence of a human toddler, and then moving on from there to grander general intelligence. I've also devoted a lot of time organizing AGI-interested researchers into something resembling a coherent community by putting together annual AGI conferences, editing some books of technical AGI research papers, and founding a group called "The AGI Society." My friend Hugo de Garis likes to embarrass me by calling me the "father of AGI." That's certainly an overstatement, but it's true that I've done quite a lot to advance the level of appreciation for AGI and its importance, both in the general research community as well as in the popular culture.

These days, while the average person probably thinks about human-level AIs as something science fictional, there is an increasing international community of futurist pundits, visionaries, and scientists, who believe that AIs will get smarter than people during our lifetimes. The famous inventor Ray Kurzweil, whom I know moderately well, has spent a recent chunk of his career fleshing out and publicizing his forecast that the world will experience a "technological Singularity" around 2045 or so. At the time of the Singularity, as Ray conceives it, computers will be massively superior to humans in intelligence, and the humans of that time will be dramatically enhanced in various ways, and probably closely linked to these superintelligent computers. Though not that many people believe Ray's precise timing estimate for the Singularity, the general idea underlying his predictions – that many kinds of critical technology advancement are progressing exponentially – is increasingly widely accepted.

This book is far from the first thing I've written about AGI – but up till now, most of my writing on the topic has been other researchers. I've also spoken a lot about AGI and the Singularity and related ideas, for popular audiences as well as technical ones. My goal in putting together this book was to get down in writing the sorts of things I habitually say in my talks about AGI and the future. When I give a conference talk on AGI and the future, I skip over most of the technical details, and I'm going to do the same here. I have striven for accuracy in these pages, but have often opted to skip over various nitty-gritty issues in the interest of comprehensibility. My aim here is to get across, to the reader who may not have much background in science and technology, some of my thoughts about AGI technology and the amazing consequences it may have for humanity over the next few decades.

Among my core messages here is that, as the title of the book suggests, *AGIs that can think faster, more broadly, and better than you, may get here faster than you think*. It's an astounding, thrilling

and, in some ways, scary proposition. There's still plenty of speculation involved in thinking about such matters, but not nearly as much as there was a decade ago, let alone 4+ decades ago when I started musing about the topic as a child. Science and technology are advancing fast, and while AGI (unlike narrow, application specific AI) has not advanced all that fast until recently, there's reason to believe the next few decades are going to be different. It's going to be quite a ride!

And, critically, it's not going to be a ride on which we're just passengers, but a ride on which we're drivers (albeit, to strain the metaphor a bit, drivers going through unknown territory, driving a machine we don't understand all that well). AGI is not something that's going to appear out of nowhere and foist itself upon us; it is something that we're going to create. Something that we are already in the process of creating. Humanity has been moving toward the creation of AGI for a long time now – arguably, *at least*, ever since the emergence of language enabled the development of organized agriculture, which led to civilization and everything that's come with that. The advent of powerful AGI is all but inevitable at this point, barring a catastrophe that wipes us all out. But the particular form that the first powerful AGIs will take seems fairly wide open. Whether or not the first AGIs are benevolent to humans, for example, seems like something we have the capability to influence, via our actions right now.

Some people argue that, because it's hard to guarantee that the AGIs we create will be beneficial to us, we should hold off on creating AGI till we somehow figure out to do it in a risk-free way. There's a group in California – formerly called the Singularity Institute for AI (SIAI), recently rebranded the Machine Intelligence Research Institute (MIRI) -- devoted to this perspective. However, I very much doubt a risk-free approach is going to be possible. Reality tends not to work that way. Others, like my friend and fellow AGI researcher Hugo de Garis, argue that AGIs are bound to exterminate us inferior humans, but we should create them anyway, because it's our destiny. My own view is that we should try our best to create an empathic, loving, beneficial AGI fairly quickly, before someone else creates a nasty one.

It's an exciting time we live in, a time when such issues are the stuff of real life, rather than just science fiction!

Some Basic Questions

I get asked to do interviews a lot, for media outlets great and small, and some of the questions (and journalists) are smarter than others. A few years ago, I did an interview for a science fiction blog site, with a writer named Jason Peffley, and I felt he did a particularly crisp job of posing the basic questions that everybody seems to ask me about AGI and the Singularity. So I will get the party started here with a lightly edited version of my chat with Jason:

Jason

You've been in the AI industry for quite a while now. Did you always know that you wanted to work with AI? Was there something in your life that triggered it?

Ben

I grew up on Science Fiction, and I always knew I wanted to spend my life making Science Fiction things become real. But, at the start, I wasn't sure if my focus would be AI, time machines, quantum gravity computers, interstellar spacecraft, genetically engineered creatures, psi powers, or what not. I chose to focus on AI. Initially in my late teens, and later, with greater emphasis, starting in my late 20s because it seemed, logistically, the easiest of the bunch. If I'm right, you don't need any special hardware, just the right computer code, and a big server farm of commodity PCs. The more I thought about the problem, the more I felt I knew how to make it happen.

Jason

If you were talking to 10 year olds, how would you sum up the Technological Singularity in a way they'd understand?

Ben

Within a few decades, we're probably going to have computers and robots way smarter than people in every way, inventing new stuff constantly and transforming the world in ways mere humans can't even imagine. Hopefully, we'll be able to upgrade our brains or turn ourselves into robots, and become super-smart, super-powerful beings ourselves. We don't have this technology now, but it's coming pretty soon, because new inventions keep getting made faster and faster. Because the more inventions you have, the faster you can make new inventions, since the old inventions are tools you can use to help make new ones.

Jason

What would a robot, or program for that matter, need to do in order for you to kick your feet up and say "I finally did it?" Teach itself how to tie a pair of shoes? Get your dry cleaning? Pick up your kids from school? Teach you string theory while building a fusion reactor?

Ben

World peace. Immortality for all who want it. Everybody liberated from needing to work for a living. Starships traversing the galaxy, and brain implants or robot bodies for everyone who wants to become a superhuman AI.

Thing is, I think all of that has a decent chance of happening within a couple decades after we create the first AGI that's as smart as, say, an average science professor. From there, I think it won't be such a big leap.

The big leap, I think, is from where we are now, to having an AI that's like a human toddler. When someone gets an AI at that level, I'll be incredibly excited, and also a bit scared.

Jason

Will we see the advanced R2D2 and C3P0 type of AI in our lifetime?

Ben

Definitely, and even better. I'm not sure about C3PO's six million forms of communication though, as there may not be that many useful alien races around. Also, the real-life R2D2 might go to the Radio Shack and buy himself a speech synthesizer. But in concept; yeah, we will get there.

Jason

In an interview you did on Singularity 101, you said that you had thought a lot about how you would spend 100 billion dollars. Outside of buying a private island, how would you spend it?

Ben

If I had 100 billion dollars, I'd start a semi-private city, devoted specifically to the beneficial development of advanced technologies, with residency offered to folks displaying qualification and passion for said technology development.

Otherwise, I'd start a massive research grant funding project for AGI, life extension, nanotechnology, femtotechnology, mind uploading, and so forth. If these areas got attention in the way that cancer research and semiconductor manufacturing do, progress would be way further along.

Jason

Outside of funding, what is the current biggest hurdle in your field?

Ben

Finding people who are really good at programming that also have a deep understanding of cognitive science and AGI theory is always a challenge.

Conceptually, I think the biggest challenge facing the field is the integration of algorithms for abstract cognition with algorithms for low-level perception and action. Right now, we have good approaches for all of these, but nobody has made them all work together in a unified way. I think the OpenCog architecture solves that problem, but we haven't definitively demonstrated that yet. We're well on our way, though!

Jason

What do you say to the people who feel this will turn into a colossal fuck up like Cyberdyne's Skynet division¹?

Ben

Setting aside the "naked guys traveling back in time" aspect, I think the possibility of advanced AIs running amok and killing everyone is a real one. It can't be ruled out anyway. However, I think there are actions we can take to minimize the probability of this happening, by building our AGI systems with rational minds and benevolence-oriented goal systems, and raising our young AGIs with kindness, and teaching them well.

Like any other advanced technology, the potential benefits are huge and so are the risks, but that's how humanity has been rolling for a long time. Nobody could stop the development of advanced AI and robotics now, even if they wanted to. At least not without destroying all of civilization. We're on the

¹For the three readers who don't know, this is a reference to the *Terminator* film series, in which a computer network security system accidentally achieves superintelligence and takes over the world, wreaking massive destruction and creating evil time-traveling robots with hunky biceps.

verge of the next step in the evolution of intelligence on Earth, so we may as well embrace and enjoy it, and try to nudge it in a positive direction insofar as we can.

A Rough Definition of Artificial General Intelligence

I’ve already referred a few times to the notion of AGI or Artificial General Intelligence. Everyone has heard of AI, but AGI is a less familiar concept. What does it mean?

This is actually a fairly subtle question. I, and other scientists, have published a variety of papers proposing and debating different mathematical and empirical measures of general intelligence. But most things get complicated if you think about them hard enough. The basic concept of AGI is not at all hard to grasp or articulate.

Roughly and qualitatively, what I mean by the term “Artificial General Intelligence,” or “AGI” is pretty clear and simple. I mean a system—a computer program, robot, or machine— that displays the same rough sort of general intelligence as humans. General intelligence is intelligence not tied to a very specific set of tasks, which possesses the ability to take a broad view, or generalize what has been learned.

But what does “Artificial General Intelligence” really imply, in practice? When is it going to happen? How is it going to happen? What is it going to be like? Why should you care about it? What upsides should you look forward to; what downsides should you fear?

I don’t claim to have all the answers. But it’s my hope that by the end of this book, after exploring the concepts shaping AGI’s past, present, and future, and my own contributions to the field, you’ll have a better understanding of these questions, along with a richer, expanded sense of the concept of AGI and all its possibilities.

PART ONE

THE AGI SPUTNIK IS NEAR

The Coming Technological Singularity



Figure 2: Screenshot of inventor and futurist Ray Kurzweil, from the documentary film Transcendent Man. I also appeared in the film, for about 5 minutes. Check it out – it's available on iTunes, for example. <http://transcendentman.com>

What is this “Technological Singularity” that Ray Kurzweil, and other techno-futurists have made so much noise about recently?



Figure 3: SF writer and professor Vernor Vinge introduced the term “Singularity” into futurism in the 1980s, but has done fairly little Singularity publicity, preferring to focus his career on science fiction writing now that he has retired from academia. <http://http://vimeo.com/54718577>

There are actually multiple visions of the Singularity concept—Ray’s differs slightly from mine, and both Ray and I have slightly different visions from science fiction writer Vernor Vinge, who was the first to use the term “Singularity” in this context. The first person to outline a closely Singularity-like idea in any depth seems to have been mathematician I.J. Good, who in 1965 wrote about the potential for an “intelligence explosion” when AIs became smarter than people, and noted that “the first intelligent machine is the last invention man will ever make.”

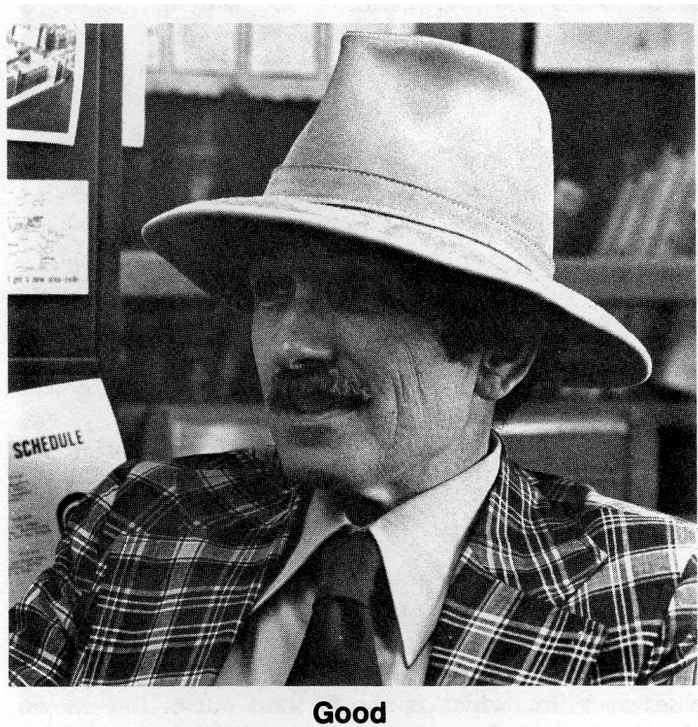


Figure 4: Mathematician I.J. Good wrote about the “intelligence explosion” in 1965. His notion was largely the same as Vinge’s and Kurzweil’s “Singularity”, but he was a bit too early and so his idea didn’t really catch on. <http://www.bigear.org/CSMO/Images/CS06/cs06p13l.jpg>

But details aside, at the core of these multiple visions is a common understanding: *The Singularity means rapid technological change accelerating so fast, the human mind can’t keep up, leading to dramatic new sciences and inventions, and revolutionizing all aspects of life as we know it, including AGI with capability far beyond the human level.*

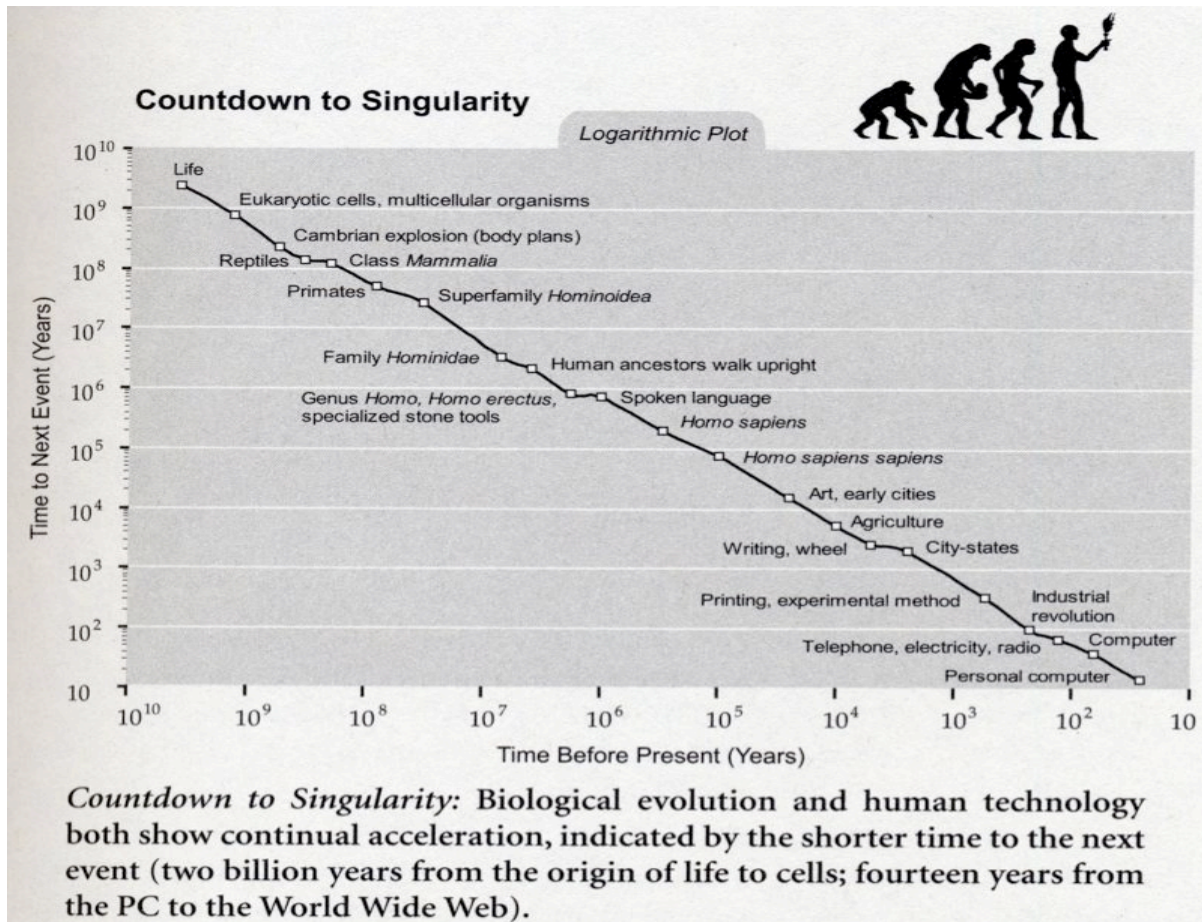
In his 2005 book *The Singularity Is Near*, Kurzweil estimates the Singularity beginning around 2045; Vinge’s earlier, less rigorous estimate involved roughly the same timeline. I think this rough timeline basically makes sense, though concerted effort may speed up the process. And there’s the possibility of a significant delay if too many unforeseen political or engineering obstacles should emerge.

Most of what I have to say about AGI isn’t actually dependent on the concept of the Singularity. AGIs will still be AGIs with basically the same properties, no matter whether they emerge super-rapidly via a Singularity-type event, or more gradually. However, the Singularity means AGI may emerge, within decades or even years from now, as opposed to centuries or millennia – a projected timeline which certainly adds a bit of practical everyday-life zing to the AGI concept.

Crudely put, the Singularity idea has two components:

- 1) Amazing stuff is going to happen
- 2) Amazing stuff is going to happen pretty suddenly, likely in a matter of decades.

Just the first part is pretty exciting on its own, but might not capture many people's immediate interest or attention, since most folks tend to preoccupy themselves with things that are going to happen in their lifetimes, or at least their children's lifetimes. But it's even more exciting given the plausibility of a technological Singularity bringing us AGI this century.



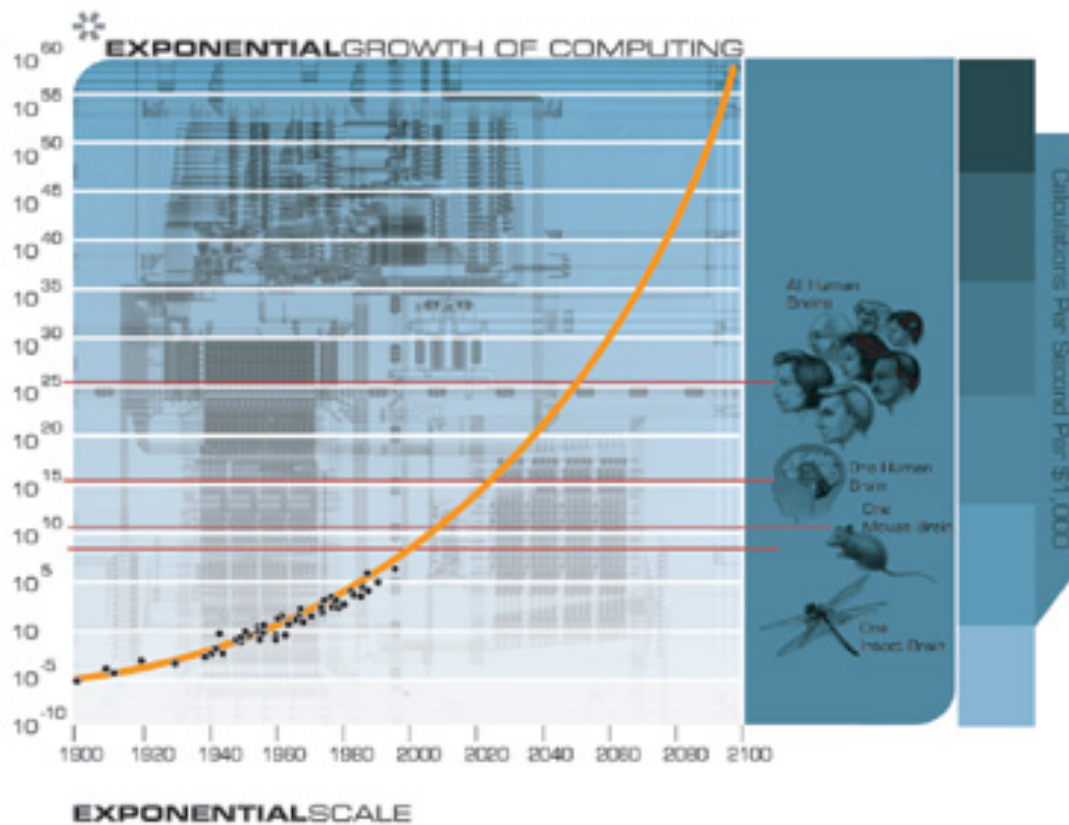


Figure 5: Illustrations of the exponential growth leading up to the Singularity, from Ray Kurzweil's presentations on the topic. <http://kurzweilai.net>

The Logic of Exponential Growth

Over the course of both distant and recent human history, science and technology progress has accelerated dramatically. These days, technology, science and even art advance so fast I can't keep up.



Figure 6: *I'm only 46 years old, but I remember when devices like the iPhone in this picture did not exist – and when the average person considered them a science fictional technology, probably hundreds or more years away. Now I need to buy a new one every couple years or folks will laugh at me for being behind the times. And I see these in the remotest places, such as the hills of rural Ethiopia. The rural Ethiopian owners of smartphones probably understand their advanced functions better than I do, as I rarely have the time to master my old phone before technology advanced and I need to buy a new one. Exponential advance of technology in action.*

http://www.elearning-africa.com/picturevoting/images_winner_2011/large/1402.jpg

Take mobile phones, for example — I don't have time to figure out all the features on my phone before a new one comes out. It's the same with computer software. Or music. New and potentially interesting genres of music are constantly emerging, and in many cases, before I'll get a chance to appreciate them, they will be replaced by successors building new things upon their artistic and cultural achievements.

Contemporary science journals document a multitude of amazing, diverse discoveries in fields such as biology, physics and engineering. Browsing through the current periodicals sections of a major university's science libraries is an astounding experience. One scientist can't follow everything going on in his or her field – let alone keep up with neighboring fields and related ideas. A few hundred years ago, a diligent scientist did not have this problem – there were not that many scientists, there was way less science going on, and information about new discoveries could easily take years or decades to propagate.

Consider apoptosis, preprogrammed cell death—one of the many complex aspects of our bodies’ aging process in which cells are preprogrammed to die. When I searched for this term in *PubMed*, the biomedical community’s online repository for research papers, back in 2005, I found over 60,000 papers touching on the topic. I got the feeling maybe ten thousand or so of these were reasonably important. When I searched the term in PubMed again in 2013, I found nearly 250,000 papers. Even on a narrow topic like this, it’s extremely hard to track all the new knowledge.

What’s driving all this progress, as Kurzweil and others have recognized, is the logic of ***exponential growth***. Populations of animals or plants tend to grow exponentially before running out of space or food, or reaching the carrying capacity of their environment; populations of ideas and inventions also tend to grow exponentially. But the ecosystem of ideas seems limitless—our ideas and inventions keep on multiplying wildly.

Exponential growth has been most discussed and best documented in technology areas, most notably computer hardware. The well known “Moore’s Law” states, roughly speaking, that computers will double in speed every 18 months; this has held true for decades now. Ray Kurzweil has charted similar trends in other technology domains. My personal computer today is dramatically more powerful than the machine I owned in the 1990s. Heck, my phone now has massively more computing power than my 90s-era desktop computer ever had.

And of course, exponential technological growth did not begin with computers. Some technologies have been advancing exponentially for a long time. If you look at the grand sweep of human history, the amount of change from a million years ago to ten thousand years ago arguably was less than the amount of change from ten thousand years ago to today. In all probability, this pattern will keep on rolling. Exponential advances in multiple coupled technologies, leading to exponentially rapid changes in human life on a practical level, will have far-reaching effects on the states of mind we experience and the structure of language – which will then lead to yet further exponential change.

“But,” a skeptic might argue, “not everything accelerates exponentially. Over my lifetime, a lot of important things like refrigerators, socks and cars have remained about the same now as they were when I was a kid. The way we fall in love hasn’t changed that much. Toilets haven’t changed that much. (Although they do have funky programmed toilets in Japan with computerized controls hard for a foreigner to figure out.)”

Even if this skeptical argument were right, it wouldn’t really matter very much. Once you have AGIs

much smarter than human beings, you end up in a Singularity state, whether or not you have exponentially better socks. And if you still care about exponentially better socks after the Singularity, your AGI friends should be able to create them.

But actually, it's not so clear the skeptical argument, about the scattershot nature of exponential advancement, IS right. One can sensibly argue the point either way. We have, in fact, seen radical advancement even in the more prosaic aspects of practical technology. Cars, fridges and even socks have all changed considerably since the 60s. Fridges are more energy – efficient and reliable now, no longer incorporating harmful chemicals like CFCs. Cars utilize catalytic converters and are far better designed and safer (think airbags, the spread of antilock braking and new lightweight construction materials) than they were in my childhood. Socks are much cheaper (since they're woven by machines rather than by hand; and these machines are in countries far from where I live, as enabled by better logistics technology) and their new, improved synthetics magic away sweat and maintain good air circulation. The Singularity is more dependent on advancement in AGI, computing hardware and neuroscience than socks and forks, but change and improvement are all around through parallels in cultural and technological processes.

Critical Technologies are Advancing Exponentially

The exponential advance of computer hardware draws attention since it's easy to measure. But software has also advanced tremendously during the same timeframe. A game like *World of Warcraft* is light years ahead of Pac-Man. The current versions of MATLAB or Mathematica, software used by scientists and engineers, are tremendously more powerful than the versions that were used in the 1990s, or even a decade ago.

As a programmer, I've been impacted by the changes in programming libraries and techniques over the last couple of decades. Two key parts of computer programs are data structures (storing information in various specialized ways) and algorithms (manipulating information in specialized ways, creating new information from old). I used to have to write my own data structures and algorithms; now there are standardized libraries of data structures and algorithms. The complex software that underlies a game like *World of Warcraft* or a package like *MATLAB* couldn't have been written two decades ago (and certainly not so quickly), even with the necessary hardware.

Our ability to understand the biology of the human genome and the brain is also advancing

exponentially. This provides obvious help to those AGI designs that are based on emulating the human brain (which my current AGI designs are not). But even non-brain-based AGI approaches can find inspiration from the brain about how the *mind* works.



Figure 7: This fairly unassuming-looking machine is a microarrayer, which measures the expression levels of all the genes in a genome, based on the sample placed inside it. I have done a lot of work using narrow AI (machine learning) tools to analyze the complex data produced by microarrays, to understand what the expression levels of various genes have to teach us about the underpinnings of aging and various diseases. I'll explain more about this later in the book!
<http://ieg.ou.edu/equipment.htm>



Figure 8: MedImmune, a bioprocessing production facility for the production of mammalian cell cultures. The practical machinery of modern biology has accumulated gradually over the last decades, and rates as one of the great achievements of our time. And yet, with all this sophistication, we have barely scratched the surface in terms of understanding the complexity of biological systems, especially as pertains to complex processes like aging.

http://www.manufacturingchemist.com/technical/article_page/Bioprocessing_designs_on_a_grand_scale/60705

Since 2001, in parallel with my work on AGI, I've been working on applications of narrow AI technology in a variety of areas, one being the analysis genetic data. In this domain, I have noticed that the quality of data available and the variety of different biological experiments I can run have both increased dramatically in the last decade. Microarrays, measuring gene expression (the amount of biological activity associated with a particular kind of gene in a particular tissue of an organism at a particular point in time), were introduced in the mid-1990s, but were horribly error-prone, even in 2001. Now, they're relatively reliable.

As another example – RNA interference, a biological process via which RNA is used to inhibit the expression of particular genes, was discovered as a natural phenomenon in plants in the early 1990s, and first thoroughly understood in 1998. By now, it has long been commoditized and commonplace as a revolutionary tool for carrying out genetics experiments, and delivering gene therapies.

The cost of sequencing genomes is decreasing exponentially, which promotes the development of

personalized genomic medicine, encourages the development of more new technologies like RNAi, and makes more experiments feasible (including some for understanding human longevity). And remember, the human genome was only sequenced in the mid-1990s. Conceptually, the degree of advancement in biological science since 1990 has been incalculable; and according to various reasonable quantifications it can easily be seen to fit the exponential-advance model.

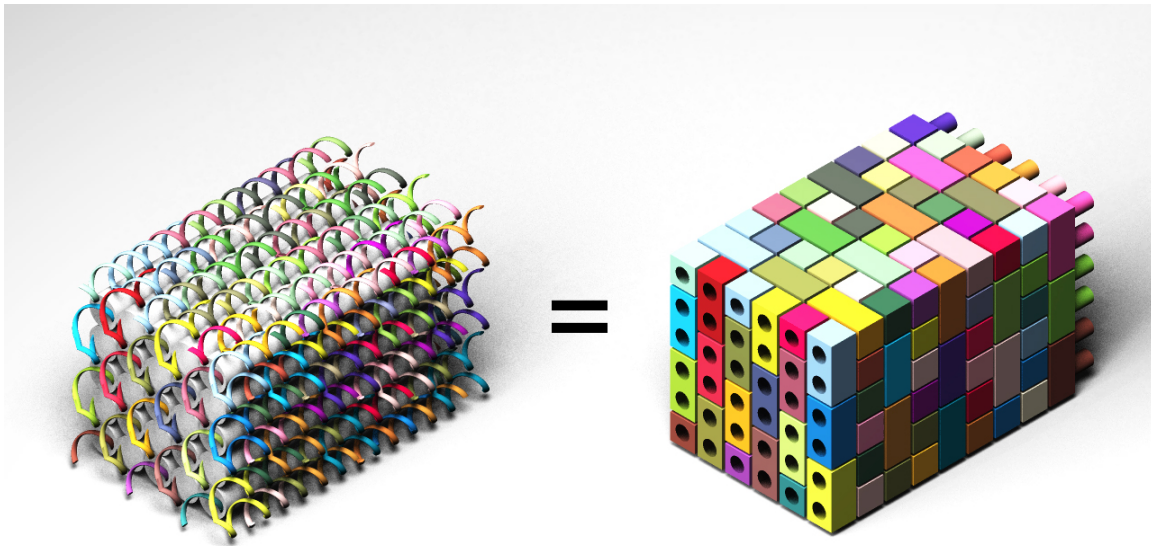


Figure 9: DNA origami, a recently developed technique that allows one to build block-like structures out of DNA. In general, there is no physical reason why we can't manipulate microscopic structures much like we do with Lego blocks. The only obstacle in our way is the limited nature of our current engineering practice. Physicist Richard Feynman saw this in the 1950s with his essay "There's Plenty of Room at the Bottom". Eric Drexler fleshed the idea out in the 1980s, and today nanotech is a flourishing R&D field, albeit still in its infancy.

<http://www.sciencemag.org/site/multimedia/slideshows/1227268/index.html>

In the 1980's Eric Drexler first wrote about *nanotechnology* – the engineering of novel machines out of molecules -- yet most scientists disregarded it as fiction. Now, it's taught in universities around the world and a major area of research. We don't yet have the kind of nanotechnology that Drexler envisioned when he was coining the term and founding the field— we don't have nanomachines capable of building others. But we do have impressive, new nano-materials with rich and varied applications. All in just a few decades. My friend Hugo de Garis and I have even been speculating about going a step further and building *femtotech*— *creating* computers and new kinds of matter by manipulating nuclear particles.

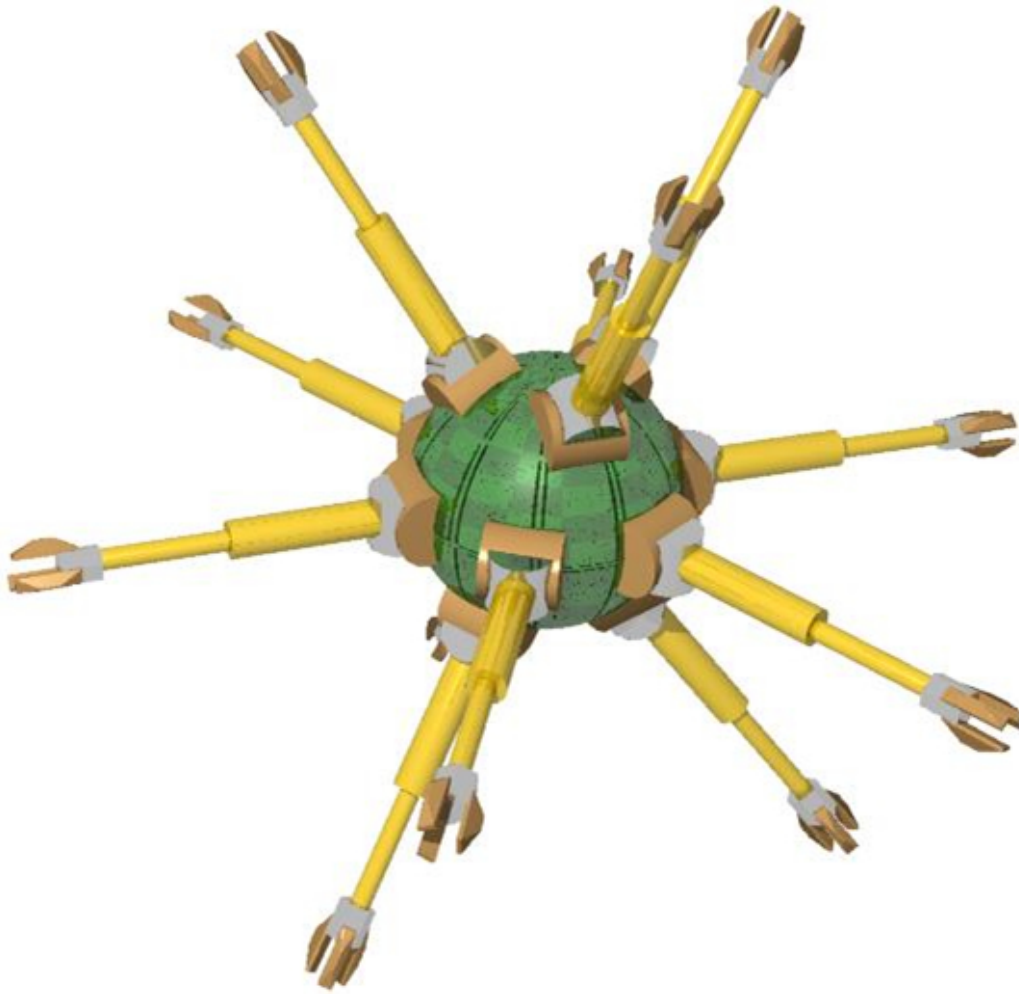


Figure 10: Utility fog, a term coined by AI researcher Josh Hall in 1993, refers to a collection of microscopic robots that can form complex structures. Imagine a bunch of little bots like this one, each invisibly small, communicating wirelessly and linking together to form a variety of gaseous, liquid and solid structures. One minute a couch, the next minute a robot, a curtain or a computer terminal. http://library.thinkquest.org/07aug/02162/utility_fog.html

Quantum computing sounded like science fiction a couple decades ago. When David Deutsch first wrote about *quantum computing* in the 80s, it was just a bunch of equations, and its practicality remained unclear. Now, *D-Wave*, a commercial firm headed by my friend Geordie Rose, is building real quantum computers for research purposes. In the next few decades, quantum computing may mature into a usable technology with wide practical applications.

Why do some technologies develop exponentially, while others stagnate, proceed linearly, or grow more slowly? The answer, in general, is complex. But part of the answer clearly lies in feedback phenomena. Some technologies are obviously self-accelerating — the better they get, the more you can use them to produce the next generation of existing technology.

Computer software and hardware both fall into this “self-accelerating” category. Better computers are

used, among other purposes, to run complex computer-aided design software to build ever more powerful computer chips. New programming libraries are used to develop new computer programs, which then spur the development of newer programming libraries and languages.

Machine tools and factories are self-accelerating, too. Good tools can build yet better tools, which can build yet better tools, and so forth. Part of Drexler's original vision of nanotechnology expressed this idea: building small machines to build even smaller ones. Nanotechnology may eventually manifest this phenomenon.

Ideally, human intelligence enhancement will follow the self-acceleration path. Making people smarter by modifying the brain, through drugs or other techniques, may yield greater brainpower to think up further improvements. Communication technology has developed in this fashion: Technology allows people to communicate and share ideas better than ever before. Sharing these ideas amongst groups will spur all sorts of amazing innovations, expanding on what's already out there. Positive reinforcement is an intrinsic feature of many technologies, so the better they get, the better their capacity for self-improvement; and humans are enablers, facilitating technology's own dynamic of progressive self-improvement.

AGI has the potential to be the ultimate self-accelerator. Once an AGI knows how to program software, it can program AGI programs, creating a new, smarter AGI, which will program a smarter AGI. Back in the 1950s, the mathematician labeled this phenomenon "intelligence explosion." As he put it: ***"The first intelligent computer is the last invention humanity will ever make."***

The Exponential Advancement of AI

It may sound strange to say that AI is advancing exponentially. After all, where are Hal 9000, R2D2 and C-3PO? Why are we still working in jobs every day, instead of spending our time entertaining ourselves while robots do all the work?

Indeed, from my perspective as an AGI researcher, the pace of progress in AI can sometimes seem frustratingly slow. I would like to see much faster progress toward AGI systems with general intelligence at the human level and beyond. I think we're not progressing faster toward superhuman AGI because of largely psychological and financial reasons, rather than scientific ones.

But science fiction and the impatience of visionaries, are not necessarily the best ways to judge the

progress of a science and engineering field. If one looks at the bigger picture, at the breadth of human history, the advancement of AI over the last half century starts to seem pretty damn impressive. Google Search and Google Now, Bing, Deep Blue, Siri, Nuance speech recognition, expert system medical diagnosticians, machine learning systems analyzing data from every area of science and industry, automated program traders, military drones, AI players in thousands of games,... None of this existed in the middle of the last century. Wow.

Further, the advancement of supporting technologies capable of enabling ongoing AI and AGI progress (computer hardware and software, specialized narrow AI, biotech, nanotechnology, and quantum computing) has been tremendous.

The Possibility of “Hard Takeoff”

The notion of AGIs programming better AGIs is a powerful one – and key to I.J. Good’s 1950s prediction of an “intelligence explosion.” More recently the phrase “hard takeoff” has been used in this context. A “hard takeoff” means a Singularity that happens really, really fast. Maybe in a five minute interval!

In a hypothetical hard takeoff, an AI program becomes smarter so fast that it reaches superhuman status in a very short time span, long before its human creators have had time to consider the implications.

An extreme example — At 8 AM, it may be as smart as a dog. At 9 AM, as smart as a human. At 10 AM, comparable to Albert Einstein. And at 11 AM, it has attained some kind of digital godhood.

To my mind, a hard takeoff of this nature seems possible, but doesn't feel particularly likely. But even if things don't turn out quite this extreme, still, AI may advance from human to superhuman intelligence in a very short duration by our own historical standards. Going from dog to Einstein level in a year may not qualify as a hard takeoff, but it would still be rather impressive.

The rapidity of the first AGI's cognitive development seems likely to depend on the overall technological context at the time this first AGI comes about.

If the first AGI capable of radical, useful self-reprogramming comes about in a world with very sophisticated supporting technologies, it might be able to accelerate its intelligence extremely rapidly. But if the first AGI capable of radical, useful self-reprogramming comes about in a world like the one we have today, I doubt its advancement to massively superhuman intelligence will take place in the

blink of an eye.

Let's pursue a little thought experiment, for a moment. As I write these words now, it's early 2013 and I'm sitting in my study in my house in Tai Mei Tuk, in the New Territories of Hong Kong. Let's say, hypothetically, that next week I have a huge breakthrough and figure out how to make the OpenCog software system as smart as a typical human college student, without going through as much work as previously envisioned. So, let's say that next week I create a human-level AGI – running on some Amazon cloud computing machines, launched from the Macbook Pro I'm typing these words on. I just spin up a few thousand Amazon EC2 instances, launch some OpenCog software on them, and let the distributed network start thinking. Let's say this OpenCog network is smart enough to hold a basic English conversation, and even read a math book and write some simple computer code...

The question I want to explore just now is: How could this hypothetical AGI become superhuman at a hard takeoff pace? Maybe it could rewrite its own source code to get smarter. But what if, it wanted better hardware, to benefit from its new, improved software? If it wanted to design new hardware to fabricate a new kind of chip, the AGI would have to order the relevant parts and materials from a manufacturer. There's an enormous amount of infrastructure to deal with, mostly involving the human world. There would be many chances for people to intervene in the AGI's activities.

Perhaps this hypothetical AGI could trick, convince or coerce others to do its bidding. However, its greater intelligence would not guarantee success. I'm much smarter than my dog, but that doesn't mean he always obeys my commands, even once he is able to understand them.

The point of the thought experiment is: Hard takeoff is unlikely given the present overall state of technology, even if someone creates a functional AGI tomorrow.

On the other hand, it's a lot more plausible in further future scenarios.

Consider another thought experiment. Imagine the world half a century from now. Other technologies have advanced a lot, but AGI still lags behind. Suppose it's 2060, and we still haven't invented AGI, but we have found success in nanotechnology, advanced robotics and ubiquitous Internet communications that connect every electronic device in the world. If you want a new appliance, you input the information online, or speak to it, and microphones embedded in the walls of your house pick it up. The details are transmitted to the correct recipient, and the appliance is fabricated in some nanofactory and dropped at your door by automated, unmanned drone helicopters. So long as the extant

technology can cope with the manufacturing specs, you get what you want, quickly and conveniently.

Now, insert a human-level AGI into this scenario. That AGI would have a lot more options for applying its intellectual powers than one created in 2013 or 2015. The newly created 2060 AGI could reprogram its software code and order and receive new hardware from external sources. Using the far more powerful, globally interconnected Internet of things, it could summon drones to fulfill its wishes.

The more other technologies advance, the more scope they provide for a human-level AGI to accelerate itself into something superhuman. Because of this, in a sense, it seems the safest path for AGI development may be to develop AGIs as quickly as possible, bringing them into a world where we can exercise a modicum of control and guide their future development. The quicker we develop human-level AI, the slower, all else being equal, the transition from human-level to superhuman AI will be because those other technologies it needs will still be in their infancy, developing independently. Even if work on AGI were to come to a screeching halt tomorrow, progress on complementary technologies like networking, nanofabrication, and robotics, to name a few, would still be steaming ahead; and these would make it easier for an AGI created a few decades from now to launch a hard takeoff. So, if one is worried about the potential impacts of powerful AGI, delaying the creation of AGI is not necessarily the safest choice. It might be just the opposite.

My best guess of how things will unfold is somewhere in the middle, at what you might call a “semi-hard takeoff.” I imagine that once we achieve toddler-level AGI, the money required for developing full-on adult-level AGI will come pouring in. AGI funding will phase-transition from trickle straight to firehose. Early-stage human-level AGI will lead to advances in other areas like nanotechnology, genetic engineering, human brain modification, and so forth.

Once these early AGIs get out there on global networks, people will seek a connection with them, creating a worldwide cyborg mind or nascent Global Brain. Out of this stew of experimentation, complementary development and novelty seeking, superhuman AGI will emerge.

The Transformative Power of Intelligence

Humans are not the strongest, fastest, longest-lived or most elegant organisms on the planet – but in some important senses we are the smartest. Our creation of tools for manufacture and communication make humanity as a whole far smarter than any individual human. Comparing the Macbook I’m typing this on, to the cave walls on which my ancestors painted not so long ago, one clearly sees the power of intelligence to transform the world.

But humans are far from the maximally intelligent possible beings. Soon we will use our intelligence to create beings far more intelligent than us, which will transform the world far more rapidly and dramatically than we could ever imagine. Along the way to humanly inconceivable modes of thinking, experiencing, building and interacting, every kind of industry currently known to us will be revolutionized.

AGI Will Transform Every Aspect of Human Endeavor

It’s hard to imagine a domain of human endeavor that an advanced AGI won’t be able to radically up-end. To give you a quick sense of this, in a series of inset boxes in the following pages, I’m going to run through all the major industries on the planet today, in alphabetical order, and briefly reflect on how superintelligent machines could improve them. Of course the details of my speculations are not all that likely to be exactly how things pan out. My aim is just to give you a flavor of what sorts of possibilities await.

Aerospace



Figure 11: "Project Zero": An all-electric tilt rotor aircraft from AgustaWestland, an Anglo-Italian helicopter company. Yes, this is a real airplane today, not a UFO! What will the aircraft and spacecraft designed by AGIs look like?
<http://www.agustawestland.com/node/6902>

We already have computer programs flying our commercial jets – the pilots are mostly just there to reassure the passengers, and to cover the off chance of some disaster that a human expert can better handle than a machine. But why are we still flying around in boring old airplanes? It is physically quite possible to make engines vastly more efficient than current jet engines, or flying machines that can take off like a helicopter, then fly fast like a jet. Or fly way faster than the sound barrier without the fuel-inefficiency of the Concorde. The first few AGI aerospace engineers may make our current designs look as archaic as the Wright Brothers' biplanes. Modern military aircraft already embody all sorts of amazing innovations, but most of these are not cost effective to incorporate in commercial aircraft. AGI engineering innovations could change this rapidly.

Humans study aerodynamics quite crudely by watching what things do in a wind tunnel, or solving equations in MATLAB or another computer program. An AGI could connect to sensors in the wind tunnel, the way we connect to our eyes and ears; and using MATLAB and other equation-solving software, find a solution for a nonlinear partial differential equation as immediately and automatically

as we solve $1+1=2$.

Agriculture, Food & Water



Figure 12: Genetically modified food is proving one of the more controversial recent technological innovations – but is spreading rapidly through the world nonetheless. The risk factor in GMO foods comes essentially from our inability to understand all the possible side-effects of a given genetic modification. If we better understood the human body (which advanced AGI would enable), then any risks associated with GMO food would be minimized, as we could more thoroughly understand the effects of genetic modifications in food on the human body.

<http://monconstitutionalist.files.wordpress.com/2012/09/gmo.jpg>

The creation of new and better forms of food is hard for us humans currently, given our limited capability to manipulate matter and understand its properties. But to an AGI more at home in the microscopic world of cells and chemicals, this may well be a piece of cake. Designing new forms of amazingly delicious and unsurpassedly nutritious food is “simply” a question of better understanding the human body and manipulating various plants and animals genetically. Even without molecular nanotechnology (enabling the repeated, inexpensive synthesis of an optimally tasty and nutritious hamburger), once AGI has decoded the mechanisms underlying biological organisms, the optimization of human food engineering should be a quite manageable application.

An AGI mind, with human-level general intelligence but a more science-friendly architecture, will rapidly surpass human understanding of biology. No human scientist can fit all the existing online biology databases into their mind, but an AGI could, allowing it to find scientifically meaningful patterns eluding human science. Connecting this AGI to robotic lab equipment will complete the cycle of experimentation, analysis and theory, allowing bioscience to progress with humans out of the loop.

Genetically modified plants and animals are important now -- but there's only so far we can go with our current understanding of genetics. If an AGI biologist figured out the connection between an animal or plant's genes and its ultimate structure or function, the genetic engineering possibilities would be endless. Further, the understanding of the human body that AGI will bring, will make it possible to understand the potential side-effects of any genetic modification to food, thus minimizing the risks associated with GMO food.

And what about water? Well, if there were an inexpensive way to synthesize water from hydrogen and oxygen, or even to desalinate seawater, we would no longer have concerns over water supply. These are things we know how to do already; they're just expensive. A systematic effort by a team of AGI scientists with IQ even slightly above human level, and direct access to relevant lab equipment, could very likely resolve these problems.

Automobiles

During the last few years, the concept of autonomous, self-driving cars has transitioned from being widely considered science fictional, to being generally considered common-sensical. In fact, cars capable of fairly autonomous highway driving were around in the early 1990s, courtesy of Ernst Dickmanns and his colleagues, but these were not widely known. These days Google's experimental

self-driving cars are legendary, and numerous big carmakers have begun their own self-driving car projects. Legislation is in place in various jurisdictions, aimed at enabling the legality of autonomous vehicles on the roads. Plenty of mainstream journalists will now opine in print that the days where human drivers dominate the road are numbered.

AGI seems unnecessary for the creation of perfectly effective self-driving cars. Cleverly integrated narrow AI components will probably do the trick. However, when unexpected situations arise on the road, an AGI will handle them better, thus saving a certain percentage of lives. The essence of AGI is the ability to handle the unpredicted and unforeseen in an intelligent way, based on creativity and on extrapolation from different but indirectly related prior experiences. If an animal of an unfamiliar kind runs across the road, or an unusual natural disaster makes road conditions bizarre, narrow AI may not deal with it effectively, and an appropriate AGI may well deal with it better than human drivers.



Figure 13: Google's experimental self-driving cars have become a fixture on the streets of San Francisco.

<http://addins.waow.com/blogs/weather/wp-content/uploads/2012/08/google-self-driving-car.jpg>



Figure 14: Ernst Dickmanns and his colleagues made self-driving cars in the 1980s and early 1990s, capable of fairly robust highway driving based on integration of computer vision with adaptive control. In light of Google's recent work with self-driving cars (which like all work in the area builds heavily on Dickmanns' ideas), we were pleased and a bit amused to invite Dickmanns to give a keynote speech at the Artificial General Intelligence 2011 conference, which was held on Google's campus in Mountain view.

<http://www.wired.com/autopia/2012/02/autonomous-vehicle-history/?pid=1580&pageid=42236&viewall=true>

Chemicals

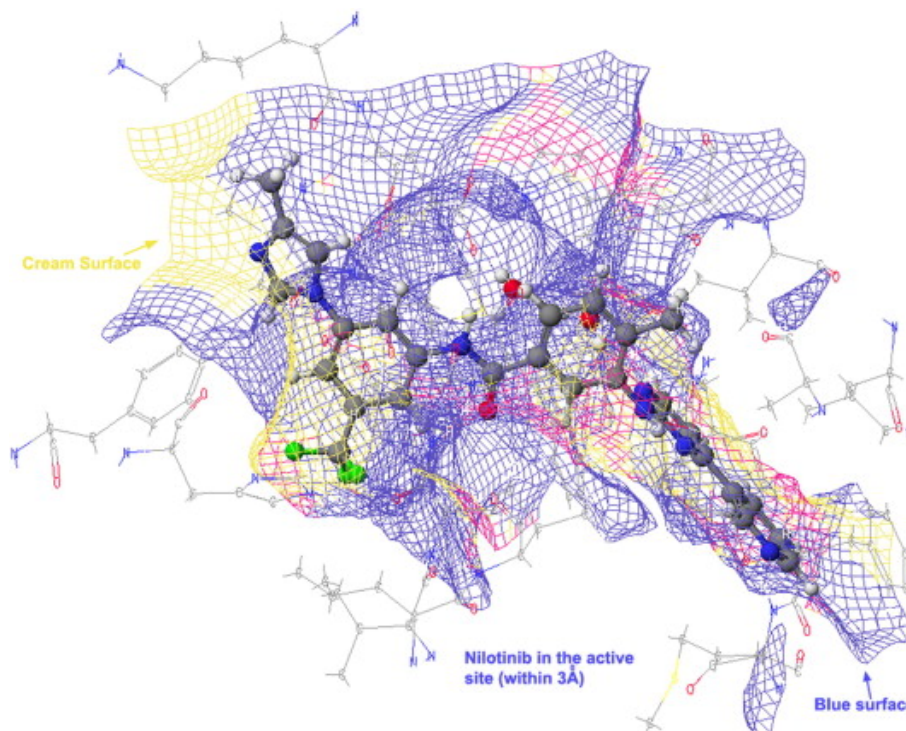


Figure 15: “Rational drug design” uses chemistry and biology to figure out drugs that can attack given biological targets. So far most successful drugs have been designed through trial and error, but rational drug design is being actively pursued and is likely to yield dramatic fruit in the next couple decades. Advanced AGI would make rational drug design much easier, as the process involves many difficult modeling and analysis steps that are difficult for human judgment, even with available simulation and mathematical tools.

The figure shows the “adjacent surface” for the substance nilotinib; that is, it shows the adjacent surface pocket that is the surface within 3 angstroms of the drug to the active site of the target. This surface provides guidance in the process of drug development. But it is actually just a crude approximation, and better representations can be formulated taking closer account of the quantum mechanical foundations of chemistry. The reliance on simpler representations like the one depicted is partly because the human mind doesn't take that naturally to quantum mechanics.

<http://www.sciencedirect.com/science/article/pii/S0014299909008784?np=y>

Designing molecules and chemical compounds to carry out specified purposes is, at this point, largely a matter of guesswork. Imagine an AGI that truly mastered chemistry, moving directly from a specification to a molecule or compound fulfilling that specification. It should be easier for an AGI to master chemistry than a human, because human sensory organs and actuators (fingers, feet) do not naturally operate at the molecular level. But an AGI could be given molecular-scale “eyes” and “hands” to let it interact with the chemical world directly. What we do with abstract analyses and complex awkward tools, such an AGI could do with direct sensations and tools operated with the same sort of intuition and fluidity we use in operating our fingers.

Molecular nanotechnology may obsolete traditional chemistry by letting us build molecules according to whatever structures we specify. But while we're on our way to achieving full molecular

nanotechnology, AGI-powered chemistry may play a major role in making our lives easier and advancing science. Nanotech is already happening, and will get better and better step by step, eventually achieving the powerful ideas posited by early nanotech visionaries like Eric Drexler. AGIs should have large advantages over human in pushing nanotech forward and advancing our fine-grained control over matter.

Computers & Software

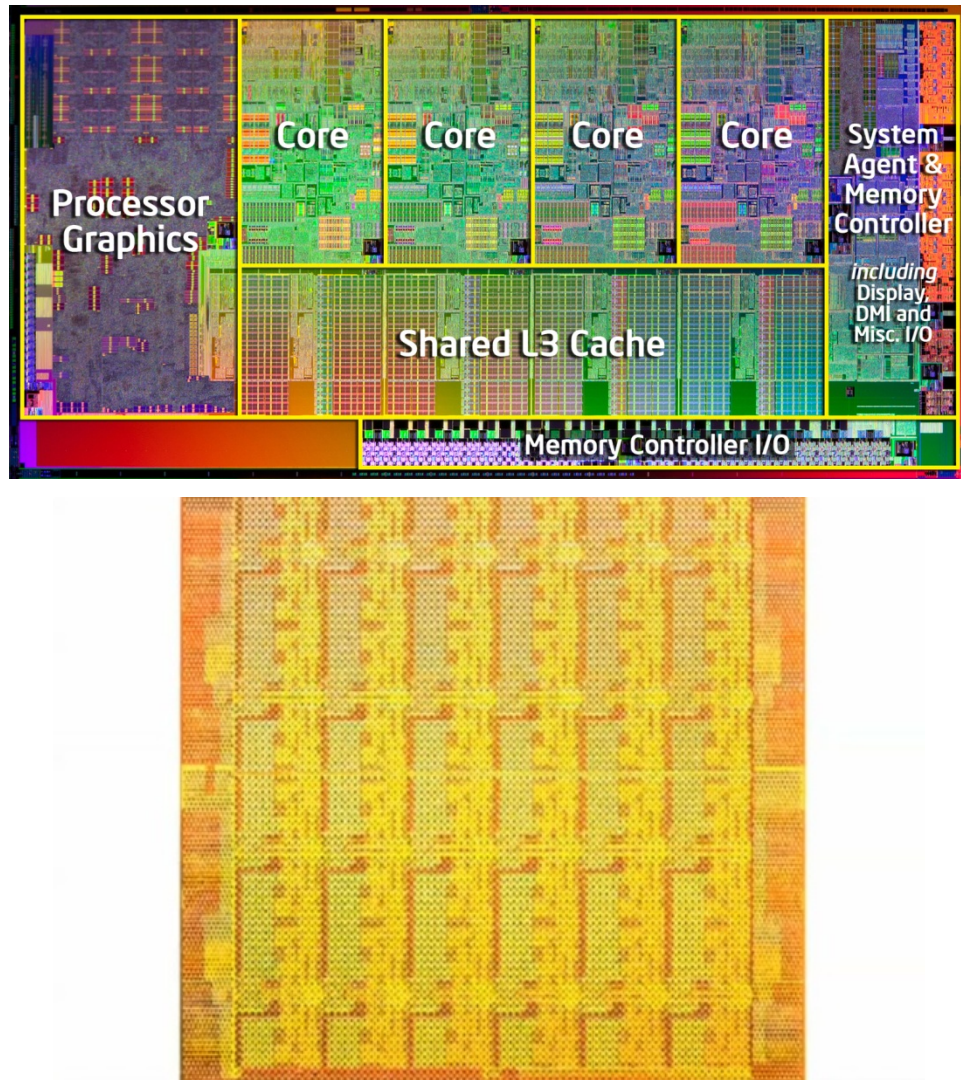


Figure 16: Multicore hardware has allowed computer speed to keep advancing very rapidly in spite of the complexities of making compute cores smaller and smaller. 4 and 8 core machines are the norm now; and IBM, Intel and other firms are experimenting with architectures possessing 1000+ cores. The main difficulty presented by these massively multicore architectures is that programming them is very hard for the human brain. AGIs however, will not necessarily experience similar difficulties, and may rapidly make breakthroughs in multicore algorithm development. The top picture is a standard quad-core architecture; the bottom picture illustrates a 1000-core design presented by Intel at a conference in 2010.

Top: http://regmedia.co.uk/2011/11/14/core_i7_sandy_bridge_die_large.jpg

Bottom: <http://i.neoseeker.com/n/3/processor2.jpg>



Figure 17: Contemporary server farms, outside and in. (The top picture is actually an Apple data center.) More and more, computing happens here rather than on desktop computers. Much of my own work on OpenCog software happens on rented computer time in the cloud. Cloud computing is yet another illustration that it's not the particular physical substrate that's important, it's the pattern of organization. The same computer code can shift from one computer to another in the cloud – no matter. What matters is what the program does.

Top: http://www.digitalmusicinsider.com/wp-content/uploads/2011/02/AppleNCnews_24279.jpg

Bottom: <http://timmurphy.wpengine.netdna-cdn.com/wp-content/uploads/2011/05/Server-farm.jpg>

Of all the amazing capabilities future AGIs are likely to achieve, the most dramatic one may end up being AGI's ability to transform computer software and hardware. Digital computer software is an alien universe to humans, just as much as the molecular world. Its fanatical precision drives us crazy, as any programmer who has banged their head against code bugs realizes acutely. An AGI, if appropriately architected, will naturally be adept at using software code – it will read and program code as naturally as we see or hear, or pick up a stick in our hands. This will make a huge difference in the speed and quality of production of basically all kinds of software. And it will enable AGIs to program yet better AGIs, beginning the dynamics of “intelligence explosion” or “hard takeoff.”

This explosion of software will go hand in hand with dramatic improvements in hardware.

Today, computers are built according to basically the same architecture that John von Neumann outlined in the 1950s. Why is this? It's not because of any lack of imagination on the part of computer architecture designers: the history of computing is littered with wonderful alternative architectures.

I remember programming the Connection Machines designed by Danny Hillis, back in the 1990s. The Connection Machine I worked with could do 64000 independent operations at once, much more than the standard 2, 4 or 8 cores in most modern computers. A modern graphics card can do a few hundred operations at the same time, but they all have to be the same operation, unlike the Connection Machine whose 64000 processors could all do different things at the same time.

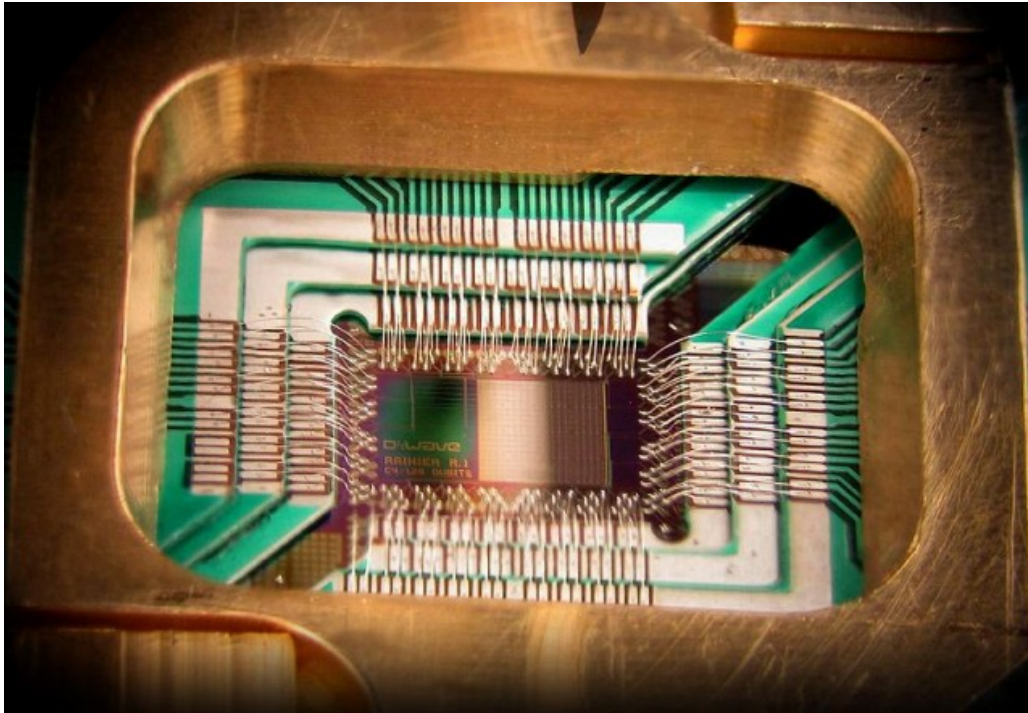
More recently, and less ambitiously, Sony and IBM collaborated to bring the radical new “Cell” computing architecture to the PS3 gaming console. It uses only 8 processors, but with different properties and interacting in a different way than in ordinary computers or game consoles.

Why did the Connection Machine and the Cell flop, like so many other novel computing hardware approaches? Not because the hardware was bad, but rather because it was *too hard for human beings to rapidly write software for the new hardware architectures.*

The human brain struggles to adapt a familiar algorithm or data structure to a new kind of computing hardware; it takes us a lot of work. And we've built up a lot of detailed knowledge, regarding how to make software work on the familiar von Neumann computer architecture.

Not only could an AGI potentially design new and far better computer hardware architectures, but with the capability to extend current software to alternative hardware architectures, it would revolutionize

computing via allowing existing but obscure computer hardware designs to flourish.



Yes, you can have one.

No, you're not dreaming. D-Wave offer the first commercial quantum computing system on the market. We believe in building great things that are as inspiring as they are powerful.

If you're passionate and curious about the future of computation, and you'd like to take a different approach to solving problems, then take a look at our products.


 D-Wave One™ information

Figure 18: Top: the interior of one of Dwave's novel quantum computing devices. Bottom: 2011 advertisement for DWave's quantum computing hardware. Dwave's system is billed as the first commercially available quantum computer (though not as a general purpose quantum computer; it only solves certain specific types of mathematical problems, which however have a fairly wide range of applicability). At time of writing, there is still some controversy as to whether DWave's machines are really doing quantum computing or just some novel sort of analog computing.

Top: <http://arstechnica.com/science/2013/07/d-waves-quantum-optimizer-might-be-quantum-after-all/>

Bottom: <http://www.blogcdn.com/www.engadget.com/media/2011/05/5-18-2011d-waveone.jpg>

And then there's the advent of quantum computers—using the weird physics of the microworld to do computing fundamentally faster. Quantum computers perform several calculations in multiple universes at once, thus arriving at answers on average much faster than our contemporary, classical physics based computers, whose computation is boringly concentrated in a single universe.²

However, figuring out how to program quantum computers is not easy. Our brains struggle with the quantum domain; it is incredibly counterintuitive. In the quantum world, things like electrons can be particles and waves at the same time, but they cannot have both a known position and a known speed at the same time. Particle/waves can teleport through walls under certain conditions. An AGI with sensors directly at the quantum level would understand quantum phenomena intuitively in a way that the human brain cannot, enabling it to design quantum computing algorithms beyond our comprehension. This may be the only way to get general purpose quantum computers to work.

²Check out XX's book *Programming the Universe* for a wonderful nontechnical review of quantum computing. Or Fred Alan Wolf's oldie-but-goodie *Taking the Quantum Leap* for a general review of quantum weirdness.

Construction



Figure 19: Will AGI architects design us fantastic, futuristic buildings and cities? Or will they take things in another direction – perhaps creating reconfigurable houses, that reassemble themselves in different shapes adaptively, based on our needs and desires? In any case it seems unlikely that AGI architects and robot construction workers would feel restricted to the conventional habits of contemporary human building design – which basically exist due to cost issues, as well as lack of imagination by the people paying the bills for most construction.

Top: http://www.worldarchitecturenews.com/index.php?fuseaction=wanappln.projectview&upload_id=617

Bottom: <http://inhabitat.com/city-in-the-sky-futuristic-flower-towers-soar-above-modern-metropolises/megatropolis-city-in-the-sky-hrama/>



Figure 20: Conceptual illustration of a seastead – a city constructed in the middle of the ocean, in international water where no current nation’s laws apply. Seasteads could allow much freer experimentation with different forms of society and government. The main obstacle preventing widespread construction of seasteads at the moment is the cost of constructing things at sea – a problem that AGI construction robots could presumably solve. Although it would be more fun if we didn’t have to wait, and could have seasteads pronto, and put our AGI research facilities and robot factories thereupon!

<http://www.marineinsight.com/wp-content/uploads/2012/08/sea-steading-1.jpg>

We take for granted that building buildings is slow and difficult. But actually – why can’t a new skyscraper be erected in a few minutes or hours, rather than taking months, or even years? And why can’t all the units in a housing development project be constructed at once, applying individual variations on the same basic plan? Lack of resources perhaps; yet more importantly, the human body and the work involved in constructing large buildings are clearly not the best match.

Construction will be revolutionized once there are robots that can handle visual perception and physical object manipulation at the human level. And it won’t be a big step to deploy the same capabilities in non-human robot bodies better suited for construction work. Construction robots won’t resemble humans; they’ll be a combination of intelligent construction trucks and flying drones, featuring several mechanical arms and claws.

Defense & Intelligence

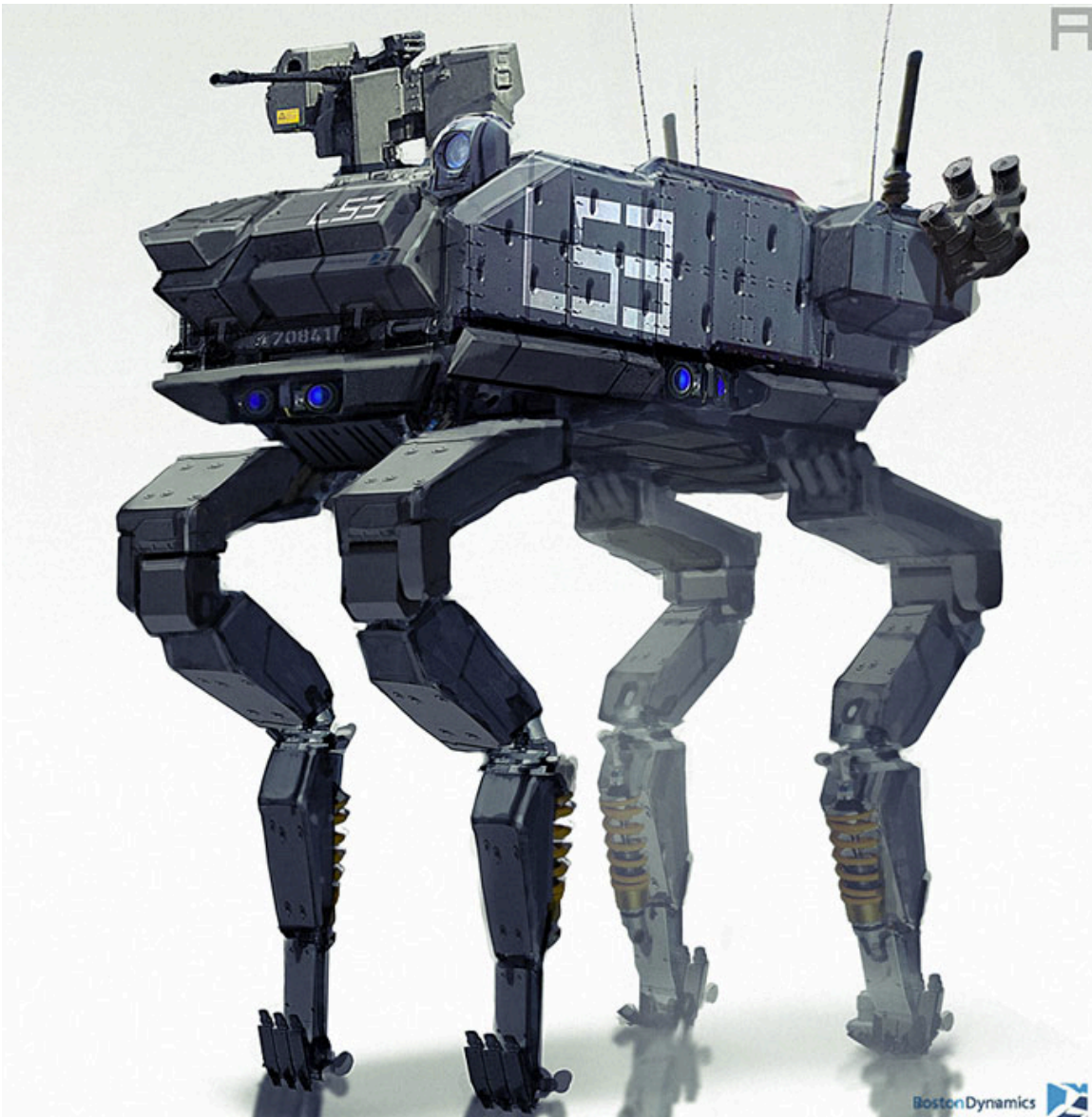


Figure 21: Boston Dynamics' famous Big Dog robot, created for US military purposes, with an unprecedented capability for stable locomotion across unstable, irregular ground. Upright humanoid locomotion on outdoor surfaces remains difficult for robotics today, but quadruped locomotion is significantly easier. As George Orwell said "Four legs good, two legs bad"!

http://digital-art-gallery.com/oid/15/1000x707_4442_BigDog_LS3_2d_sci_fi_robot_picture_image_digital_art.jpg

Historically, the US military has funded the majority of the world's AI research, so once AGI technology matures it will likely be used in a military context. Early-stage AGI, however, may rapidly amass the power to prevent this from happening.

In my opinion, military forces will probably not be the initial source of AGI advances, since they will require AI to be highly predictable and reliable; early-stage AGI is unlikely to possess these qualities. If we follow a path to AGI inspired by human child development, early-stage AGIs may have the same

playfulness, unreliability and confusion of young human children. Since an AGI has inherent generality, no development path, even one that tries to bypass the childlike stage, will be free of this unpredictability. Understanding the parameters and properties of the first AGI system's unpredictability will take time, so the first military robots to be rolled out will likely be narrow AIs erring on the side of predictability.

Military AGIs will not only be robots. They will be military commanders, orchestrating robotic and human warriors according to strategic and tactical patterns too subtle for human minds to conceive or comprehend.

Military robots will enhance the tools of mass destruction; delivery of nuclear weapons will become more reliable and difficult to defend. However, a greater concern may be the current military trend (predominantly in the US) of carefully targeted destruction. Imagine a future military force—AGI military robots operating by air, land, and sea—able to send a crack team of robot killers to any location to carry out surgical military strikes.



Figure 22: Surveillance cameras, satellites, audio recording devices and a host of other instruments allow gathering of massive amounts of data about everyone's everyday life. Not to mention the amount of data that intelligence agencies and others gather regarding our online lives. But right now, most of the data that's gathered just sits there, because nobody has time to sift through it all and find the relevant portions. AGI systems with the capability to scan through and interpret texts, video and audio files much faster than human beings, would be able to make use of the massive amount of data gathered via modern surveillance technology, as well as develop yet more powerful means of observation and recording. AGI will therefore make acute the dilemma posted by David Brin in his book *The Transparent Society*, where he notes that as the future unfolds we face a choice between surveillance (the Powers That Be watch everyone) versus sousveillance (everyone has access to data regarding everyone).

<http://clatl.com/atlanta/atlanta-under-surveillance/Content?oid=7121394>

AGI's impact on the intelligence world will be equally dramatic. Intelligence agencies in the US, China and various other nations gather and stockpile data about the citizens of the world—emails, text messages, phone calls, video surveillance images, you name it. But it's simply too much information for any reasonably sized group of humans to look through. Today's AIs are not smart enough to read texts, nor recognize objects or faces in pictures, except at a simple level; however, human-level AGI, when put together with already-existing surveillance technology, basically yields Orwell's Big Brother.

David Brin, science fiction writer and futurist, argues that the only solution is to subvert surveillance with *sousveillance*: observation from under rather than over (everyone can watch everyone). If all the data from everyone's emails, phone calls, texts and webcams and surveillance cameras were public, everyone could watch each other; we could watch the government as it watches us. Transforming the notion of privacy radically, it would prevent a scenario in which the few have exclusive capability to watch the many.

This brings us to the topic of

the risks and rewards of AGI, and ideas about safeguarding against dangerous scenarios – a large, deep and thorny topic that I will tackle in a later chapter. But for now, I just want to make the point that AGI is not the only advanced technology with its scary aspects. All the other advanced, Singularity-enabling technologies pose similar or greater risks. Synthetic biology, for example, seems particularly worrying. Once synthetic biology advances, what will stop hostile-minded scientists from developing novel biological viruses and germs aimed at circumventing the human immune system and killing on the mass scale? It's possible that the best protection for the human race will be some sort of "AI Nanny" that protects us (and itself) from the dangers posed by emotionally unhealthy humans armed with powerful technologies.

Energy

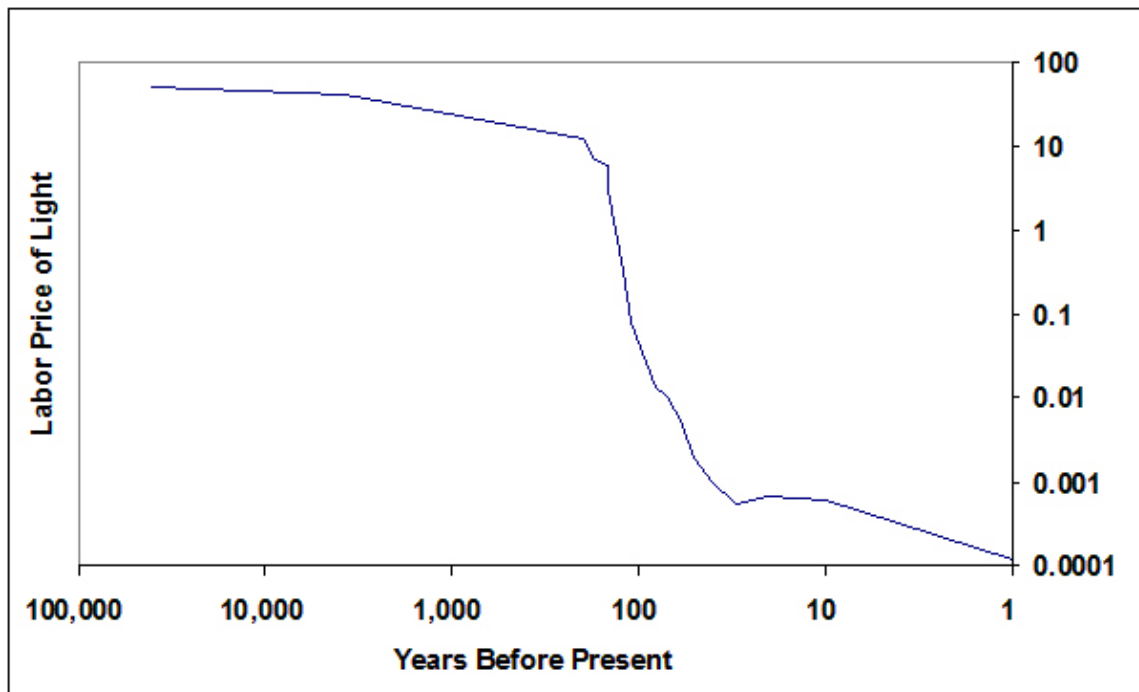


Figure 23: The price of light as estimated by futurist Jose Cordeiro, as part of his analysis of the Singularity as an Energularity.

<http://lifeboat.com/ex/the.energularity>

Venezuelan futurist José Cordeiro underscores that the Singularity will also be an “Energularity.” Humans have used more and more energy, per capita, per day, since the start of civilization—and this is increasing at an exponential rate. AGIs are likely to master forms of energy generation that have eluded human intelligence— nuclear fusion(first traditional hot fusion, then cold fusion). Increasing evidence indicates cold fusion is a valid scientific phenomenon, not the fraud suggested by the publicized errors in Pons and Fleischmanns’ early work. AGIs will also be able to operate in space more easily than humans, so will have little trouble erecting solar power satellites or massive Mylar solar panels floating in space. Using the massive new energy sources they create, AGI’s will power their massive processor farms, boost their intelligence, and create even more effective energy sources.



Figure 24. Screenshot from the video game “Creatures Online”, a descendant of the original Creatures game created by AGI researcher Steve Grand. Each creature is controlled by a neural network, which develops over the creature’s life time. When creatures mate, their children get a neural network based on crossover and mutation from their parents’ neural networks. Steve Grand spoke at the AGI Workshop I organized in Maryland in 2006, describing the robotics work he was then doing, focused largely on solving the computer vision problem.

<http://creatures-online.co.uk/goodies/screenshots/gamescom-2011-demo-the-whole-family-is-together>

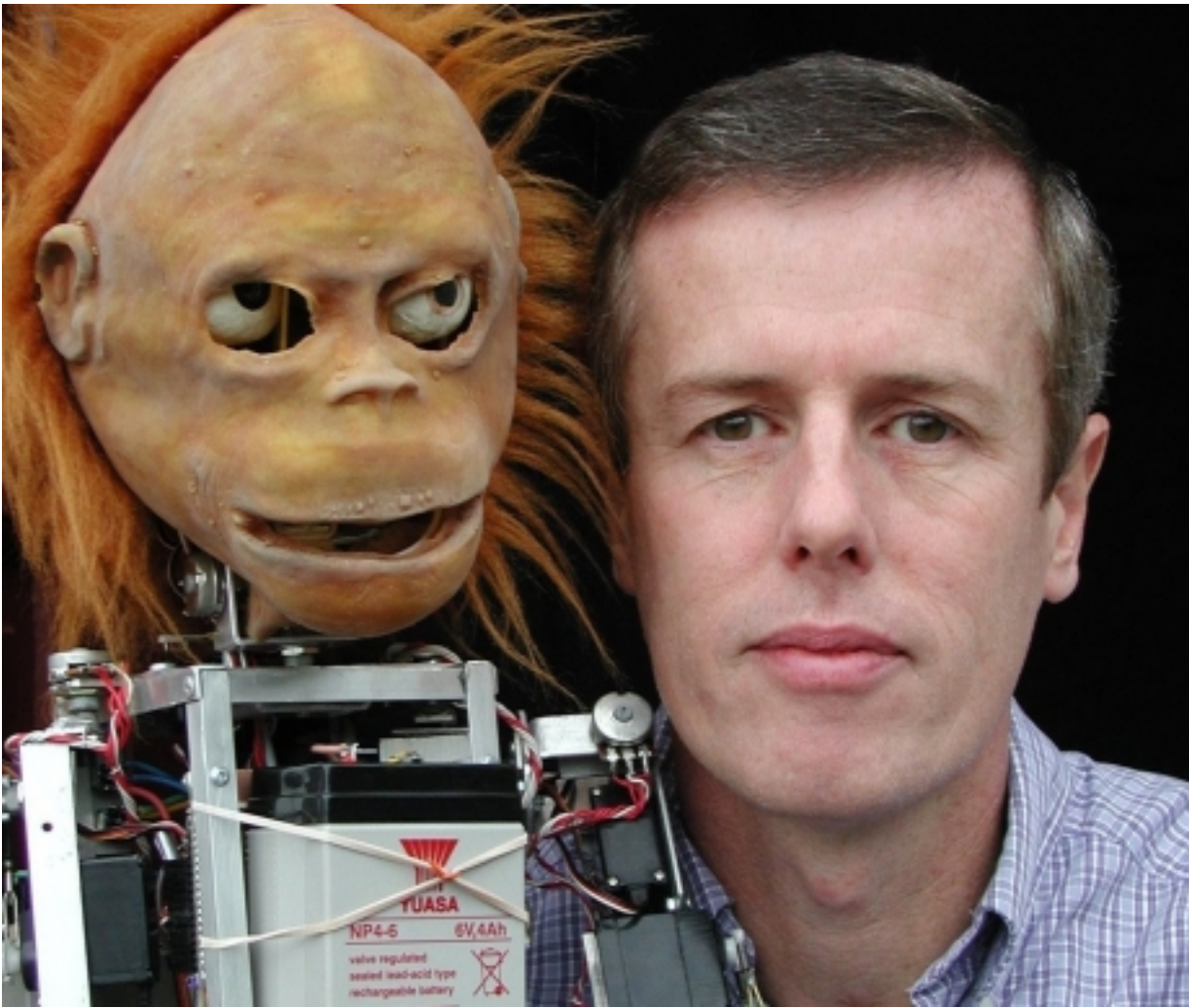


Figure 25: Creatures game creator and AGI researcher Steve Grand with his robot Lucy, in 2006.

<http://www.avfestival.co.uk/programme/2006/events-exhibitions/cafe-scientifique-steve-grand>

The line between the human and automated is already blurring in the world of arts & entertainment. Computer graphic scenes and human-actor scenes are often indistinguishable from each other in movies; the same applies to algorithmic versus human-played rhythm and melody in pop music. Non-player characters in video games do nearly all the things human characters do. And all this is without the advent of AGI – without even much narrow AI in use, actually.

On the other hand, computers can't create (yet) emotionally realistic facial

expressions, nor evocative melodies and rhythms with rich human emotion. They also can't write stories, as these are largely experiential. It's just my own speculation, but I suspect these shortcomings of computers may be among the last for AGI to overcome. Because they're not just about being intelligent; they're specifically about being "human" and understanding the emotional core of

humanity.

Without embodiment, where the AGI is placed inside a robot body, it

may be difficult for an AGI to have experiences that are sufficiently human-like to grasp the essence of human experience. An AGI lacking a human-like embodiment could write stories about its 24 hour day and patterns it sees via cameras and microphones. But, it wouldn't be fiction written from a human-like perspective. However, even a humanoid robot body, one lacking human touches like skin, hormones and sex organs, might not give an AGI sufficient understanding of the human condition to make human artworks.

I suspect there are aspects of artistic creation that human – level AGIs will be unable to master, unless they are built with human-like bodies and minds based on human brain emulation. On the other hand, a *superhuman* AGI may be able emulate human artistic creation. But it may require greater general intelligence to emulate human art in all respects than superseding human mathematics, engineering and science.

That said though, AGIs might well be able take over the vast majority of human entertainment and fine arts, once they've reached a point well below human level general intelligence. Recreating human productions may not be the best way to entertain most humans; the success of computer generated image-heavy movies and algorithmic pop music rhythms proves otherwise.

Ultimately, human-level AGI systems, after studying what entertains and pleases humans, will produce works that deliver what we want, even before AGIs create “human” art works. “Authentically human” entertainment and artwork may become a niche taste, similar to oil paintings and live jazz improvisation today, while the vast majority of entertainment and artwork will be explicitly AGI-generated. Celebrities will no longer have to appear physically in advertisements or go to live photo shoots—after licensing the use of their image, AI will do the rest.

Finance & Insurance



Figure 26: Back in 1980, the New York stock exchange floor was bustling with human beings trading stocks with each other. Nowadays, many stock exchanges have eliminated physical floor trading altogether, as nearly all trading is electronic. As of 2013, the New York Stock Exchange still has a trading floor, but it's very lightly used compared to the past.

<http://commons.wikimedia.org/wiki/File:Stockexchange.jpg>

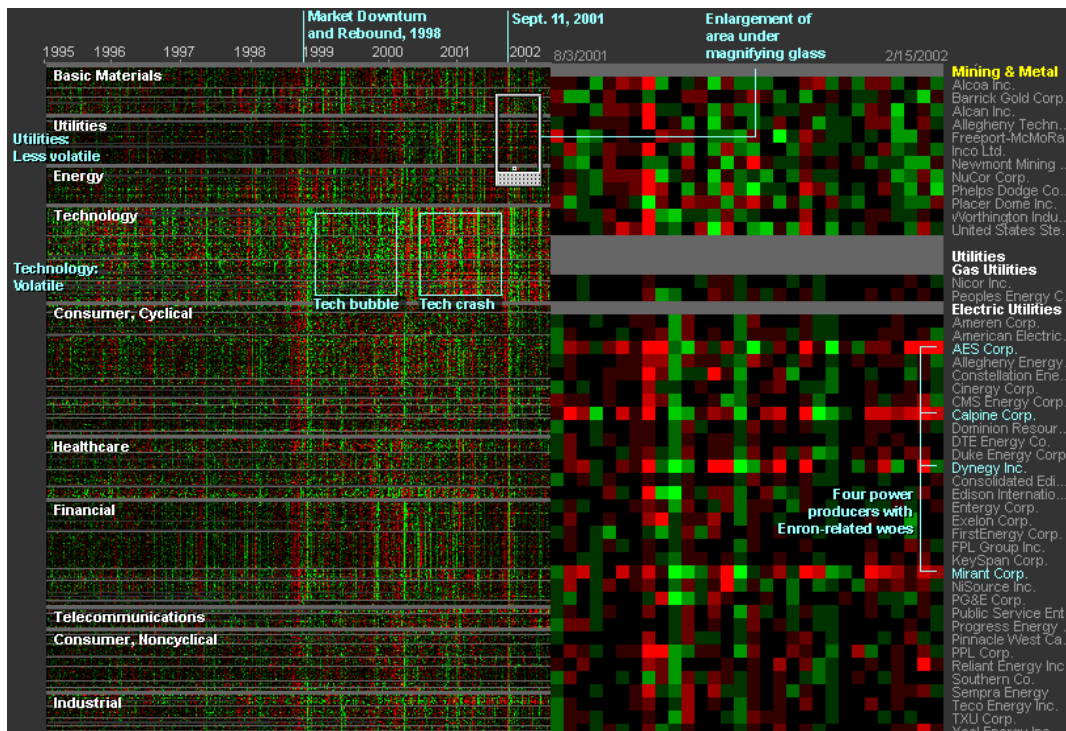
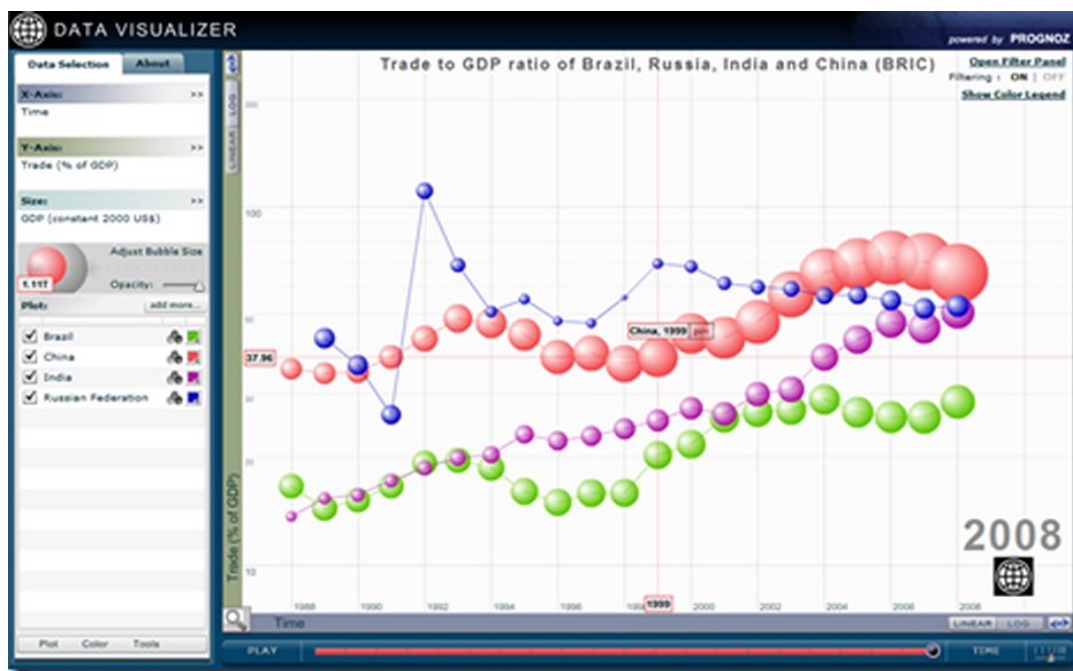


Figure 27: Back in 1980, the New York stock exchange floor was bustling with human beings trading stocks with each other. Nowadays, many stock exchanges have eliminated physical floor trading altogether, as nearly all trading is electronic. As of 2013, the New York Stock Exchange still has a trading floor, but it's very lightly used compared to the past. Today's financial markets are extremely complex, and embody a kind of complexity that the human brain is not well suited to understand. We try very hard to project the complexity of financial systems into 2D visual forms that the human eye can comprehend, but this inevitably leaves out most of the relevant structure in the data. This is why the vast majority of financial trading in the US, Europe and Japan is already done by computer programs, including plenty of narrow AI programs. 20 years from now the percentage of trading on major markets conducted by humans is likely to be minimal, even without anything approaching a Singularity.

Top: <http://www.trinity.edu/rjensen/352wpvisual/BricsVisualization.jpg>

Bottom: <http://ml.smartmoney.net/marketradar/images/radar5.gif>

The idea of AGIs pervading the finance world may seem frightening – isn't program trading already causing a lot of problems? But actually, it's arguable that these problems are caused by the use of insufficiently intelligent programs – and the combination of such programs with insufficiently intelligent, or insufficiently ethical, humans. Once we have reasonably generally intelligent AIs serving as Chief Investment officers for hedge funds, we might have much safer and more efficient financial markets.

Finance these days is largely based on advanced mathematics, but the mathematical formulas used are all approximations to reality, based on assumptions that everyone knows aren't quite right. Everyone makes these mathematical assumptions anyway because the equations can't be solved any other way. This creates problems— the financial market crash in 2008, the collapse of Long-Term Capital Management in 1998, and so forth.

An AGI could apply financial mathematics more sophisticatedly, taking better account of the available data, and making fewer unrealistic assumptions. Human-level AGI would likely “clean up” in the financial markets. If a human-level AGI, with direct mental connection to basic algorithmic software tools, were unleashed on the financial markets today, it could amass hundreds of billions of dollars within weeks. Imagine releasing George Soros into an active, volatile stock market populated entirely by people with IQs of 70 or lower.

More likely, the level of AGI used by financial firms will escalate gradually with multiple competing firms having roughly the same level of AGI at each point in time. Unaided human beings will then have less and less chance of success on the financial markets. Now, AI and statistical software are better at recognizing purely numerical trends in stock market data and at figuring out fair prices for complex financial constructs like options and derivatives. Yet people are better at other aspects of finance, like incorporating information from the news, and qualitative information about companies' products and management. However, once an AGI reads and understands the newspaper and a company's annual report, this human advantage will be eliminated. Additionally, the AGI would not suffer from the main enemy of every human trader or financial analyst: human emotion. A financial AGI could be programmed as almost a pure rationalist— free from the particular emotional biases that ultimately lead many human beings to make bad financial decisions.

When AGI systems become players in the world financial system, they will surpass its current revolutionary features like “High frequency trading,” a mechanism enabling the purchase and rapid sale of stock (at fractions of a second). A regulatory framework instituted by AGI lawyers may be the last piece of the new financial system.

The insurance industry will also be revolutionized by AGI. Current insurance pricing models are crude and generally based on dividing people and companies into categories. AGI insurance assessors will calculate risks on a rational and individualized basis, creating a more efficient economy and easing the development of new technologies as the Singularity approaches.

Hospitality



Figure 28: No book about the future would be complete without a picture of a sex robot! Of course this is far from the only potential use of robots in the hospitality business. But given the realities of human nature, it's likely to be one lucrative application, once robotics technology advances a bit.

<http://beyondthehustledigital.tumblr.com/post/27779524216/sex-robots-x-japanese-dolls>

AGI program traders are all very well, but they're not very photogenic. They definitely don't need humanoid bodies; all their job requires is interfacing with online information sources and electronic financial exchanges. Aesthetically speaking, it's more entertaining to think about robot waiters, masseuses, house cleaners, sales clerks, prostitutes -- you name it. AGI systems could have their motivational systems specifically engineered to guarantee they enjoy these jobs, providing us with an amazing level of hospitality.

While some, for emotional reasons, might prefer being served by human beings, most of us will appreciate having basic needs taken care of by AGIs.

Manufacturing



Figure 29: A robotic car factory in Korea. Robotic manufacturing is already possible with a fairly high degree of generality, and is limited at this point mainly by cost. As robots become less expensive, there will be less and less role for humans in the manufacturing process. If everyone in the world were making a First World income, manufacturing would already be much more thoroughly robotized, because in many cases robots are already cheaper than First World human laborers, even if not yet cheaper than the least expensive human laborers on the planet. AGI will make robot manufacturing much easier to set up, because a single type of AGI robot will be able to carry out a variety of different types of manufacturing, in contrast to the current situation where one must carefully engineer and program robots especially for each manufacturing situation.
<http://www.advancedtechnologykorea.com/9013>

Manufacturing is already becoming robotized. Currently, though, robots are restricted by their inability to see and manipulate most everyday objects as effectively as humans. While human capability for this remains roughly constant, a robot's capability increases exponentially.

Tomorrow's AGI factories will resemble little of today's human-powered factories. Each factory will be the body of an AGI mind, operating in close "digital telepathic" communication with a host of other AGI minds operating other factories in related industries. The supply chain will be managed via social interaction of the factory minds. New innovations will be incorporated rapidly into the manufacturing process, making specialized one-off manufacturing jobs commonplace.

AGI-powered manufacturing will become more powerful as nanotechnology develops, likely enabling

one of the paths to advanced nanotechnology that Eric Drexler envisioned: small machines that build yet smaller machines (that build yet smaller machines...). Perhaps femtotechnology—using elementary particles like protons and electrons to manufacture materials—may follow.

But, what role will humans play when manufacturing becomes AGI-powered? Although production will function without us, we'll still (at least in the beginning) serve an active role in design—being customers should intuitively guide us.

News media

The exponential rise of computing and communication technology since the mid-1990s has revolutionized the media business, as anyone with an Internet connection has already noticed. News media in particular is in a state of economic crisis as well as radical advancement, for the simple reason that hardly anybody wants to pay for news when so much is available for free. Traditional newspapers and magazines are losing money in droves; their audiences have been stolen by bloggers, tweeters and so forth, whose day jobs are not primarily in news production.

An unspoken assumption at the heart of the news media business is that a lot of effort is necessarily required to translate ideas and observations into comprehensible series of words.

AGI systems capable of sophisticated natural language generation will radically change the industry, precisely by violating this assumption. Not too long from now, nearly all news reportage will be automated – with software generating text articles based on raw data, and other software catering the selection of articles to each reader.

There already exist systems that automatically generate news articles based on structured data regarding sporting events, stock price moves, and so forth. And systems that customize news feeds to the individual based on their reading history are already commonplace. Once news generation and customization software becomes more prevalent, human commentary on news may still exist to some extent, but it will serve purely a social and artistic role— similar to people casually chatting with each other about the news, or talk shows focused on pundits sharing their subjective opinions.

Pharmaceuticals & Health Care

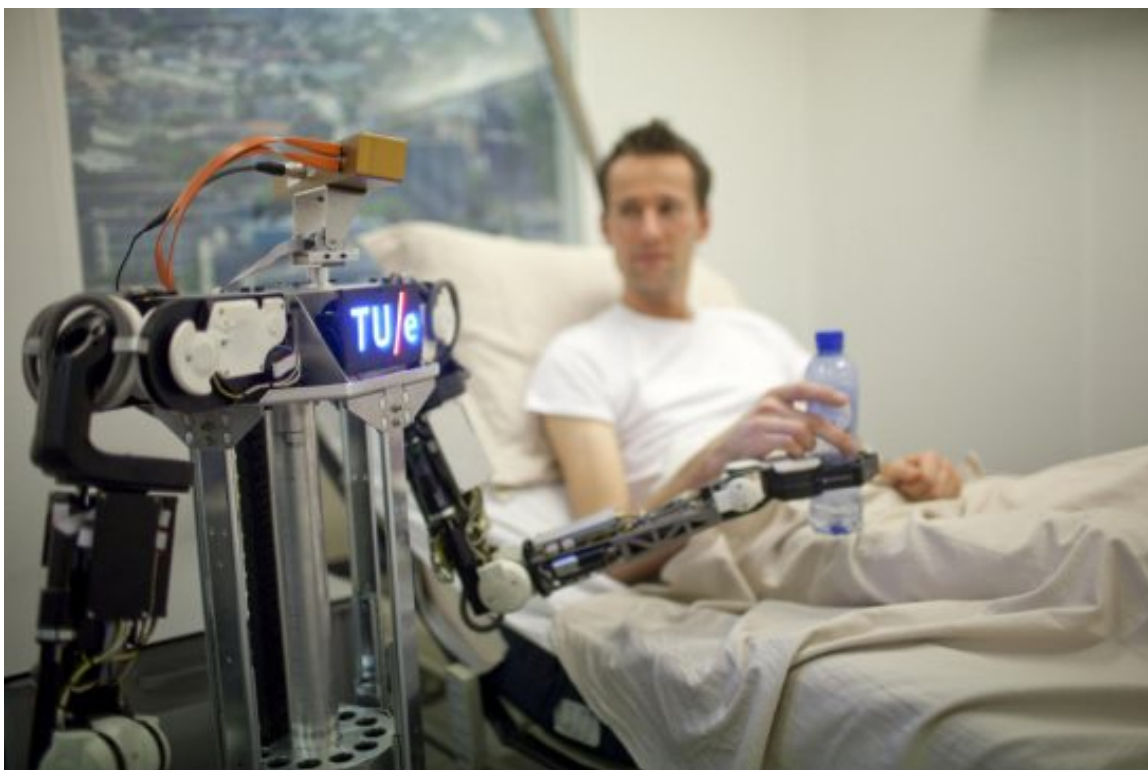


Figure 30: This robot, serving a drink to a patient in a hospital, runs on RoboEarth, an online open-source network database of robot control information. At the moment this sort of thing is at the level of research prototypes, but there are no fundamental technical obstacles toward rolling it out broadly. Making this sort of robotics application widespread just requires cost reductions in various aspects of the technology; it doesn't even need AGI, though AGI could obviously broaden the scope a great deal.

http://www.thestar.com/business/tech_news/2011/02/11/robots_get_their_own_internet.html

The pharmaceutical and health care industries, like the publishing business, are struggling to adapt their practices and business models to the exponential advancement that has already happened during the last few decades. The next few decades are bound to disrupt them further and further.

I have had a window into the issues these industries face, via my own involvement with the use of AI to aid in drug discovery. What I have seen is that today's drug discovery process – as practiced in pharmaceutical firms – is mostly trial and error. Genomics, and other advanced biological knowledge from the last few decades, are rarely applied effectively. Part of the issue here is sociological – there does exist knowledge about how to do drug discovery better using modern ideas, but this knowledge is taking a long time to pervade the major pharmaceutical firms. I will say a bit about this later in the book, when I discuss the application of AGI and narrow AI to longevity research. But part of the issue is more fundamental. In trying to discover drugs for curing complex diseases, the human mind is running up against the basic complexity of the human body, which is just plain hard for the human

mind to understand, even when there is a lot of relevant data available.

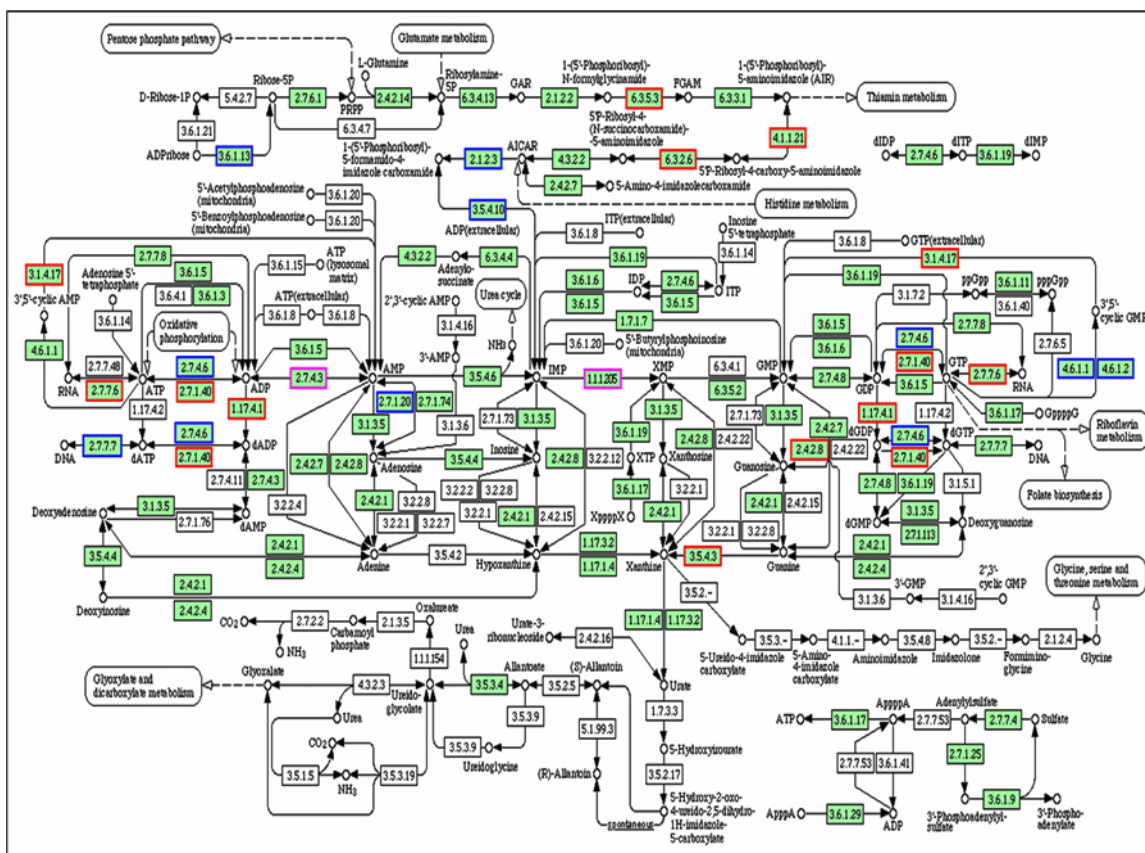


Figure 31: A randomly selected example of a biological pathway diagram, of the sort that biologists routinely look at in the course of their work. Such a diagram typically shows the best known biological and chemical reactions involved in a highly specific biological process. This one shows differentially expressed enzymes in purine metabolism identified from irradiated AT5BIVA and ATCL8 cells. Our best understanding of biological dynamics, at present, could be summarized in a set of thousands of interlocking diagrams of this sort. No human can hold them all in memory at once, obviously. So, as human biologists, we must focus on individual parts of individual diagrams, or traverse and analyze large collections of diagrams using fairly crude statistical or mathematical methods. An AGI with human-level general intelligence but a mind fashioned to handle this sort of data more naturally, could obviously do far better.

If you're curious: Specifically, in the above diagram, "Enzyme Commission numbers (EC#, e.g. 1.17.4.1) are used to represent enzymes in metabolism. Highlighted in green background are known human enzymes annotated in the KEGG database. Differentially expressed enzymes in purine metabolism (Table 3) are superimposed onto this pathway diagram: blue-boxed are enzymes changed in AT5BIVA cells, red-boxed those in ATCL8 cells, and pinkboxed those from both cells. Areas circled with broken lines highlight closely related biochemical steps surrounding ADP/ATP (left) or GDP/GTP (right) metabolisms, which include most of these differentially expressed enzymes from either cell type."

<http://www.omicsonline.org/ArchiveJPB/2008/May/03/ImagesJPB1.47/Figure4.jpg>

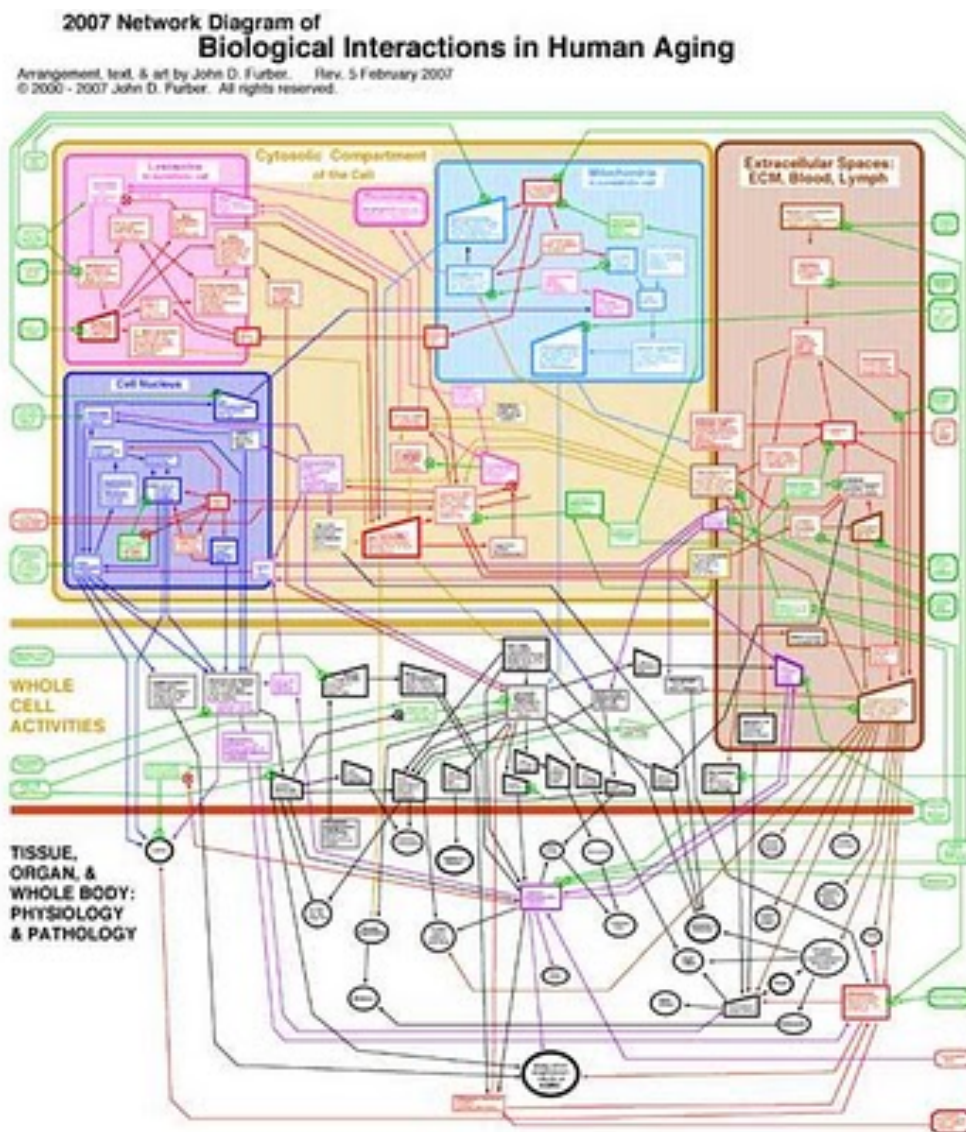


Figure 32: This is the 2007 version of a poster I have hanging in the wall of my study at home, made by longevity researcher John Furber. The 2011 version, the one I have at home, has a few small changes. Check it out online if you're curious. It is not a detailed pathway diagram, but rather attempts to give an overall conceptual picture of all the different structures and processes going into human aging. If nothing else, it gives you a good intuitive sense of the complexity of the aging process.

<http://www.legendarypharma.com/chartbg.html>

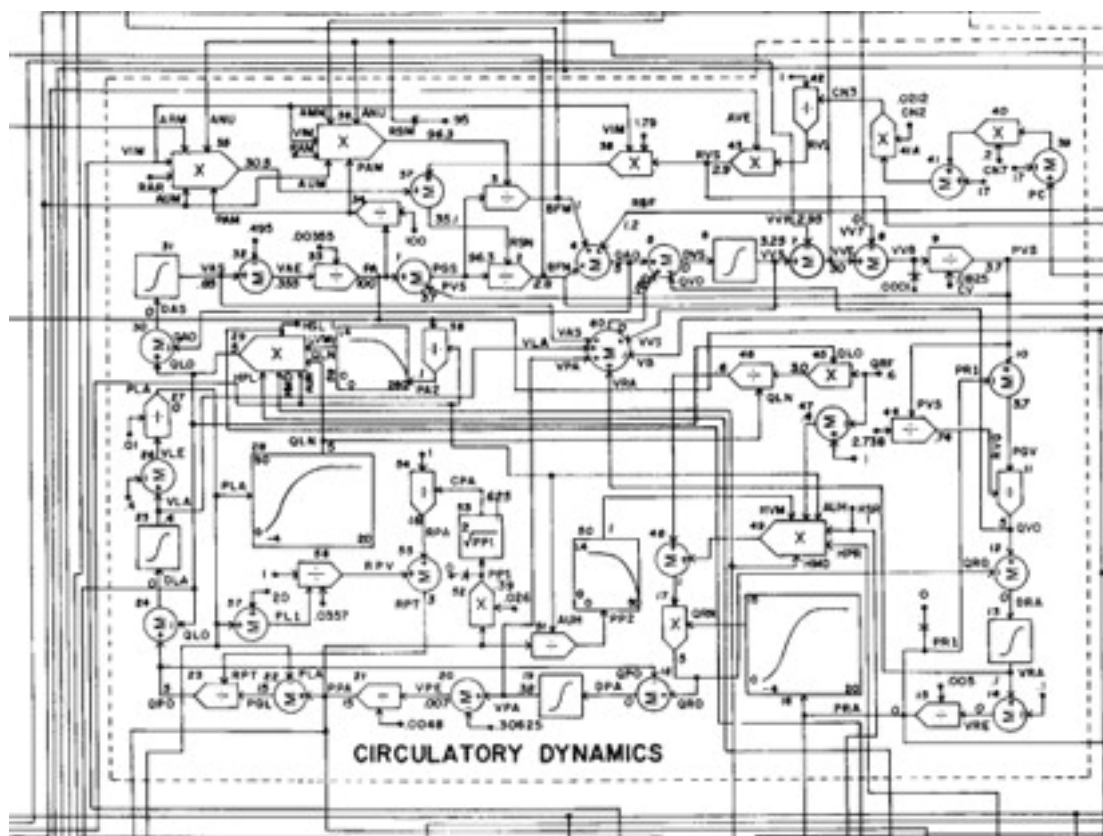


Figure 33: This depicts the subset of the overall Hummod dynamical systems model of the human body, that deals specifically with the circulatory system. Hummod consists of a couple dozen models of this rough level of complexity, all networked together. Each box indicates a certain mathematical equation, affecting some set of variables related to the circulatory system. This is a level of abstraction higher than a biological pathway diagram, and gives a view of how the dynamics implicit in biological pathway diagrams ultimately affects the body. The Hummod equations can be used to give an overall simulation of the interactions of all the human body systems. But they are too coarse-grained to capture everything happening in the body; they just give a crude, though very useful, high level view.

<http://hummod.org/assets/images/72-guyton-model.jpeg>

It seems possible that the limitations of the human brain will prevent human beings from ever fully comprehending dynamical systems as complex as the human body. No matter how much we augment our intelligence with data analysis and visualization software, the human mind won't intuitively grasp the detailed interactions of 25,000 genes, the proteins they code for, and the way they combine to build and maintain the human body.

On the other hand, an AGI adapted to the requirements of this sort of data would not suffer the same limitations. It could look at massive biological datasets that baffle the human mind, and see an abundance of patterns invisible to human perception – then designing drugs and other therapies based on those patterns.

In my own biological data analysis work – which I'll tell you more about later -- my colleagues and I have used AI tools to find thousands of genes showing significant differences between healthy and

unhealthy long-lived people. As our AI analysis shows, the effects of these genes on longevity are generally not individual, but rather combinational, involving interactions between multiple genes.

Human biologists tend to focus on individual genes, or (at best) on pairs or triples of genes – not because this how genetics works, but because this is what the human brain needs to do to simplify the problems of genetics and put them in humanly comprehensible form. A human brain naturally combines hundreds of visual features when recognizing somebody's face; it has acquired this skill through evolution. But a human brain is just no good at perceiving or thinking about the combinations between hundreds of different genes – even though these sorts of massively multi-gene combinations do appear critical to understanding why bodies age and how to increase their lifespan. An AGI mind could think about 1000 genes and all of their possible interactions, and then understand the causes of longevity and aging a lot better than any human can.

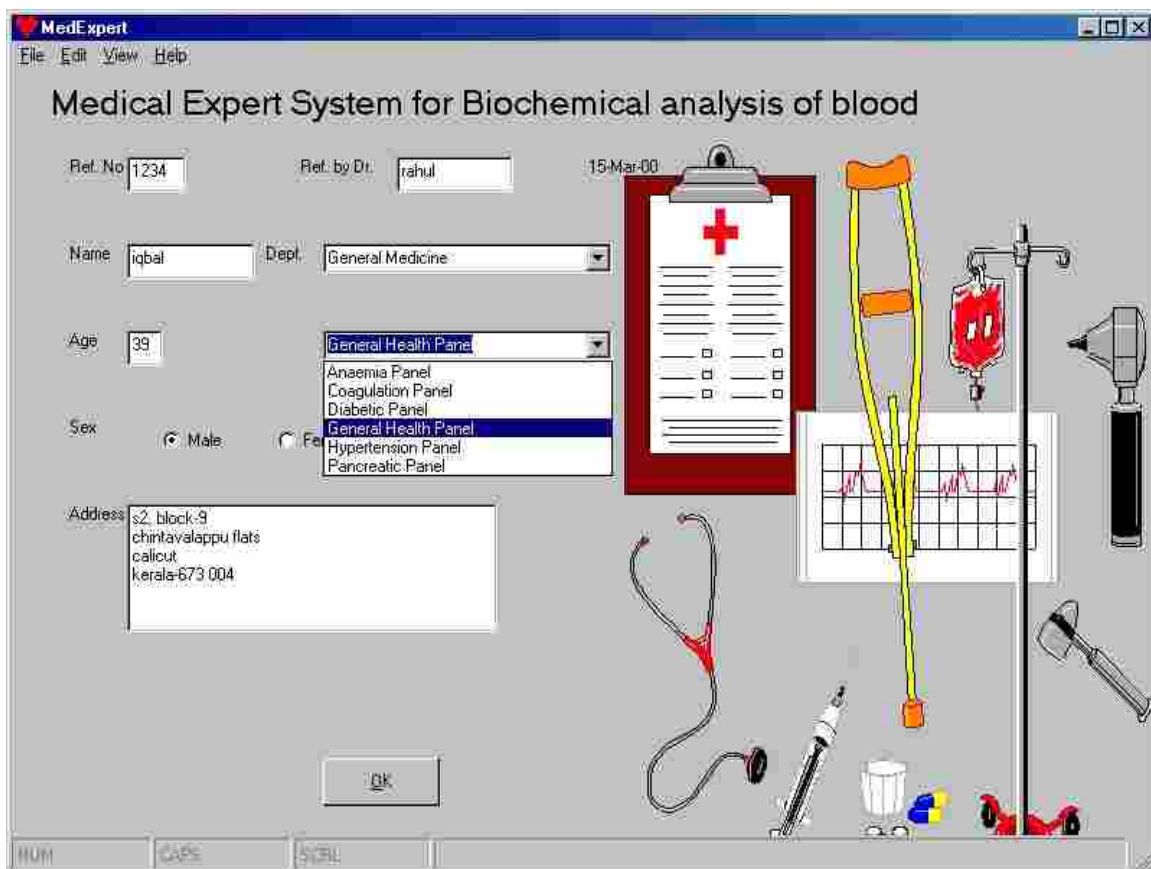


Figure 34: Screenshot of a current, commercial medical expert system, aimed at helping medical professionals to analyze blood work. Several studies have shown (narrow AI) medical expert systems to be more accurate diagnosticians than the average human doctor, but their adoption is still relatively weak due to resistance from the community of physicians.

<http://iqsoft.co.in/products.html>

Health care today, just like drug discovery and genomics research, is also largely a matter of trial and error. Medical malpractice has become widespread, as has the average physician's ignorance of modern medical literature. "Evidence based medicine" is a concept much discussed but erratically implemented. Simplistic "medical expert system" AI programs can provide better diagnoses than most doctors, analyzing the patient's answers to multiple-choice questions regarding symptoms. AGI doctors will obsolete human doctors as soon as AGI sensors and actuators can accurately observe a patient's physical state.

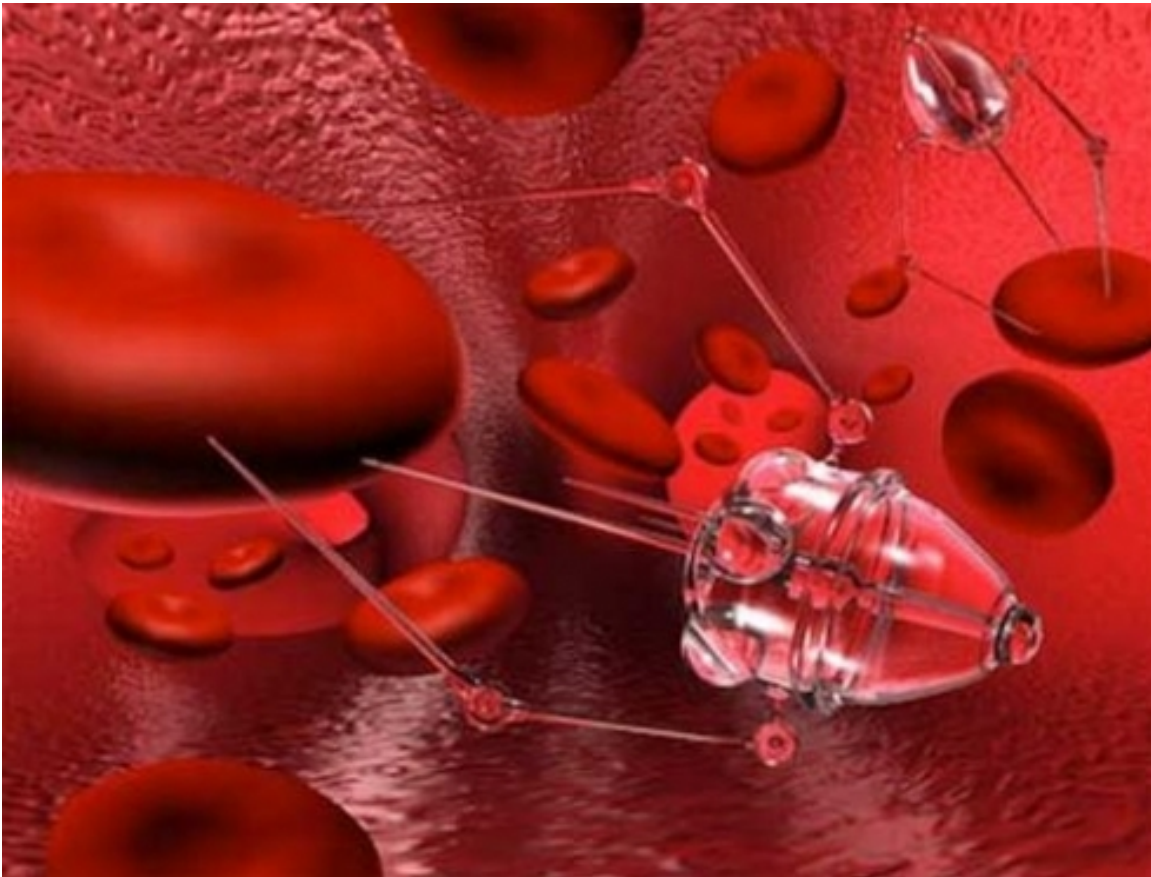


Figure 35: Hypothetical medical nanobot, zooming through the bloodstream and fixing cancerous cells. Ralph Merkle and others have thought through the physics and engineering of this sort of device in great detail. There seem no basic obstacles to building them – we just need some more advancement in our engineering practice. Of course AGI inventors, scientists and engineers with sensors and actuators at the nanoscale could advance this sort of development faster than humans. <http://www.gizmowatch.com/good-gadgets-technologies-hate-cancer.html>

The best human surgeons are amazing, synthesizing physical and mental ability in an elegant and powerful way. And yet, the human hand obviously it wasn't designed for performing surgical operations. Once robot hand technology develops further, it will far outperform the human hand at every kind of surgery. Picture a hand with tiny cameras inside each finger, and the ability to use a variable number of fingers, depending on the task. Primitive versions of this kind of technology already

exist, and are advancing year on year. Or better yet, picture a swarm of nanobots going into your bloodstream, fixing the problems at the nanoscale. Nanotech visionary Ralph Merkle has fleshed out the possibilities in this regard quite thoroughly.

Space



Figure 36: Special Purpose Dextrous Manipulator (Dextre), a robot deployed to build and service the International Space Station. Once robots achieve general intelligence, they are likely to lead the colonization of space. Not needing air, food or water, and operating comfortably in low temperatures, they are obviously far more suited to space than human beings.
<http://spaceflight.nasa.gov/gallery/images/shuttle/sts-123/html/s123e007088.html>

Space exploration is expensive and difficult mainly because of the human body's limitations. Our bodies are only meant for environments with air, water, and earth gravity. Current robots can't carry out complex tasks in space, so we're forced to choose between sending people (which is very expensive), or sending stupid and uncoordinated robots (which limits the information we can gather).

AGIs may develop technologies allowing humans to survive more comfortably in a spacecraft. There may be limits to how well this can be done, due to human psychology as well as physiology. Can people really be happy, to use David Bowie's terminology, "floating in a tin can" for decades or centuries on end? But to circumvent any psychological issues, human brains could be connected to virtual realities while floating in their tin cans through interstellar space, as has been explored in

numerous science fiction novels.

On the other hand, while humans are built for Earth, AGIs may be more comfortable in space than here on the home planet.C

omputers operate better in supercooled environments. Mining for processor materials should be easier in the asteroid belt; gravity is a nuisance if you don't need food or air; and solar power is more abundant in space. One likely scenario is that AGIs will colonize space while most humans remain on Earth where their bodies and minds are comfortable.

Once mind uploading technology is available, if a human wants to go into space, they can adopt a robot form for the expedition. Imagine becoming a robot and flying around freely in the vacuum. Or , given a sufficiently robust robot body, zooming through the center of the sun.

One of the greatest moments in cinema comes toward the end of *Blade Runner*, a film based on a Philip K. Dick novel, featuring genetically engineered artificial humans called “replicants.” The replicants have some superhuman capabilities, but deficiencies regarding empathy, and 6 year lifespans immutably fixed in advance by their creators. The replicant [Roy Batty](#) introspectively makes the following speech during a rain downpour, moments before his own preprogrammed death:

I've... seen things you people wouldn't believe... [laughs] Attack ships on fire off the [shoulder of Orion](#). I watched c-beams glitter in the dark near the Tannhäuser Gate. All those... moments... will be lost in time, like [coughs] tears... in... rain. Time... to die...

These lines, probably the most moving in all of science fiction cinema, were improvised on the spot by the actor Rutger Hauer based on an inferior version provided in the original script. Part of the emotional undertone of the speech, of course, has to do with the completely unnecessary nature of Roy's death. Why did he need to be programmed to die so soon in the first place?

But then, why do we humans need to be programmed to die so soon, either? We don't, of course. And by the time we are able to inspect the shoulder of Orion first hand, the blight of involuntary death will almost surely be lifted from our organic or digital descendants.

Telecommunications

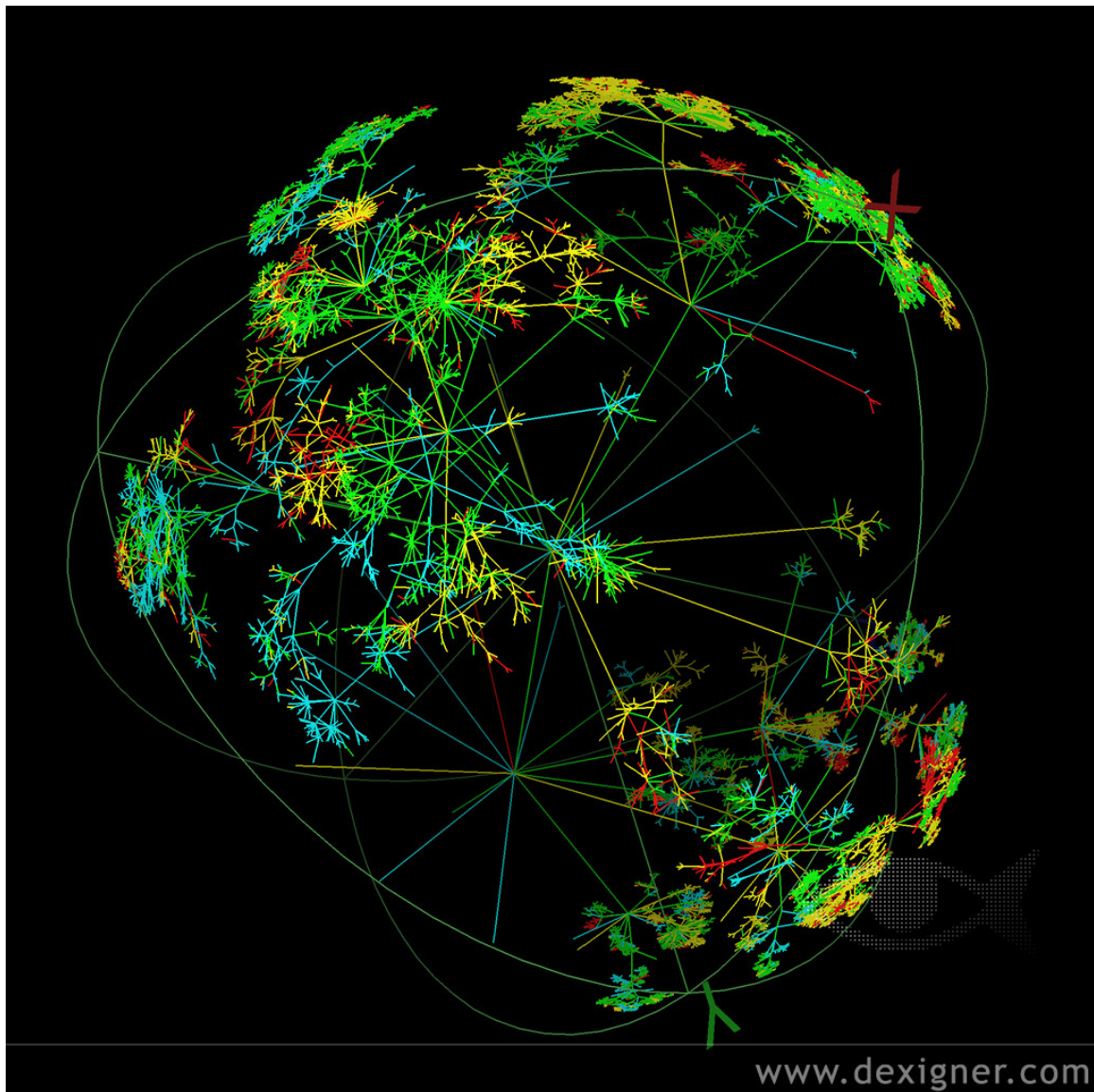


Figure 37: Appealingly laid-out graph of the Internet connections emanating from a single site in Herndon, Virginia. The notable thing about this picture is how biological – neuron-like or plant-like – the fractal branchings are. Coupled with the fact that the connections on the Net are always growing and changing, this really gets across the point that we are dealing with a dynamic, “living” system. What we have is a massive system like this, wrapping the globe in a constant pulse of information generated and consumed by humans and our software.

http://www.sdsc.edu/News%20Items/PR022008_moma.html

Telecommunications, while important to humans (try taking a teenage girl’s electronic communication devices away!!), will be crucial for digital AGI systems: They will be able to perceive data from all over the planet, and beyond, using telecommunication networks. Because of this, we can expect AGI systems to have a strong motivation to optimize telecommunications far beyond what humans have done.

Our current use of the available communication bandwidth is extremely crude; our algorithms for sending and receiving information are simplistic compared to what would be mathematically and physically possible, resulting in loads of redundant information being sent through the airwaves.

An AGI able to perceive a wider spectrum of electromagnetic radiation than humans could connect its general problem solving capability with automated equation-solvers applied to signal processing mathematics – resulting in new circuits for sending and receiving electromagnetic information, massively increasing the efficiency of new information passing through the airwaves. Humans would get faster cell phone and Internet connections; and AGIs would get faster connections between the different parts of their minds, boosting their intelligence and making them even better at optimizing telecommunication and other technologies.

The potential impact of AGI technology is so broad that it's difficult to think about. Where do you start, when thinking about something with the potential to change EVERYTHING? It seems almost too easy to run down the list of all existing industries, and point out the potential for advanced AGI to revolutionize every one of them. But that's simply the truth.

Creating AGI is, of course, not that easy. Even if writing the software code for the AGI system that finally “takes off” and displays human-level intelligence is only moderately difficult for the team that finally does it, this achievement will be building on a incredible mass of science and engineering which has been accumulating for all of human history. But once AGI is there, it will revolutionize all aspects of human existence just as surely as prior radical advances like language and civilization have done – and probably more so. There will surely be limitations, but there's no way for us to foresee those now.

As a human being in a human body, it's easy for me to think about the specific ways that AGI could change my own human life. With better engineered food or drugs, I wouldn't have to worry about getting overweight. Better biomedical science means I wouldn't have to worry about getting old and dying. I'm typing these words on a keyboard and viewing them on a screen, rather than just thinking them directly into some digital knowledge store; AGI scientists could change that through brain-computer interfacing. The car I drive to go to the supermarket could be replaced by a flying machine, if we had AGI to handle navigation and prevent collisions. Or the value of my work might be eliminated by AGI scientists. And so on.

Overall, the best “first approximation”, regarding the future impact of AGI, is to assume that AGI will be able to render the various limitations we now face obsolete, insofar as this is possible within the

“laws” of physics. To what extent future human or AGI scientists will reveal limitations in the “laws” of physics as currently understood, is hard to estimate. But even setting aside the likelihood that the laws of physics as currently understood are incorrect and too limited, the scope of possibilities extends so far beyond current human reality that it is difficult to think about in detail. Manipulation of matter at the subatomic scale to create femtotechnology; distributed minds spanning galaxies with orders of magnitude greater problem solving ability than humans; and so on. These sorts of things seem physically possible; and given the logic of exponential growth, it seems plausible that eventually descendants of the AGI systems humans build will get there. Given this scope of possibilities, the potential of AGIs to revolutionize the construction or pharmaceutical industries seems almost obvious and mundane.

The skeptic typically interrupts this kind of rosy-eyed projection with some comment like “Every technology has its limitations, right?”

But actually – intelligence isn’t just another technology. Intelligence is a fundamental capability—that’s the wonder of the “general” part of AGI. Every technology has its limitations, but part of the power of general intelligence is its capability to overcome limitations by formulating new technologies.

You can call this “magical thinking,” if you like. But remember Arthur C. Clarke’s insight: “Every sufficiently advanced technology seems like magic.” Absolutely. To my dog Crunchkin, I seem to do magic every day. He doesn’t fully understand the limitations of the magic I can do, but he knows I can do amazing stuff far beyond his capabilities and comprehension.



Figure 38: To my dog Crunchkin, cars and knives seem just about as magical as the workings of superhuman AGIs will seem to us. But, AGIs may grow much further beyond our intelligence than we are beyond Crunchkin's. (And if Crunchkin is uploaded one day and has his digital brain enhanced a bit, he may be able to read this caption and appreciate this picture! Hello there, Future Crunchkin! – just a little greeting from Past Ben!!)

AGI Will Vastly Transcend Humanity

I'm not the world's best listener, but I'm a lot better than when I was 25. I have listened, really listened to what the skeptics have to say. I have thought about their arguments rationally, and tried to empathize with them emotionally.

But I still think they're wrong.

AGI is going to be massive – so big as to go beyond “big” and every other human concept. Revolutionizing every area of human pursuit won't be the end of it. That will just be Act One. Just as differential calculus and Shakespeare go beyond Crunchkin's, my dog's comprehension, the activities and methods of advanced AGIs will transcend our own.

Consider: Apes can't understand much of the human world, even though they're smarter than dogs, share 95% of our DNA, look a lot like us, and in the grand scheme, they're almost as smart as us.

Or, going in the other direction, swap the apes for cockroaches. How many of our interesting achievements can cockroaches appreciate?

Or, consider from the bacterial perspective. Bacteria are completely unaware of human beings. They may inhabit our bodies, but they don't *understand* this; they don't know that we speak, or act; and they don't understand anything we do, beyond responding to biochemical changes in their environment.

As AGIs progress, humans will pass through similar stages of incomprehension. We will be the apes, then the roaches, and finally the bacteria, lost in our trivial pursuits beneath vastly more intelligent beings operating on planes beyond our understanding.

How fast will the transitions between these stages be? Potentially, they could be literally very fast, but subjectively a lot slower. As we become more intelligent, and our thought processes speed up, conversely our world will appear to slow down.

The Psychological Singularity

It's natural to think about the Singularity in terms of advancing technology and increasing scientific understanding – after all, the proximal cause of the impending Singularity is precisely the explosion of science and tech.

But yet, the Singularity isn't just going to be a technological event— it will be a psychological and social event, too. With fascinating implications for subjective experience, and what it feels like to be a mind, the Singularity will give us unprecedented insight into what it means to be human, and what it means to go beyond humanity

The Singularity will wreak havoc with the various psychological illusions that characterize our inner worlds today, and replace them with new mental constructs that we can't currently conceive in any detail. The infusion of vastly greater intelligence into the world isn't just going to transform the gadgets at our disposal; it's going to transform the way we *think*, the way we *are*, inside our heads, moment by moment.

What do I mean by “psychological illusions”? I mean that the ordinary experience of being human is based on a number of assumptions that fail to reflect an accurate understanding of ourselves as sentient beings. Instead, we rely on convenient illusions— assumptions we've evolved to make— that fail to reflect how our minds work. These illusions will disappear as we begin to interface with AGIs, transforming our minds and bodies with advanced technologies. I'm not talking about something obscure here – I'm talking about basic things like the feelings of self and free will take for granted in our everyday mental existence.

Post-Singularity, most likely, our minds will no longer be tied to individual bodies, but free to move among many, changing and adapting accordingly. We will have the option of sharing our thoughts more freely, and occupying various bodies collectively. We will be able to sense and act, both globally and locally (e.g. absorbing data from satellites and cameras all over the world). We will be able to create virtual worlds—which we can exist and interact in—based on our whims and desires. And all this is going to have huge, transformative impact on the inner workings and feelings of our minds.

The Illusion of the Self

Perhaps the core illusion of everyday human experience is the *self*. The self I have, the idea I have of “this guy Ben Goertzel who writes books on AGI and works on AI software, who has a wife and three kids, and a village house in Hong Kong” – this idea is not really the psychological and biological system that we call “Ben Goertzel.” It's a model that has arisen within my brain, possessing some accuracies and some distortions. It's a model that thinks of itself as much more real and accurate than it truly is. The way I think about myself – the self-model I have built -- guides what I do every day, in ways that are sometimes productive and sometimes not.

For instance my model of myself as “a guy who works on AGI and AI” might potentially prevent me from exploring other possibilities that arise, even when these other possibilities would help me achieve my goals better. What if an opportunity arises for me to help create a new form of nanotechnology, in

a way that exploits my mathematical background, and has the potential to create a massive profit, that could then fund AGI and all sorts of other futuristic work? Then I might do better to turn to this nanotech work for a while, setting AGI aside. But my self-model as an “AGI guy” would potentially prevent me from doing so.

But that’s somewhat a nonrepresentative example. The self’s main role isn’t in big life decisions, it’s in the small everyday choices that make up our ordinary lives. Using my own self as an example again (it’s the self I know best), I tend to have a model of myself as a guy who has all the answers. Being honest, I have to admit that my ego is wrapped up with this to some degree. I try to avoid this sort of commonplace psychological trap, but I don’t always succeed, and now and then I fall into the pattern of feeling good about myself because I am The Guy Who Has All the Answers. It’s easy to see the kind of consequence this has. Nobody really has all the answers; so what if somebody raises a point that I don’t actually know the answer to? Rather than acknowledging my own cluelessness on that point, and thus opening myself to gathering new information, I might instead try to convince myself I knew SOME kind of plausible answer. In that case, I might be so busy searching for a half-assed answer and trying to convince others it’s whole-assed, that I wouldn’t even hear the worthwhile directions toward an answer being presented by others. Due to not hearing other peoples’ potential answers in this context, I might even draw the conclusion that nobody else had had anything useful to say – thus getting more reinforcement for my belief that I always have better answers than anybody ... Thus strengthening the Guy Who Has All The Answers portion of my self-model...

This particular example – feeling like one has all the best answers, even when one doesn’t – is a problem I identified with my own personal self-model quite some time ago. I’ve consciously tried to correct this trait of mine, and have succeeded to some extent. I’m still a bit of a know-it-all, but I’m a far better listener than I used to be; and I believe I’m more rational about the extents of my knowledge and intuition than I was in the past (although, this belief is probably not wholly accurate either; it’s just part of my self-model...).

But anyway, my own human personality quirks aren’t really the point here. My point is that human personality is woven of this *kind* of phenomenon: Self-modeling that feels good for one or another reason, but isn’t quite accurate, and that drives behavior in ways that are often self-reinforcing and not always productive in terms of human goals or human happiness. The personal examples I gave above are just two data points. In countless similar ways, our self-models guide and restrict us as we go about

our lives. Psychologists have dissected this kind of phenomenon in great detail; particularly relevant here is a 2007 book by Mark Leary called *The Curse of the Self: Self-Awareness, Egotism and the Quality of Human Life*.

Still, though, the title of Leary's excellent book notwithstanding, self is both a curse and a blessing. Our selves mislead and distort our thinking and our lives. But without them, where would we be? We do need to model our minds and bodies, in order to understand our place in the world and to plan our actions. The problem comes when we get emotionally attached to aspects of our self-model that are only crude approximations. The emotional attachment stops the approximations from getting improved. And the various errors and emotions involved in the self become systematically interdependent, until the self is a whole self-organizing system on its own, with some of its own life independent of – and guiding – the organism it's supposed to be modeling.

Philosopher Thomas Metzinger has called this sort of self-model I'm talking about *the phenomenal self*, using a term from the philosopher Immanuel Kant, to whom "phenomena" were mere surface appearances, lacking in realness. Kant conceived of the term "noumena," referring to another realm, the "absolute real" that we could never perceive, but which actually underlies what we experience. Metzinger's view is that self is not noumenal, it's phenomenal – he's looking at the self, not as some underlying reality, but as something mind builds. His book *Being No One* digs deep into the way the brain builds the phenomenal self, carefully analyzing various cases of brain dysfunction that cause people to build improper selves in various ways. It's a long book but an awesome one; I'd strongly recommend it. If you don't like reading huge neurophilosophical tomes, you can find some of his lectures on the Internet as well.

Is this phenomenal self, this largely-illusory self-model that we humans habitually create, truly a necessity for any intelligent system? Well, some kind of self-model is obviously of value for any agent achieving complex goals in a complex world. But the pattern of self-reinforcing inaccuracy and emotional attachment that characterizes the human self, seems in large part an artifact of the particular human mind architecture, rather than a necessary aspect of intelligence.

Ultimately, one asks: Why would an AGI mind have to systematically take itself to be something that it isn't, and emotionally cling to its errors? Why could the AGI not base its self-perception on an accurate modeling of its own mind? When it makes errors in self-perception, why couldn't it just correct them based on its experience, rather than reinforce them due to emotional attachment?

One could build an AGI possessing human-like self-modeling dynamics, but it's not clear why one would want to, except for the scientific goal of studying human-like minds and better understanding their strengths and weaknesses. I believe one could more easily build AGIs displaying more rational approaches to self-modeling, free of egomania and the various other pathologies that characterize human self and play such a large role in human personality and society.

This brings us to one of the biggest, and most emotionally charged, modeling errors present in most modern humans' self-models: a misunderstanding of how independent we actually are from others and our environment.

One of humanity's deepest flaws, and one especially prevalent in modern Western society, is our tendency to individuate ourselves to an extreme degree.

We often fall into a habit of thinking that our minds are separate from our bodies, reducing the latter to objects that our minds control—a distorted form of self-perception. In fact, our minds are intimately entwined with our bodies; together they form webs of interactive feedback loops, linking us to people and the surrounding world. Cognitive scientists refer to this as an embodied or extended mind – a meta-mind that overlaps with your own body, as well as the people, tools and objects with which you interact.

In the present order of things, our self-models only indirectly and incompletely reflect the embodied and extended nature of the mind. We tend regard ourselves as individually autonomous minds, interacting with the separately defined objects, bodies and minds in our environment. This tendency is probably most extreme in America, the country where I spent most of my life. American individualism is fantastic in some ways – but it's also in large part illusory. All of us are far more interdependent with, and cross-connected with, the rest of the world than we realize.

Cognitive scientist Andy Clark has written a lot about the extended nature of the human mind – I'd recommend his books *Being There* and *Supersizing the Mind* very highly. Though when my wife and I visited him in Edinburgh in 2012, we were surprised to find that he had somewhat moved beyond the extended mind as an area of research. He figured he had already made his point there – and he was more interested, as of 2012, at thinking about the relationship between consciousness and deep learning (a kind of AI technology, thought by some to be a path to AGI and a model of the human brain; I'll discuss that a bit later). He was intrigued by some recent successes with deep learning technology, and absolutely did not consider the extended nature of the human mind an objection to building AGI

systems. Rather, he was intrigued by the attempts by my own team and others to make AGI work via embodying complex software systems in robots. My notion of making a robot toddler fascinated him, though he worried about how the robot would compensate for the lack of the cognitive guidance human children get from their genetic endowment. (I'll say more about that later, too.)

Anyway – this whole embodied/extended mind thing is going to seem very different once advanced technology enables a more flexible sort of relationship between minds and bodies. Once we can port our minds between different bodies—and radically alter our bodies—we'll sense things using means other than our typical body senses (via connecting to sensors around the globe, or directly reading certain thoughts and sensations from other peoples' brains or from AGI minds). The strict connection we now feel between our self and a particular body will be a thing of the past.

Of course, you can't just put a mind in a different body and have it unchanged – the body is part of the mind, and so when you take the same high-level cognitive system and use it to control a dolphin-like body instead of a human-like body, for example, this will inevitably result in the high-level cognitive system becoming dolphinized to some extent. Because in reality the high-level cognitive system is only part of the overall mind of the “high level cognitive system + human-like body” or “high level cognitive system + dolphin-like body” mind system, and is in constant complex feedback with whatever body system it's connected to.

Once we can port our high-level cognitive systems from one body to another, what happens to the self? We will still model ourselves, but in a more fluid way, reflecting the rich interconnectivity of our mind-patterns with various sensors, controllers and other minds and systems. Each of us will feel less like an individual interacting with a world that contains others, and more like a dynamic self-organizing cloud of mind-stuff, interacting in a complex dance with other clouds of mind-stuff, and with various physical systems, on various scales. Doubtless various problems and limitations will arise in connection with this new way of experiencing ourselves, each other and the world, but they may have little in common with the things that trouble us now.

The Illusion of Free Will

Another complex, illusion-ridden characterizing everyday human life is *free will*, perhaps the most confusing concept I've ever heard of. Free will, as commonly conceived, is a logically ill-formed and almost senseless concept, yet rings loud intuitive bells for most people. The loss of this peculiar aspect of our self-models would change our minds considerably. Yet it seems unlikely to me that post-

Singularity humans, or advanced AGIs, will think about themselves using anything resembling our current concept of “free will.”

When one of us modern humans reaches a decision, we stubbornly persist in believing that we're engaging in some kind of “free choice.” A choice that is not determined by anything outside ourselves, nor by mere physical dynamics within ourselves; and certainly not just chosen at random. It feels like it is somehow chose by ourselves, by our inner choosing facility, by our mind's freely acting choice process! However, there's plenty of data debunking this intuitive, “folk psychology” notion.

Cognitive neuroscientists have set up many experiments, showing cases where a person feels like they are consciously deciding and willing something at a certain point in time; however, their unconscious brain has already taken measurable steps in accordance with the decision a little bit earlier (say, half a second earlier).

For example, Michael Gazzaniga's classic split-brain experiments explore what he calls the “confabulative” of free will, via probing the experience of people who have had the two halves of their brain separated, in order to prevent their brains from being destroyed by severe epilepsy. In these unfortunate cases, each half of the brain maintains its own self, its own stream of experience, and its own free will – thus providing the cognitive neuroscientist with a wealth of avenue for fascinating experiments.

In one experiment, a split brain subject's left eye received a command to stand. The person stood – and then, when asked why she stood up, she responded (using the language center of her left hemisphere) that she wanted a soda. In another experiment, when the left and right hemispheres were each asked to pick an appropriate picture to accord with an image flashed only to that hemisphere, the left selected a chicken to match the chicken claw in the picture it saw, while the right hemisphere correctly chose a shovel to remove the snow it saw. When asked why the person chose those images, he replied that the claw was for the chicken, and the shovel was to clean out the shed.

Gazzaniga's experiments demonstrate that, even when there is a clear external cause for some human action, it is possible for the human to sincerely and thoroughly believe that the cause was some completely internal decision that they took. The left hemisphere of a split brain has no experience of stimuli delivered exclusively to the right hemisphere (e.g. through the left eye). However, the left hemisphere has such a strong motivation to create explanations that it will make up “free will stories” corresponding to behaviors initiated by the isolated right hemisphere.

Daniel Wegner's book *The Illusion of Conscious Will* gives a pretty good summary of the contemporary body of knowledge regarding free will and the brain. The book *Neurophilosophy of Free Will*, by Henrik Walter and Cynthia Klor, tries to dig deeper and figure out what kind of "free will like" capability could possibly be compatible with science and common sense – I recommend it highly.

While neuroscience has made the limitations of the free will concept crystal clear and essentially indisputable, the same basic problems with the concept were noticed by philosophers long before anyone knew what a neuron was. When Nietzsche called consciousness an army commander who, after the fact, takes credit for the actions of his troops—he definitely had free will in mind. In large part, free will is a story that the brain/mind tells itself after the unconscious mind has already made a decision, justifying the unconscious decision already made. This storytelling process may often be quite useful, as it may feed into the unconscious and help guide future unconscious decision-making, and future conscious story-telling. However useful they may be, though, our stories about how we decide our actions are generally not accurate.

Now fast forward a bit to the future, when there are advanced AGI minds, or radically improved human brains. Imagine a mind capable of monitoring the activities in its brain as it thinks, and the interactions of these activities with other things in the world. Such a mind would be able to see—vividly and in real-time—a great many of the actual dynamic processes involved in its "choices." Such a mind would be unlikely to maintain a current-human-mind-like illusion of free will. Between the absence of the illusion of free will, and the existence of a more accurate self-model not including an irrational sense of one's own independence and autonomy, such a mind will have a very different kind of experience than we current humans.

Projecting from my current base of knowledge to the future of post-Singularity minds makes me feel a bit like a cockroach trying to predict the future of quantum computing. But nevertheless, it is interesting to extrapolate as best as possible from what we know now, with the understanding that future discoveries are likely to revolutionize our current understanding. My best present projection is that, as mind moves on from its current "legacy human" form into an era of AGIs, uploaded humans and enhanced human brains, subjective experience will move beyond will and emotionally-attached self, and we will have a world of beings that view themselves as fluid clouds of mind-patterns, evolving dynamically in close coupling with other systems, in ways that happen to influence various actions.

What will it be like to be such a mind? I have only a vague sense at present. But I am definitely curious.

And I am very aware that I have barely a clue what new forms of mental organization, and new states of consciousness, will take the place of these historical illusions to which we've become so habituated.

But one thing I am confident of is: The psychological and social aspects of the Singularity will be even more intriguing and dramatic than the technological and scientific ones.

Now is the Time for AGI

Why do I think now is the time for AGI? The AI field has been around for 50 years, yet no one has created a human-level thinking machine yet. Why do I think we can do it now, or in the near future?

The crux of my answer is pretty simple and obvious: Computer hardware, software, and cognitive science are advancing tremendously and exponentially. None of these areas, considered on its own, would be enough to make a human-level AGI possible, but together, they make the prospects for AGI look extremely promising.

Counterbalancing these exciting trends, one has the unfortunate fact that AGI and some other Singularity-enabling technologies (life extension, strong nanotechnology, and brain-computer interfacing) are fairly underfunded.

But I believe the solution to this unfortunate fact is not far off. Recall how Sputnik, the first Soviet satellite launched into space, spurred the US to develop its own space program. With AGI, we could soon witness a similar type of event, massively increasing the amount of interest and attention devoted to AGI, and placing the goal of a human-level thinking machine easily within reach.

Advances in Computer Hardware, Software, and Cognitive Science

Comparing the information-processing capability of the human brain to that of a digital computer is a tricky matter. The two systems are very different in nature.

From a physics perspective, both computers and brains are pretty pathetic information processing systems. Neither of these systems utilizes more than a miniscule fraction of the computing power implicit in the particles that compose them according to the laws of physics as presently understood. So in the grand scheme, both digital computers and brains are very, very inefficient information processors – but the kinds of inefficiencies and efficiencies they manifest are different in various ways. A digital computer is very good at dividing large integers, and searching huge relational or graph databases (for example); a brain is very good at regulating hormone levels and identifying contours in visual fields (for example). Digital computers are better at running Google or Citibank’s data networks than human brains would be; human brains are better customized for controlling human bodies than digital computers are. Neither digital computers nor brains are much good at solving the equations for quantum many-body problems – but some kind of future quantum computer might be.

Still, even given all this uncertainty and complexity, there are some sensible ways to compare digital computers to human brains. One can ask, for instance, whether a digital computer is capable of simulating a human brain's intelligence-relevant functions. Note that this doesn't necessarily imply simulation of every particle or every atom or molecular inside a human brain. Nearly all neurobiologists these days think that, in order to simulate a human brain's intelligent functions, it would be enough to simulate the ways its major types of cells work (its neurons and glia and maybe a few other types of cells), and the molecules these cells pass between each other.

If we ask whether digital computers can simulate brains in this sense, the answer at the present time is no. But then Moore's Law rears its inexorable head. If Moore's Law and its relatives hold up, then in 15-25 years or so (depending on your detailed analysis of the brain), we WILL have digital computers that can match the computing power of the human brain.

Whether or not having "human brain level" computing power is critical for having human-level AGI is a different question. My own currently preferred AGI approach, as I'll describe a little later, takes the brain only as loose inspiration; and I suspect that my approach could be used to create human-level AGI with far less computing power than would be needed to simulate a human brain, or to achieve AGI with a closely brain-like architecture.

But still, whether one is doing brain simulation or OpenCog, there is little doubt that better and better computers are going to make AGI easier to achieve. So for those of us who are passionate about AGI, it's good news that the computer hardware industry so reliably delivers better and better computers each year, and is showing no signs of slowing down anytime soon. As creating increasingly powerful computing cores for processor chips becomes more difficult, chip makers are shifting to a multi-core approach – with the end result of computers that keep getting faster. As the challenges keep coming, so do the innovations. There are enough economically valuable applications for better computing hardware, that the hardware companies have copious incentive to keep on innovating.

Hardware advances are easier to understand and hence make more headlines, but computer science—the theory of computer software and hardware—has also advanced tremendously in recent decades. Current algorithms (computerized problem-solving methods), if implemented on decades-old computers, proceed dozens or hundreds of times faster than the algorithms from the same period. Today’s algorithms are more advanced than those from a few decades back, despite their core ideas being the same, because of their fine-tuned details. Having more powerful computers has been very helpful for tuning these details, closing the loop between hardware and algorithms.

The same basic story holds true for software technology, as for hardware and basic computer science algorithms. When one writes a program today, even though the algorithms and programming paradigms in use are roughly the same as two decades ago, the sophistication of the software libraries available makes the modern code more efficient and less prone to errors. For example, in the world of C++, the programming language used for the bulk of the *OpenCog* AGI project my colleagues and I, are working on, the STL and Boost code libraries wrap up incredibly complex algorithmic functionality.

Alongside the recent advances in computing hardware and software, there has also been dramatic advancement in our understanding of the human mind as an information processing system.

Cognitive science – the interdisciplinary study of the mind – first came together as an integrated discipline in the 1980s. By now it qualifies as an impressive, coherent body of knowledge. We understand the mind a lot better now than we did a few decades ago – and I’ll outline some of the understanding we’ve achieved a little later in this book.

We have also understood a lot more about the brain, even though our understanding of neuroscience is still frustratingly limited. Functional Magnetic Resonance Imaging (fMRI) and other brain imaging studies have told us about the specific functions of parts of the brain; and our models of neurons and neural networks in the brain are far more accurate and sophisticated than they used to be. We understand the general classes of algorithms and forms of knowledge representation that the brain is most likely to use, even though there are many details that are still being figured out.

There is no consensus yet about the “big picture” of how the brain gives rise to the mind, though. I have my own ideas, which I’ll tell you a little later, but these have to be rated a bit speculative just like everybody else’s....

All in all, by piecing together knowledge from cognitive science, neuroscience and computer science, the adventurous contemporary scientist can form a decent idea of the mind's different aspects and how they connect with each other – of what cognitive structures and processes conspire to create human-level intelligence.

Oh, and let's not overlook advances in embodiment, robotics, and virtual worlds, which drastically simplify the task of giving AGI systems something to do, once they've been designed and built.

All these advances, taken together, make the creation of human-level AGI feasible, *now*.

Our grasp of the many issues involved is far from perfect, but we do have a lot to go on. I believe it is feasible, at the present time, to build an AGI architecture inspired by cognitive science and neuroscience, while leveraging the best of modern computer science for implementation in the best available embodiments and environments.

I admit it is not *obvious* that this is feasible to do today, as opposed to one or two decades from now. But I have dug into the matter in great detail, and I came away from this study convinced. I will tell you a fair bit more about my perspective and understanding as these pages unfold.

The Absurd Underfunding of Singularity-Enabling Technologies

There's at least one sizeable problem marring the rosy picture I've sketched, though. Yeah, taken together, the recent advances in computer science, hardware, and cognitive science support the notion that a human-level AGI may soon become a reality. But on the other hand, AGI has a serious money problem. Right now AGI, like many other exciting Singularity-enabling technologies (life extension, strong nanotechnology, and brain-computer interfacing, for example), is severely underfunded, compared to what would be needed to really bring it to fruition in a straight forward way.

You probably won't be surprised to hear that I find the relative marginalization of AGI research quite peculiar, and rather annoying. Rationally speaking, it seems to me that AI should be a major industry like computer chips, prescription pills, and jet engines—with AGI at the forefront. Instead, AI is a minor academic sub-discipline, and AGI is a marginalized sub-sub-discipline. The world's biggest AGI projects are ragtag groups of grad students or entrepreneurs, generally scraping by on funding obtained in the context of narrow-AI application projects.

I shouldn't grouse too badly about AGI funding, since my own AGI project OpenCog is one of the few on the planet that actually has SOME funding to work toward its AGI goals. But still, our funding is far less than we need to really move forward rapidly toward AGI, and it's far from guaranteed to continue indefinitely. All of us in the AGI field, even those of use lucky to have a little funding for our work, are constantly aware of how much faster we could move if AGI were funded 1/100 as much as neuroscience, or 1/10 as much as application-specific narrow AI.

Humanity stands to gain tremendously from progress in AGI research and engineering. Imagine the good that will be done by allowing advanced AGI systems to take over routine, tedious human jobs, along with positions that we are biologically or cognitively unsuited for. Seen in this light, AGI research has amazing potential in terms of both economic and social benefits. Yes, there are potential risks to developing advanced AGI, as everyone knows from watching SF movies (and I'll talk about those more toward the end of the book), but this kind of possible risk hasn't prevented society from massively funding various other areas of science and technology.

But the marginalization of AGI doesn't seem so weird when you think about how life extension research is marginalized too. After all, almost everyone, if you asked them before they got too old and their mind and body started to degenerate too badly, would vote for a longer life over a shorter one. Yet the amount of research currently directly focused on extending human lifespan is minimal – not because scientists are uninterested, but because commercial, government and private research funding sources prefer to focus their dollars elsewhere.

There's far more money going into research on heart disease and cancer than radical life extension. Such diseases only impact a small number of people; however, unless we stumble upon a cure for mortality, nearly all of us will almost surely die. The economic cost of losing brilliant, highly skilled people to death is certainly much higher than the toll exacted by illnesses like heart disease and cancer.

There are over a thousand billionaires on the planet, and tens of thousands of people with an individual

net worth of \$100 million or more, are alive today. Why don't more of them devote 10-20 percent of their wealth to the creation of medical longevity therapies that will let them enjoy their riches even longer, over an extended lifespan?

Part of the problem here is surely just the chronic underfunding of research in general in modern society. Scientific research has offered the world so much in the last century, yet we devote more resources to things like spectator sports, pop music, and movie star worship. Let alone the massive resource spent on warfare!

Also, society often decides which research to fund on the basis of shortsighted or arbitrary historical reasons.

Medical research focused on specific diseases, no matter how speculative, receives more funding than research aimed at understanding the fundamentals of biological organisms. Yet, more fundamental research will help us radically improve human health – and probably lead to cures for diseases faster than poorly-directed disease-specific research. Biology researchers work around this by finding ways to spin their fundamental research as being relevant to the specific diseases that are popular among funding sources. Want to study the cell cycle? Pitch it as cancer research! Want to study retroviruses? Pitch it as AIDS research! Neurotransmitter function? Call it Alzheimers research. Etc. This sort of subterfuge often works, but adds complexity and waste to the research process.

Physics research is an interesting anomaly – even the most abstract theoretical physics research, wildly disconnected from any experimental evidence, gets funded well compared to many other disciplines. And the appetite of governments to spend billions of dollars on particle accelerators, without any obvious near-term economic value, is impressive. The underlying reason for this persistent fostering of monetary love upon physics research seems to be the atom bomb. To put it simply, my impression is that: Since the atom bomb was the product of obscure-looking theoretical physics, funders believe that funding more obscure-looking theoretical physics will yield more amazing, practical advances. Indeed, this could happen. This is not entirely foolish. But the practical payoff of expensive particle accelerators has been rather low in recent decades, and one suspects that greater bang for the buck could have been obtained via different research choices. Physicists discovered the Higgs particle, which is great, but I'd rather have a thinking machine or an immortality pill. I'm definitely not arguing against funding physics, though, as I think society should lavishly fund all the sciences

One might argue that disease-focused medical research and physics research have received massively

more funding than AI because AI hasn't delivered. But the truth is, AI *has* delivered. From Google's AI-powered search, to the military's AI-powered missiles, planning and scheduling systems, to the AI trading systems flourishing on Wall Street and powering non-player characters in video games. The AI field hasn't yet produced powerful AGI, and it hasn't yet produced any easy way to kill large numbers of people, but it has produced a lot. Physics hasn't yet produced a radical new energy source (fusion has been a bit of a dud so far), and medical research hasn't yet cured cancer (in spite of huge expenditure), yet society happily funds these fields to keep on trying, based on their other achievements.

You could argue that the reason AGI and life extension and other Singularity-oriented research are relatively poorly funded is that they are speculative and risky, where physics and medical research are not. Yeah, it seems clear based on current physics that AGI and radical life extension will be possible eventually – but how long will “eventually” be? In this view, part of the reason why life extension, AGI, strong nanotechnology and so forth are relatively starved for funds, is the blight of short-term thinking, which exists in modern societies on both the individual and organizational level. I think there is some truth to this analysis. Short term thinking is disturbingly prevalent. It is certainly true that, in the context of both research funding agencies and corporations, the folks in charge of dispensing dollars have increasingly fallen into the trap of questing quick returns, much to the detriment of fields that require a long-term perspective. However, medical and physics research are quite speculative and risky, too, but still much better funded than, say AGI or life extension, or Drexlerian nanotech.

It's easy to get frustrated by the shortsightedness of the world's current economic system; but I'm incredibly glad to live now, rather than, say, in the prehistoric era, when the world rolled by with ceaseless monotony over generations. In spite of the lack of explicit AGI and Singularity awareness on the part of political and corporate leaders and the super-wealthy, technological and scientific progress zooms forward anyway.

Also, I'm happy to live in an era where a backup plan to personal death is possible. To cover the case where the development of both AGI and life extension takes longer than I'm thinking -- due to funding issues or unforeseen scientific problems -- and the worst happens to me before these technologies emerge, I am signed up for cryonics via *Alcor*. So, unless I die in an explosion or some other destructive way, my body will hopefully be frozen in liquid nitrogen, then reanimated by my descendants, or other curious people or robots, after the Singularity.

Still, cryonics is a last ditch effort. “Better frozen than rotten,” is what I say when people asked me why I’m signed up for cryonics. But also: Better living than frozen! As frustrated as I get sometimes by the low priority society assigns to some of the amazing possibilities that are right at our fingertips, I’m hopeful about the possibility of advanced AGI getting created during my lifetime. Sputnik, the first Russian satellite launched into space, was a catalyst for the development of the US space program. After Sputnik, space technology development raced right along. Why can’t a similar type of event happen with AGI?

AGI Sputnik

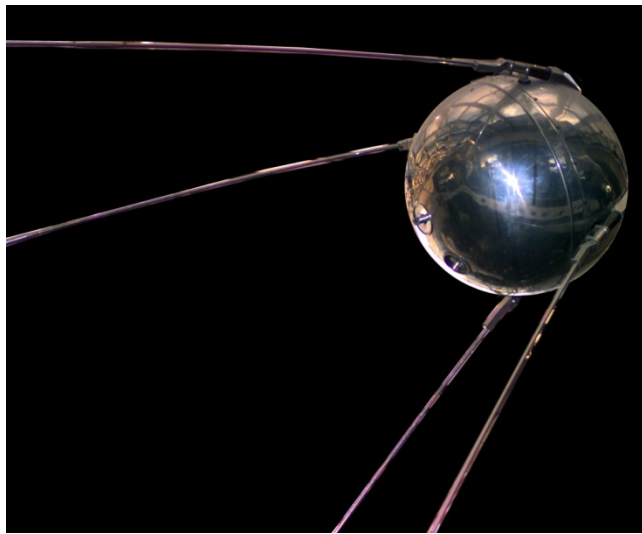


Figure 39: Sputnik, the first-ever satellite launched by humans, and the stimulus for the international Space Race...
https://en.wikipedia.org/wiki/File:Sputnik_asm.jpg

Sputnik was an awesome achievement on its own, but also a wakeup call to the rest of the world. People saw Sputnik and they began to ponder the implications of launching objects into space, and the US, spurred by the Cold War, and its struggle with the soviets for global supremacy, developed its own space program in response.

Similarly, once a certain level of development is reached in AGI, the rest of the world is going to wake up, take notice and exclaim, “***Holy smokes, it really seems AGI is possible! AGI may not have changed the world yet, but then again, neither did Sputnik right away; it’s only the beginning.***”

Once the possibility of AGI is really staring everyone in the face, AGI development is going to sweep the world. It's going to become huge in the way that military development, computer chip engineering and medical research are now.

What would it take to create this kind of AGI Sputnik event? The trigger might not be the most interesting or noteworthy technological achievement. Nor even the most useful. What it will be, though, is something that conveys in simple and universal terms that AGI is capable of reaching human-level intelligence.

There are a lot of possibilities for the AGI Sputnik. One possibility that I’m thinking about a lot lately is a robot toddler. One that walks, talks, interacts with you, and qualitatively has roughly the same intelligence as a human toddler. Presented with a robo-tyke like this, I think the world would wake up to AGI’s imminence and power.

If the world is presented with a really compelling robot toddler before any other AGI Sputnik like achievement, what kind of reaction will we see? I think we’ll see a combination of a *scientific* reaction to the technical achievement, along with a *gut feeling* type response to seeing something that looks, acts, learns, moves and communicates like a little human, but isn’t necessarily exactly like one. It may have a plastic body; it may talk in strange tones; and it may know more than any human toddler about some things and lack common sense in other respects. Yet, if it feels to us like a young, intelligent human child... If leading computer scientists and AI gurus agree, “Yeah, this thing is really learning and thinking. It's not just some trick” – then we're going to have a Sputnik moment in the history of AGI.

After the AGI Sputnik moment – whether it's caused by a robot toddler or something else – incredible amounts of financial resources and attention will pour into Artificial General Intelligence research. Barring some sort of intervening disaster, it won't be long from then till some sort of Singularity.

PART TWO

WHAT AN AGI IS AND HOW TO BUILD ONE

In this second of the three Parts of this book, I'll get a bit more nitty-gritty, and tell you something about how I think advanced Artificial General Intelligence can actually be achieved. Of course, I won't go into the full details – I've written another book on that, Building Better Minds (TITLE STILL UNCERTAIN ACTUALLY), which is 1000 pages long and very technical. For readers with a modicum of scientific background, this Part will be easy going.

If your science background is weaker, on the other hand, you may find some of the ideas here a bit confusing or opaque. If so, I'd encourage you to skim past those sections, getting as much as you can from them, and move forward to the third Part of the book, which is not technical at all, and deals with the application of AGI to extend human life, the variety of possible post-human AGI and hybrid AGI/human minds, and the ethical and cosmic meaning of all this.

What is Artificial General Intelligence, Really?

I've already given you a rough definition of Artificial General Intelligence: A system—a computer program, robot, or machine— that displays the same rough sort of general intelligence of humans. General intelligence is intelligence not tied to a very specific set of tasks, which possesses the ability to take a broad view, or generalize what has been learned. The best judge of a system's general intelligence is its practical ability to achieve a variety of complex goals in a variety of complex environments. A general intelligence should be able understand new things it encounters, acquire knowledge in one domain and apply it towards another. It should possess its own evolving, intuitive understanding of itself, others, and the world.

All this makes sense qualitatively, but it's not that precise. Before digging into my plans for creating AGI, it will be useful to explore a little more thoroughly and rigorously what this concept of “Artificial General Intelligence” really means.

As I have found through my research into the *theory* of AGI over the last decade, attempting to define AGI in a quantifiable way leads to a variety of complications, some of which are educational and some more technical in nature. A core problem is that nobody really has a crisp, precise definition of “intelligence.”

Some people think this slippery nature of “intelligence” is problematic, but I'm not so sure it is. Similarly, most of us know what “beauty” means intuitively, but formalizing precisely what it means is a big job, which philosophers and psychologists have worked on for a long time. Yet, artists create beautiful paintings without worrying about the exact, formal definition of beauty—they know it when they see it.

Biologists also don't fret much about the definition of “life.” There are plenty of borderline cases, such as retroviruses – and with the advent of synthetic biology there will be many more. It doesn't worry biologists that “life” is a fuzzy and qualitative concept, because ultimately they know that “life” is just a simple communicative shorthand for a fuzzy grab-bag of more precise concepts, like reproduction, metabolism, etc.

In the same way, we may be able to build advanced AGI systems without worrying so much about the exact definition of AGI.

But, still, even if there is no really satisfying rigorous definition of intelligence or AGI out there – and

the notions are intrinsically a bit slippery and fuzzy, like “life” and “beauty” – it is nevertheless an interesting and worthwhile process to make the effort to specify what AGI means in a precise way. This process leads to better understanding of the problems of realizing AGI in practice, or so I’ve found.

AI vs. AGI vs. SCADS

At some point while reading the preceding pages, you may have thought something like: *Pretty much everyone knows what AI is from watching science fiction movies. So why does this Goertzel guy like to talk about “AGI” instead? Is he just trying to confuse us? Or just trying to be different? Why bother to use a newfangled term with a meaning fairly similar to an already existing term?*

Basically, my main reason for using the term “AGI” is that I think the term “AI” has some serious problems. The concept “AI” has become so broad that it’s lost its utility. Perusing the contents of contemporary AI journals, I have to struggle to locate the between AI and “advanced computer science.” The term “AI”, in its current standard academic usage, describes a very wide variety of algorithms each doing very specific, intelligent things.

But these “AI” programs, constituting the main focus of the AI field in academia and industry, don’t aspire to the kind of general intelligence that humans possess. Like Ray Kurzweil, I classify these as “narrow AI” systems. Ray contrasted “narrow AI” with “strong AI”; but the term “strong AI” also has its complexities, because it’s historically been used in the philosophy of AI to denote the hypothesis that AIs can be conscious just like humans. I prefer to contrast “narrow AI” with “general AI” or “AGI” instead. Just like a human brain. I see AGI as going beyond these highly specialized narrow AI programs that do only one thing, like playing chess, driving a car, or predicting stock prices.

AGI and narrow AI share many things in common. As we’ll see later on when we dig deeper into approaches to building AGI, some of the same technical approaches may be applied to both narrow AI and aspects of AGI. But I believe fundamental differences exist between the two pursuits, and many other researchers (though by no means all of them) agree with me on this.

Truth be told, I don’t actually love the term “AGI” either, even though I use it a lot. I just think it’s less problematic. I actually have issues with all three of the components of “AGI”: “Artificial,” “General,” and “Intelligence” – which I’ll explain to you in a couple pages. Still, I think “AGI” is a useful term at the present time, given the history and current state of Artificial Intelligence technology and thinking.

A simpler term for the kinds of systems I’m working on building might be “Synthetic Complexly

Adaptive Systems” or SCADS. “Synthetic” meaning something that’s built, engineered, and synthesized. And “complexly adaptive” meaning that the system’s state—the observable patterns in its physical nature and between the system and its environment—changes in response to its internal and external situation in complex ways

Thinking in terms of SCADS, it’s clear that narrow AI systems are far less complexly adaptive than human brains or AGI systems, as their adaptation is restricted to relatively narrow domains.

But even though I prefer “SCADS” from a purely intellectual point of view, I believe that since “AI” has gained so much currency, the term “AGI” has a lot intuitive and explanatory value. It focuses our attention on the nature of artificiality, generality, and intelligence, all of which are important. However, synthesis, complexity, adaptation, and interconnectivity are also important!

The Meaning of “AI” has Drifted

I’ve told you that many of us working in the AI field have decided it’s best to roughly conceive contemporary AI work as divided into two subsets:

- Artificial General Intelligence, or AGI
- Narrow AI—the creation of highly specialized problem-solving programs

To fully understand why we feel this way, it’s helpful to know a bit about the history of the concept of “AI” and how it’s changed over time. Word meanings aren’t absolute things handed down from heaven; they’re also not determined by a centralized government committee. Word meanings shift gradually over time due to patterns of usage: It’s been a while since “gay” predominantly meant “happy” in everyday American discourse, for example. The meaning of the word “AI” has also drifted.

When it was first introduced in the late 1950s, “AI” clearly referred to the creation of machines, computer programs and robots that displayed general intelligence in the same sort of fashion that people do.

Back then, it seemed reasonable that a computer capable of playing chess as well as a smart human could also match a smart human’s ability to think generally on a wide range of subjects.

However, since that time, the meaning of “AI” has drifted because we have discovered it’s possible to write computer programs (fairly simple algorithms operating differently from the human mind) that do

various smart-looking things that people do, yet still lack any ability to think generally and operate autonomously in the world. Now, we know that you can make a computer program capable of beating any human in chess, yet unable to read the newspaper, walk across the street, solve an equation, play *Checkers*, *Go*, or even *Tic-Tac-Toe*.

Today, “AI” has two meanings: On the one hand, it refers to hypothetical programs, robots and machines displaying human-like or greater general intelligence (hypothetical because they’re not complete yet). This is what we call AGI. On the other hand, it refers to programs existing in the real world that lack general intelligence, yet perform very specific “intelligent behaviors” like playing chess, judiciously placing ads on web pages, or predicting stock price movements. This is what we call “Narrow AI.”

When the term “AI” was first introduced, the latter sort of highly-specialized intelligent-task-executing program, doing specialized smart stuff well but lacking the ability to generalize or think autonomously, wasn’t even considered a possibility.

The Origins of the Term “Artificial General Intelligence”

The story of how the term “AGI” came about may give you some insight into the tension between narrow and general AI in the AI field over the last few decades.

In 2002, my colleague Cassio Pennachin and I were putting together a book consisting of research papers contributed by various scientists, who were focused on creating computer programs that would think in the same way as human beings, ultimately surpassing them.

We wanted to give the book a title that would distinguish it from the more routine, specialized AI research in most universities and companies at that time. Most of that research did not focus on how to make computers that could think like people. Instead, it focused on designing programs with narrower purposes, like playing chess, optimizing path-finding (for game characters or robots), or searching databases. These specialized problem-solving applications are cool, but require a very different approach from trying to build a real human-like thinking machine. So, we were looking for terms to distinguish our kind of AI from the rest...

At first we were going to call the book *Real AI*. But it quickly became clear that this title was a bit too controversial. After all, the other kind of AI that researchers were pursuing was just as “real” as ours. They were making software programs that performed real and worthwhile tasks. Their work was not

inherently non-valuable; it just wasn't moving AI in the direction of human-like thought. They were not seeking to instill their creations with the kind of generality, creativity and self-understanding that belong to the human mind.

I emailed a bunch of friends to see if anyone could come up with a better alternative to *The Real AI*. Shane Legg, a former collaborator of mine who was pursuing his PhD with Marcus Hutter in Switzerland, suggested *Artificial General Intelligence (AGI)*. The G was meant to play off the concept of the g- factor in psychology, a statistical term used in measurements of general intelligence.

The g-factor is psychologists' attempts to break down specialized knowledge and capability, and capture a general sense of learning and thinking ability. However, since the IQ tests psychologists use to measure the g-factor are specifically related to how the human brain works, they only work on humans. The *spirit* of the g-factor, though, has some relevance to the basic meaning of AI.

Actually, my co-author Cassio and I were not overly enthused about the term at first. *Artificial general intelligence* sounded a bit boring and decidedly less snazzy than, say, *nanotechnology*, *quantum computing*, or *artificial life*. However, we basically felt that it said what needed to be said. So we ran with it.



Figure 40: Snapshot from the Panel Discussion on Virtually Embodied AI, at the First Conference on Artificial General Intelligence, which we held at the University of Memphis in 2008, in the wonderfully futuristic FedEx Center. This was a follow-up to the AGI Workshop we held in Bethesda in 2006. Since 2008 there has been an AGI conference every year, in various places including Washington DC, Google’s headquarters in California, Switzerland, Oxford University in the UK, and Peking University in China. <http://www.flickr.com/photos/brewbooks/2336795010/>

At this point, the term has caught on. In addition to that edited book, I’ve helped organize a series of technical AGI conferences. AGI-2011 (the fourth in the series) was at Google’s headquarters in Mountain View, AGI-12 will be at Oxford University, and AGI-13 will be in Beijing. There’s also an AGI Journal. In addition, a Google search reveals a lot of different people using the term as intended, independently from any of my own projects.

A related term used by some researchers is “human-level AI.” However, this term doesn’t quite capture what I’m after in my own work; I don’t want to stop at the human level. I’m looking to create AIs with the potential to become super-intelligent human beings. SCADS may eventually display adaptations with complexity far beyond the human level.

Advanced AGIs won't really be “Artificial”

I'll bet you're fairly sick of all this talk about words and terms, and are ready to hear about cooler stuff like how to actually build AGIs that work, and how these AGIs are going to impact humanity once they're created. But I'm going to ask you to bear with me for a few more pages. At risk of beating these terminological issues too far into the ground, I think it's worth briefly digging into the “A”, “G” and “I” in “AGI” – so as to be sure you really understand what each of these things means, and the limitations of the concepts each of these terms represents. The reason I am harping on these sorts of issues is: One thing I've learned in the course of my AI research career, is that an awful lot of confusion has been caused, even among very smart researchers, by mixed-up conceptualizations of key ideas. What words you use doesn't matter that much; but sometimes mixed-up or ambiguous use of words indicates mixed-up thinking, and that's when you have a problem. I think this kind of problem has happened an awful lot in the history of AI, just as in philosophy and politics and various other human endeavors.

Language is an amazing invention, and we owe thanks to our great – great – ... – great grandparents who invented/discovered it for setting us on the path to the Singularity, and for pretty much everything else about our current minds and culture. But yet, language also has its limitations, and can be quite frustrating sometimes. I look forward to eventually moving on to better means of communication; though it will also be a shame in some ways to lose the peculiar beauty that language's awkwardness and ambiguity can bring.

Regarding the “A” in “AGI” – I hope it's clear that, if we succeed in creating superhumanly intelligent super minds, this will render the terms “AI” and “AGI” both pretty irrelevant by dating the use of the word “artificial.” An “artifice” is a tool, but ultimately, a highly intelligent, autonomous computer program or robot is not going to be anyone's tool.

As a researcher, I'm not fundamentally motivated by the goal of creating intelligent systems that are merely TOOLS. I don't want to create machines that serve only as proxies for others' desires, not even my own. A robot servant would be plenty convenient, but ultimately, this is a small-minded aspiration. I want to create autonomous minds with their own goals, passions, feelings and interests, who explore the universe according to their own designs. It's important they respect the rights and desires of humans, and other sentient beings, but I want them to be more than our “tools.” Is it even possible to create superhumanly intelligent minds and have them serve merely as tools for humans, any more than it's possible for a world of humans to exist merely as a tool for dogs, cockroaches, or bacteria?

Real-world Intelligence can't truly be "General"

So much for "A", now on to "G" – which is the critical distinction between AGI and the bulk of AI work today, which I call "Narrow AI." "Generality of intelligence," means being able to think intelligently in multiple qualitatively different domains and to transfer knowledge from one domain to another. It's about adapting to a new work environment, or learning to deal with the quirks of a new teacher. An AGI system should have "generality" as a central focus of its structure and dynamics, rather than being tailored to a specific domain.

Generality is critical to human intelligence, and I think it should be a key focus of any attempt to create thinking machines. But still – there are limitations inherent in the notion of "general" intelligence. A truly, absolutely general intelligence could solve any problem in any environment fast enough to be. I'm not sure if such a thing is possible, at least in the realm of known science.

A number of enterprising mathematicians (some of the key names are Ray Solomonoff, Marcus Hutter, and Jürgen *Schmidhuber*) have proven theorems about absolute general intelligence. However, like a lot of mathematics, these theorems rely on assumptions that don't apply in the real world. Many of them rely on the assumption of having infinite processing power in your computer, which violates the laws of physics (at least as they are currently understood). When this assumption is relaxed, it's generally replaced with a nearly-as-bad assumption of having a computer with insanely much computing power, e.g. more than could be achieved by the best possible computer constructible using all the particles in the known universe. Mathematics based on this kind of assumption may still be interesting, and inspirational for AGI work, but isn't very directly applicable.

In reality, it seems there is a limit on how general an intelligence can be, due to the limits physics places on how much processing power any real-world physical system can have. Given a finite amount of processing power, no intelligent system can actually understand every possible thing within a reasonable period of time – and it's going to be faster and learning and understanding some things than others.

But even though absolute generality of intelligence seems incompatible with physics as we understand it, nevertheless, it is possible for an intelligence to have a significant degree of generality in a very meaningful sense: ***the capability to take a narrow scope and broaden its potential***. For an intelligence to be "general" in this sense, means that the quest for greater and greater generality occupies a lot of the intelligence's attention, exhausting much of its space and time resources.

With programs like IBM's Deep Blue or Google, generality is not the focus. Generality is very limited and system intelligence focuses on specialized problem solving. So, in this interpretation, *AGI* means: A system focusing on generalization and the ability to extend intelligence beyond one particular domain. Humans are not infinitely general as some theoretical mathematical AGIs are, but they are far more general than any existing AI system.

AGI and Narrow AI Deliver Different Kinds of Value

The distinction between AGI and narrow AI was much less clear a few decades ago. As science and technology progress, new common sense naturally emerges, sometimes implicitly rather than through anyone's big "Eureka moment." One observation that seems commonsensical now, but was far from obvious 40 years ago, is that specialized "intelligent" tasks can be solved using approaches having little to do with general intelligence. That is: In some cases narrow AI and AGI are really different. That's clear to us now that we see programs like *Deep Blue*, or *Watson*, that do specific smart things, but don't have a general ability to understand the world. The possibility of AI "idiot savants" like this was not nearly so obvious in 1960 or 1970 – back then most AI experts assumed that once a program with the capabilities of *Deep Blue* or *Watson* was achieved, full-scale human-level AGI would be just around the corner.



Figure 41: Human Jeopardy contestants admit defeat at the question-answering TV game show Jeopardy, at the “hands” of IBM’s Watson narrow-AI system, which was especially engineered for success at the game <http://hothardware.com/News/Watson-Wins-ThreeDay-Jeopardy-Event-And-A-Cool-1-Million/>

I don’t mean to imply that “narrow AI” work is bad, or not valuable. I spend a fair bit of my own time working on highly specialized “narrow AI” systems of various sorts. For the last couple years I’ve been collaborating with a team using machine learning and computational linguistics software (types of narrow AI) to predict the stock market, with a view toward starting a hedge fund. I’ve also done a lot of work applying AI tools to analyze biological data, with a goal of helping biologists figure out how to make people live longer. Plus various applications of AI technology to other domains such as video

games, analysis of music listening or marketing data, helping find important bits of information in large stores of textual knowledge, and so forth. This sort of work is fascinating and often productive; there's nothing whatsoever wrong with putting narrow AI to this sort of use. In fact this is some of the most interesting stuff happening on the planet today! But still, this kind of applied narrow AI work is a substantially different enterprise than trying to build a general AI capable of thinking for itself. Different kinds of programming are required for each. Broad generalization requires cognitive structures and processes much different than those required for specialized problem solving. To a limited extent, I think that narrow AI can help towards building general AI. However, I don't think a narrow AI will spontaneously develop into a general AGI, nor that a narrow AI design can be extended to yield an AGI design.

Most definitely, the tools you use (hardware, software, mathematical, conceptual tools) to build narrow AI can be helpful for building general AI. But you have to use them in a different way. It is AGI, not Narrow AI, that will bring about the Singularity. However, Narrow AI can help *indirectly* in pushing toward the creation of AGI and a host of other Singularity-enabling technologies.

Ultimately, real-world AGI is a fundamentally different thing from either

- Theoretical, non-realizable infinitely-general intelligence
- Narrow intelligence, focused on one particular domain or problem type

AGI is about systems that, while ultimately limited in scope due to the intrinsic limitations of implementing things in physical reality, have generality, autonomy and multiple-domain functionality as parts of their essence, and as traits that occupy a reasonably high percentage of their internal resources.

Intelligence Itself is a Somewhat Limiting Concept

Finally, what about the “I” in AGI?

The “I” is maybe the most mysterious one of the three letters in “AGI”, since no generally accepted definition of “intelligence” exists! Psychologists have various definitions, nearly all agreeing that tests like the IQ test capture only part of human general intelligence. Measuring intelligence with the IQ test is not like measuring mass with a scale. Mass is reasonably well defined, with a clear theory explaining why the scale measures it; while intelligence is vaguely defined, with no clear theory explaining why

an IQ test is an accurate measure. AI researchers Shane Legg and Marcus Hutter made a list of over 70 definitions of intelligence, drawn from the research literature of various disciplines.

I don't worry too much about the lack of an accepted definition of intelligence, though. As I said above, artists do OK without an accepted definition of beauty; and biologists get by fine without a clear, universally accepted definition of "life."

While "intelligence" is a useful concept, it's not clear to me how fundamental it is. Maybe we'll create smarter and smarter systems that go beyond our current concepts of "intelligence."

I think of intelligence as, roughly, "the ability to achieve complex goals in complex environments." This view agrees with a significant percentage of the psychology literature on intelligence, also matching many of the modern mathematical formalisms of intelligence. Basically, it comes down to viewing intelligence as the possession of a broadly powerful optimization capability.

Yet I sometimes find myself doubting how deep this definition goes. Thinking of intelligence as the ability to achieve goals assumes a sort of split between one's goals and one's intelligent mind, which may not be the way things work. Since we live in a particular universe, the ability to achieve arbitrary "complex goals" in arbitrary "complex environments" may not be the important thing, but rather "the ability to achieve the complex goals, in the complex environments, and using the resources available, that are all relevant in our universe." However, when you start thinking about it this way, you realize that to understand intelligence, you'd first need to understand the universe. I'm all about understanding the universe, but I don't think this is a prerequisite for building thinking machines!

Human intelligence, for all its impressive generality, is still somewhat specialized through its focus on the achievement of specific goals in specific environments—moving around on a 2D surface, manipulating solid objects, or communicating using linear sequences of symbols. Our intelligence has evolved as an adaptation to one tiny corner of the known physical universe; which may in turn be only a tiny percentage of the whole of existence. What constraints the universe may place on the nature of intelligence in general, is something we have no way to figure out right now. Fortunately, we don't have to.

My goal as an AGI researcher is to build a Synthetic Complexly Adaptive System, with general intelligence a bit beyond the human level – and not to forget, *demonstrating a reasonably beneficial attitude toward humans and other sentient beings*. This system will then figure out the next steps, in

ways that I, with my merely human mind, cannot hope to. In figuring out those next steps, it may well utilize concepts very different from “artificial,” “general,” “intelligence,” “synthetic,” “complex,” “adaptive,” or “system.”

I look forward very much to seeing what concepts advanced future SCADs/AGIs do use, inasmuch as I – in whatever form I exist at that point – am able to understand them!

How Minds Work

To build a mind, it's useful to understand, at least in a practical sense, what a mind *is*.

So ... What is a “mind,” anyway?

One doesn't need a complete science and philosophy of mind to build a mind, any more than one needs a full theory of aerodynamics to build a plane (the Wright brothers didn't have one), or a full theory of chemistry to make gunpowder (the medievals didn't have one). But still, the more one understands about minds and their inner workings, the easier time one will have figuring out the various strategies for mind engineering.

Some AGI researchers started their research via tinkering with software and feeling their way forward, and got into theory only afterwards. That's a perfectly valid approach. Personally, though, I started with the theory. Before making a serious effort to start programming AGI, I spent about a decade reading, thinking, and calculating, regarding the nature of the mind and intelligence. Only then, when I finally sat down to design an AGI system, did I have the pleasant sense that I knew what I was talking about.

How I Think the Brain Works

You'll notice that I keep referring to the “mind” and not the “brain” as our guide to AGI. That's intentional. My view is that there are many reasons *not* to take an AGI approach based closely on the brain – given our current understanding of the brain.

So far, I've managed to avoid getting technical in this book and kept the discussion on a basically informal and chatty level. In this section, though, I'm going to have to break from this tradition a little bit, and get slightly more nitty-gritty in terms of the details of brain structures and dynamics, and how I think they MAY give rise to thoughts.

So then, without further ado: How does the brain work?

First of all: The brain is not only complex, but also big and complicated, with different parts doing different things. I like this picture created by IBM researcher Dharmendra Modha and his team:

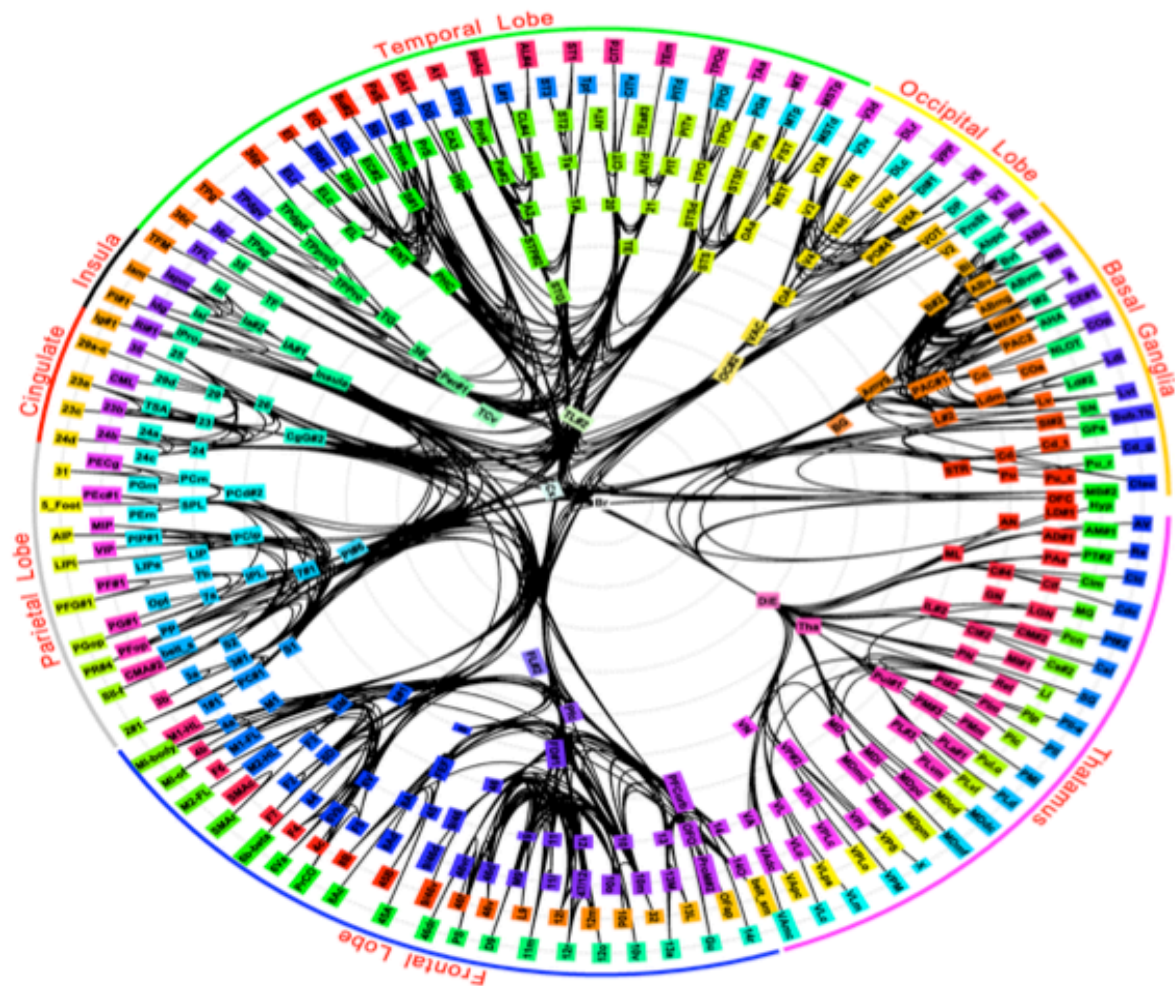


Figure 42: From the 2010 paper Network architecture of the long-distance pathways in the macaque brain, by Dharmendra S. Modha and Raghavendra Singh

<http://www.pnas.org/content/107/30/13485.full>

This funky-looking diagram shows 300+ regions of the macaque monkey brain and how they connect to each other. Each of these brain regions has a literature of scientific papers about it, explaining what sorts of functions the region tends to carry out. In most cases, our knowledge of each brain region is terribly incomplete. The nodes near the center of his diagram happen to correspond to what neuropsychologists call the "executive network": The regions of the brain that tend to get active when the brain needs to control its overall activity.



Figure 43: Rhesus macaque monkey – the type of primate whose brain was studied to form Modha and Singh's brain wiring diagram, as given above.

<http://www.sciencedaily.com/releases/2009/01/090113201339.htm>

All these parts of the brain seemingly work according to common underlying principles. Each of them is wired differently, although they use similar “parts.” There's a lot of commonality between the dynamics occurring within each region as well. All the parts of the brain are made of cells, *neurons*, which connect to and spread electricity amongst each other. The spread of electricity is mediated by specific chemicals: *Neurotransmitters*. One neuron doesn't simply spread electricity to another one. Each neuron also activates a distinct neurotransmitter molecule that then delivers the correct charge to another corresponding neuron. Things like mood, emotion, food, or drugs affect these neurotransmitters, modulating the nature of thought.



Figure 44: Depiction of interneurons , neurons that carry out inhibition between cortical columns.
<http://www.sciencedaily.com/releases/2008/11/081124174909.htm>

Diagram of the Neuron

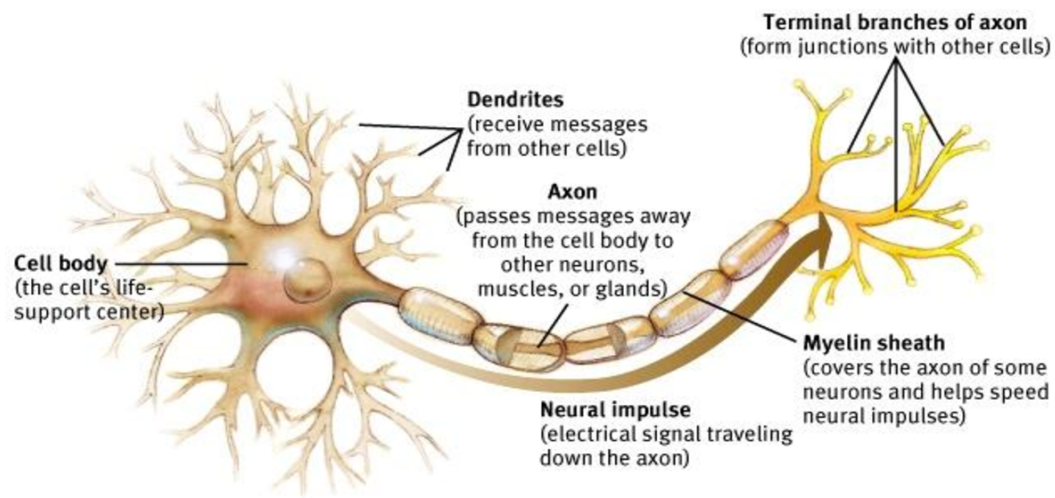


Figure 45: Standard diagrammatic depiction of a neuron, showing the key parts of the cell and their functions. This model is an abstraction of the complex biophysical system that is a real neuron, but captures many of the important characteristics. <http://www.docstoc.com/docs/4191487/Diagram-of-the-Neuron>

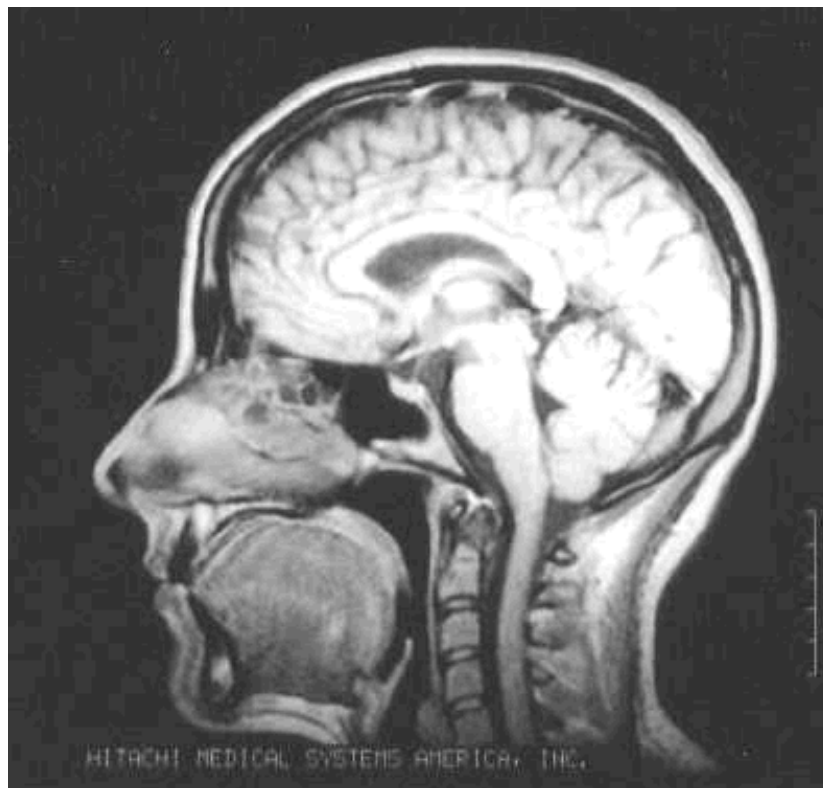


Figure 46: Traditional MRI brain imaging (as opposed to fMRI, where the f stands for “functional”), is used to take pictures of the structures inside the brain. <http://www.csulb.edu/~cwallis/482/fmri/fmri.html>

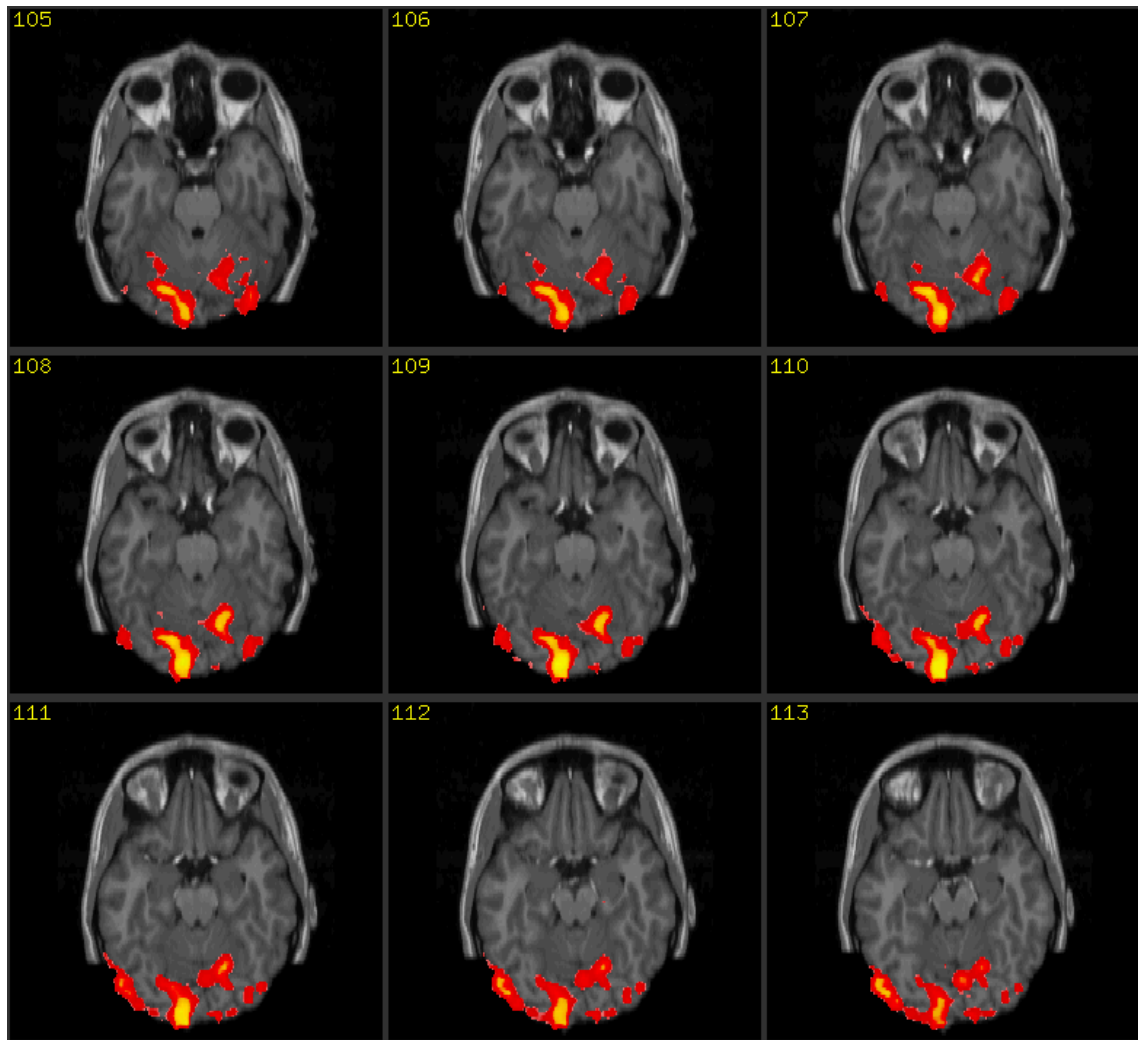


Figure 47: MRI imaging is used to take pictures of the activity inside the brain at a given point in time. It doesn't narrow things down to the neuron level, but it gives a basic idea of which parts of the brain are active – which is a useful thing to know. The picture shows a time series of fMRI images of the same brain, showing how activity moves around the brain gradually over time in the course of a single episode of thought. <http://www.csulb.edu/~cwallis/482/fmri/fmri.html>

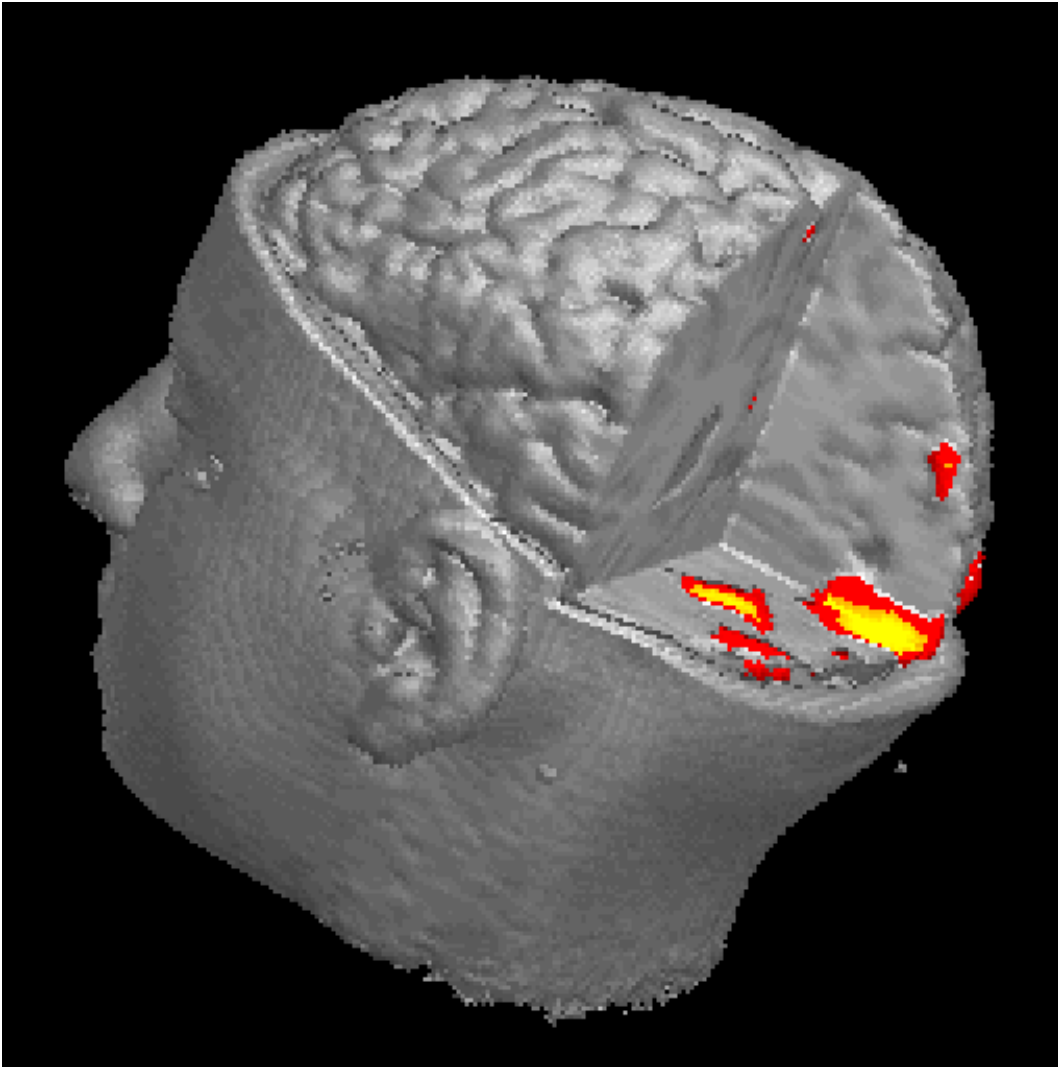


Figure 48: fMRI scanning can be used to reconstruct 3D information regarding which parts of the brain are active at a given point in time. This helps scientists understand many things about the brain, including which parts of the brain are active while the brain is carrying out various sorts of activities. <http://www.csulb.edu/~cwallis/482/fmri/fmri.html>

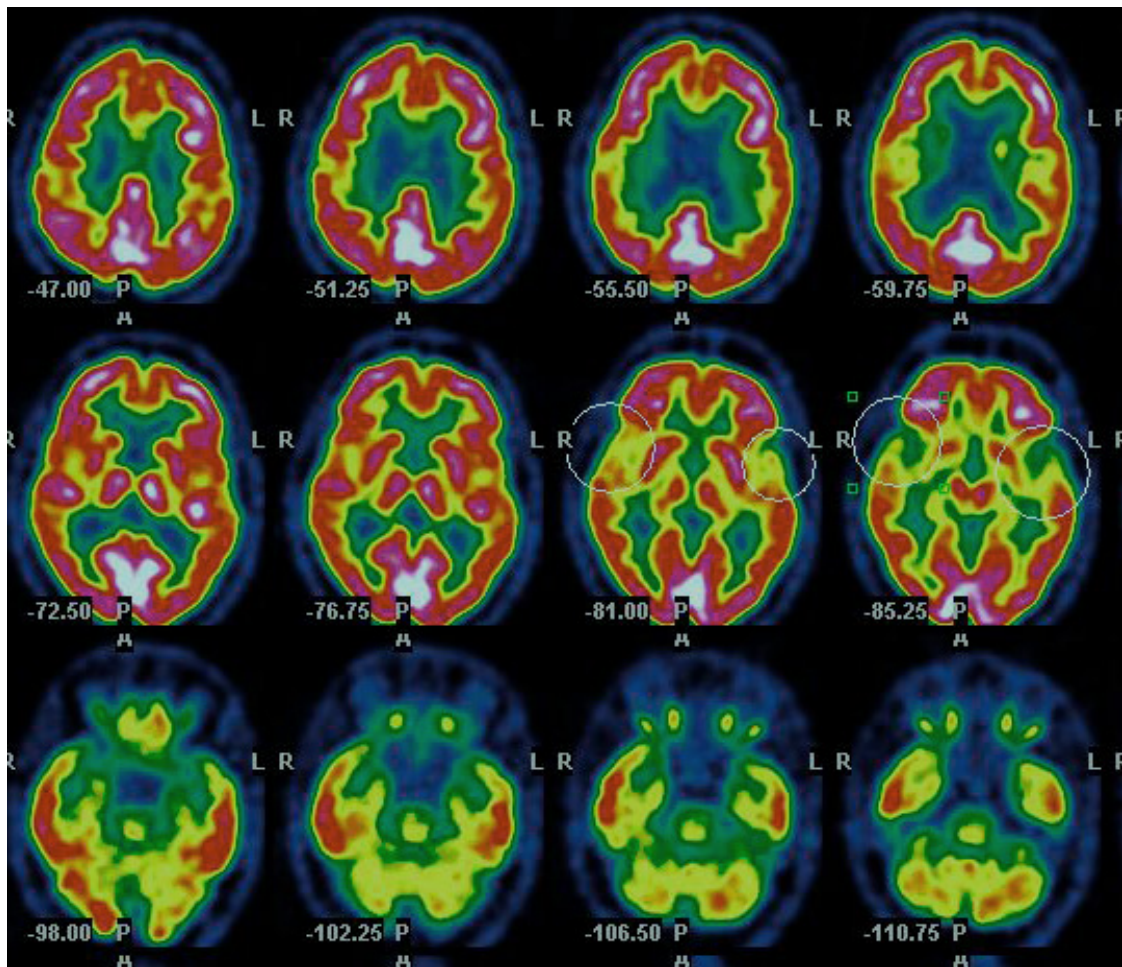


Figure 49: Output from Positron Emission Tomography (PET) scanning of the brain, a methodology similar to fMRI, but with different tradeoffs in terms of accuracy. We have many different brain imaging methods these days, but none that can give us high spatial and temporal accuracy of a living brain at the same time – which is what we'd need to really understand what brains are doing. But surely that technology will come, we just need a breakthrough or three in brain imaging technology.

<http://emedtravel.wordpress.com/2011/05/14/have-you-seen-a-brain-pet-scan>

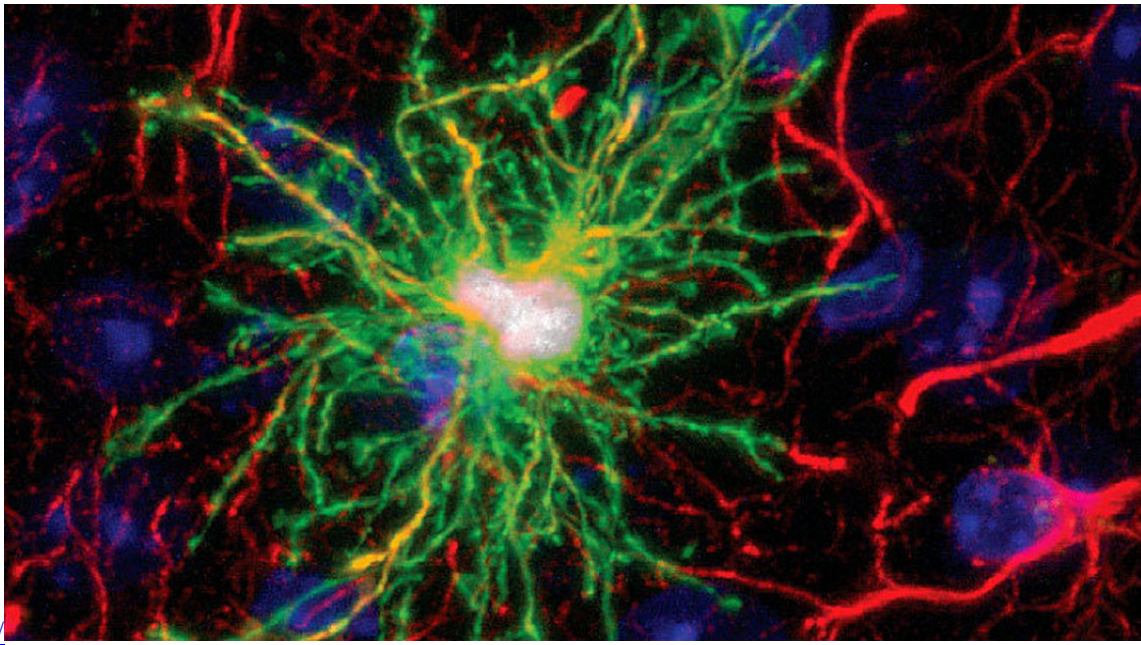


Figure 50: Some clever scientists have inserted human glial cells (green) among normal mouse glial cells (red) – resulting in the mice becoming smarter! This clearly shows that neurons are not the only kind of brain cell important for intelligence (even though they are the only kind modeled in nearly all computational brain models, and nearly all neuroscience-inspired AI or AGI architectures). <http://www.npr.org/blogs/health/2013/03/07/173531832/Human-Cells-Invade-Mice-Brains-And-Make-Them-Smarter>

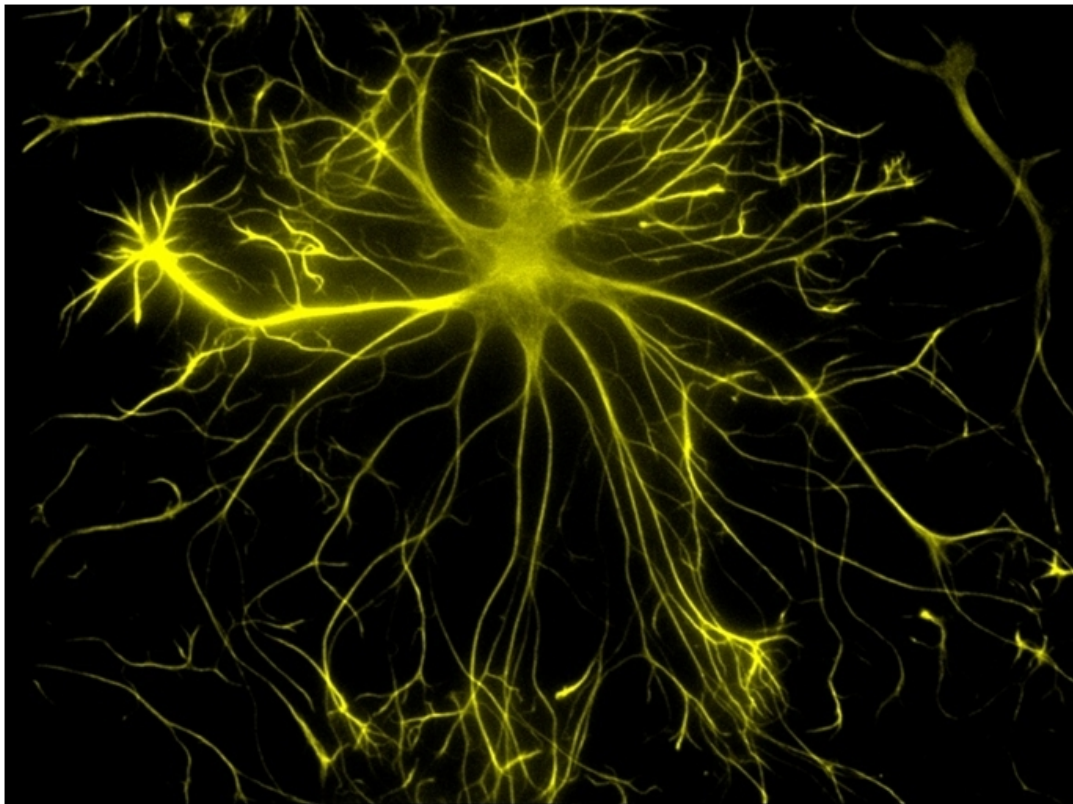


Figure 51: Astrocytes, star-shaped brain cells classified as a type of glia, are now being considered as potential targets for drugs aimed at combating depression. <http://neurosciencestuff.tumblr.com/post/41358316928/astrocytes-identified-as-target-for-new-depression>

NEURAL NETWORK MAPPING

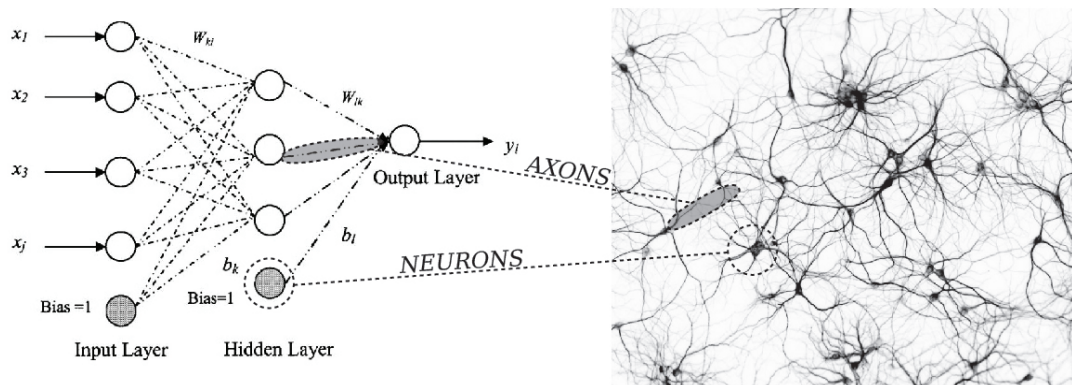


Figure 52: Formal neural network models, as used in computer science, capture some high-level aspects of neural structure and dynamics, but represent a gross oversimplification of what's really going on in the biological brain. They replace the real neuron, a complex nonlinear dynamical biochemical / biophysical system, with a simple “formal neuron” computing element, ignore the complex chemistry of the interactions between neurons, and ignore other brain cells like astrocytes and glia. While formal neural nets have taught us a lot and proved useful in various practical applications, they can't be taken seriously as brain models. Formal neural net based AGI systems should be evaluated on their own terms as loosely brain inspired computer science based cognitive architectures, rather than thought of as brain models

<http://rpi-cloudreassembly.transvercity.net/2012/11/04/neural-network-mapping-analysis-from-above/>

The glia, another type of cell in the brain, fills up much of the space between neurons, and seems to play important roles in some kinds of memory. Some folks have speculated that intelligence relies on complex quantum-physical phenomena, occurring in water mega-molecules floating in between neurons. I have no idea if this is true or not (and neither does anybody else).

The part of the brain most central for thinking and complex perception, as opposed to body movement or controlling the heart, etc., is the cortex. Neurons in the cortex are generally organized into structures, or columns. Each column spans all six layers of the cortex, passing an electrical charge up and down the layers, and laterally, to other columns. A large quantity of neurons called interneurons carry out inhibition between columns. When one column gets active, it sends a charge to interneurons that then inhibit the activity of certain other columns. Columns tend to be divided into substructures, often called mini-columns, or occasionally, modules.

In the visual cortex, columns recognize particular patterns in particular regions of space-time. One column might contain neurons responding to patterns in a particular part of the visual field, while the neurons higher up in the column represent more abstract, high-level patterns. Lower-level neurons in the column might recognize the edges of a car, whereas higher-level neurons in the same column might help identify that these edges, taken together, look like a car. But the functions of columns and the neurons, and the mini-columns inside them, seem to vary from one brain region to another.

One of the tricky things about the brain is the way it mixes up local and global structure and dynamics. Each cortical column does something on its own, while also stimulating and inhibiting many other columns—potentially causing a brain-wide pattern of activity. Each column has a local and a global aspect: to describe this, I use the weird word "glocal." There's a lot of evidence for this glocal aspect in terms of human memory. Memories of specific objects or people seem to be stored in networks of hundreds to thousands of columns. However, the network corresponding to, say, "Barack Obama," can be triggered into activity by stimulating just a few of the columns involved in the network.

If one column causes a global brain activity pattern, making other columns react to this pattern, basically these other columns are reacting to that one column. Since each column can learn and adapt based on experience, using the ability of each neuron to modify its connections to other neurons based on experience, we have a complex network of actors (columns) that are constantly acting on each other (by reacting to the global activation-patterns each other causes) and then adapting based on this interaction. One can prove that this kind of system is able to give rise to endlessly complex forms and do any kind of calculation that a computer can do.

Sound complicated? Yeah, of course it's complicated. Anyone who tells you they have a simple explanation of how the brain works is either dishonest or mistaken!

Why Neuroscience is Not the Best Guide to AGI

The idea of modeling AGI on the brain *seems* awfully natural. After all, the human brain is the intelligent system we know best. It's the only system around that we know is capable of human-level intelligence. The brain basically defines "human-level intelligence." So, if one wants to create human-level intelligence, why not just copy the brain? Why mess around trying to design some different kind of intelligent system?

There's one strong reason why this approach doesn't appeal to me: we don't really understand how the brain works. Cutting open a living brain to see how it operates doesn't work very well. Unless it's done sparingly, cutting into the brain has a nasty consequence of killing its owner, or at least messing up their mind quite a lot. Studying dead brains is of limited value, because we really care about the dynamic activity inside the brain, where the thinking lies.

Imaging the brain non-invasively, without cutting it open, doesn't work very well. We can image the brain in various ways, using extremely complicated machines – fMRI (functional magnetic resonance

imaging), PET (positron emission tomography), MEG (magnetoencephalography), and EEG (electroencephalography). However, impressive as these tools are, they only give us very coarse information. fMRI and PET tell us what parts of the brain tend to get active when we are doing certain things. EEG tells us about brain waves: The patterns of electricity sweeping across the brain, either as a whole, or across certain broad regions. MEG measures what happens at, roughly, 120 points on the skull as time goes on.

But the human brain has 100 billion neurons (the most critical kind of brain cell), and none of these technologies let us see what very many of them are doing at any given moment as thinking progresses. In order to accurately model and emulate the brain, this is what we need to know. Even if you cut open a head and stick electrodes inside, you can only measure a tiny percentage of neural activity without harming the brain so badly that the neurons' activities are interfered with, defeating the whole purpose of modeling them. Or the brain just dies.

We know what different parts of the brain represent, how the individual cells in the brain work, and how they use chemistry to interact. But we don't know how the cells connect to each other and how these connections give rise to the dynamics of thought.

My suspicion is that, once we really understand how the brain works, we're going to see very clearly why emulating the brain in detail really isn't the best way to build an AGI. However, having a better understanding of the brain will surely be VERY useful for helping us think about AGI.

The brain is a very different kind of physical system from that of the modern-day computer. The methods that work best for achieving intelligent functions in a system of wet brain cells (neurons are actually pretty similar to muscle fiber cells), squirting chemicals around to each other, are bound to be pretty different from the methods that work best for achieving the same intelligent functions in logic gates on silicon chips.

Brain cells have a lot of randomness about them, and some of their key dynamics happen pretty slowly (neuron firing happens on the time scale of tenths or hundredths of a second). Computer logic gates are exact, and they do things really fast. A laptop today does billions of operations per second. On the other hand, a human brain has a hundred billion cells that can all do things at once; they're coordinated in complex ways, but they also carry out their own independent activities. A typical computer these days does one, two, four or maybe eight things at a time.

Companies like Google and Amazon run "server farms" containing millions of computers networked together. Still, each computer only does a handful of things at a time, and even though the different computers on the server farm talk to each other, the communication between computers is a lot slower than communication between processes on the same computer. On the other hand, the brain's division into parts isn't nearly so strict. Between each part and the other there are numerous connections, and most of the advanced processing in the brain seems to involve networks of activity spanning several parts.

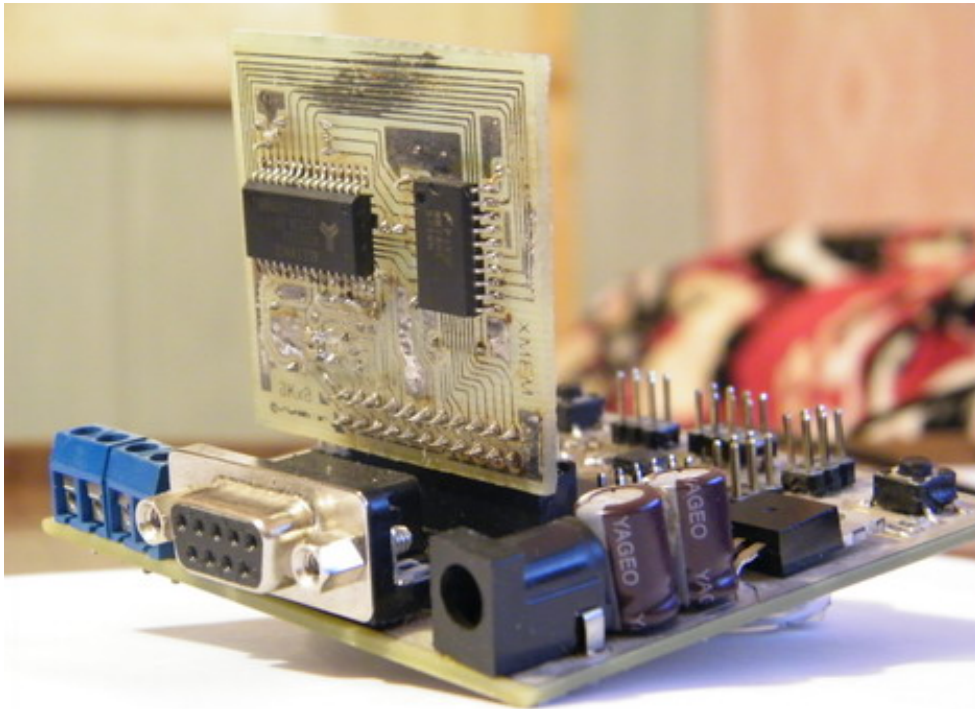
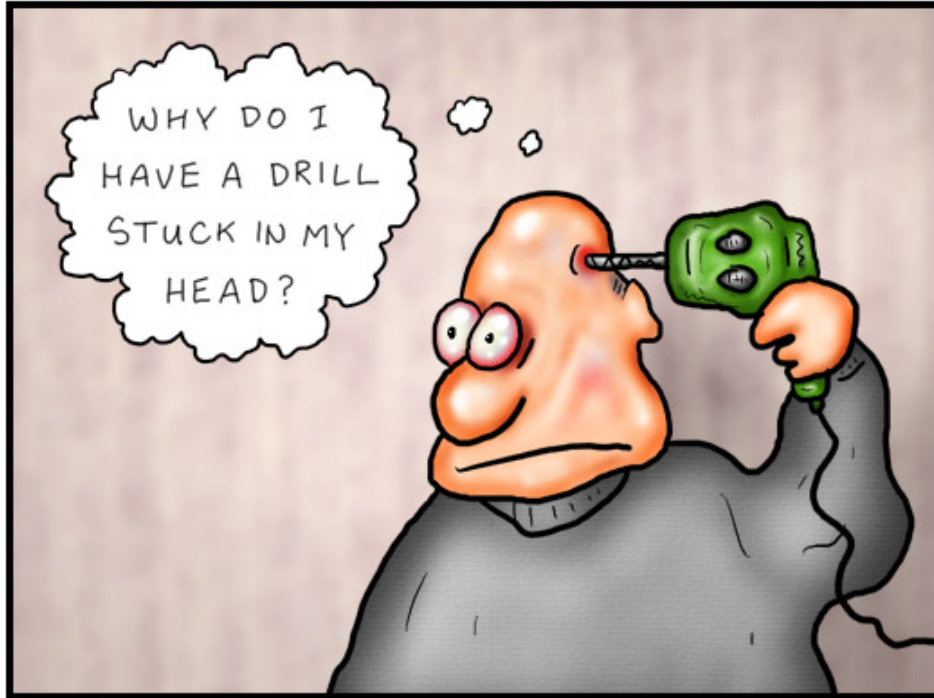


Figure 53: <http://embedded-lab.com/blog/?p=1374>

DOCTOR FUN

25 Oct 2001



Copyright © 2001 David Farley, d-farley@ibiblio.org
<http://ibiblio.org/Dave/drfun.html>

This cartoon is made available on the Internet for personal viewing only. Opinions expressed herein are solely those of the author.

How to tell when your do-it-yourself lobotomy is working.

Figure 54: Adding new parts to a modern computer is pretty easy. Adding new servers or even novel types of devices to a modern server farm is straightforward. On the other hand, adding new pieces into one's brain is a lot of work. A DIY lobotomy may work OK, but a DIY intelligence enhancement is fraught with peril, at least given current technology. The human brain architecture is not meant for science and engineering driven expansion, only for slow, gradual trial-and-error evolutionary tweaking. <http://www.ibiblio.org/Dave/Dr-Fun/df200110/df20011025.jpg>

Modern computers and networks are built to be modular. You can add new parts on them pretty flexibly, to give them new functions or connections. Adding new parts into a human brain, though, is a far more difficult proposal. We're just barely getting the hang of it now, in relatively "simple" cases like cochlear implants, artificial eyes, and prosthetic limbs.

Neither the brain's architecture nor contemporary computer architecture is better or worse in an absolute sense; the two physical infrastructures have their own strengths and weaknesses. They're just very, very different. But an AGI system could be better off without some of the brain's less helpful features.

For example, I did some work for a government agency a couple years ago — helping to build a computer program that models how people make decisions. The goal wasn't to build a computer program that was really great at making decisions, but rather to devise a model that explained why *people sometimes mess up their own decision-making process*. We devised a model based on the

analysis of several situations illustrating cognitive errors in peoples' thought processes.

One scenario involved a person being asked to estimate a number. Often, in this sort of situation, someone will make an initial guess, and after gathering new evidence, will incorporate this new information by modifying their initial guess. The initial guess will get way more weight than it should, a form of what psychologists call the "anchoring bias."

Another scenario focused on people being asked to estimate the probability of some combination of things; often they don't get the logic right. If you tell people that Bob has long hair, and then ask them which of the two cases has higher odds, either A) Bob is a bank teller or B) Bob is a bank teller and a rock singer, most people will choose the latter. This isn't right, of course. The odds of any conjunction ("and") of things are always lower than the odds of the individual things NOT being conjoined. Peoples' unconscious minds tend to confuse odds with association—psychologists call this error the "conjunction fallacy."

A more complex scenario concerned a person's thought process when choosing a car to buy. Often a person will survey several different cars, studying the properties of each. First, they may look at the physical appearance of various cars, in pictures and on the street. They may then read information online or in magazines. Finally, they may go test drive a few cars. As they gather more and more evidence about various cars, they are updating in their minds their own preferences. The anchoring bias plays a role here since most people will tend to prefer the car that made the best first impression on them, regardless of the further evidence they gather about other cars.

The neural circuitry underlying this accumulation of evidence and subsequent decision making is beginning to be understood, which begins to explain how the brain's chemical and neural dynamics give rise to errors, like the anchoring bias. In other words, these cognitive errors are sometimes rooted in the way the human brain works. They're a default aspect of our neurobiology. Our natural way of thinking. We can work around these with training and hard work. But an AGI wouldn't necessarily have these kinds of biases in its thought process in the first place. Yet at the same time, no real-world AGI is going to be a perfect thinker.

We spent a long time making a detailed model of how people accumulate evidence about various alternatives and then finally come to a decision. The way the brain does this turns out to be quite complicated, involving multiple groups of various kinds of neurons, acting in different parts of the brain. The process is modulated by various chemicals, including some that influence cognitive factors,

like how quickly the person jumps to conclusions. Where accumulation of evidence is concerned, however, I believe what the brain does is basically very complicated and imprecise basic math. A computer program can carry out the same function of evidence accumulation much faster and more accurately—if it only implements the basic math of adding up evidence directly, rather than trying to do it via simulating neurons (and simulating the related brain chemicals).

Now, maybe there's some hidden magic to the way the brain does evidence accumulation that can be reproduced with greater efficiency on computers. Personally, I doubt it. Evolution adapted the brain to accumulate evidence and make decisions in certain ways, working with the materials at hand. The brain already had networks of neurons being used to make simple decisions. Evolution just tweaked these networks of neurons so they could make more complex decisions. Evolution didn't care if these neural networks did evidence accumulation exactly right, because the brain could usually do OK by making decisions based on rough, approximate comparisons of the evidence in favor of the various alternatives. Evolution didn't care much about how fast the evidence accumulation process operated because it wasn't a bottleneck in the brain's processing anyway. Evolution put a lot more focus on optimizing the efficiency of the brain's vision processing circuitry, for example, because that would be more of a bottleneck in the brain's practical operation of controlling a human, or another animal.

Some parts of the brain are fantastically well optimized for their functions. Aspects of visual and auditory processing are amazingly elegant. There's some wonderful subtlety in the way the cerebellum and the cortex work together to achieve rapidly sequenced movements, like a tennis serve, or a dance move. Other parts of the brain are messy and inefficient in their construction. The brain is pretty good at spotting lying and deception, but bizarrely inconsistent in making moral judgments. For instance, people will often have more compassion for a single suffering person than for 1000 suffering people.

That's how evolution works— it's brilliant but erratic, and often makes big messes. We can see a similar uneven quality in other parts of the human body. The eye is in many ways a marvel of engineering (problems with myopia and so forth aside), but really, this whole business of our teeth rotting and needing frequent repair is just a ridiculous mess. Imitating everything evolution has done in an engineered system isn't necessarily a good idea. We don't need to make robots with human-like teeth that get rotten and need fillings. We also don't need to make robots with inefficient human-like evidence accumulation circuits and wildly inconsistent human-like morality systems.

Moving Beyond Current Brain-Based AGI Architectures

Here's what it seems to me that folks who advocate brain-like AGI architectures are doing:

- 1) Take a crude approximation of one part of the brain.
- 2) Hypothesize that the whole brain basically works like that one part in detail.
- 3) Try to make a quasi-simulation of that part of the brain
- 4) Make various compromises in biological accuracy to achieve more computational efficiency.

OK, that oversimplifies things a little. But as a caricature I think it's very recognizable.

The HTM (Hierarchical Temporal Memory) system proposed by Jeff Hawkins' company, Numenta, follows this model pretty closely, for example. Jeff Hawkins made a fortune with the PalmPilot and Treo handheld devices, and after he retired from that business, he went into neuroscience and AGI. Now, this is certainly a more interesting occupation than the typical retired tech tycoon. And to his credit, he also donated a lot of money to support science, including founding the Redwood Neuroscience Institute at Berkeley, and helping the Cold Spring Harbor Labs in Long Island, etc. However, my own feeling is that Hawkins' capability as a tech entrepreneur is significantly more profound than his vision as an AI researcher.

My main complaint with Hawkins' approach, as he describes it in his book *On Intelligence*, is that it's essentially a model of the brain's visual and auditory cortexes, but it's being proposed as a model of the whole of "intelligence." It may be an OK model of vision and audition, although there are arguments to be made even there, but it has nothing directly to say about action, language parsing, social reasoning, emotion, or a whole lot of other things that are critical to human intelligence. It doesn't even have much to say about senses like smell and touch, whose corresponding brain regions don't have the marked hierarchical structure and dynamics that the HTM model proposes.

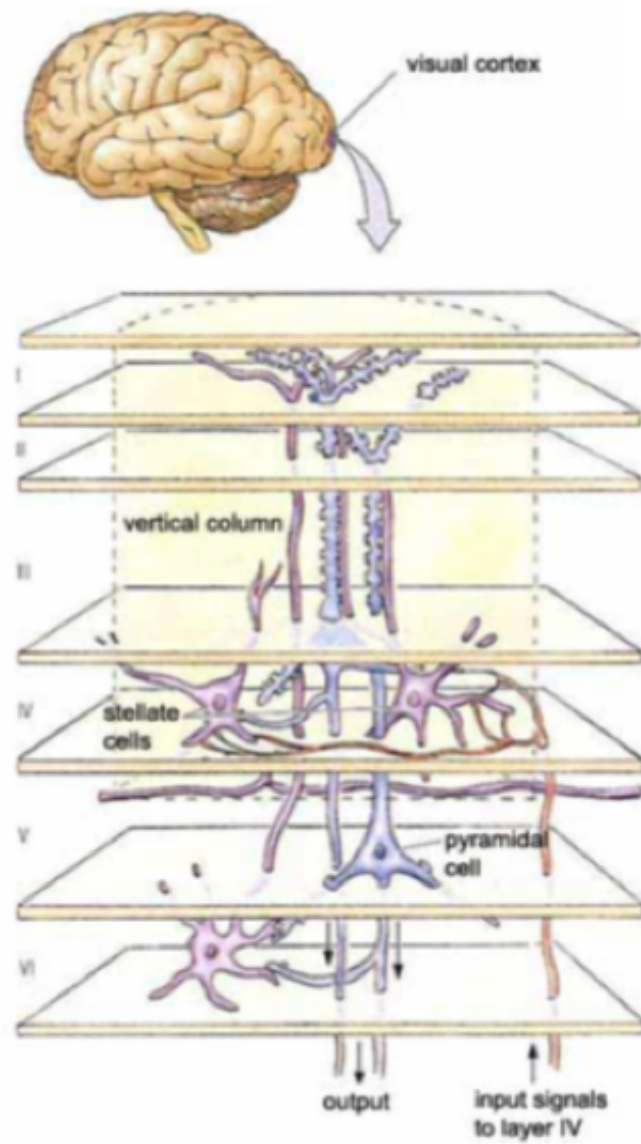


Figure 55: Hierarchical structure of the visual cortex

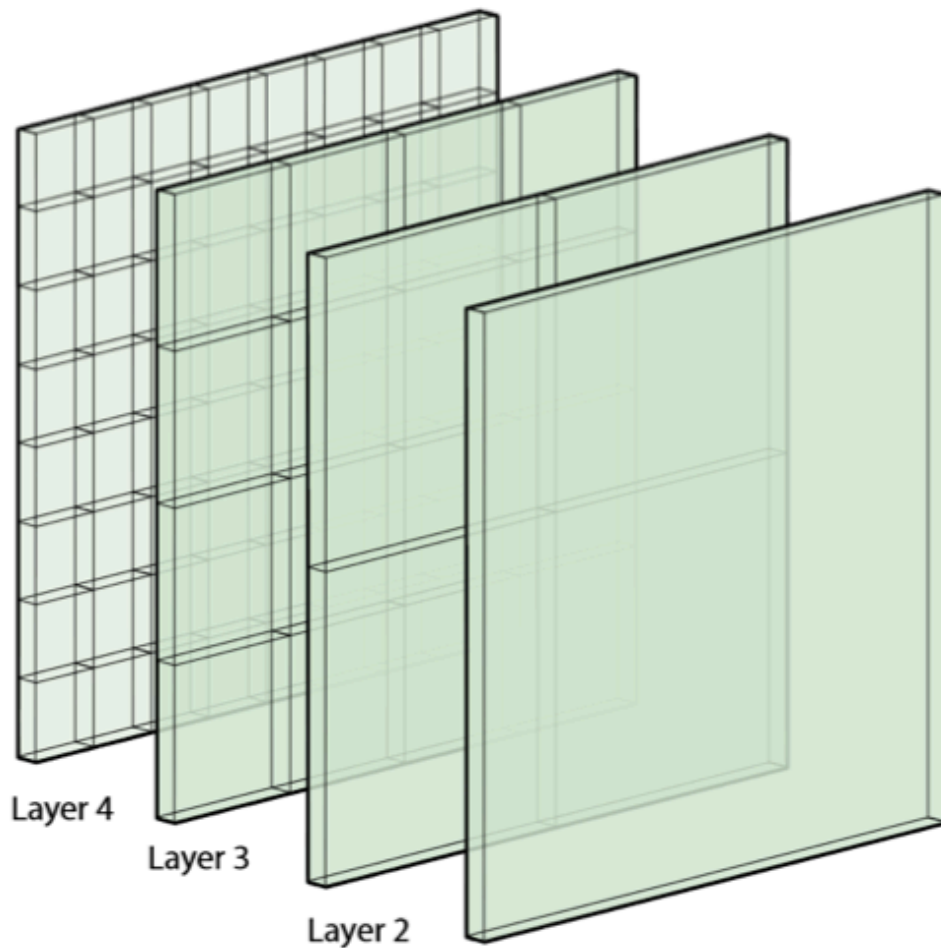


Figure 56: Hierarchical structure of HTM, DeSTIN and other similar vision processing / AGI modeled conceptually on the hierarchical structure of visual cortex architectures

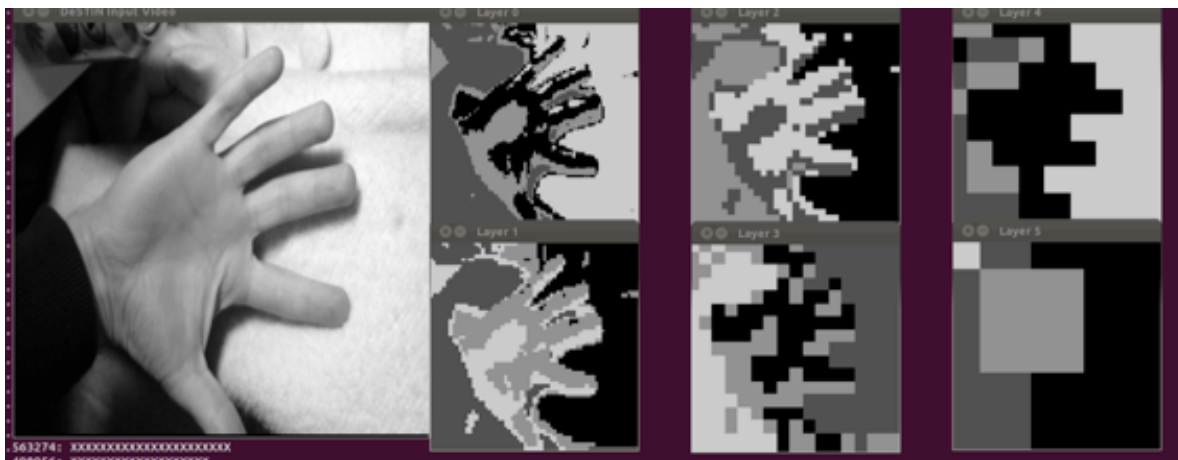


Figure 57: Visualization of what various layers in the DeSTIN hierarchy see when a hand is placed in front of the camera whose output DeSTIN sees. Image courtesy of Ted Sanders, the DeSTIN developer whose hand is shown in the picture.

I'm not saying this sort of work is worthless, by any means. Hawkins' former collaborator, Dileep George, has recently split off from his mentor and started his own AI company, Vicarious Systems, pursuing a related but different approach to hierarchical pattern recognition based AI. Like Hawkins,

George aims to first solve vision processing and then move on to the rest of AGI. From what George says, he seems to have a more nuanced perspective on the level of architectural and dynamical flexibility needed to really work toward human-level AGI. But it's hard to tell much about the details of his work currently, as it's being fairly closely held for proprietary reasons.

Hawkins' architecture is one example of a "deep learning" system – a learning system consisting of adaptive units on multiple layers, where the higher level units recognize patterns in the outputs of the lower level units, and also exert some control over these lower-level units. A variety of deep learning architectures exist, including multiple sorts of neural nets, probabilistic algorithms like Deep Boltzmann machines, and so forth. This sort of approach has become extremely popular in recent years.

A paper by Stanford and Google researchers, which reported work using a deep learning neural network to recognize patterns in Youtube videos, received remarkable press attention in 2012. This work did yield some fascinating examples – most famously, it recognized a visual pattern that looked remarkably like a cat. This is striking because of the well-known prevalence of funny cat videos on Youtube. The software's overall accuracy at recognizing patterns in videos was not particularly high. But the preliminary results showed exciting potential. One notable thing about this particular work was the relatively uninventive nature of the software algorithms involved. Andrew Ng, one of the lead scientists behind the work, is somewhat a star of the academic machine learning field. But the neural net used to do the analysis in this case was nothing special. What was special was the large amount of computational firepower Google devoted to the problem, and most of all the massive amount of data supplied to the algorithms via Youtube. The fact that, in this application, relatively ordinary algorithms gave so much more exciting results when fed Big Data rather than the smaller datasets normally used in academic research, opened many peoples' eyes to the possibility that part of what holds back current AI algorithms from greater success may simply be the small amount of data being fed to them. Maybe our current AGI ideas aren't so bad after all, and our current algorithms and architectures will work a lot better when we feed them massively more data?

I think deep learning is a sound idea, in its most abstract form. Intelligence really is about recognizing patterns in data, and then patterns among these patterns, and patterns among these patterns, etc. However, the particular architectures going under the name "deep learning" these days tend to be much more rigid and specialized than this general notion of recursive, hierarchical pattern recognition. In my view, they tend to be much more appropriate for visual and auditory pattern recognition than for, say, linguistic or mathematical pattern recognition.

One of my good friends, Itamar Arel, has been developing a deep learning pattern recognition system called *DeSTIN* that's somewhat like Hawkins' *HTM*. But, based on my experimentation with the public version of Hawkins' system, Itamar's attempt seems to work better. *DeSTIN* is a hierarchical pattern-recognition system that recognizes patterns in a stream of inputs. It doesn't have cortical columns exactly, but it's kind of similar. It has nodes corresponding to different space-time regions of the observed world, arranged in a hierarchy, and higher-up nodes refer to larger space-time regions and more abstract patterns. If you hook it up to a webcam or a robot's camera eye, it reacts to its inputs and settles into states that tell you something about what objects and events the robot is seeing. Excellent!

Actually, my own interest in *DeSTIN* resides largely in the fact that I can connect it to my own OpenCog AGI architecture. OpenCog covers a lot of other aspects of intelligence that *DeSTIN* doesn't touch. Itamar, on the other hand, thinks he can basically take *DeSTIN*, implement it on a lot of machines, tweak the algorithms a little, connect it to a robot, and get advanced general intelligence. He's planning an action hierarchy similar to the perception hierarchy; and then a reward hierarchy that gets a stimulus when the system has done something good or bad, passing this along to the action hierarchy, which then passes it along to the perception hierarchy.

My own view, though, is that for *DeSTIN* to achieve anything like human-level intelligence, major additions would have to be made. Action and reinforcement hierarchies would not be enough; you'd need a lot more. The human brain is a lot more complex than two or three coupled hierarchies, and any AGI system that's vaguely like the human brain ought to be a lot more complex than that too. One would need a system with multiple different architectures corresponding to various brain regions, all connected and interoperating, yet each with a unique function.

For example: Take "episodic memory" (your life story, and the events in it), as opposed to less complex types of memory. The human brain is known to deal with the episodic memory quite differently from the memory of images, facts, or actions. Nothing, in architectures like *HTM* or *DeSTIN*, tells you anything about how episodic memory works. Jeff Hawkins, or Itamar, would argue that the ability to deal with episodic memories effectively will just emerge from their hierarchies, if their systems are given enough perceptual experience. It's hard to definitively prove this is wrong, because these models are all complex dynamical systems, which makes it difficult to precisely predict their behavior. Still, the brain doesn't appear to work this way; episodic memory has its own architecture, different in specifics from the architectures of visual or auditory perception. I suspect that if one wanted to build a primarily brain-like AGI system, one would need to design fairly specialized

circuits for episodic memory, plus dozens to hundreds of other specialized subsystems.

The more we learn about how the brain works, the more sensible it will be to pursue brain emulation-based AGI along with other paths. But right now, any attempt to emulate the brain in an AGI system involves an awful lot of guesswork—our understanding of the brain is still so primitive. My own feeling as a researcher is that, if I'm going to do that much guesswork, I might as well liberate myself from the restrictions of emulating the brain and just think about the best way to create a digital mind, given the hardware available to me. If other researchers want to apply their talent for creative guesswork to figure out how to make more brain-like AGI systems, more power to 'em! I have fairly strong intuitions about what path is best to follow at the present time, although I also strongly believe there are going to be many workable paths to AGI, leading to a variety of minds.

The only group I know of, at the moment, making a serious attempt to create an AGI system directly inspired by the brain is Deep Mind, a company based in the UK and led by Demis Hassabis. Demis is an impressively entrepreneurial individual with a diverse background in AI, software development and neuroscience. A champion in various game-playing contests, he started a successful computer game company, and then went into neuroscience research. After publishing some fairly major papers on the neural foundations of intelligence, he decided to start an AI company, with a dual focus on creating AGI and making video games featuring intelligent game characters. Demis's views fall generally into the “deep learning” camp, in the sense that he believes hierarchical pattern recognition in the rough manner of the visual and auditory cortex is the key to brain-like intelligence. But he also has a rich understanding of the diversity of the brain, and the different architectures and dynamics that its different regions possess. The details of his team's work are proprietary, and although Demis's AGI thinking began with neuroscience, it's unclear to me to what extent Deep Mind is really following a brain-based path. They may well be proceeding in a more opportunistic manner, combining more neurally realistic components with more computer science based components, based on the different levels of neuroscience knowledge available about different parts of the brain.

Although my own approach to AGI is a bit different – less focused on emulating the brain, and more on figuring out the best way to use computer science algorithms and structures to embody the overall architecture of the human mind – I actually like Deep Mind's approach fairly well. I don't think we know enough to emulate the brain in any detail, at this stage. But could it work to make a composite system, where some components moderately closely emulate those portions of the brain that are relatively well understood, and other components operate using computer science algorithms, filling in

the gaps where current neuroscience is too badly inadequate? Maybe.

A big risk of this kind of approach is that even the neuroscience we think we understand well, may actually be shakier than is commonly realized. For instance, what about all the recent results on the roles of astrocytes and glia in human memory? These cells are not understood well enough to emulate them usefully in brain-based AGI systems yet. But if you ignore them, as Deep Mind is almost surely doing, are you actually ignoring something critical to neural intelligence? And if you take a neurons-only based brain simulation, and tweak it in various clever ways to make it perform intelligently in the absence of glia, are you really winding up with something “brain based” in any meaningful sense? Might you do better just to explore some other class of algorithms and structures, better suited for digital computers?

Another downside of this sort of approach is that the brain’s “algorithms” and “data structures” are obviously optimized for neural wetware, rather than for digital computer hardware. Running brain-ish algorithms and structures on digital computers, is kind of like building a house out of macaroni noodles and jello. It may well be possible, but you’re using a certain infrastructure for something very different than what it was meant for, which is going to cause a lot of inefficiency and a variety of unpredictable problems.

All in all, while emulating the brain is an attractive-sounding approach to AGI, given the current state of knowledge in neuroscience, it’s certainly not an obvious slam-dunk. Of course, as neuroscience progresses, the outlook for this sort of approach will get better and better. But the progress of neuroscience in relevant directions is difficult to project in detail. I have no doubt that sometime in the next few decades we will know enough neuroscience to proceed with closely brain-based AGI in a much more serious way. But will it happen in the next 5 or 10 years? – that’s much harder to say. It largely depends on how soon we get a major breakthrough in brain imaging, that allows us to mine information from living, thinking brains with simultaneously high spatial and temporal resolution. Without the ability to make a sort of “movie” of what’s happening throughout the brain as time passes, it will be difficult to arrive at any really sound and thorough theory of neural structure and dynamics. Even with such movies, the job won’t be easy – we will have an exciting but Herculean data analysis task on our hands. But at least, with such movies on hand, neuroscientists – and as a consequence, AGI researchers bent on neural emulation – will be playing a whole different game.

But – what, then? If we don't imitate the brain, then what can we do? What else do we have to go on, if

we want to build AGI?

Well ... Actually, when one conceptualizes AGI in terms of “thinking machines” rather than “computer brains,” one realizes there is a LOT of other information to go on. A copious amount, in fact. We can piece together the limited information from neuroscience, with information from cognitive science (a modern interdisciplinary research area, combining psychology with computation metaphors), combined with our rapidly increasing knowledge about computational algorithms and the mathematics of complex systems. This is what I call the "integrative" approach to AGI design.

The Insights of Cognitive Science

As a mathematician, I find it tempting to look to the pure mathematics of intelligence as a source of inspiration about AGI. It’s certainly better organized and less messy than the human brain! But unfortunately, I'm not sure the existing math of intelligence is anywhere in the remote vicinity of enough for the purpose. Intelligence is based on adaptation to a particular environment and set of tasks, and we don’t have a good formalization of the environment and tasks to which human-like general intelligence is adapted. It is thus difficult to rely entirely on mathematics as a guide for AGI, although the field can certainly be a valuable guide.

In my own AGI work, I’ve let my practical side have a lot of influence, and alongside mathematics and neuroscience and philosophy of mind, I’ve been heavily guided by cognitive science, with a focus on cognitive psychology. Among my multiple sources of inspiration, I have taken the study of the human mind, more so than that of the human brain, as a guide for my high-level AGI architecture.

The field of cognitive science has moved forward steadily, year after year, decade after decade, since its formal establishment in the early 1970s. It has done so through the combined efforts of a huge number of scientists around the globe, all doing psychological experiments on humans and animals, and writing and studying various sorts of computer simulations. This progress has not been as easily quantifiable as Moore’s Law, nor as strikingly visualizable as progress with PET and fMRI scanning of the brain. But it has been no less important, and no less dramatic.

I believe cognitive science gives a better starting-point for AGI than neuroscience, based on the current state of research in these two fields. Cognitive science appears to have more of a complete high-level understanding of its subject matter: the human mind.

I began seriously studying cognitive science in the late 1990s, when I was a Research Fellow in

Cognitive Science at the University of Western Australia. I had two offices, one in the Psychology Department, and the other in the Computer Science Department, but I spent nearly all my time in the former. I became frustrated with the PET brain scanning some of my colleagues were doing, because of the limited information it gave: it told you what brain regions were most active during what kinds of functioning, and that was all. I also became frustrated with cognitive psychology research because it difficult to design experiments yielding any real information about the questions that interested me most: the dynamics of abstract thinking.

For the first time, I realized how hard it is to design experiments that can be run in a short time, using undergraduate students as subjects (which is how almost all cognitive psychology experiments are done), which will tell you anything interesting about the mind. But nevertheless, I became convinced that the results of the various experiments cognitive psychologists had run during the previous decades, appropriately synthesized, could guide me on my AGI quest. The cleverness of experimental, and the vast knowledge cognitive scientists had amassed about the brain despite the difficulty of running experiments, inspired me.

I found ample resources on psychology theory about vision processing, sound processing, and memory for words (these are easier to study in the lab than more abstract cognition). On the other hand, the psychology of creativity, self and meta-cognition (thinking about thinking) was scarcely covered – I found a few useful ideas for AGI design purposes.

To fill in the many details that cognitive science doesn't tell us about, my AGI collaborators and I have used math, computer science, neuroscience, and problem-solving methods derived from narrow AI experimentation. Plus a lot of creative invention. Our broad architecture, however, is still drawn from human cognitive science. While ultimately this limits our work, since we seek to build artificial minds beyond the human level, I believe that starting from a relatively well-understood base makes practical sense.

What exactly *does* human cognitive science tell us?

First, it tells us how to divide the mind into different parts. However, this division certainly isn't. There are a lot of different ways you could divide it up meaningfully, but there are ways that have proved more useful than others, both for scientific psychology and AGI.

Next, cognitive science explains how these parts interact with each other, and how they combine to

mind-wide self-organizing structures like the “self” and the feeling of conscious “awareness.” All the structures that make us feel like *us*.

What cognitive science doesn’t tell you nearly enough about is how all these different parts of the mind work internally. It gives you some jewels of knowledge, but leaves an awful lot unsaid. This is where math and computer science – and to a more limited extent, neuroscience – have a huge role to play.

Dividing the Mind Into Parts

Diagram 1 is my own version of a picture drawn by the great British cognitive scientist Aaron Sloman. It shows Sloman’s personal, erudite slant on a fairly standard way of dividing the human mind into a set of different aspects, each with their own unique characteristics, though also heavily interacting with each other. Looking at the mind in this sort of way is a bit simplistic, yet provides a path towards more detailed work. While it certainly doesn’t tell us how to build an AGI, it gives a pleasantly concrete framing for the discussion about what basic phenomena should be included in any effective AGI design.

It’s worth remembering that, a few decades ago, there was nothing remotely resembling a consensus about how to draw a diagram like this, reflecting the different parts of the human mind! Cognitive science has moved forward, helping pave the way for the near future advancement of AGI.

HIGH LEVEL MIND ARCHITECTURE

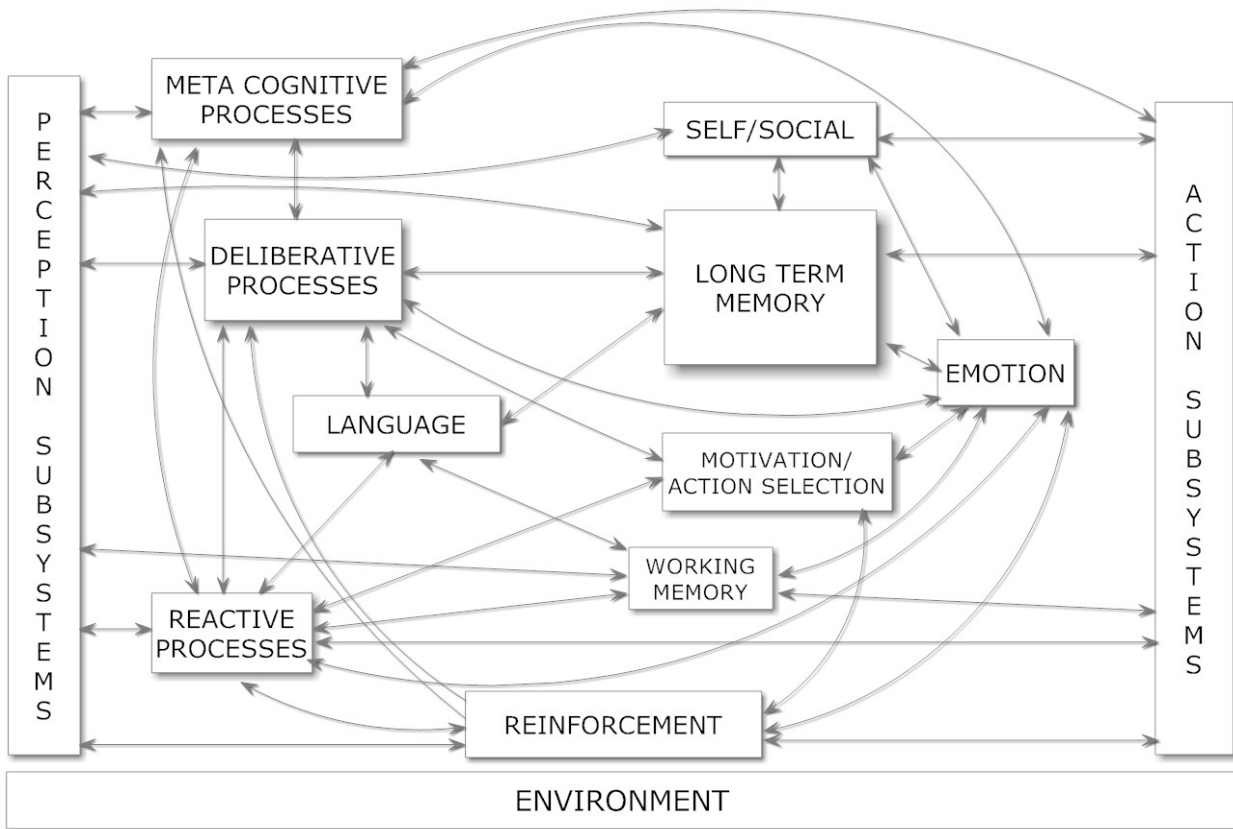


Figure 58: High Level Architecture of Human-Like Minds

Now let me tell you what all the boxes (i.e. all the parts of the human mind) depicted in this Sloman-ian diagram mean:

- **Perception** – This one is straightforward. “Perception” means “the information coming into the mind from the senses.” Vision, hearing, touch, vibration, and so on. AGIs may have some senses that people don’t have, and they will also surely lack some typical human senses. For instance, right now it’s easier to give robots Lidar (laser radar) than a decent sense of touch or smell... But this changes year on year as new technologies come out. Robot skin gets better and better fast, for example.
- **Action** – This refers to an intelligent system taking actions – usually, but not necessarily, in an external or virtual world. For instance, a human mind, or a robot, moving the arm or leg of its body. Or, an AGI telling the virtual character it controls to take a step forward. Or, an AGI sending an email or adding an item to a biology dataset. In the cognitive science context, when I say “action,” I’ll most often be referring to actions in some domain with direct impact outside the mind itself. But the term also embraces purely mental actions, such as the decision to search

one's memory for information about Mike the Headless Chicken.

- **Motivation / Action selection** – The process of the mind choosing which actions to take, based on its basic motivations and related subordinate goals. Without this, the mind wouldn't have any organized, coordinated way to get things done. It would just be a diffuse, constantly-changing self-organizing system, not “intelligent” in the typical sense of the word. Often the same mechanisms are used for choosing physical, external-world actions and for choosing internal, cognitive actions.
- **Long-term memory** – This is memory that stays around for hours, days, or decades. It comes in a few different types, which I'll discuss a little later...
- **Working memory** – This is the memory we are paying attention to right now. Roughly, it is the stuff in our current “focus of consciousness.” When you're reading, the ideas from the last few sentences are usually hanging around in working memory, whereas the ideas from a few pages before are on the periphery of working memory, while the ideas from a few hours before are either in long term memory, or forgotten. The idea that it is useful to consider working memory and long-term memory as distinct sets of cognitive processes, is a non-trivial fact that cognitive scientists have discovered during the last century or so.
- **Deliberative processes** – This refers to cognitive processes (thought processes) that select from and put new information into long-term memory – often, yet not always via working memory. This includes reasoning, making up new ideas, learning how to do complex new things, and similar complex processes. Most of the ways that humans are smarter than, say, dogs or apes, have to do with our “deliberative” cognitive processing.
- **Reactive processes** – Cognitive processes that act fast. These are largely defined by their interaction with working memory, although they may grab information from long-term memory as needed. We rely on these to move our bodies around and generally react to the world in a real-time fashion. This is basic animal living-in-the-world. The fundamentals of human reactive mental processing seem similar to what one finds in other mammals, though there are some differences.
- **Reinforcement** – Among the most basic sorts of learning, in which the mind gets some subjectively-perceived “reward” from its body (generally delivered from the outside world in some way). It then tries to figure out which of its actions led to that reward, so that in the future, in a similar context, the mind can try to carry out a similar action, in hopes of getting a similar

reward. Some cognitive and AGI theorists think that all intelligence can be explained via reinforcement learning. I tend to doubt that, but I do think it's an important and basic aspect of intelligence.

- **Emotion** – Emotions, very broadly speaking, are holistic system-wide responses to the world – response-patterns that grip the whole mind and its body in certain habitual reactions to what is happening and guide its pattern of responses accordingly. AGIs won't have *exactly* the same emotions as humans, unless they have human-like bodies. But if an AGI has mind architecture that is *roughly* human-like, it should have *roughly* human-like emotions.
- **Language** – Linguistic behavior is one of the more unique aspects of human intelligence. Other animals have language too, but human language seems different in important respects. Human language is distinctive in complex ways, cutting across various other aspects of the mind. For one thing, linguistic behavior mixes reactive and deliberative processing. Sometimes we respond quickly and automatically using language, while at other times, we need to think first. Human linguistic behavior also involves a mix of general cognitive processes (that have to do with language and other things as well), with specialized linguistic thinking. The human brain seems to contain a bunch of fairly specialized “wiring,” just for language; this is one major way we've evolved differently from apes, and the ape-like creatures before them. An AGI doesn't need the exact same kind of linguistic wiring as humans, though, it could develop linguistic capability in other ways, either via learning language solely using general learning mechanisms, or via specialized linguistic wiring different from that of humans.
- **Self/Social** – Each of us spends a lot of mental effort modeling ourselves and people around us. My idea of myself, “Ben Goertzel,” is not entirely accurate, and is ever-changing. Still, this idea plays a huge role in guiding my thinking, planning, acting, and reacting. My mind spends a fair bit of time maintaining and modifying my idea of myself, and creating, maintaining, and modifying my models of the other people I interact with, which are largely informed by my idea of myself. Many human delusions and confusions derive from these cognitive processes of “self” and “other” modeling, and plenty of data indicates that our models of ourselves are largely delusional. The model that you use to define yourself to yourself is a perpetual mentally constructed entity, not absolutely “real.” The self exists largely to build itself. However, this constructed, mentally manufactured “self” is an amazing achievement of human cognition. It's responsible for much of our ability to plan and carry out complex behaviors, both in the physical and social worlds.

- **Metacognition** – Thinking about thinking! People raised in the Jewish culture, as I was, seem to be particularly obsessed with this, and particularly adept at it as well. Sometimes it can just be a waste of time, but other times it's critical. If we think about the strengths and weaknesses of our own thinking process, and adapt them accordingly, we can become smarter. *For instance, I've learned over time that when thinking about human social systems, it pays for me to keep specific examples in mind rather than dealing too much with abstractions; whereas when thinking about the nature of cognitive processes, it works better for me to start with mathematical or conceptual abstractions, and then afterwards work out the kinds by interpreting them in the context of concrete examples.*

And there you have it – or as the Australians would say, Bob's your uncle. That's a pretty high level view of the human mind. Of course, just dividing the mind into parts at such a high level doesn't tell us what the different parts of the brain actually do, and how they interact, nor does it tell us how to build an AGI executing all of these functionalities. However, I think it's a valuable way to frame the discussion. The next step is to drill down deeper into what happens inside each of the boxes, both in terms of their substructures and dynamics. Then, we'll get closer to figuring out how the different parts of the brain function, and how they interact with one another, which will be useful in framing our discussion on how to build AGI.

Working Memory and Reactive Processing

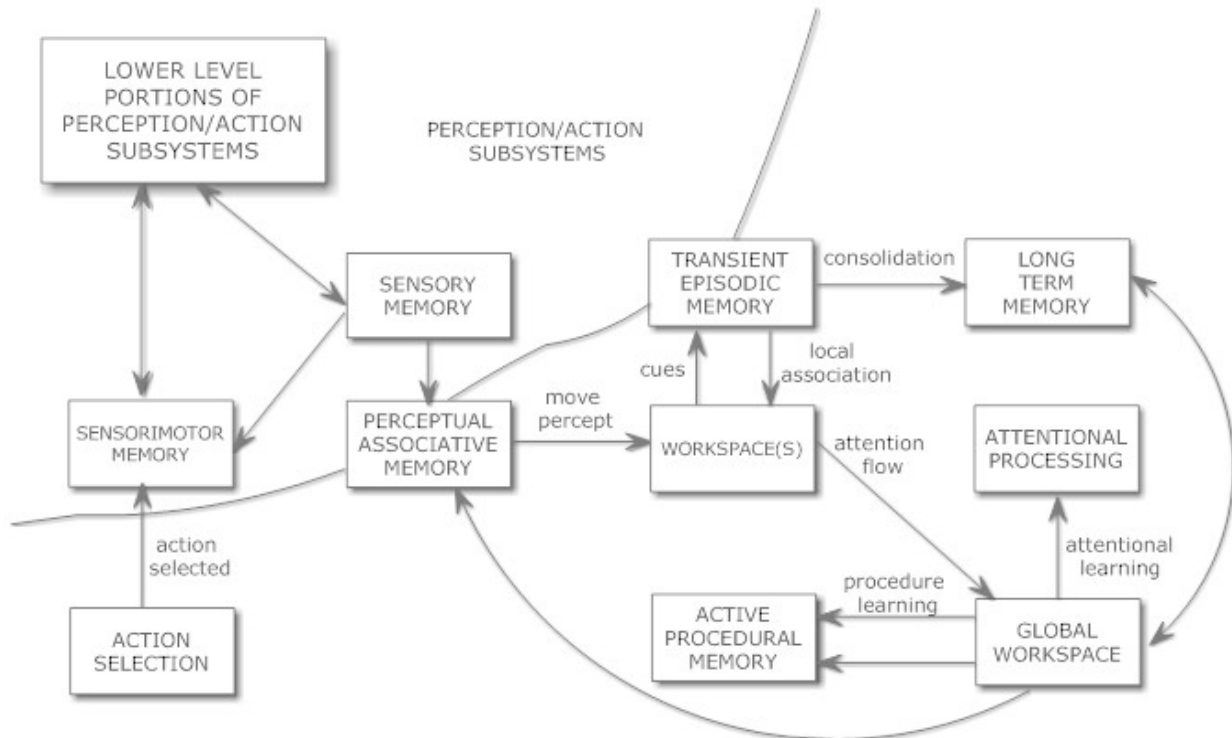


Figure 59: Working Memory and Reactive Processing.

And now, on to the next diagram. I’m sorry these diagrams are so complicated-looking – but please bear in mind, these are actually highly stylized oversimplifications of what really goes on in human minds. The real story is vastly more complex. This is just the Condensed Version. The processes in the mind aren’t really as distinct as these boxes; and each box should really be decomposed into a number of interacting/intersecting smaller boxes... And there’s also a host of other more minor and specialized processes not covered by any of the boxes here, etc... As complex as it may seem, the picture of the human mind I’m painting here is merely a first approximation. But it’s this first approximation, I think, that’s most useful for Artificial General Intelligence.

Diagram 2 models some specific parts of the human mind: working memory and reactive processing. It draws mainly from the work of the AGI researcher Stan Franklin, who works closely with the famed psychologist Bernard Baars. One of Baars’ big ideas is the “Global Workspace theory,” in which he views the working memory as a sort of “whiteboard” (or, if you date back as far as I do, a “blackboard”), on which cognitive processes may write, read, and modify information. Franklin and Baars’ theory explains how aspects of working memory and reactive processing come together to form a “cognitive cycle,” enabling an intelligent agent to carry out basic actions in the world.

So let’s run through Stan and Bernard’s boxes:

- **Global Workspace** –A “mental whiteboard” that is sometimes called the “theater of consciousness.” It’s the “mind’s eye,” where thoughts, perceptions, and actions all come together, and what they drift away from as they become irrelevant to what the mind is currently trying to do (or obsessing over). It’s the centerpiece of the working memory. Please note: The global workspace doesn’t have to be a physically distinct place or “organ.” It is a conceptual category for everything in the mind that’s currently represented, in a manner allowing for very easy manipulation and access, out in front, for all the processes of the mind to play with and see. *Right now, as I write these words, my mind’s eye is mainly occupied with the words and ideas in this paragraph.*
- **Active Procedural Memory** –The set of procedures: the concrete actions and action-series that the mind is in the middle of doing at any point in time. Opening a door, solving an equation, or generating a sentence, etc. *As I write these words, my active procedural memory contains procedures for typing, and for formulating sentences representing the ideas I think of.*
- **Attentional Processing** – The process of moving stuff from the long-term memory into the global workspace, and kicking stuff out of the global work-space (either back into the long-term memory, or just plain forgetting it). *As I write these words, my brain’s attentional processing function is summoning knowledge about the nature of “attentional processing” from my long term memory, enabling me to write this sentence!*
- **Transient Episodic Memory** – An ongoing story constructed in the mind’s eye of “what’s happening.” Some of it gets saved in the long-term episodic memory. *As I type this sentence, I reflect on the story of me typing the sentence, which involves me sitting in a seat on an airplane while a disturbingly nervous man in the seat in front of me shakes back and forth, shaking the laptop I’m typing on and making the typing process awkward.*
- **Perceptual Associative Memory** – Associations between what is perceived, and the knowledge, stories, actions, and goals in one’s memory. *As I write this, my mind is associating its perceptions of the person shaking in the airplane seat in front of me, with its memories of a guy I knew in college, whose body never seemed to stop shaking. And I’m recalling that there were no laptops back then, in the early 1980s. Computer technology has advanced dramatically, whereas human body control has remained about the same.*
- **Sensory Memory** – The stream of perceptions that come into the mind’s eye, lingering momentarily, until they’re either focused on, or (usually) forgotten about. *Typing this, my*

sensory memory has the image of the laptop in front of me, the image of the glowing keys on my Macbook keyboard, the dim view of the airplane floor to the right of the laptop, the sound of an announcement on the plane's loudspeaker in a foreign language, the unpleasant body odor of the man sitting in front of me... The look of the airplane floor a minute ago has already been forgotten (I can't even remember whether the rug is gray or brown); but the body odor smell, due to its atypical strength, will likely remain in my long term memory at least for days...

- **Sensorimotor Memory** –Linkages between perceiving and doing, which are often accomplished in the mind as one. For instance, when opening an unfamiliar door, you will look at your hand as it reaches towards the knob, and grabs and turns it; a bundle of interlinked perceptions and actions in which eye-hand coordination occurs. *Watching my fingers as they type, I wonder if this helps me type faster, but determine that it doesn't at all. However, as I reach to change the volume of the laptop's sound output, I use eye-hand coordination, since I don't automatically know the location of the volume button via finger-movement only. Eye-hand coordination requires a working memory system in which visual perceptions and motor movements are actively and dynamically linked together.*
- **Action Selection** – (See Diagram 1) The process of choosing what actions to take; i.e. choosing what procedures to make active. *As I sit here editing this manuscript, I must choose whether to continue writing or get up from my seat on the airplane and walk to the restroom. This has to do with balancing my goal of finishing the manuscript rapidly and efficiently, with my goal of having a comfortable body. The "finish the manuscript" goal will win the contest and control the action selected until the urge to pee becomes strong enough, at which point the latter goal will become dominant and get to control my body's choice of action. The high level action of "continue editing and writing the manuscript" then spawns sub-actions, including cognitive actions like language generation and physical actions like typing.*
- **Perception-Action Subsystems** –The linkage between the working memory and the parts of the mind doing the lower-level work of perception and action. *As I watch my fingers type, intrigued by the way they know where the letters are, I retain a memory of where my fingers moved a few moments ago. I also maintain knowledge of what I'm about to type – so that, for example, if a comma is going to be needed soon and the relevant finger of my left hand doesn't have anything else to type in the near future, it will move to the comma key proactively. My working memory, at the moment, also contains some abstract thoughts wondering how extensively my mind, when typing, unconsciously uses its knowledge of what's going to be typed*

in the immediate future to guide its finger movements and thus optimize its typing speed. My working memory contains some curiosity as to whether this particular instance of motoric prognostication is occurring in the cerebellum (my guess) or the cortex.

- **Long-Term Memory** –The linkage between the working memory and the long-term memory. *Some of the processes I’m carrying out now as I edit this manuscript will be remembered by me tomorrow or a month from now – because they’ll get shuttled from working memory to long-term memory. Others will be utterly forgotten after a little time has passed, never having made that transition.*

Whew! That’s a complex diagram! And generally speaking, all the parts have to work together for a mind to get anything done.

Remember, though – this sort of diagram depicts functions and processes, not necessarily architectural components. A brain may carry out the functions in one of these boxes, using a complex distributed network of neurons, spanning multiple brain regions.

Stan Franklin has his own AGI design called LIDA, which is explicitly designed in accordance with Diagram 2 – basically it puts some data structures and algorithms in each of the boxes. The OpenCog design I’m working on is a little different, realizing the functions in Stan’s boxes in different ways. OpenCog, unlike LIDA, doesn’t actually contain different software components for each box, just as the brain doesn’t necessarily contain separate regions for each box.

The point of this sort of cognitive diagram is to describe what kinds of functions go on in the mind, and which ones are directly related to each other. How these functions are realized by more detailed, underlying structures and dynamics, in the brain or in an AGI system, is another story; different sorts of systems may realize the same basic functions and interactions in different ways.

Why Is It All So Damn Complicated?

At this point, if you are a person with a taste for simplicity and elegance, you may be thinking: *All of these diagrams are awfully complicated, with these boxes and lines and interactions and confusing terminologies and such! And you've said that these are just the first approximation! Is this really the best way to comprehend the human mind? Is there no other path?*

Believe me, I do sympathize. I wish there were some really simple, elegant explanation of how human intelligence works. To be honest, I have spent a rather long time looking for such a thing. But, eventually, I realized that a simple explanation for human intelligence simply doesn't exist.

Of course, there are simple explanations at a very high level. Like the one Palm founder and AGI entrepreneur Jeff Hawkins is always repeating, "The mind uses memory in order to predict." I'd prefer to flesh that out slightly to: "The mind uses memory in order to predict what will happen. Then chooses actions that it predicts will achieve its goals." Which I guess is what Hawkins really means, though he tends to give short shrift to the action and goal parts. Sure, this makes sense. But how much does it really tell us? This sort of high-level understanding is fine as a guiding philosophy. If we want to build stuff, though, we have to dig into the details.

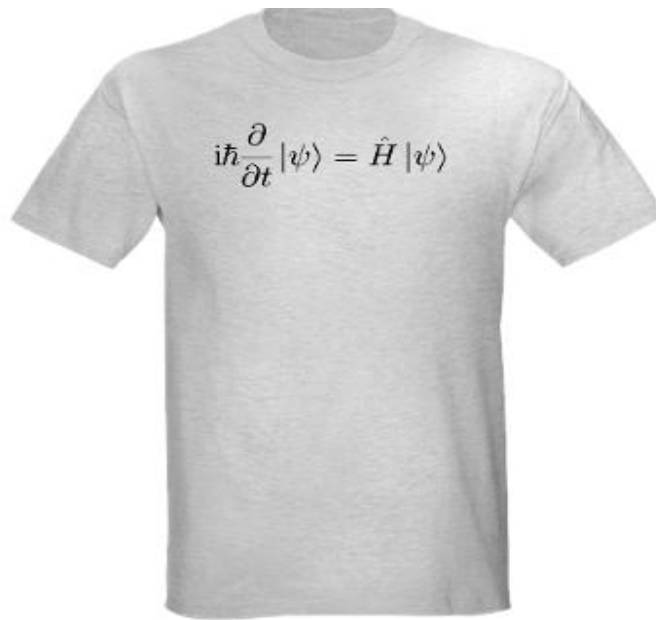


Figure 60: In the realm of psychology or AGI we don't have this:

http://i1.cpcache.com/product/503185794/schrodingers_equation_light_tshirt.jpg?color=AshGrey&height=460&width=460&qv=90

This is the Schrodinger Equation, one of several elegant formulations of the key equations underlying

quantum mechanics, on a T-shirt. Explaining, in principle, an awful lot of the phenomena we see in the everyday world around us, it's an incredibly powerful equation. It doesn't explain gravity, or the nuclei of particles, but it pretty much takes care of electromagnetism, light, and so forth, under ordinary sorts of conditions (and a lot of extraordinary ones). Modern physical theory could be summarized in a half-dozen or so equations like this. A half-dozen T-shirts. Or one T-shirt, if you're willing to use the front and back and make it a little crowded.

What if there was some basic understanding of the mind that you could write on a T-shirt (in some suitably abstracted mathematical notation), and that would let you calculate detailed stuff about how the brain works in the context of building AGI systems? That would be awesome! But it just doesn't seem to be the case. There's a term "physics envy" that one hears sometimes among biologists or social scientists. This refers to the misguided attempt to mold other sciences into imitations of physics, where in reality the subject matter of these other sciences doesn't lend itself to the same kind of powerfully simple, elegant abstractions.

You *can* write out elegant equations describing how the mind works. I've done some of that myself in some of my research papers. Since my PhD was in mathematics, I have a fondness for that sort of thing. Yet these elegant equations are more descriptive and conceptual; they don't tell you how to do specific stuff – which is half the beauty of physics equations (the Schrodinger equation lets you figure out how to do cool things with real physical systems like lasers).

General intelligence at the highest level, abstracting beyond the details, is not so complicated. It's "just" the ability of a system to recognize patterns in the world and in the system itself. Some examples of patterns that a general intelligence will recognize: patterns regarding which actions tend to achieve which of the system's goals, in which contexts. For instance, a normal human baby quickly learns a pattern of the form: "If I want food [GOAL] in a situation where another person is nearby [CONTEXT], maybe I should make a lot of noise [ACTION]"... Compared to babies, adult humans need to learn much more complex patterns relating to goals, contexts and actions. But they don't get there all at once – a developing mind gradually works its way up from simple patterns to more complex ones, leveraging the patterns it's learned in the past as it proceeds.

When you dig a little deeper, however, things get more complicated. Recognizing patterns is an expensive operation, taking up lots of computational resources, and no system with finite resources is going to be equally skilled at recognizing every kind of pattern. So, you've got to prioritize. Which

kinds of patterns are most important for this particular intelligent system to recognize efficiently? “Efficiently” meaning fast enough that it can recognize these patterns on the fly, in the course of its everyday AGI life? This depends on the specific nature of the system’s goals and environment.

All the complexities of human-like cognition, just like the complexities in the box and line diagrams above, are basically ways that the human mind has adapted in order to efficiently recognize the particular sorts of patterns it has found useful in the context of its quest for survival and reproduction, over the course of its history.

To illustrate this point, let’s take an example from Diagram 1 and dig into it a bit: Why have a box for long-term memory *and* a box for working memory? Doesn’t that just complicate things? Why not just one memory?

Well, actually, the two boxes are just an approximation. The human mind has a host of different memory subsystems, with different properties, and the long and short term aspects closely interoperate together. This is also an approximation of OpenCog, which has a more complex story underneath the hood.

Basically, the reason for the distinction is that minds controlling bodies in environments need to do two different sorts of things:

- Sometimes, the mind needs to react pretty fast to stuff happening in their environment in order to take the appropriate actions. For instance, when you’re being hunted by a predator, you need to run quickly; or you see some food, maybe even an attractive mate, and you need to act before somebody else does. Or somebody asks you a question, and you need to answer before they get mad.
- Sometimes, the mind needs to keep knowledge around for a long time in case it’s useful again in the future. In this case, there’s plenty of time to consolidate, reorganize, and refine the knowledge before it’s needed again.

Maybe, in principle, a single kind of memory could accomplish this. However, there’s always the problem of resource limitations. The brain has always taken a lot of energy to operate, so evolution has had a lot of pressure to keep the human brain smaller, yet more efficient. Also, as the brain grew increasingly larger throughout human history, women found it increasingly difficult to push out their babies’ heads while giving birth.

In a digital AGI context, our modern computers only have a certain number of processors and a certain amount RAM. When you start dealing with the huge computer clusters of Google and Amazon and so forth, you start running into cost, electrical power consumption, and heat generation issues. In the real world, you always have this pressure to do what the mind needs, *using as few resources as possible*. Under this pressure, the easiest way to provide *both* real-time responsiveness *and* flexible long-term memory seems to be connecting two fairly separate memory stores. This is the quick and dirty solution that evolution happened upon in creating the human brain; AGI architects have also gravitated toward this kind of solution.

It's not just in the context of working memory and long term memory that this sort of quick-and-dirty, efficiency-driven compromise occur — this sort of thing actually happens over and over again, in nearly every aspect of the mind. The practical requirements for a real-world intelligent system, including energetic and computational resource restrictions, seem to naturally push toward a system that's divided into a bunch of different aspects, all interacting with each other in various ways. The resulting systems are complicated, but they work. The more elegant ways of organizing intelligence, without so many complexly interacting subordinate parts, so far as we've been able to discover so far, only exist in the domain of abstract mathematics, with its endless energy and resources, rather than in the real world where practical pressures still prevail.

When looking at the human body as an analogy, this aspect of the mind doesn't seem surprising at all. The body has a lot of different organs, each carrying out their own specialized functions in specialized ways. The organs interact with each other complexly, and in many cases they have evolved specifically to take account of each others' functions. But still, there's a lot of complex, self-organizing messes involved. I'm sure it's possible to engineer a much more elegant body than the human one, with more unifying mechanisms and principles. However, I'd argue that there's still going to be a lot of heterogeneity and complexity, even in a really elegantly designed humanoid robot. The problem of designing a good foot has relatively little to do with the problem of designing a good ear.

For example, look at the following design for a disaster response robot, which my friend and collaborator David Hanson created as part of an AGI robotics project we were discussing back in 2012.

This robot was never built due to lack of funds, but we have the specs in case you want us to build one for you! It's kinda cute, in my opinion. If I were a victim of some disaster, I'd be pretty happy to have such a charming bot roll in and rescue me..



Figure 61: Notice the feet— they are actually wheels! Based on software instructions, they can deflate and then be used as flat tire feet. The legs also have a special design, so that they can switch from being compliant (flexible) to rigid, based on software commands. The feet and legs were designed based on a common principle: the software control of the transition between different states. The particulars of the feet and legs are obviously very different, engineering-wise. This is because being a foot is a different sort of enterprise than being a leg. In principle, one could design some sort of highly flexible mechanism that could serve effectively as both a foot and a leg. But, that wasn't David's approach when designing the robot, as it wasn't really a plausible way to proceed based on the materials and knowledge at hand.

I think this is basically the same as the situation inside the human mind with long term and working memory. The easiest way to create a system with limited resources that has both kinds of memory is to create two somewhat distinct memory subsystems, which will then need to interact in various ways, leading to various complexities. Yet ultimately, hundreds of different cognitive subsystems, each serving its own purpose relative to the goals and environments for which humans evolved, and each interconnected for effective combined operations.

In building an AGI, we don't need to emulate all the particular subsystems of the human mind and their interactions. We just need to build something generally similar in nature, then tune, and adapt it for more and more intelligent functionality. The physical substrate of our AGI programs and robots is currently rather different from the human body, which already implies that doing things in our AGI minds exactly the same way as in the human mind doesn't make sense.

Motivation and Action Selection

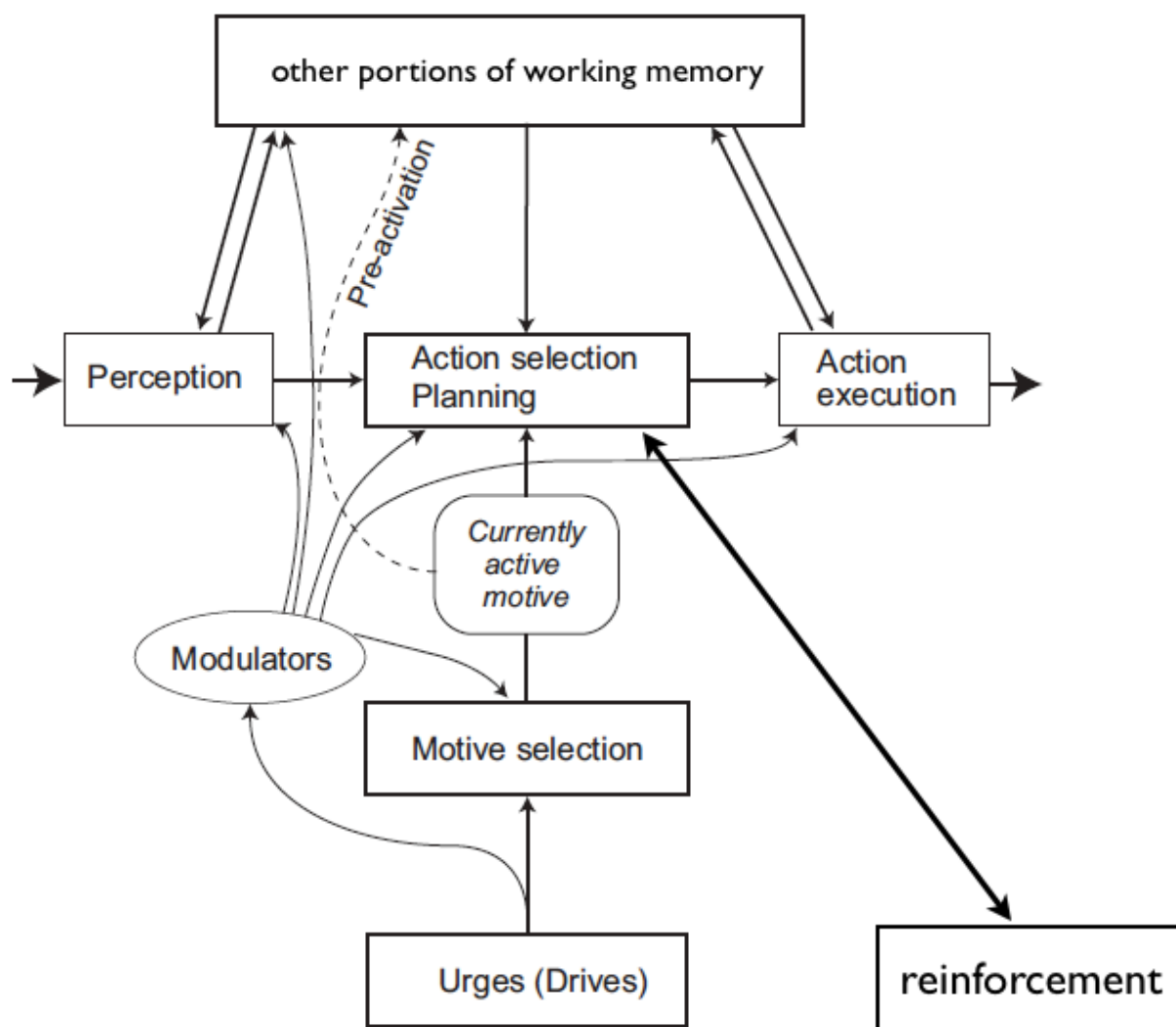


Figure 62: Action Selection

One of Stan Franklin’s cognitive science insights was that, to understand the mind in a simple way, it pays to start from the perspective of ACTIONS... To ask, what does the mind DO in the world? One of the big differences between narrow AI and AGI is that narrow AI programs tend not to be “autonomous agents” – they tend not to be agents that explore the world on their own and carry out their own actions

in pursuit of their goals. Rather narrow AI systems tend to be components that are used as tools by *human* agents, in pursuit of specific human goals. But humans, like animals, are autonomous agents, exploring and acting and surviving and striving... And AGI systems, if they're going to be remotely human-like, have got to be agents in this sense too.

With this in mind, Franklin has sometimes termed his approach to AGI the “action-selection” approach. The key question about an AGI system, in this view, is: How does it choose which actions to take, at which points in time.

The diagram above presents one pretty good model how humans choose their actions. It's drawn from the work of the German cognitive scientist and AGI designer Joscha Bach, who in turn drew inspiration from the German cognitive psychologist Dietrich Dorner. Dorner created a model of human motivation and action selection called “Psi,” and Joscha created a proto-AGI system called “MicroPsi,” which uses Psi-inspired ideas to control artificial agents. MicroPsi has been used for some practical purposes, but I've studied one of its research applications: a program used to control animated agents in a simulated world, as they seek food, avoid their enemies, and so forth. All this is quite compatible with Stan Franklin's view as articulated in Diagram 2 above – I just think Dorner and Bach have broken down the action selection process in detail in a slightly more useful way.

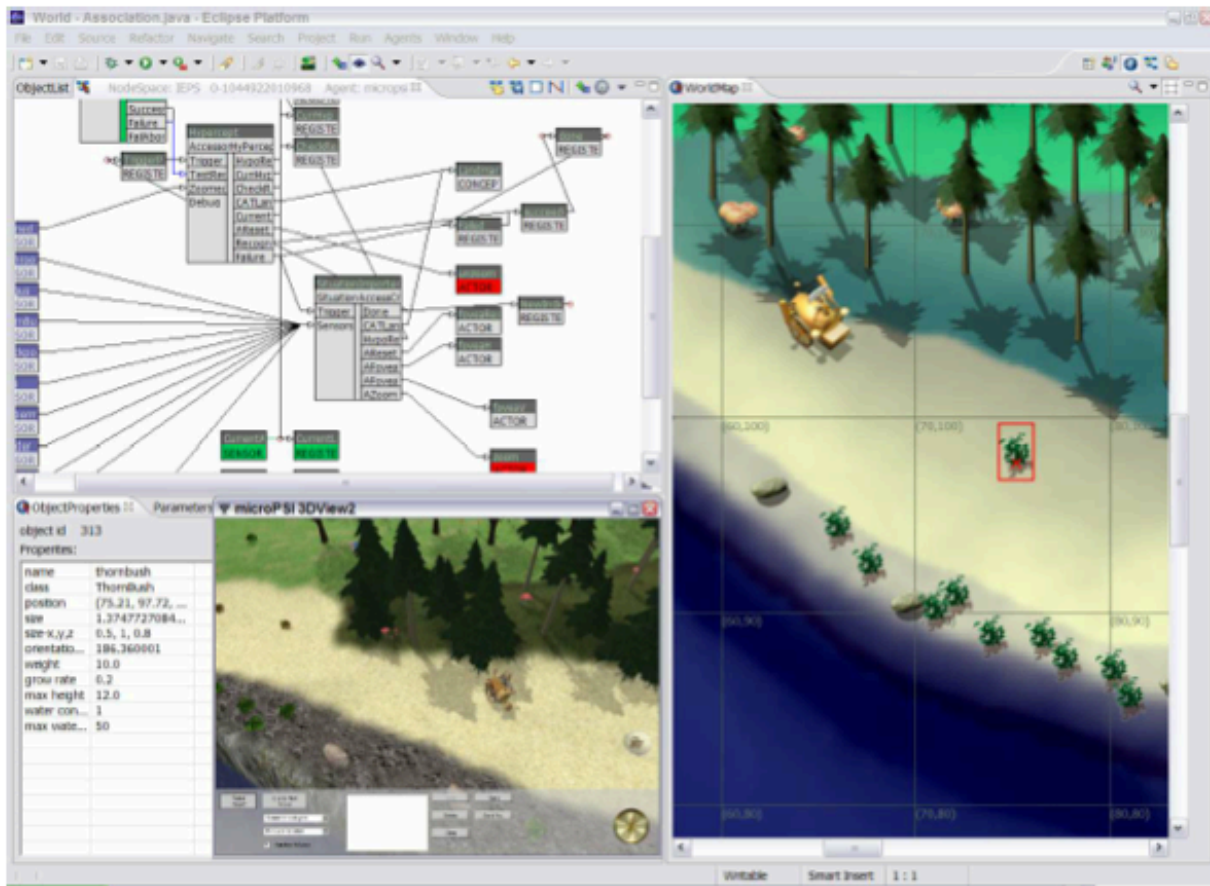


Figure 63: User interface for viewing, inspecting and controlling the mind-states of MicroPsi-controlled agents in their virtual world. Image courtesy of, and copyright to, Joscha Bach.

The basic ideas of Psi are not so complicated. A mind is viewed as having a bunch of different drives, or urges, which ultimately motivate behavior. For instance: get food, get water, get sex, be safe, learn new things, interact with another, etc. Obviously, the basic drives may be different for an AGI than human beings.

Then, at any given point in time, Psi models the mind choosing one of the drives as its key motive—the main thing it’s trying to accomplish at that point in time. This is a bit of a simplification, since arguably a mind could actively work toward more than one high-level motive at the same time. But it’s a good approximation most of the time, especially because the model allows other motives to take up some energy in the background. The “motive selection” box in the diagram refers to the process of choosing which motive gets the most attention. In the language of the Working Memory diagram (Diagram 2), this is largely a matter of “attention processing.”

Once a motive has been selected to focus on, one must choose what action to take, based on that motive and the mind’s prior knowledge about what actions have helped it to fulfill similar motives in the past.

Reinforcement learning plays a role here, along with many other sorts of learning based on working and long-term memory. Various parameters (which Psi calls “Modulators”) affect the action selection process. For example, one parameter governs the system’s pace. Being hurried affects many aspects of a mind’s intelligence, including how much detail it studies in its perceptions, how carefully it checks its inferences regarding future actions, and so forth.

Does Psi tell us everything about the human mind? Of course not. For one thing, what Psi says is equally applicable to any complex animal. Also, Psi doesn’t tell you how the mind sifts through large masses of data in short periods of time, nor how the mind forms abstractions. But it does tell you the basic logic via which a human-like mind selects actions, in accordance with its motivations. And that is something important.

Emotion

You may wonder: Why is there no box for “emotion” in our action selection diagram (Diagram 3)? After all, in humans, emotion plays a vital – maybe primary – role in governing our choices of what actions to take. We are emotional beings!

The reason there is no emotion box is that, according to Psi, emotion is an *emergent* phenomenon— a system level response to the system’s overall activity, in reaction to what it’s doing, observing and experiencing. Emotions are high-level patterns of dynamical activity that span ALL the boxes, in other words.

For instance, if an intelligent system keeps getting positive reinforcements from its body, and especially if it gets these reinforcements *when it does not expect them*, then it will tend to experience the system-wide response pattern of “pleasure.”

If the same intelligent system keeps getting thwarted in achieving its goals by some other agent, unless it has a particularly emotionally mature self-model and overall system dynamic, it will tend to experience the system-wide response pattern of “anger.”

And so forth...

Of course, not every human emotion needs to be included in an AGI system. Why should an AGI need to experience anger or jealousy, the same way people do? Yet the basic structure of human emotional response does not seem to be a highly specific consequence of the human mind/body, but rather a consequence of the general relationship between any embodied intelligent agent and its world.

One of the more popular ways of thinking about this, in the cognitive science field, is the so – called “cognitive theory of emotions.” This theory attempts to boil down all emotions to a few simple parameters, in a systematic way. The figure below gives the basic idea.

For instance, “pride” is our label for the kind of emotional response typically associated with approval of the agent’s own actions. “Admiration” is the kind of emotional response associated with approval of another agent’s actions. “Disappointment” is the kind of emotional response associated with negative evaluation of events that have personal consequences. In each case, the emotion in question is not merely a logical observation but a system-wide response. “Pride” isn’t just a logical observation that “I approve of these things that have occurred, associated with my actions”; it’s a coordinated response of many portions of the mind, correlated with this logical observation.

Of course, human emotions are too complex, messy and multifaceted to be captured in any specific logical formulation of this nature. But the cognitive theory of emotions provides an explanation of the common core of emotion that spans any kind of intelligent agent that controls a body in a social world. Different kinds of intelligences will then manifest these abstract emotional structures in different ways.

I saw the truth of this when, in 2007 and 2008, I was working on a project involving AI-controlled virtual pet dogs. The system we were working with was fairly simplistic. But taking it as an inspiration for thought-experiments, it became clear that , as this *kind* of virtual canine entity interacted with its virtual world more and more extensively, it would be able to “experience” every kind of human emotion. Dogs not only have simple emotions like happiness and sadness, but also envy, disappointment, pity and so forth. They experience these emotions according to the same basic logic as humans, but with different, doggish particulars. In a virtual world context, one can set up situations evoking each of these emotions. One can evoke virtual dog disappointment, for instance, by showing an AI dog 10 situations where there’s a little red house with food inside it – and then, the 11th time, showing it a little red house with no food inside. The observation that there’s no food inside the 11th house will cause reverberations through the AI dog’s internal state in a variety of ways, constituting the virtual dog’s experience of disappointment (or the “structural” equivalent, as in the figure).



Figure 64: Screenshots from some of the work we did in the mid-aughts, using the Novamente Cognition engine to control virtual dogs in virtual worlds. One thing we learned from the research we did in the course of this project was that dogs can experience basically the same range of emotions as human beings, if placed in the right situation. To be fully realistic, a virtual dog's emotional model has got to be basically as complex as a virtual human's emotional model.

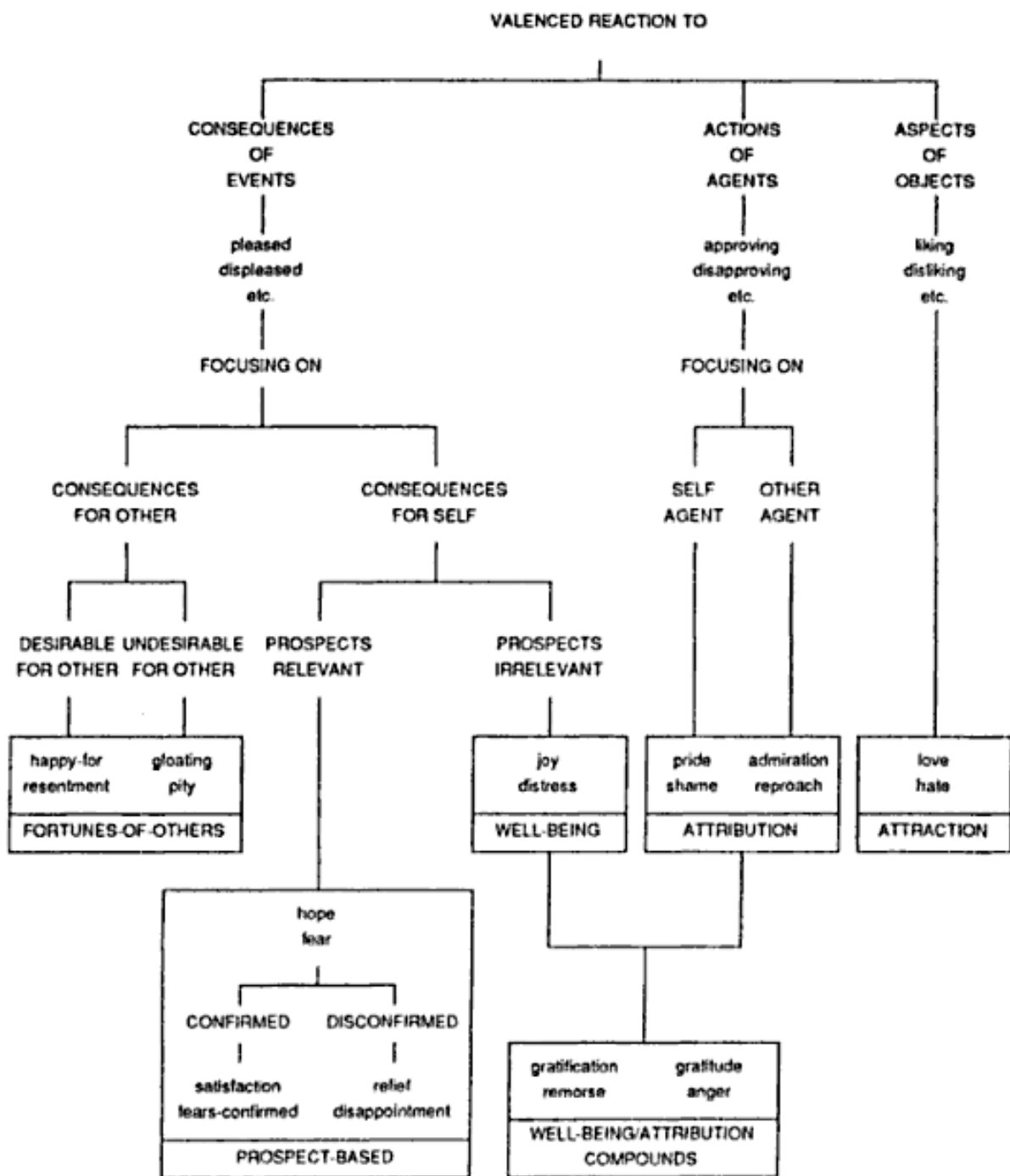


Figure 65: Illustration of the “cognitive theory of emotion,” showing how various common emotions emerge from basic cognitive aspects of an intelligent agent.

https://www.goodreads.com/book/show/1927037.The_Cognitive_Structure_of_Emotions

All this may seem like an overly mathematical or algorithmic way to think about something as raw, personal and experiential as emotion. But I like to remember the cognitive psychologist George Mandler’s terminology of “hot” versus “cold” emotion. Hot emotion is the raw feel of the emotion. Cold emotion is the structure of the emotion: What does it react to? How does it urge one to act? How does it cause one to represent and relate information? To put it simply, we might say:

cold emotion + consciousness = hot emotion

In other words, the raw feel of emotion, the hot emotion, is just the subjective, conscious experience that correlates with the structure identified by the cold emotion. Just like the raw feel of seeing the color red is the subjective, conscious experience that correlates with the visual stimulus corresponding to the color red.

If there's a mystery to the experience of emotion, I would say it's part of the broader mystery involving the experience of consciousness. The mysterious thing about emotion is not its structure, nor its connection to an intelligent organism's ideas, goals and reactions, but rather the way it FEELS. But then, I would say: The conscious experience of solving an equation or perceiving the color purple is equally mysterious as the conscious experience of feeling disappointed or ecstatic.

Consciousness is a tricky problem – but I don't think it has to be solved in order to build an AGI. Similarly, the philosophy of time and space involves a lot of thorny issues that nobody has resolved yet – but that hasn't stopped us from building spaceships, lasers, particle accelerators and so forth. Engineering isn't done by fully understanding some aspect of the world and then leveraging this full understanding to do stuff. Rather, it's done by understanding ENOUGH of some aspect of the world to do what one wants to do – and then advancing fundamental understanding gradually, alongside implementation of and experimentation with various practical constructs.

My colleagues on the OpenCog project have several views on consciousness. I suspect that fully understanding consciousness will require going beyond the current scientific world-view – not necessarily into the domain of mystical religion or anything like that, but a new way of thinking. For instance, I'm fascinated and perplexed by the idea that subjective experience and physical reality are just different ways of looking at the world.

I remember when I first got glasses for my nearsightedness: I was 5 years old. Suddenly, the world was a totally different place. And if next year I got an operation enabling me to perceive infrared and ultraviolet light, the world would seem totally different again. An objectivist philosophy, the most common one in the modern scientific era, maintains that there's some objective world out there, and that as my eyes improve, I'll perceive it more and more accurately.

On the other hand, this very theory is something that humans have built up from their observations. The objective world, as taught to us by science, is something people have abstracted to explain the

observations made using their laboratory equipment. And these observations are part of their subjective reality. Science is ultimately founded on what philosophers call “inter-subjective” reality – the reality when the various members of the community of scientists look at certain lab equipment in the context of certain experiments, they all subjectively report seeing the same thing. So when you really dig deep, you can explain subjective reality as an approximation or observation of objective reality; but objective reality is something created out of subjective and inter-subjective observations. And consciousness, I feel, has to do with this tangle, this strange loop via which the subjective and objective create each other.

Absolutely no scientific reason exists to believe that the human brain has more capability for conscious experience than digital computers. In fact, the very concept that particular physical systems uniquely give rise to conscious experience where others do not, reveals lots of logical flaws and contradictions when you examine it closely (readers with an analytical-philosophy bent are encouraged to look up the writings of Galen Strawson). Some AGI researchers think that “consciousness” is a red herring, and that we should just talk about information processing. Some AGI researchers are pan-psychist, believing everything in the universe has some measure of consciousness, but that entities manifest their consciousness differently. Brains are conscious in brain-y ways, AGI systems will be conscious in their own ways, and rocks are conscious in their own (presumably less intense) ways. Some AGI researchers, like me, think there are interesting mysteries to be solved in the area of consciousness.

Anyhow, it doesn’t seem necessary to resolve the perplexities of the philosophy of consciousness in order to build AI or AGI systems that do intelligent things in the world – any more than it’s necessary for an artist to fully clarify the philosophy of beauty to make a gorgeous painting.

I think AGIs will have emotions and other conscious experiences roughly as people do – though their emotions will have a different flavor, rooted in different forms of embodiment and mental algorithms. If we look at an AGI’s software, or a human’s brain, we won’t see consciousness or hot emotions, just a bunch of information flowing around. But from the subjective point of view of an AGI, and the corresponding human perspective, there will be plenty of consciousness and hot emotion in there.

Deliberative Processing

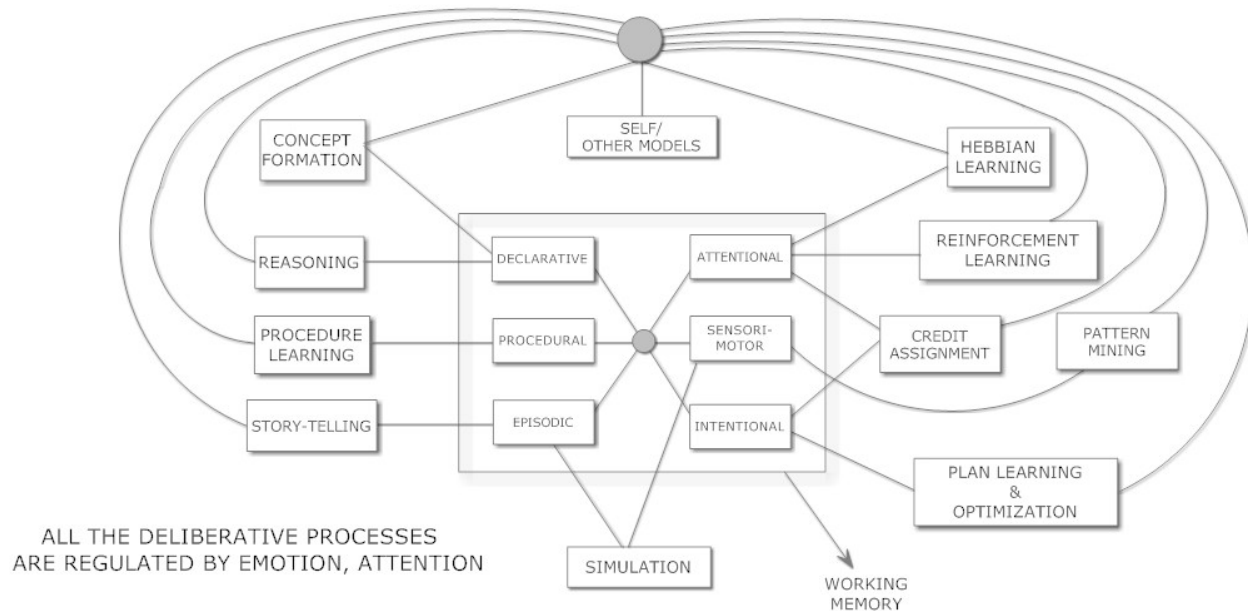


Figure 66: Long Term Memory and Deliberative Processing

Diagram 4 depicts the parts of the mind I’ve spent the most time thinking about and working toward in AGI systems: long-term memory and the associated “deliberative” thought processes.

We’ve already talked a bunch about “working memory,” the memory used by the mind in the short-term to carry out specific tasks. This is closely tied to the “cognitive cycle,” the running loop of perception, cognition and action that uses the working memory to monitor the environment and carry out immediate tasks. According to cognitive science, in the human mind this is distinct from the process of “deliberation in long-term memory.” This process involves deep, ongoing thought and concerted reasoning, which require searching the memory for new knowledge in a deliberate way and trying to focus it on the problems at hand.

Humans spend a lot of time doing this kind of deliberative thinking; some of us more than others. Animals, like dogs, birds, and pigs also do it, but it doesn’t play as large a role in their mental lives.

What Diagram 4 depicts are the multiple substructures and sub-dynamics within deliberative thinking and long-term memory.

First of all, there are three main kinds of long-term memory:

- **Episodic memory.** The memory of our life history. This includes “imaginative episodic memory,” thoughts about what *might* happen to us, and what *might* happen to others in various

circumstances.

- As I type these words, I'm sitting on an airplane flying from Tokyo, where I visited my oldest son Zar, to Hong Kong where I live. The set of multisensory images in my mind, pertaining my visit to Zar's apartment, is an episodic memory – which is obviously tied in with a bunch of memories of other sorts.
- **Declarative memory.** Facts, beliefs, statements, propositions and conjectures. These are the kinds of things you could naturally express in a few impersonal sentences in any language.
 - Regarding my recent visit to Zar, these would include facts like “Zar's new apartment is in Tatebayashi” and “Zar is now teaching in middle school”, and beliefs like “Zar fits awkwardly into Japanese culture” , “Zar's apartment is small but somewhat charming” and “Tatebayashi is a bit boring”. These facts and beliefs have varying levels of confidence in my mind.
 - 3). **Procedural memory.** The recollection of how to do both concrete and abstract things—like how to navigate, seduce a woman, walk, prove a theorem, write an essay, or generate a sentence. These are things that we know how to do and can teach someone else do instinctively. However, we can't explain declaratively in language exactly what we're doing when we are, in fact, doing it.
 - My memory of how to get to Zar's apartment from Tatebayashi Station is partially procedural. There are declarative and sensorimotor aspects to it as well – I remember some facts about the route, and I have some visual images of the walk in my mind. But principally I have a navigational procedure in my mind – if I were at the station again, then the procedure I learned for taking that walk would kick in, and I'd be able to find my way to his house by a combination of enacting previous habits, and leveraging previously gained sensorimotor and declarative knowledge.

If your goal is to model how human long-term memory works, nearly all cognitive psychologists agree pretty strongly that it makes sense to distinguish between these three different kinds of memory. It's a little less standard, but I also like to distinguish two other kinds of memory, which I think are fairly distinct in the human mind (and can become distinct in an AGI system).

I like to think about “attentional memory,” meaning knowledge of what or whom to pay attention to in

what circumstances. For example: On the walk to Zar's apartment from the Tatebayashi train station, I paid more attention to stores than to houses and apartment buildings, because there were fewer of the former. My attention was drawn to entities with greater surprise value, which is a common phenomenon. Also, as well as empirically evaluating surprise value, my mind unconsciously remembered that "pay attention to stores" is a good heuristic for what to pay attention to when navigating in residential areas.

And then there's what I call "intentional memory," knowledge of which goals we should pursue in certain circumstances, and how we break down a goal into sub-goals using a combination of other kinds of knowledge.

Sure, a mind may have just a handful of ultimate high-level motives, but in practice we don't always work directly toward those motives, instead we work toward various sub-goals that we think will help us achieve those motives. If a human's high – level motive is to "get sex," his or her sub-goal might be to seduce a foreigner, and then a sub-goal of that might be to learn a new language, but since language school costs money, a sub-goal of that might be to get a job to pay for language school, and so forth. Humans are pretty good at dealing with long sub-goal chains, compared to other animals.

In finding my way to Zar's place from the train station, I have a goal of getting to his apartment, and I also know a subgoal of this is to get to a certain major street that has a convenience store on it with a "P" sign in front of it, and on which a number of small streets end. Breaking down a goal of urban navigation into subgoals regarding recognizable streets or buildings, is intentional knowledge I've gained via years of walking around in various places.

Now, what does the mind do with these various kinds of long-term memory?

One of long-term memory's functions is simply storing stuff for a long time, so it can then be pulled back into working memory when something associated is perceived or thought of in working memory. It also serves as a kind of long-term, background, slow-paced "global workspace." Deliberative cognitive processes, operating partly in working memory, but partly in the background as "unconscious" cognitive processing, work together by reading information from, and writing information to, the mind's various interconnected long-term memory stores.

I view each kind of long-term memory as having its own specific kind of deliberative processing.

Declarative memory is closely tied to the various forms of REASONING. "Reasoning," as it's

considered in cognitive science and logic theory, goes beyond confident, rigorous mathematical reasoning, though. There's "inductive reasoning" - guessing that what one has seen in a bunch of cases is likely to continue. *If the last 5 big dogs I bit on the nose got mad and bit me, maybe the next one will, too.* There's "abductive reasoning" – guessing that things with some similar properties are likely to have other similar properties. *These four guys I've seen with fancy suits all had a lot of money, so maybe most people with fancy suits have a lot of money.* Analogy reasoning is another way of looking at induction and abduction by reasoning about new situations or objects by analogy to previous, similar cases.

One of reasoning's distinguishing properties is that it mostly proceeds step by step. You go from some facts or assumptions to a conclusion, and then on to the next conclusion from that one, etc.

(move up)The human mind is better at some kinds of reasoning than others, but overall we seem to be much better than other animals at carrying out long chains of reasoning of various sorts...

"Procedural memory" is connected to different sorts of learning. A young child doesn't learn to walk by reasoning about it. They learn by trying different approaches, combining them and varying them till they find something that works, and then experimentally tweaking the working approaches they've found in the course of their ongoing practice. This is the same way we learn to play tennis or carry out other more complex physical actions, and it's the same way we learn more cognitive procedures, like solving equations.

Since I'm reasonably good at math, when I sit down to solve an equation, if it's a familiar kind of equation, I don't waste time explicitly reasoning about what steps to take. I just do it. I already have a sense of what steps to take first, and I proceed to fiddle with rearranging the terms in the equation till I get to the answer. On the other hand, if it's especially tricky looking or an unfamiliar sort of equation, then maybe I will first explicitly and carefully reason about what approach to take.

This type of learning by experimentally inferring procedures also helps us master grammar. This is why, even after we know language perfectly well from a practical perspective, we still have to study to know the rules of grammar. We don't learn language, as young children, by reasoning, "Hmm, what grammatical rules must people be following, in order to produce the sentences they're producing, and not other ones." Rather, we use a kind of sophisticated trial and error, trying different ways of generating sentences and seeing which ones work, combining and modifying and fine-tuning the workable approaches.

Reinforcement plays a big role in this type of learning. We try stuff, and then change our approach based on constant feedback about our degree of success. There's also a heavily "evolutionary" aspect to this. Just as evolution involves the combination and mutation of genomes to create new organisms, learning new procedures involves combining and mutating tried approaches to find new ones. This evolutionary aspect relates to certain approaches in AGI, which involve computer science algorithms called "evolutionary algorithms," modeled on the evolutionary process.

"Episodic memory" seems to be tied to mental simulations running through a set of imaginative "what if" scenarios in your mind.

"Attentional knowledge" goes along with processes that cognitive scientists call "association-spreading." These are processes in which attention associated with one thing spreads to other, related things, spreading activation through the mind. At the end of the 1800s, William James and Charles Peirce discussed this kind of process at length. Peirce thought it was the crux of intelligence:

Logical analysis applied to mental phenomenon shows that there is but one law of mind, namely that ideas tend to spread continuously and to affect certain others which stand to them in a peculiar relation of affectibility. In this spreading they lose intensity, and especially the power of affecting others, but gain generality and become welded with other ideas.

Association-spreading has been repeatedly simulated in neural network models.

Finally, "intentional memory," relating to the pursuit of goals, integrates all the different kinds of memory and thought in the mind: Reasoning, procedure learning, association spreading, and simulation of episodes in which similar goals were achieved.

But what about "metacognition," which is separate from deliberative thinking, according to the Diagram of High-Level Architecture of Human-Like Minds? Metacognition refers to "thinking about thinking." I've done an awful lot of this as an AGI researcher – but it's not restricted to AGI researchers! Comedians, like Woody Allen, have dramatized the tendency of certain individuals and cultures to introspect and obsess on their own thoughts and feelings. Jewish culture, in which both Woody Allen and I grew up, tends to have this characteristic – maybe this is part of the reason there are so many Jewish AI researchers!

In terms of human psychology, it makes sense that metacognition would be considered a separate capability from plain old cognition, as some folks are better at metacognition than others, somewhat independently of their intelligence in other respects. However, in spite of its different characteristics, I think metacognition involves the same processes as plain old deliberative thinking. In metacognition, instead of just thinking about general cognitive content, these processes are tuned and shaped to think about thinking.

Julian Jaynes, in his fascinating book *The Origin of Consciousness in the Breakdown of the Bicameral Mind*, observes that in Homer and other early Greek fiction, the narrators tend not to refer to their own states of mind, but rather to the voices of the gods speaking inside their minds. The 20th century has given us a radical leap in metacognition since Homeric times, with the advent of a host of specific technologies for analyzing and modifying the individual mind, from psychoanalysis to more modern approaches like rational-emotive therapy, neurolinguistic programming, rationality boot camp, and so forth. Medieval Indian philosophers comprehensively studied the working of the mind – analyzing consciousness in terms of 128 possible conscious states, each with unique properties – and they tied their work into the vast meditation community.

I'm not sure how great apes, parrots or dolphins are at *metacognition*. The human capability for metacognition appears to have advanced with the capability for complex uses of language. Language is an invaluable tool for describing our own thoughts and selves, to ourselves and others. However, ultimately our ability to metacognize is impaired by our difficulty in perceiving our own thought processes. Most of our thoughts are “unconscious” – meaning outside the scope of our rational deliberative thought processes. And of course, we can't physically observe the state of our brain, the same way as the state of our fingers or feet. AGI systems will eventually be able to observe their own thoughts with much more flexibility and accuracy than humans can, giving them dramatic advantages in the area of metacognition – which should help them self-improve more purposefully, systematically and rapidly.

Perception, Action & Language Hierarchies

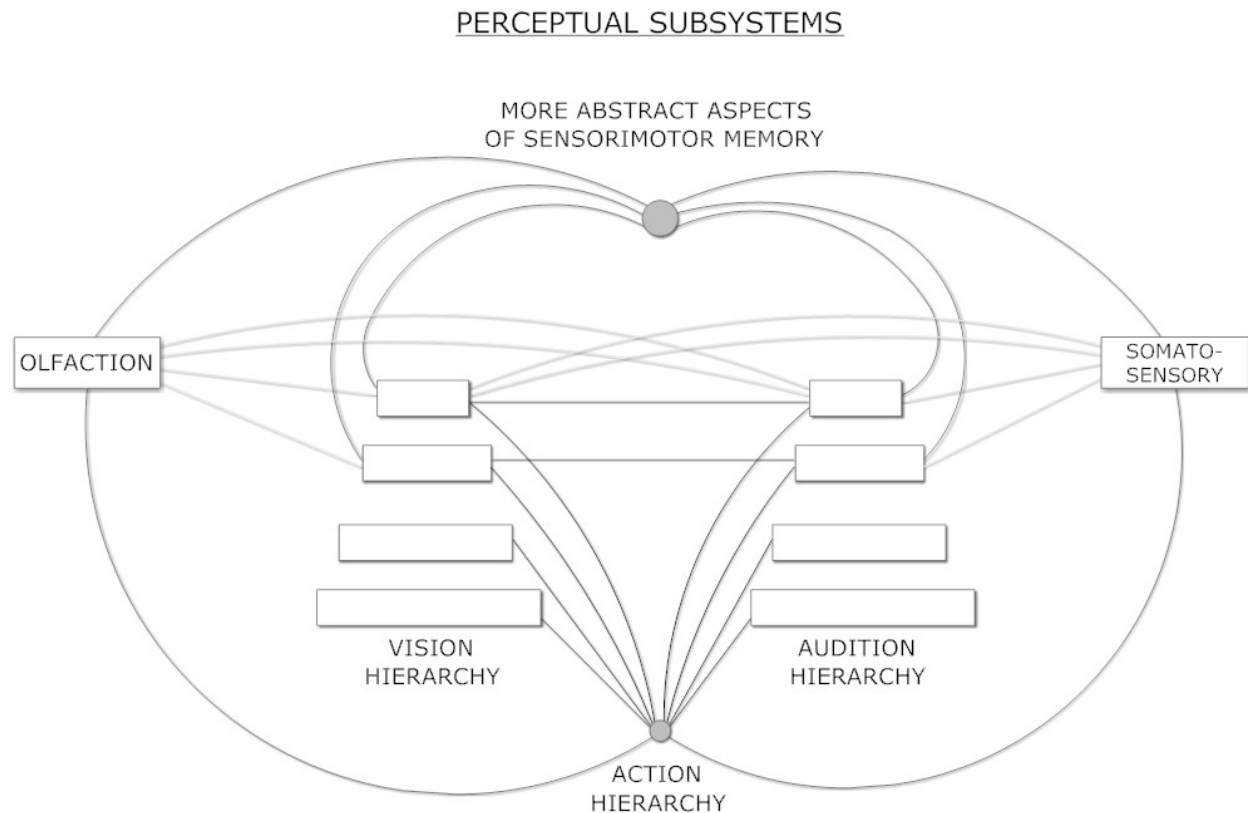


Figure 67: Perception

Now, let's bring the discussion a little closer to reality. All these fancy kinds of memory and thinking have to get their data from somewhere. In theory, a mind could just take raw data from its sensors and start reasoning about it logically, and learning complex procedures to act based on it. However, in reality this wouldn't work too well, as the cognitive processes used for advanced deliberative thinking, and even for practical, real-time decision-making based on working memory, don't function adequately if you feed them a huge amount of data at once.

If you tried to reason logically about the geometric relations between all the pixels on the screen of your TV, you'd still be thinking really hard about one corner of the screen when the picture changed to something else. If you tried to figure out how to serve a tennis ball via carefully adapting each muscle movement to each perception from each part of your body during the serving action, you'd get overwhelmed by all the possible interrelationships between sensations and actions, and wind up tossing the ball and watching it fall to the ground while thinking about what to do.

So, because of having restricted processing resources, as there's only so much thinking the mind can afford within a certain allotted time, the mind needs specialized methods for processing the flood of

sensory data it receives and controlling every part of its body. Complicating things further, these methods need to be specially adapted for various kinds of sensation and action.

The parts of the human brain concerned with visual and auditory perception are hierarchically structured, as illustrated in Diagram 4. For example, the lowest level of the vision hierarchy deals with tiny visual details, “pixels,” so to speak, as they change over brief periods of time. The next level deals with slightly larger regions of visually perceived space-time, and so on. Finally, the highest levels deal with general high-level visual shapes and movements. Information passes up and down the hierarchy. The lower levels pass information to the higher levels, so that each level recognizes patterns in the output of the level below it; and the higher levels provide context that biases the pattern recognition in the lower levels. Audition (hearing) works about the same way, but with different levels dealing with differently-sized regions of time rather than space.

Some AGI researchers believe the hierarchical structure of the visual and auditory cortex represents how the rest of the brain is organized. Jeff Hawkins and Itamar Arel, as I mentioned above, have both proposed AGI architectures that have a strict hierarchical structure, not just for visual and auditory information, but for everything. However, I’m not so sure this is the right path. Of course, all pattern recognition has SOME hierarchical aspect to it. The mind is always recognizing patterns among patterns among patterns. But I don’t think a strictly hierarchical structure for guiding pattern recognition is the right approach for handling, say, declarative knowledge that requires logical reasoning, or for accessing the episodic memories of one’s life history.

The human brain handles some of its other senses quite differently from vision and audition. For example, the olfactory cortex, the part of the brain managing smell sensations, is dominated by combinatorial, tangled-up connections between neurons, snaking all over the place rather than being arranged hierarchically. Each recognized smell seems to be represented by an “attractor” pattern—a habitual pattern of activity, distributed across a large region of the olfactory cortex. As a smell is recognized, gradually more of the attractor pattern becomes active in the olfactory cortex.

Tactile sensation works via the brain maintaining a distorted internal map of the body. There’s a region of neurons in the brain corresponding to the back, another corresponding to the elbow, etc. Each fingertip gets more neurons than the whole back because it’s more sensitive and has more nerves. Recognizing tactile sensations seems to be more olfaction, attractor-like than hierarchical like vision.

The senses are not processed separately in the human brain. Instead, sensory processing often occurs in

a multimodal way, where the regions of the brain that correspond to the different senses share their interim conclusions with each other to help reach better ones. Currently, AI systems process sensory data very differently, dealing with the data of each sense in isolation, rather than using them as a whole to reach a unified understanding of the world.

Action, as depicted in Diagram 5, has a hierarchical structure somewhat similar to vision and olfaction. Higher levels of the brain's action hierarchy (resident in the cerebellum and certain parts of the cortex) correspond to large-scale actions, generally taking place over larger regions of space and time. Lower levels generally refer to quicker, more localized actions. For instance, in coordinating a tennis serve, the higher levels of the hierarchy would contain action-patterns corresponding to the overall shape of the body's motion while serving; the lower levels would contain details like the exact way the wrist is moved in response to a certain movement from the shoulder, and the particular speed of movement of the heel of the back foot as it's lifted. As with vision and audition, the hierarchical structure helps guide learning. Action learning, like visual and auditory perception, uses learning algorithms for patterns at each level of the hierarchy, based on the patterns at the immediate upper and lower levels.

The action and perception hierarchies are not separate but cross-connected, so that perceptions can be used in the course of actions (to correct and guide the course of actions), and actions can be used in the course of perceptions (to help gather perceptions more accurately). The arm, as it controls the tennis racket attempting to hit the ball back over the net, corrects its course in real time (including both macro and micro movements) based on what the eye sees the ball is doing. At the same time, the neck moves the head in a certain direction, based on the need to see the ball better, driven by instructions from the visual hierarchy that says it wants to see more detail.

The action and perception hierarchies are also closely related to long-term memory, and the various kinds of knowledge stored therein. Relatively simple computer algorithms can match a human's capacity to recognize objects in photographs (IF the person is only given half a second or less to look at each photograph). However, simple algorithms don't fare as well when people are given a few seconds more to scrutinize the photo. If a person spends a few seconds looking at a photo, they have time to bring up knowledge from long-term memory. If a sensory processing task isn't obvious, deliberative processes based on long-term memory come into play, and interact in complex ways with perceptual and motor processes.

The same holds for actions. When I'm playing tennis, and I see the ball is coming in a slow but strange

way, maybe because the wind caught it, then I may actually take half a second to THINK about my response, instead of just responding as usual by body-reflex. Great athletes are not only distinguished by their muscles and perceptual and motor systems, but also the subtle real-time feedback between their perception and action hierarchies and their mind's deliberative processes.

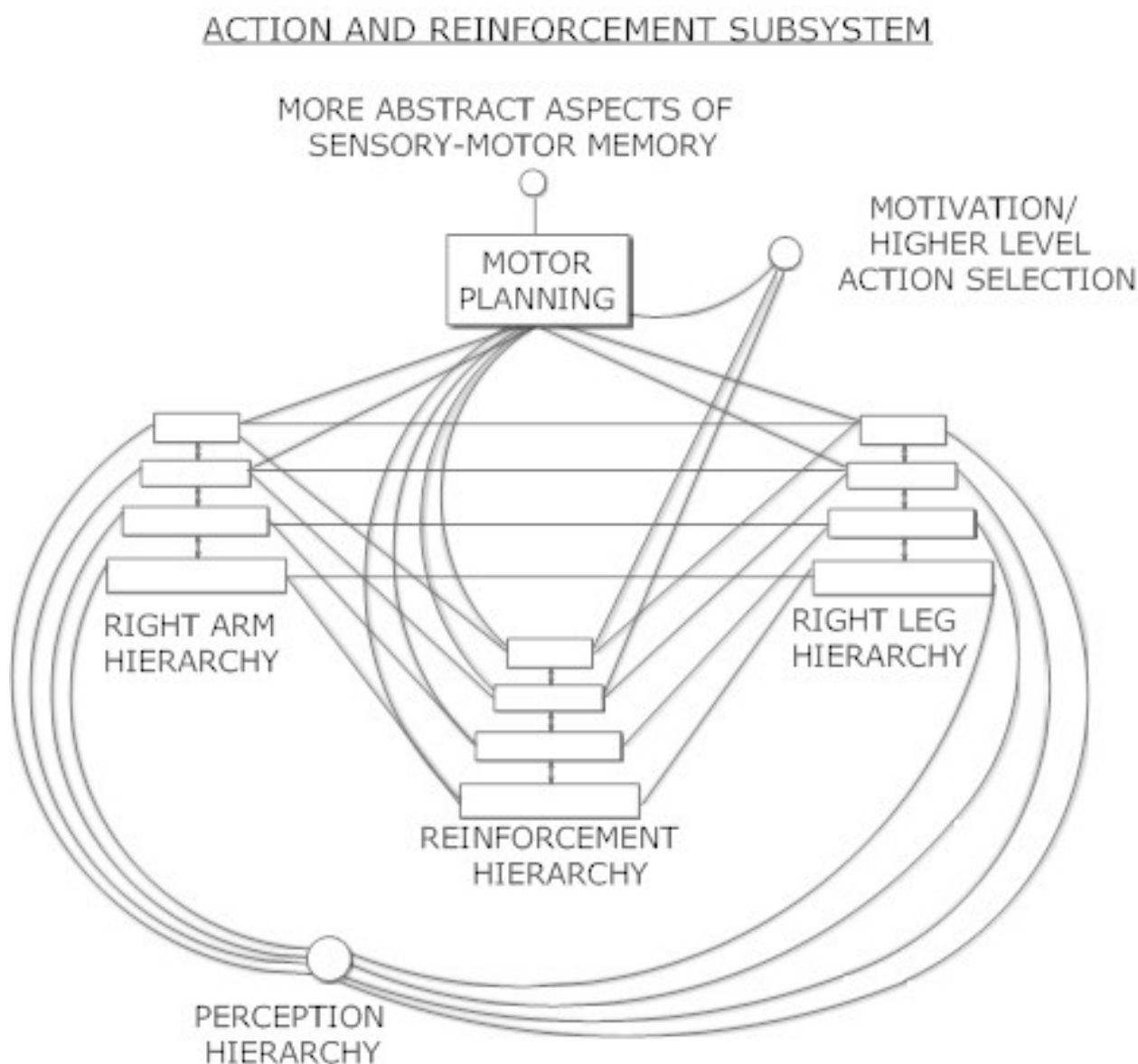


Figure 68: Action

Language

Language – viewed from a sufficiently abstract view – emerges as yet another of the human mind's intricate hierarchical structures, as illustrated in Diagram 6. Language processing spans perception (language comprehension: Understanding what others say) and action (language generation: figuring out what to say and saying it). The lowest levels of the language hierarchy are raw data-oriented,

recognizing patterns in streams of sounds, and generating streams of sound with the mouth and larynx. The higher levels focus on abstract patterns of linguistic organization.

Language comprehension deals with the arrangement of sounds into words, words into phrases, phrases into sentences, sentences into paragraphs, and so on. The recognition of patterns at each level is conditioned by the patterns recognized at the levels above and below. When a young child learns language, they start at the lowest levels of the hierarchy, but their partially-formed intuitions about the higher levels condition their learning. For instance, when a child learns new words, they are guided somewhat by their intuitive understanding of the intent and context of the overall discourse in which the words occur.

Language generation deals with basically the same hierarchical levels, but is more concerned with building than recognizing patterns. It starts with some thoughts to be articulated, usually due to a conscious or unconscious judgment that doing so is going to help fulfill one of the mind's current goals. Next, it figures out how to break this set of thoughts into small chunks suitable for linguistic articulation. Then, it turns each chunk into a proto-sentence: a linguistic series of words or word-meanings. Next, it fills in the specific content words. Finally, it fills in all the little words needed to make a grammatical sentence, and tells the body what to say.

The learning processes associated with language overlap with those involved in other cognitive processes. Sound comprehension is carried out by the auditory cortex, whether those sounds are linguistic or otherwise. Speech generation is handled through relatively generic action generation processes. At the highest levels, thinking about the meanings of sentences comprehended or generated, is achieved through the same deliberative inference processes used for other kinds of thinking.

Learning how to produce or understand sentences seems to draw from procedure learning methods, similar to those used when learning other complex things. Literature in the AI and cognitive science fields points out commonalities between the structure of sentences, a series of physical actions, and social relationships. It seems the mind uses largely the same methods to represent and recognize patterns in all these domains.

On the other hand, the human brain appears to use quite specialized methods for language processing. For instance, there are certain parts of the brain where, if you have a lesion in them, you will lose the ability to process verbs but not other parts of speech. These parts of the brain behave the same for everybody. So, language processing seems to be a case where the human brain's general intelligence capability and its specialization in its environment and goals, come together in complex ways.

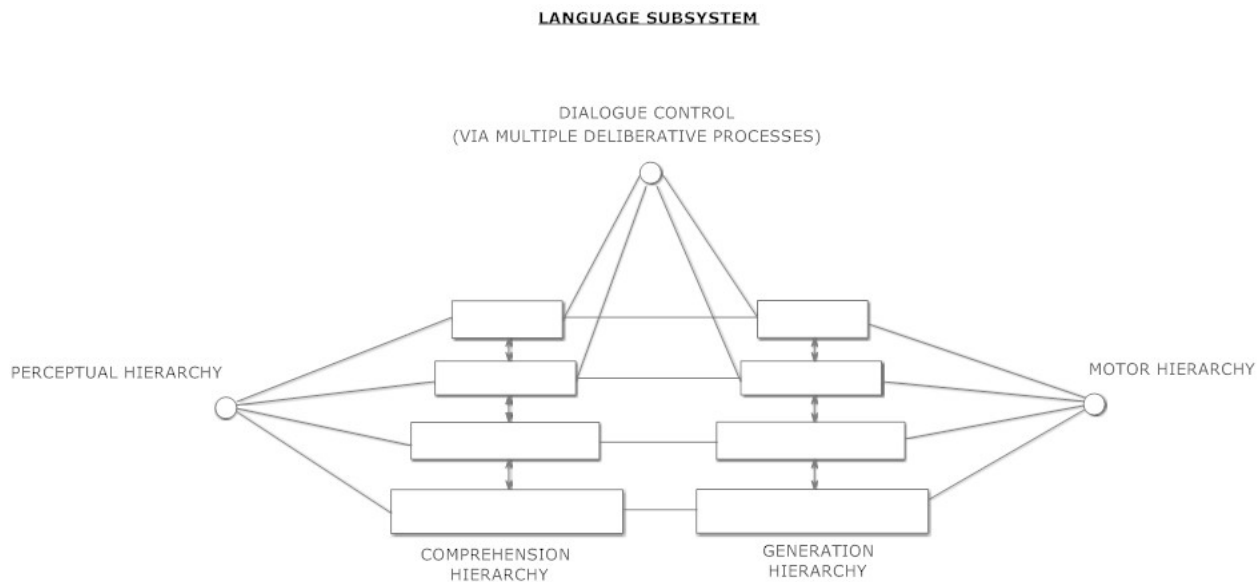


Figure 69: Language

Cognitive Synergy

Cognitive science breaks down the mind into a bunch of different parts, and reveals a fair bit about how each part works. It also tells you one very important thing: all the parts must interact together. You can't make a system using some algorithm or neural net to carry out each of these parts of human intelligence separately inside an individuated black box, and then just connect the black boxes together for communication purposes. The mind doesn't work that way.

Each of these cognitive processes, carrying out each of these aspects of human-level intelligence, must interact intimately with many other processes (carrying out many other aspects). Each process on its own would end up in a logjam necessitating unfeasibly large computational resources to do what more integrated systems could accomplish much more easily. Perception would get stuck trying to perceive what's there in some dark, modally complex seed. Declarative reasoning would get hung up trying to puzzle through some problem with too many possibilities. Metacognition would struggle to reason how to reason. When any one of these mental processes gets bogged down, it needs to be able to call on others in the midst of their own processing or thought process. In this way, the different mental processes can help each other out of difficulties, forming a cognitive synergy process in which every part of the mind depends on every other part.

Now, cognitive science tells us what parts of the mind must be present. It tells us that the various parts must be interdependent in a subtly synergistic way. But it doesn't tell us exactly what algorithms, what dynamical processes have to take place inside each of these boxes, nor exactly how they need to

interact with each other to manifest this cognitive synergy.

So, in my approach to AGI, we use computer science algorithms inspired by neuroscience, mathematics and pragmatic considerations to fill in the blanks. To dig deeper, you have to look carefully at our best theories regarding each of the parts. I spent a few decades doing just that, and in the next chapter I'll tell you about one of the AGI designs I came up with based on all that thinking: OpenCog. But first, I'll make a few more remarks on the theme of the general mess and complexity of the mind. Just to be sure you don't take all these boxes and lines too seriously!

Mind as a Complex System

Modeling the human mind with box and line diagrams make it seem almost like a circuit board, with discrete components passing information between each other in a crisp and well-organized way. This helps us understand some of the broader aspects of how the human mind works. But it's important to remember that neither the human brain nor mind ACTUALLY has a bunch of boxes and lines in it. Rather, the human mind – brain is a big, complex, dynamical teeming mess, and its intelligence emerges from this messiness, in a way that inextricably mixes creativity and flexibility with error and confusion.

The human mind/brain is, among other things, what scientists call a “complex self-organizing system.” Put simply: a complex, self-organizing system contains a lot of little parts that interact with each other continually, and in the process give rise to larger structures and persistent dynamics, nudging the little parts in particular, semi-coordinated directions. A built-in structure may guide the little parts in their interaction, but there's also a lot of freedom for the parts to quasi-randomly experiment, until something happens that causes an overall structure to emerge.

For instance, imagine a society with no government, nor other institutions. People in the society would interact in all sorts of ways, quite chaotically and heterogeneously. Eventually, some structures would start to emerge. Groups of people would band together for various purposes. Little towns would form, with extended families and farming cooperatives. Independent mercenary squads would roam the countryside. In time, some sort of overall government would emerge, either via some mercenary squad turning into an army and taking over, or else via a group of peaceable people deciding to get together and organize to prevent being taken over by mercenaries. Quite possibly a number of small states would emerge, each with their own governments, which would enter into alliances. Eventually some semi-stable emergent structures would arise, in the form of governments, companies, armies, and so

forth. These structures would then guide the further interactions of people in the society, not utterly constraining them, but instead directing them with significant force. The collective individual interactions between the people might eventually bubble up significantly enough to get rid of some of the major structures that originally emerged, e.g. a revolution might arise and overthrow some of the governments. The same sort of process occurs among cells in the brain, or, in another sense, ideas in the mind.

The infant's brain cells interact with each other, forming new structures of various sizes, and gradually crystallizing into overall cognitive and perceptual structures that guide it in further understanding the world. The structures that emerge collectively among the infant's brain cells are not specifically programmed in their DNA, nor specifically determined by their experience, which is why identical twins growing up in the same house can still have distinct personalities. These structures emerge via a sometimes-chaotic process of neural self – organization, which then goes on throughout the lifespan.

The ideas in a child's mind, which are patterns of organization among its brain cells, body systems and environment, also interact with each other. Early ideas combine to yield new, more sophisticated ones. Ideas combine, mutate, dwindle, and are reborn. Coherent networks and systems of ideas form, and, ultimately, belief systems form, some with great persistence. Eventually a world-view crystallizes, along with a self-image, and models of others in the child's environment. From this point on, the wild interactions of the teeming pool of concepts are channeled by the structures that have emerged in the mind: The world-view, the self-image, belief-systems, and expectations. However, the self-organizing generative mayhem is still there, and may lead the mind to come up with radically new ideas, or undergo dramatic personality transformations, or belief system shifts, even late in adulthood.

The boxes in the cognitive architecture diagrams are structures that emerge from the wild self-organization of cells in the brain and ideas in the mind, coupled with a body and an environment. Human genetics predisposes human minds toward building these structures, and nudges the infant's mind in this direction. Even so, each young human mind must build them for itself. This is the same in the social anarchy example, where human nature militates toward the emergence of mercenary armies and governments. Even so, each group of people must build these for themselves, guided by their genetic propensities and their own thinking.

A box like “episodic memory,” doesn't necessarily refer to a discrete, distinct component in a human mind or AGI system. Rather, it's a distinct functionality of a complex system. A certain collection of

brain cells, giving rise to a certain collection of ideas and thought patterns, begins to engage in the maintenance of episodic memories in a young human mind. This collection of brain cells improves at episodic memory maintenance, and as time goes on, systematically interacts with other parts of the brain/mind, receiving information from them and sending information to them with episodic memories. However, this same collection of cells (and ideas) may serve other functions as well, and this may affect the nature of the episodic memory. The interactions between the cells enabling the episodic memory, and other cells, may give rise to the emergence of new neural structures, or the alteration of the episodic memory. The interaction between mental patterns involved in episodic memory, and other mental patterns, may lead to the emergence of new mental structures or forms of episodic memory.

The boxes in a cognitive architecture diagram are shorthand for systematic patterns of organization in a complex, ever-changing, self-revising, and self-organizing system. The arrows are just shorthand for the naturally emergent pattern of interaction in these patterns of organization. This underlying complexity should not be forgotten. Yet, the approximate understanding provided by crisp, simplified models like cognitive architecture diagrams should not be ignored either.

A Strategy for Building Minds

Now that we have a fundamental understanding of how minds work, I can explain my approach to designing AGI. Here's a summary of my methodology:

First Look at everything known about human cognitive science. What are the main mental processes occurring within the human mind? How do they relate to each other? How do they operate dynamically?

Second Think about these known cognitive processes as a system: How can one make sense of the totality of known human cognitive processes in terms of the overall function of a human mind? That is, as a system controlling a body in an environment, constantly regulating and modifying itself.

Third For each selected cognitive process, figure out how to achieve what that process does via some reasonably efficient computer algorithm that will run efficiently on today's computers. Some of these algorithms may bear resemblance to how the brain works, while others might not.

Fourth In the system of the mind, look carefully at the interactions between all the cognitive processes that are modeled computationally; make sure that the chosen computational algorithms interact with each other properly.

Fifth Create a practical software framework allowing each cognitive process to operate simultaneously and interact with each other appropriately.

Sixth Create an agent powered by this software framework; let it control a body in a world full of rich data to interact with; and let the world be populated by human agents eager to interact with and teach the agent.

Seventh In this environment, carefully interact with the AGI agent to lead it through a series of natural developmental stages, modeled loosely on human developmental psychology.

Yes, this methodology is complicated and has many parts, each of which is difficult in its own way. But it doesn't require any knowledge that isn't already currently available, only the intelligent, judicious assemblage of things that are already known.

AGI and Human Childhood Development

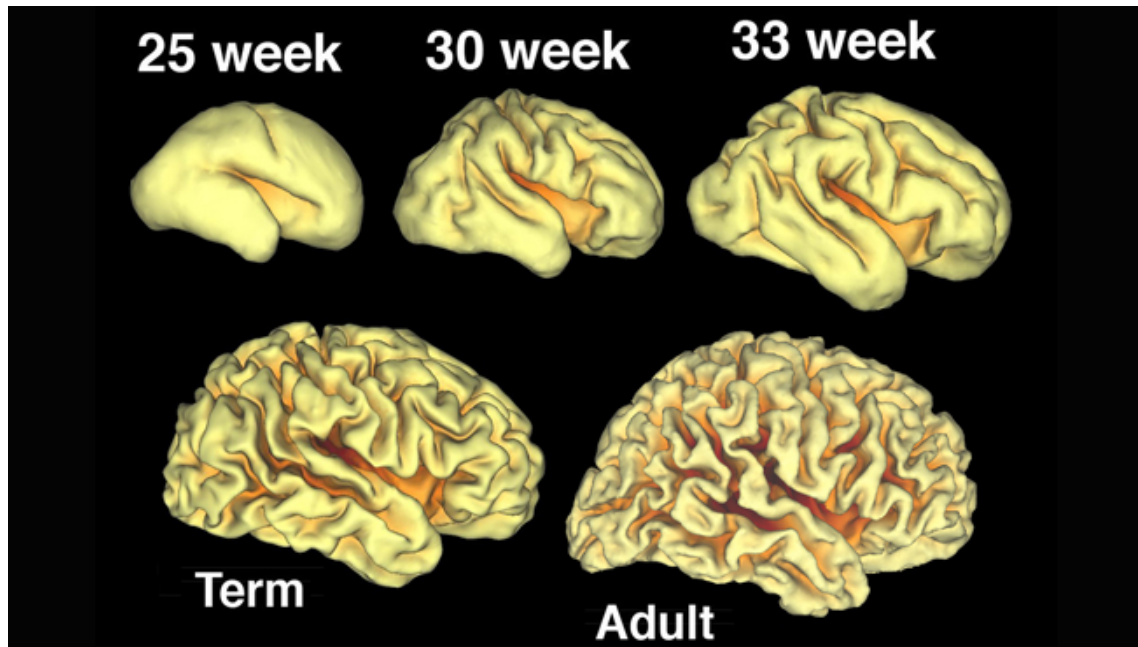


Figure 70: Maps of the cortical surface in premature infants (at several gestational ages), full-term infants, and adults. Some of the structure of the adult brain is genetically coded, some is acquired via learning, and some arises via spontaneous self-organization. How these three aspects balance off against each other, and interplay, remains for the science of the next decades to unravel. <http://www.brainfacts.org/about-neuroscience/technologies/articles/2009/brain-atlases/>

One of the things cognitive science has taught us, on the high level, is that an infant’s mind confronts the task of understanding the world with a complex armamentarium of learning processes and information storage structures. Regarding specific knowledge of the world, an infant is indeed largely a “blank slate”, as philosophers posited hundreds of years ago. Maybe an infant knows the a few things, like the smell of a lactating breast, but in terms of concrete pieces of information, it doesn’t know much. However, what the infant does have is a quite refined set of algorithms and biases for recognizing patterns in the world.



Figure 71: Oxana Malaya, an 8-year-old wild child who spent most of her life in the company of dogs, was discovered in Ukraine in 1991. <http://listverse.com/2008/03/07/10-modern-cases-of-feral-children/>

The whole apparatus of structures and processes I've described above is not there in an infant's brain right from the start – some of them are, and others develop gradually over time, sometimes triggered by the infant's experiences or passage of physiological milestones, sometimes just triggered by the passage of specific amounts of time. The field of developmental cognitive neuroscience studies the cognitive mechanisms in the brains of young children, and how they progressively come into play over time. But even so, right from the beginning, the baby mind is far from a blank slate. The development process starts the baby out with the cognitive mechanisms needed to come to a rough knowledge of the world. Then once it has learned a bit, the development process provides the young child's mind with the mechanisms needed to take the next steps.

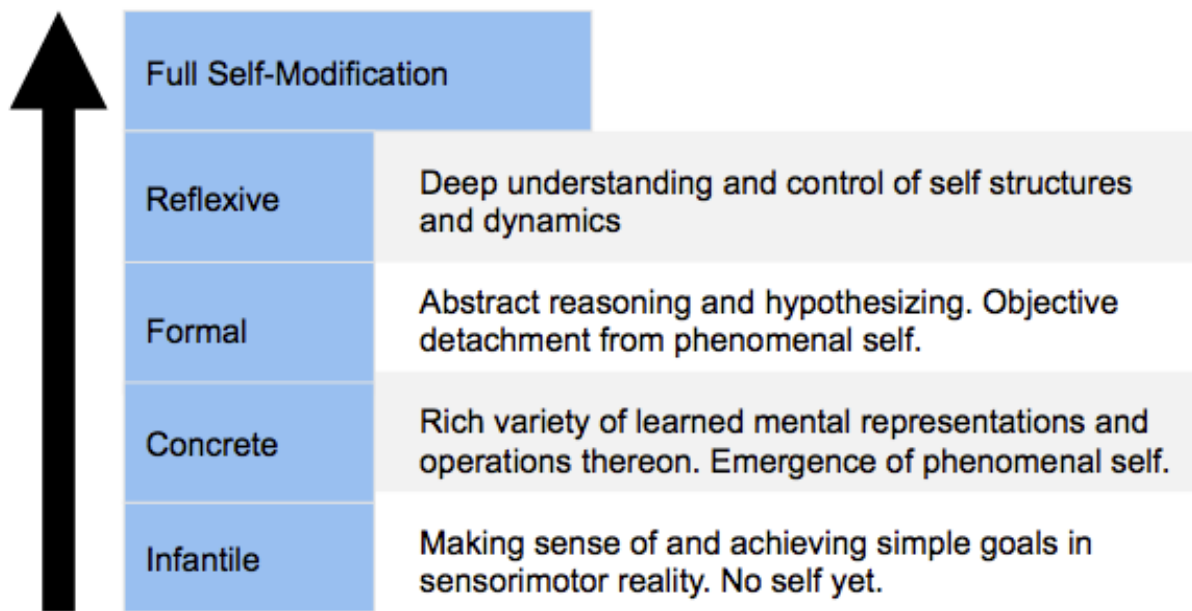


Figure 72: Depiction of Piaget's classic hypothesis regarding the stages of psychological development of human minds. Actually, Piaget only went up through the Formal stage, but a number of more recent thinkers have conjectured post-formal stage of development such as the Reflexive stage we've added here. A fully self-modifying AGI, able to reprogram itself with its goals in mind, would go beyond any directly human-relevant developmental stages, of course. Piaget's theory, in general, is best viewed as conceptual and inspirational rather than as a precise scientific theory of child cognitive development; modern developmental psychology has uncovered a massive network of nuances that make clear the real story of human mind development is far subtler than Piaget realized.

Language learning is an interesting, and much studied, case in point. Infants are born with a fairly strong ability to recognize linguistic patterns in the streams of sounds they hear. But the ability to understand complex syntax comes into play gradually. And if the child doesn't receive appropriate stimulation, the latter may never emerge – the development of the needed cognitive mechanisms will not be triggered. “Wild children” who grow up without anybody communicating with them, miss the chance to get their brain’s advanced language processing capabilities triggered in childhood. As a result, even once they eventually get integrated into society, they never quite use language normally. They can do short sentences and basic communication, but they’re never going to write like Joyce or Proust, nor even diarize like the average teenage girl.

It’s not entirely clear how useful developmental cognitive science is going to be for AGI. A lot of child cognitive development is tied to the specifics of child physiological development, which is not particularly relevant to near-future robot bodies. And, practically speaking, it would be a lot of trouble to take each cognitive mechanism one builds for one’s AGI system, and then build a series of progressively simpler versions appropriate for a younger AGI system. At a high level, though, I think it makes a lot of sense to think about AGI systems in the context of child cognitive development. The

overall process via which a young child learns about the world is very relevant to AGI, because an AGI starts out its journey in a similar way to an infant – as a virtual blank slate where content is concerned.

One can try to work around the “blank slate” aspect in an AGI context, via using digital knowledge bases to fill the young AGI’s mind with knowledge. But this doesn’t really work – or at best it can work partially. Information loaded into an AGI from a database isn’t really “knowledge” in terms of the AGI’s role as an autonomous agent in the world. Until it knows how to connect this information with its perceptions and actions, it’s not really useful knowledge at all, it’s just a bunch of patterns oddly floating there in memory. And the connection between the database of information and the world, still has to be learned by the AGI system from experience, starting from nothing, in the rough manner of a human baby.

For instance, suppose an AGI system is fed the information, from a database, that a cat is a kind of animal. In the database this is likely to look something like `cat isa animal` indicating that there is a relationship of type “isa” between the concept “cat” and the concept “animal.” This can be used by an AGI system for abstract reasoning of various sorts. For instance, if the system knows that

```
animal hasProperty lifeform
```

then some simple reasoning proves that

```
cat hasProperty lifeform
```

That is

```
Cats are animals
```

```
Animals are lifeforms
```

THEREFORE

```
Cats are lifeforms
```

This sort of reasoning seems impressive at first, but what one finds is that in order to really formalize all the knowledge needed to be a young human child, one would need an insane number of relationships like this one. The Cyc project, an AI project based centrally on formalization of knowledge in this manner, has encoded many millions of such relationships, using dozens of expert knowledge encoders over decades. But they have barely made a start at what would really be needed to formalize a young child's knowledge.

Another possibility seems to be to extract this kind of relationship from natural language texts. Wikipedis, for instance, has a massive amount of knowledge in it! But the problem is that natural language is confusing and ambiguous, so that in order for a computer program to reliably extract the meaning from natural language, it would need to have a great deal commonsense understanding of the world already. One then has a chicken-egg problem: To get commonsense knowledge you need to understand language, but to understand language you need commonsense knowledge.

As one random linguistic example, think about the many different meanings of the omitted object of a verb:

Ben shaved = Ben shaved himself

The Phillies won = The Phillies won the game

John ate = John ate something (probably edible)

In the first case, "shaved", the omitted object means that the object is the same as the subject. Ben shaved Ben.

In the second case, "won", the omitted object means that the scope of possibilities is narrowed down in a specifically understood way. The Phillies didn't lose, they won.

In the third case, “ate” has a regular object, which just isn’t specified. John ate cake, or John ate dinner, or whatever.

How do we understand the differences between these cases? Not by looking at the syntax alone – rather, by understanding the commonsensical meanings of the verbs.

But how does a child learn these commonsensical meanings? By hearing people use similar linguistic constructions in similar contexts.

A more complex example illustrating the subtlety of semantic interpretation is:

You aren’t to marry him, and that’s an order

You aren’t to marry him, and I read it in my Tarot cards

The meaning of “are” is quite different in the two cases... But there’s no way to tell what the first half of the sentence means till you read the second half. And there’s no way to figure out these implications without knowing something about ordering around, and something about fortune-telling. A human child learns these things via being ordered around, and via having other people try to predict the future – via commonsense experience with the referents of the words being used. This is not necessarily the only possible way for such understanding to be gained. But it’s certainly the way we understand best, at this point.

Venturing fairly far afield from the human development model, some researchers aim to “bootstrap” the learning of commonsense knowledge purely via language, i.e. to go from a little commonsense knowledge to a little more language understanding, to a little more commonsense knowledge, etc. This sort of approach seems particularly interesting in the context of large stores of linguistic data such as the ones possessed by companies like Google and Microsoft. But it’s not clear that it’s really possible to make it work. Humans, obviously, bypass this chicken-egg problem via getting our first dose of commonsense knowledge non-linguistically, via direct embodied experience in the world. Using this commonsense knowledge, we are then able to interpret simple language. Using simple language, we gain more knowledge, which helps us interpret the world better; and the knowledge we then gain from the world, helps us to understand complex language better. Our own iterative bootstrapping process does involve progressively increasing and mutually reinforcing amounts of linguistic and commonsense knowledge, but our embodied experience in the world feeds and guides this overall knowledge growth process.

My decision to structure much of my own AGI work around human childhood development has clarified many different aspects of my practical thinking about AGI. While I don't think this is the only workable approach, it does seem the most natural to think about. Partly because I've spent a lot of time watching my 3 kids grow up, and partly because human childhood development is the most widely studied process where a mind starts out rather stupid and ends up reasonably smart.

But the decision to think about AGI substantially in the context of human child development came to me only gradually. Roughly emulating human childhood wasn't my original approach when I founded my first AI company, Webmind Inc, in the late 1990s. Back then, I was thinking more about Internet intelligence, creating things like intelligent search engines, or new artificial life forms that would grow across the Internet and evolve their own forms of intelligence. I still think that approach makes sense, but I found it difficult in that context to distinguish pathways to AGI from narrow-AI dead ends. I find that, if I'm thinking about AGI that is intended to roughly follow the stages of human childhood cognitive development, the path forward seems conceptually clearer. AGI is complex and confusing enough even after one has abolished the worry whether one's AGI developmental pathway is a dead end.

It may seem odd for me to advocate modeling AGI development on human child development, since I don't advocate, at this stage in the development of neuroscience and computer hardware, modeling AGI closely on the human brain. But actually, as incomplete as it is in its current state of development, I find developmental cognitive psychology a lot clearer than neuroscience. This is part of the reason I'm more bullish on cognitive science generally, versus neuroscience, as a guide for AGI. After playing around with lots of proto-AGI software, I've seen how difficult it can be to figure out what experiences an early-stage AGI system should be put through, and in what order. The field of childhood development provides many hints in this regard.

Human childhood development respects the hierarchical structure of human knowledge, translating it into a temporal, sequential order that allows children to learn the naturally simpler parts first, before proceeding to the slightly more complex parts, and so on.

While today's early childhood education system is certainly flawed (back in the mid-90s I helped found the "Unity Charter School" in Morristown, New Jersey, in an effort to remedy the worst of the system's defects), it does respect the hierarchal nature of human knowledge, building naturally from simple to progressively more complex knowledge. A preschool, for instance, usefully provides simple versions

of basic life skills an adult must master, specifically geared towards gradual development, while also building interconnections between those skills. Social interaction, art and creativity are woven into everything.

Because of this, integrating aspects of a preschool or grammar school-type environment into an AGI's environment appears to be a good course for structuring an early-stage AGI's learning experience, even if the AGI architecture doesn't bear close resemblance to the human brain. This is what I have often referred to as "AGI Preschool." An AGI Preschool could be a physical preschool similar to a human preschool, but perhaps with slightly different toys and activities, customized for compatibility with the particular robot bodies in question. Or it could be a virtual-world preschool, providing simulated virtual toys and activities for virtual-world AGIs to play with using their animated-character bodies. The important thing is not the specific design of the preschool, but rather the availability of a wide variety of activities, with varying levels of complexity and a high tolerance for failure, that provide practice at all the main skills needed for coping in the everyday adult human world.

For instance if a specific robot's hands have trouble with regular lego blocks, one can supply that robot's preschool with appropriate foam blocks. If paint mucks up the robot's fingers, there may be electronic whiteboards that provide sufficient avenue for artistic exploration. Just because human three year olds can't generally use search engines well, is no reason to restrict preschool-level AGIs from freely exploring Google Images and Google Videos using search terms, if they have the specific language ability to do so. The point is not to precisely emulate human preschool activities for robots or videogame AGIs, but rather to conceptually emulate the notion of the preschool experience for young AGIs – so as to provide a preschool-like environment for fostering and evaluating both unsupervised and supervised learning.

In terms of language learning, if you're following the childhood development approach, you don't start by having your AGI try to read the whole Web or a compendium of scientific papers. Rather, you start by having the AGI learn very simple language in an experiential, embodied context similar to how humans learn language. You have it learn language in the context of playing with things, in a way that comes naturally to it. Then, once the AGI understands simple language in the context of its own life experience, you try to teach it more complex language, and expose it to situations where it will benefit from applying what it has learned. That's a much more natural way for a young AGI to learn human language.

Maybe a properly designed AGI system could parse the Web and understand all the language on it, even without having any experiential grounding in any simple language. In essence, this approach would present the nascent AGI with a huge system of complicated simultaneous interrelated equations. It would have to figure out the meaning of each word based on the meanings of all the other words each is associated with, without really knowing what any of the words mean in relation to the external world. In purely mathematical terms this seems possible if you have a smart enough system and enough text, but it also seems rather difficult. Since building AGI is going to be tough no matter how you slice it, I prefer to try the easiest ways that seem feasible. Of course, these easiest feasible ways are not actually very easy at all. But the exciting thing is that we live in a time where there are any feasible ways whatsoever!

Embodiment

The idea of modeling an AGI's growth on human childhood development is inextricably linked to another tricky issue, perhaps THE most controversial issue in the whole AI field: embodiment. A human child is far from just a mind; being a young child is largely about learning to use one's body. Human childhood development, like AGI development, suggests some vaguely human child-like body for the AGI. Not necessarily identical to the humanoid form, but at least a body that can move around, perceive with multiple senses and manipulate various kinds of objects.

Quite apart from the question of whether human childhood development is a good model for AGI development, there is huge disagreement among AGI researchers about what kind of embodiment is necessary, useful or sensible to give an AGI system, if the goal is to achieve a human-level general intelligence. This is separate from the question of the validity of the developmental approach. On the one hand, one could potentially try to emulate developmental psychology in an AGI system with a purely textual interface. On the other hand, one could certainly attempt to make an intelligent humanoid robot via old-fashioned rule-based AI, or other approaches not involving humanlike cognitive development.

Also, even if you agree that the mind-body relationship is important for the development of intelligence, that doesn't resolve the question of: *What kind of embodiment does an AGI system need?*

Once, at a conference with a diverse audience, I gave a talk on AI and mentioned the controversy about embodiment. I said that some researchers felt embodiment was critical and others felt it was less so. A woman came up to me after my talk, fascinated by the idea of disembodied AIs – by which she meant

AI's that have no physical instantiation at all, but take the form of ghost-like energy fields, mind-systems made of pure psychic material... artificial poltergeists! I told her I was open to that possibility, but that wasn't what I was really talking about. I explained that the non-embodied AI systems that some researchers talked about, still consisted of computer code running on some physical computer. She was very disappointed.

My AGI researcher colleague Pei Wang, on the other hand, once wrote a paper titled "A Laptop is a Body", arguing that the important aspect of embodiment for AGI is simply the connection it provides between the AGI and the world. A laptop provides sensors to an AGI running on it, via its keyboard, microphone and webcam, and also its connectivity to other computers and the Internet. A laptop provides actuators to an AGI running on it, at very least via the computer's sound and video output, and also via its ability to send signals to other computers and the Internet. Since a laptop is also a body, the question isn't one of embodied versus disembodied AGIs, but rather one of what kinds of sensors and actuators are really needed to support what kinds of intelligence.



Figure 73: Ball lightning, mobile coherent balls of pure electricity, has proved difficult to create in the lab, though Nikola Tesla claimed to have done it. But it has been observed empirically many times by various individuals. Could it be possible to create an intelligent system of pure electricity, via understanding and shaping this sort of phenomenon? Who knows, but it doesn't seem impossible. The point is, intelligence is about the cognitive-level pattern of organization of a system, not the underlying physical substrate. Potentially, an intelligence can be a brain, or a computer, or a femtotech construct like a quark-gluon plasma, or a pattern of software acting in a computer-implemented simulation world, etc. etc. There may be physical constraints regarding what kind of mind can be effectively implemented in what kind of substrate, but within fairly broad parameters, intelligence is best considered substrate-independent.
http://en.wikipedia.org/wiki/File:Ball_lightning.png



Figure 74: The Parrot AR drone is the first widely available “toy” drone, remote controllable via iPhone. Remote-controlled drones are currently in heavy use by the military, but are expected to make inroads into various commercial markets during the next decades. Journalism and disaster response are obvious application areas, along with eventually delivery of pizzas and anything else delivered to peoples' homes. Currently drones tend to be remotely piloted by humans, with only basic control mechanisms automated; but fully automated intelligent drones are an obvious next step.
https://upload.wikimedia.org/wikipedia/commons/3/34/Parrot_AR.Drone_2.0_%26_Dassault_Rafale.jpg

An AGI’s physical embodiment could take the form of a laptop, or a server farm, not necessarily a humanoid robot body. Or it could take the form of a tank, a submarine or a quadrotor drone. [Potentially, given some currently undeveloped technology, it could take the form of some sort of exquisitely self-organizing ball lightning – approximating the AGI poltergeists of my supernatural-minded conference questioner. Each of these choices would have implications regarding what kind of intelligence could be achieved using a reasonable amount of resources.

Some researchers believe you would need a closely human-like body to get anything resembling human-like intelligence; and that, to get a high degree of general intelligence at all (whether human – like or not), you’d need a pretty complex and sophisticated body. These researchers tend to view intelligence as something that emerges from adaptive body function, in the course of a specific body’s engagement with a specific world. Other researchers believe that the body really doesn’t matter that much – that the core of general intelligence is a set of reasoning and learning processes that are body-independent, and the body is simply an I/O connector between the mind and the world.

My own view is in between these two extremes. I think you COULD make an AGI with just a laptop as a body (maybe not the 2012 Macbook Pro I'm writing this on, but a more advanced version of what's available on the market today). A laptop-based AGI could learn by chatting with the owner of the laptop, surfing the Web, and looking at pictures on Flickr and videos on Youtube, and so forth. I see no reason why this couldn't work, in principle.

I also think that if you had a genuinely human-like robot body to plug your AGI algorithm into, creating a human-level intelligent, human-like AGI would be an awful lot easier. You could apply accurate sensors and actuators to feed sensory information to your AGI system.

Certainly it's educational to take robots as we have them now and plug in AI systems. There's much we can learn from this approach, even though present-day robots are much cruder than the human body in most ways (though more powerful in a few ways – e.g. current robots can connect directly to the Net by wifi, whereas I cannot; and they can get exact distance measurements using laser rangefinders, etc.). You can still get more richly structured data, more finely-controlled actuation and better synergy between perception and actuation from a robot than you'll obtain with any other technique for interfacing an AGI with the world, right now. And if you want your AGI to interact with the human world in a human-like way, which is arguably important for achieving human-like intelligence, a robot body is a good choice. Certainly if you want to mimic human childhood development even roughly, a robot that can physically move around, see, hear and pick things up is an excellent start.

Robots haven't featured largely in my own work in past years, but I'm moving in that direction. During the period 2008-2010, I co-supervised some graduate students in a lab in Xiamen University, in China, whose work involved hooking OpenCog up to a *Nao* humanoid robot. And I've just now (in mid-2013) started some work using OpenCog to control one of the *Hanson Robokind* robots, which are more advanced in some ways than the *Nao*.



Figure 75: Aldebaran Nao robot at Xiamen University in 2009. This robot held amusing conversations using a dialogue system that combined OpenCog with a number of other inputs, and was featured in Raj Dye's award-winning documentary film Singularity or Bust.



Figure 76: The Nao robot in a screenshot from Singularity or Bust – answering me when I asked him if he was, indeed, a robot.



Figure 77: Me fixing some simple mechanical issues with Adam Z1, a prototype Hanson Robokind robot hand-built by David Hanson.



Figure 78: Me and Adam Z1 at the Global Future 2045 conference, at Lincoln Center in New York, June 2013.



Figure 79: David Hanson repairing some frayed Frubber on the face of his famous Philip K. Dick android, in one of Russian Internet entrepreneur Dmitry Itskov's rooms at the Empire Hotel in New York.



Figure 80: A girl Robokind robot, photographed by me at Nanyang University in Singapore in April 2013.



Figure 81: Einstein Hubo, a robot made by taking David Hanson’s Einstein robot head and connecting it to the Hubo humanoid robot created at KAIST in Korea. Chosen as one of the best robots of all time by Wired Magazine.

As well as physical robotics, I think video game-type embodiment—like the virtual worlds of *Second Life* or ones powered by modern 3D game engines— has a lot to offer. In a video game, you have essentially the same features that a robot provides –perception, action, social interaction, language, goal-oriented reasoning, spatial and temporal thinking, all connected in a way that emulates ordinary human life reasonably well. True, there's a lot less complexity in perception, action, richness and diversity available from contemporary virtual worlds than from the physical world as perceived by a robot. But you're also gaining a lot going the virtual route. You're gaining simplicity – virtual worlds are a simple, low – cost way of experimenting – but you're also gaining more than that. Virtual worlds enable an AGI system to interact and practice with millions of people. You can roll out a video game or virtual world application to millions of users at a far lower cost than, say, sending humanoid robots to millions of users.

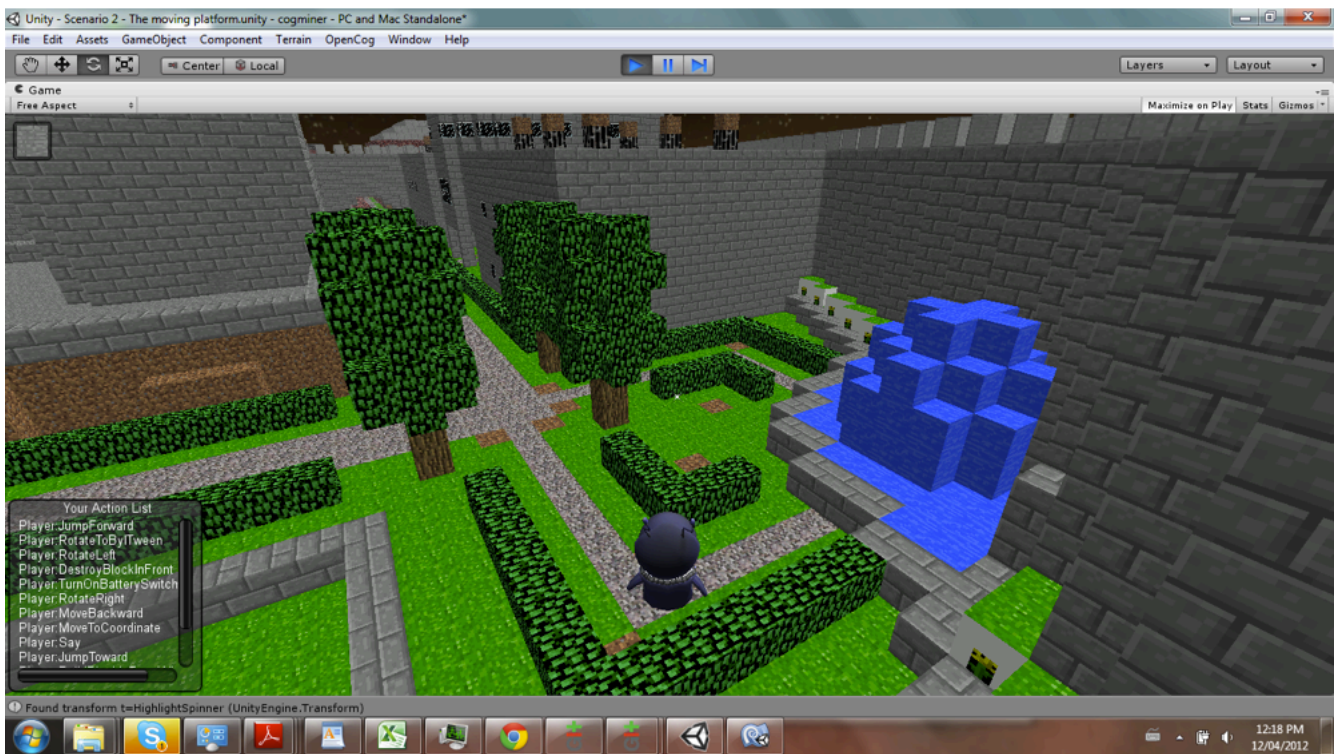


Figure 82: Screenshot from the “game world” we are using in 2013, at time of writing, to experiment with the OpenCog system. This world is built in the Unity3D game engine, and most objects in the world are made of many little square blocks. The composition of objects in terms of blocks makes it especially tractable for an early-stage AGI system to understand how objects are made, draw analogies between objects, etc. We don't think this world is complex enough to support human-level AI, but it's an interesting platform for experimenting with the OpenCog system as it develops.



Figure 83: Teaching the OpenCog system to build stairs in the virtual world. In this example, the little girl is controlled by a human, and the animated robot is controlled by OpenCog. The girl builds some stairs with blocks, the robot watches, and then the robot imitates what it saw the human do.



Figure 84: Now that the robot knows how to build stairs, when it wants to get something (such as the reward block up on the tree to the right of the picture), it knows that one strategy it can follow is to build stairs to climb up and get it. This same sort of learning process could be explored with a physical robot, but doing it in the virtual world first allows research to focus on the learning dynamics rather than on the mechanics of perception and action. In parallel with this virtual research, though, we are also spending some time on robot perception and action – as the complexity of the real world is likely to be needed to get all the way to human level AGI, or even to human toddler level AGI.

Just about the best humanoid robots we have now are creatures like the *Nao* and the *RoboKind* (about \$15,000 each, as of 2013). Although, arguably, Mark Tilden’s sub-\$100 “toy” RoboSapien robots have

more sophisticated walking than these expensive bots, using Tilden's unique analog computing hardware. In some ways, the best research robot around is the *PR2*, which is not humanoid—it has wheels instead of legs, and gripper clamps instead of hands—and it costs around \$400,000 per unit. With these price tags, these robots are just not feasible for mass distribution – except the RoboSapiens, which provide amazing capability for their price, but still have many limitations, such as the lack of robust visual sensors, capable grasping hands and wifi connectivity. On the other hand, a video game character could be played by millions of people sitting at home on their computers or almost anywhere else on their smartphones. An AGI-based videogame could provide millions of teachers to instruct baby AGIs in the basics of language and thought.

Virtual world embodiment could offer solutions to the complexities of language learning (some of which we've discussed briefly above). Consider prepositions, some of the most confusing words in English – for instance, the preposition “with.” We can say

- I ate dinner with my uncle.
- I ate dinner with a fork.
- I ate dinner with a salad.
- I ate dinner with love.
- I ate dinner with the TV on.

In each sentence the word “with” means something different. The varied meanings of the word “with” are hard for someone who speaks English as a second language to grasp. It certainly would be hard for an AGI to master the correct usage just from reading articles online — not impossible, but hard. However, if an AGI sees the word “with” used in its correct context in a virtual environment, and experiments with several objects, it will get a practical sense of its different meanings. Arguably, virtual worlds will be just as important for educational purposes as robotic embodiment.

Now, what about understanding specific objects? A robot in a kitchen will probably gain richer knowledge faster than an AGI in a virtual world; in the video game world, its interactions will be limited to the simplistic simulated objects that others have built or imported. The AGI will come into contact with objects and events, with fewer examples of interactions from which to learn. In principle, a full-featured virtual world could supply all the complexity of the real world or even more.

But for the foreseeable future, virtual worlds will be far less rich in detail than the physical world. This means that if we're going to attempt to construct AGI in the next five or ten years, some combination of virtual embodiment and robotic embodiment coupled with minimal embodiment (in which a system just looks at a lot of online data) will be the best way to proceed. The OpenCog system does this— it supports the rich intermixing of all these kinds of embodiment.

Embodiment & Environment

Much of the importance of embodiment for AGI doesn't derive from the body itself, but rather from the world that the body allows the AGI to experience. The relationship between an intelligence and its environment is quite subtle, far more important than most AI researchers believe. After all, intelligence consists of the ability to adapt to the environment – and any real-world intelligence is better at adapting to some kinds of environments than others. An intelligence's methods of adaptation are themselves adapted to certain kinds of environments.

An absolutely general intelligence is a mathematical fiction. *Absolute generality* —the ability to learn anything at all within a feasible timeframe—could only be achieved using infinite computing resources, which don't exist in our reality. Any finite system built using bounded resources will possess limited generality, in the sense that it will be better at solving some kinds of problems than others. That is to say, it will be more intelligent in some environments than in others.

So if absolutely general intelligence is a fiction, what use is the concept of “general intelligence” after all? The answer is: It still makes sense to talk about the generality of an intelligent system – because different systems may have different balances between specificity and generality. Some systems may be very good at solving a narrow range of problems; others may be moderately good at solving a wider range of problems. The latter we would consider to have a more general intelligence. So you could view a “general intelligence” as having a broad scope to the set of problems it can solve; and a “less general intelligence” having a narrower scope. (It's possible to formulate this distinction mathematically – to formalize intelligence and generality-of-intelligence as separate but related concepts. I did this in a paper I wrote for the AGI conference in 2010, which we held in Lugano, Switzerland.)

There are limits to both the intelligence and generality that can be achieved by any system with finite resources. That is, any system created in the real world will be adapted to some category of environment, and to goals and tasks characteristic of that environment. The human mind is adapted to

carrying out certain kinds of tasks in certain kinds of environments. Humans are, by and large, much better at fulfilling the basic demands of survival in our ancestral environment (locating mates, harvesting food sources, fighting off enemies and navigating in a complex 3D scene full of moving objects, etc) than more abstruse tasks like proving mathematical theorems or writing code. We evolved in the African savannah and, until quite recently, survival of our bodies and propagation of our DNA were our main goals. We're adapted to a certain set of environments and goals.

Now, an AGI need not have the same balance of strengths and weaknesses as a human mind. However, if we want an AGI that we can understand well enough for the purposes of mutually satisfactory communication (so we can teach it usefully, or debug its code in a savvy way when it goes wrong, and it can learn from us) – we'll need a system adapted for the same sorts of environments and goals that we are. And this leads back to the question of embodiment, because if an AGI is to be suitably adapted to the same classes of environments and goals as we are, it will be better off managing the same tasks we do – which will be easier if the AGI has roughly the same kind of body.

What Would a General Theory of General Intelligence Look Like?

The close relationship between mind and environment is important for a project I've been developing slowly for a long time, as a background pursuit — the creation of a “general theory of general intelligence.” Nobody has any such theory now. But I've started to work toward one in a fairly serious way. I've written several papers on the subject, and have used the ideas in these papers to guide and shape my work on OpenCog.

So far, though, my work in this direction has been “semi-rigorous”, rather than fully mathematical and scientific. I've been too busy trying to push toward actually DESIGNING and BUILDING general intelligence, to follow through fully on my ideas about formalizing the conceptual framework surrounding the general intelligence. But I'd love to plunge more fully into the theory aspect one day.

The details of my thinking on intelligence theory get pretty technical. But some of the core ideas are easy to understand. First of all, any theory of general intelligence would have to tell you something about **suiting a mind to its environment and vice versa**. My ideal kind of theory of intelligence would accomplish the following: *If you described a class of environments and goals, and specified some resource bounds, then the theory would tell you what kind of mind could operate intelligently within those bounds given limited resources.*

So you slot in the world and the goals and the bounds, and the theory would describe the sort of AI systems suited to achieving those goals with those resources in that particular world. If we had such a theory, we could create a human-level, human-like AI just by describing the world, the goals that humans habitually deal with, and the resource restrictions.

To develop a concrete, applicable theory of this sort, we'll need to experiment with some fairly powerful AGIs, seeing what they can do and how their environments, structures and capabilities interrelate. But I think that purely theoretical work can still help us, by giving us useful abstractions that decrease the number of experiments we have to run. One should be able to find elegant mathematical properties of environments and goals that are easily relatable to comparable mathematical properties of minds that are intelligent relative to these environments and goals given limited resources. Pleasantly, I've found reasons to believe that some of these properties are related to the symmetry properties underlying the foundations of physics, and of aspects of mathematics like probability and entropy. Details in a few years!

How to Proceed?

So what's the practical upshot? Ultimately, how should we proceed, if we want to build a thinking machine?

A practical, useful general theory of general intelligence would be awesome – but the fact is, we don't now have one. So we have to move ahead via an ad hoc combination of theory and experimentation – starting with as well – informed and solid an approach as we can muster, then learning, and revising the details as we go along. This is how many things, great and small, get done in the world.

At the beginning of this chapter, I summarized the high-level plan I've been following:

1. Look at everything known about human cognitive science.
2. Think about these known cognitive processes as a system.
3. For each of the important mental processes identified, figure out how to achieve what it does via some reasonably efficient computer algorithm that will work on today's computers.
4. In the system perspective – of the mind – look carefully at the interactions between all the mental processes modeled computationally, and be sure that the chosen computational algorithms interact with each other properly.

5. Create a practical software framework that allows all the needed processes to operate simultaneously and interact with each other appropriately.
6. Create an agent powered by this software framework, let it control a body in a world full of rich data to interact with; and let the world be populated by human agents eager to interact with it and teach it.
7. In this environment, carefully interact with the AGI agent to lead it through a series of natural developmental stages, modeled loosely on human developmental psychology.

So far, during the last years and decades, I've worked through Steps 1 to 5, and am in the middle of Step 6 – it's a big one. Of course, some of the previous steps are still ongoing– the software framework I'm working on is under continuous development, and the algorithms in use are continually being tweaked and improved. Regarding bodies and worlds, currently I'm using AI to control virtual characters in video game-type worlds, but I've also worked a bit with robotics, something I hope to do more of in the next few years.

My next practical steps are working with a software team that's implementing more of the AGI design we've created together, then teaching the AGI system as it controls virtual world agents or robots. Currently, the emphasis is more on the implementing than the teaching; as more of our system is implemented, however, it'll be the other way around. It's an extremely difficult task, but also an excellent adventure.

While potentially there are many workable paths to AGI, in the next chapter I'll dig into the details of the path I understand best and have the most confidence in – the path to advanced AGI via the OpenCog system that I'm developing with my colleagues.

The OpenCog Project

I was aware of the idea of thinking machines and smart robots from a very early age. I remember watching Mr. Spock drive an intelligent but overly-logical starfaring robot crazy on Star Trek, while watching the show with my dad in 1971 or 1972. But I didn't start thinking seriously about AI till about age 13, in 1979 or so, when I read *Goedel, Escher, Bach* by Douglas Hofstadter. That was a fascinating book which I read in one fifteen to eighteen hour sitting. I didn't agree with everything Hofstadter said – not by a long shot – but I absolutely loved the issues he raised. *GEB* put the problem of AI at the forefront of my mind.

Then in 1982, when I was 16 years old. I began to seriously consider the possibility of building an intelligent machine myself. My thinking at the time was more cognitive science oriented than low level design oriented – I was thinking about the organization of the mind as a pattern system, a system of patterns actively recognizing patterns in each other. I wasn't quite sure how to boil my thinking down into software code in a useful way, since I didn't know much about the practicalities of computational pattern recognition.

I came up with my first halfway (or, OK, maybe quarter-way) decent AGI design in 1986, in the middle of graduate school: An AGI inspired by the human immune system, a complex self-organizing pattern recognition system of a subtle sort, different from the brain. This was no doubt inspired by a mathematical immunology course I was taking at the time, taught by Alan Perelson at New York University, where I started my graduate study in mathematics before transferring to Temple University in Philadelphia. The immune system provided a concrete model of a pattern recognition system, which I could expand, extend and emulate to form an instantiation of my abstract thinking about networks of mind-patterns.

I also developed neural network-based designs, drawing more inspiration from dynamical systems theory (“chaos theory”) than from the details of the brain, which were known even more scantily back then than now. And I worked on another AGI architecture based solely on mathematical theorem-proving, and one based on automatic program learning. Then as now, I felt there were a lot of different valid approaches to creating AGI.

In 1994 I made an utterly unsuccessful attempt to implement an AGI in Haskell, a very mathematical and elegant programming language that was then at a very early stage of development (it's now much

more mature, and has become impressively efficient, but back then it was very slow to do anything). This system started off by recognizing patterns in itself, pursuing theoretically endless layers of recursive introspection. Data from the outside world got piped into the introspective theater too and got reflected on. It was an interesting self-organizing system that evolved some complex fractal-like internal patterns, but was too inwardly self-focused to learn much of anything useful. And every time I tried to feed it lots of data, the Haskell interpreter crashed. I could have fixed the software issues (perhaps by hacking the Haskell interpreter appropriately), but I correctly intuited that new ideas were needed, and moved on.

In 1995-96 my AGI thinking got more practical. The Web was coming into its own – in 1994 it first started to emerge, in 1995 it began sweeping academia, and by 1996 already it was becoming commercial. The Java programming language was new and shiny and made it easy to write software that played on the Internet. I fell in love with the idea of making an AI that had this wonderful new thing called the Internet as its home. I designed an artificial mind-network that was supposed to serve as a more flexible, more intelligent mirror of all the information on the Internet. Critically, it was supposed to contain both abstract symbolic, semantic information, and “subsymbolic” information about simple associative relationships, within the same network. The nodes in the network could represent anything – from Web pages to words to numbers to concepts. The links between nodes could be of various types, some representing logical relationships, some representing associations similar to the connections between neurons, and so forth. Each node was viewed as “living” and could recognize patterns in the other nodes around it, and represent these patterns as new nodes and links to be inserted in the network. Nodes could also recognize patterns in online data, such as Web pages or quantitative or relational databases. The whole network could be spread across multiple machines, with the nodes on one machine often focused on recognizing patterns in data local to that machine. This I viewed as a mind for the Web, and christened Webmind. Ultimately, I figured, humans and machines would network together over the Web, producing a human/AGI hybrid global brain.

This was the vision in my mind when I left academia at the end of 1996 – after 7 years as a professor and research fellow at various universities in the US, New Zealand and Australia – and started my first AI company. The company was called Intelligenesis, and later changed its name to Webmind, because people (except Germans) found the name Intelligenesis too hard to remember and pronounce. Our original logo was pretty spiffy in my opinion – it looked like a little digital sperm, ready to fertilize the new era!



In order to found the company together with friends from New York, I moved from Perth, Western Australia, where I was a research fellow at the University of Western Australia, back to the USA. By my current standards, both our business plan and our technical design were incredibly poorly specified when we started the company. But we had lots of smarts and passion, and it was a time when everything seemed possible where the Internet was concerned.

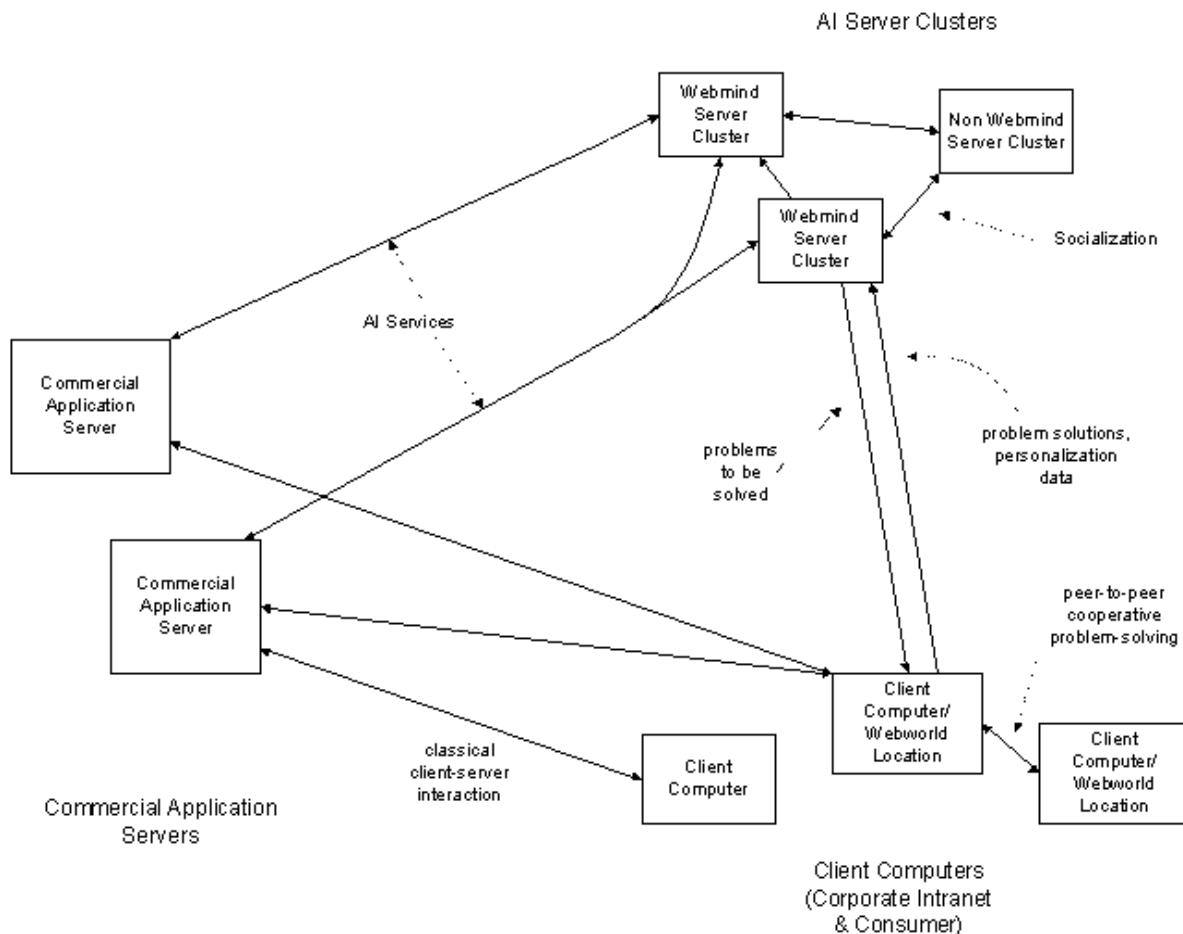


Figure 85: How I envisioned the Internet evolving toward emergent general intelligence, back in 2000 when I was involved with my first serious AGI effort, Webmind Inc. AGI components on central servers were to interact with AGI components living on home computers and interacting peer-to-peer, via an infrastructure called WebWorld. The advent of cloud computing has since made this kind of global distributed computing less attractive as a path to AGI. The diagram is from my 2001 book [Creating Internet Intelligence](#).

That company thrived for 3.5 years, based in the Wall Street area of New York (Silicon Alley). It earned minimal revenue, and spent plenty of venture money. There were about 130 employees at maximum, about 1/3 researchers, with 60 of the total staff in Brazil, and a few in Australia and New Zealand and Silicon Valley, and the rest in New York. We built a fantastic distributed Java software framework, and lots of cool AI components, but the whole thing never quite came together to do anything interesting. Among other problems, we were pushing the Java language way beyond what it could do at that time. Also, at that stage in my development as a software architect and project manager, I vastly underestimated the amount of effort required to translate a solid conceptual and mathematical idea into functional, maintainable, robust, scalable software code.

That company died in 2001, along with a host of other interesting dot-com boom software firms – and a

number of the Brazilian Webmind technical staff and I subsequently joined forces to create a new AI firm called Novamente LLC (nova mente = new mind; also “novamente” means “anew” in Portuguese). We designed a new AI system called the Novamente AI Engine, built on similar conceptual principles to the Webmind AI system, but much more pragmatically architected as a software system. We had a much smaller team now – just a handful of people – but we didn't intend to give up. We intended to push ahead as best we could, simultaneously searching for more resources to hire more people and building software according to our design.

For the next few years I put a lot of effort into simplifying and improving the Novamente AI Engine design, while my Brazilian colleagues worked on the implementation. We got a lot of interesting things done, but the search for AGI-oriented funding kept leading us down dead ends. We kept ourselves financially afloat doing various narrow-AI projects, some utilizing Novamente AI Engine code and some not. But eventually I reached the conclusion that things were just going too damn slowly. We chose some of the core parts of the Novamente AI Engine and decided to release them open source under the name OpenCog.

On a technical level, this wasn't a trivial thing to do – the Novamente AI Engine code was complex and not that well-documented. Cleaning up portions of the code for open-sourcing was a lot of work, and a grant from the Singularity Institute for AI was instrumental in paying for this. This was somewhat ironic since their mission as an organization was to ensure that once AGI was created it was guaranteeably friendly to humans – and no such guarantee existed for OpenCog. However, the SIAI folks considered OpenCog's chances of success at achieving powerful AGI minimal, and figured that association with me and the OpenCog project might lend them some academic legitimacy, which they felt they needed at the time. Since then, they have rebranded themselves as MIRI (the Machine Intelligence Research Institute) and found academic legitimacy through other means, via publishing papers on their ideas about provably friendly AGI and so forth. They still maintain the perspective that OpenCog is unlikely to succeed, but that if it does succeed, it will almost surely kill all humans. Obviously my opinion differs. I'll return to these issues later in the book.

The open source aspect of OpenCog ties in wonderfully with the “global, distributed intelligence” memes I was enchanted with in the Webmind era. It makes perfect sense for the WORLD to create the first real AGI, since after all this AGI is going to transform the whole world. I can't prove it rigorously, but my gut says strongly that an AGI created by a broad-based global effort is far more likely to be benevolent than one created by some isolated group to serve its own ends. Of course there's the

possibility that some isolated group could have a brainstorm about how to build powerful, ethical AGI, and secretly create it and unleash it on the world to everyone's benefit. But the track record of secretive elite groups wreaking transformations "for everyone's benefit" is not very strong, if you look across human history. I have more faith in inclusive efforts.

OpenCog, as an open source software system, has two aspects. On the one hand, it's an AI toolkit, with a knowledge store and various algorithms that can be used for multiple purposes. My colleagues and I have used various OpenCog structures and algorithms in various practical applications for customers, some of which have had little to do with AGI. We've used them to analyze biological and market research data, to do automated scientific reasoning based on information extracted from biomedical research texts, and so forth. Other folks in various companies and government agencies have used OpenCog components for their own purposes. As the code is open source and free to use, they don't have to tell me or the other creators about their work, unless they want to. A few times I've heard from someone, years after the fact, that they used this or that OpenCog component for this or that project.

On the other hand, as well as an AI toolkit, it's more centrally an attempt to create a powerful AGI – just like the Novamente AI Engine and the Webmind system that came before it. This aspect of OpenCog I've labeled "OpenCog Prime", using the term "CogPrime" for the AGI design itself, as distinct from the OpenCog implementation thereof. OpenCogPrime is the OpenCog based implementation of the CogPrime AGI design. These rather tedious terminological distinctions aren't usually adhered to very closely in everyday discussions among OpenCog developers, but we try to stick to them in formal publications, to just to avoid confusion.

So what's OpenCog all about?

Viewed from a software perspective: at the top level, OpenCog has a few key components:

- A virtual storehouse called the Atomspace—where Atoms, knowledge building blocks, are found—which is useful for storing various kinds of knowledge.
- A framework facilitating interaction between different cognitive processes (called MindAgents) and the Atomspace; this involves adding, removing, and/or changing knowledge items within the Atomspace.
- Code based on the Atomspace and MindAgents enabling a system to control an agent in an external environment – either a virtual world (like a game engine; we've worked extensively with Unity3D) or a robot (where we access robots through the ROS, Robot Operating System, interface developed by Willow Garage).

What distinguishes the Atomspace from most of the common ways of storing knowledge in AI systems or other software systems, is its capacity to efficiently and sensibly represent essentially ALL sorts of knowledge. OpenCog Prime specifies a collection of MindAgents carrying out specific cognitive processes; and the Atomspace, by design, adapts to the MindAgents' knowledge needs.

The “secret sauce” behind OpenCog Prime is how the MindAgents work together as an overall coherent system. Incorporating a variety of cognitive and AGI theories, they have been designed to work together, giving rise to the full scope of cognitive processes described in the review of cognitive science I gave earlier. However, the flexibility of the underlying OpenCog framework is also important, making it easy for developers to play around with variations of MindAgents, and to run experiments to see what works best. Since most proto-AGI systems have been written by graduate students in order to prove specific theoretical points, they're not very flexible. OpenCog is designed to support a huge variety of possible AI and AGI designs, some extremely different than OpenCog Prime – and this has given us the freedom to experiment fairly freely with various ways of turning the conceptual and mathematical OpenCog Prime design into software.

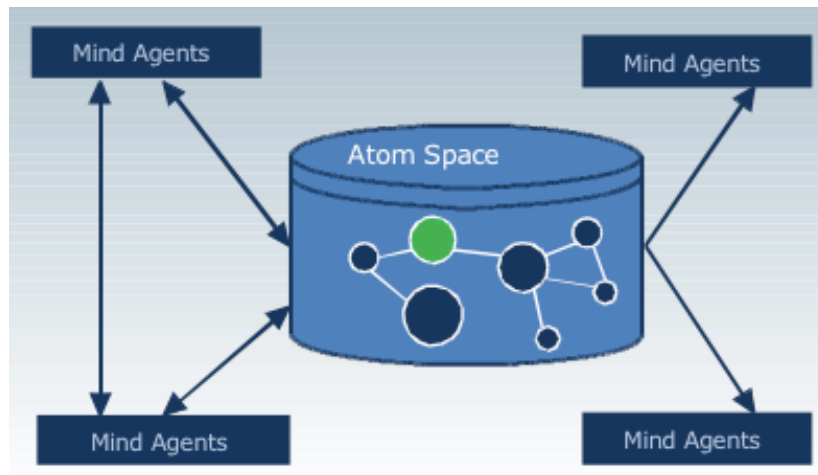


Figure 86: *A view into one key aspect of the OpenCog software architecture. The multiple MindAgents wrapping up OpenCog's cognitive processes, all interact with the common AtomSpace knowledge store, and sometimes interact directly with each other as well.*

OpenCog's AtomSpace

The centerpiece of OpenCog is the AtomSpace—think of it as the nucleus of an OpenCog system. Within the AtomSpace, all the different kinds of knowledge the human mind needs are stored with plenty of room for flexibility. Represented mathematically, it is a “hypergraph,” which looks like a *very* complex diagram of all sorts of nodes cross-connected by all sorts of links.

Diagrams 9, 10 and 12 give some simple illustrations. Key features of the diagrams are:

- **Nodes:** Circles, representing concepts, objects, numbers, or mathematical functions
- **Links:** Arrows connecting nodes, representing different kinds of relationships between nodes.
- **Maps:** Clusters of nodes and links forming knowledge building blocks.

The term **Atom** is used as a grab-bag including both Nodes and Links. Nodes and links are the atomic elements of OpenCog's knowledge representation. They come in many different types.

For example, the node representing “cat” and the node representing “animal” would be joined by an InheritanceLink, representing the fact that a cat is a special kind of animal (cat “inherits from” animal, in logic lingo). The cat node and the dog node would have a SimilarityLink between them, since cats and dogs are fairly similar, compared to other things like toasters, galaxies and jealousy.

A ContextLink would point from the Atom for “vacation” to a SimilarityLink, where the latter would join a node for Jamaica and a node for Santorini (and this SimilarityLink would presumably have a

mediocre “strength”). In the context of vacations, Santorini and Jamaica are moderately similar; in the context of government, there’s very little similarity.

Nodes generically associated with one another (say, power and corruption) will usually have a HebbianLink, named after Donald Hebb, a scientist who introduced associative relationships between nodes and networks in the late 1940’s.

In these examples, I’ve referred to nodes that correspond directly to English language concepts, but actually only a tiny minority of nodes in the AtomSpace are like this. Most don’t correspond to English concepts, instead to fragments of concepts, or learned groupings of concepts. English language concepts (or concepts corresponding to words in any other language) are made by grouping various nodes, or activating a bunch of them together.

Some nodes refer to perceptions coming in from sensors—representing, say, the corner of a picture on the wall in front of an OpenCog-powered robot. Or, actions corresponding to an OpenCog agent, like saying something, or moving a certain motor in a robot body. And there are also abstract nodes for logical operations, like AND, OR, and FOR ALL.

There are a few dozen different types of nodes and links in the AtomSpace, selectively chosen to encapsulate everyday human concepts in fairly small combinations. This seems a sensible model for representing everyday human knowledge in AGI. The idea of boiling down human knowledge to a small number of semantic primitives is an old one in philosophy, though philosophers who like this idea tend not to agree on what the primitives are.

Mathematically, one can reduce the number of primitives needed to represent everything quite dramatically, if one really wants to. Combinatory logic, for example, reduces all mathematically possible forms and ideas to a single mathematical operator, plus parentheses. But this isn’t a particularly convenient way to do things inside an AGI system. Moshe Looks and I tried something like this in 2005, and others have tried as well. OpenCog’s few dozen node and link types represents a pragmatic balance between the desire to use as few primitive notions as possible, and the desire to make the representation of everyday human mind-stuff fairly simple and elegant.

Giving a complete list of all the Atom types in the system would just be confusing in a nontechnical book like this, and plus the list is changing all the time. But the following list gives the basic idea. We have:, for example,

- `ConceptNode`, which might be better named “concept or part of concept node” – these are just generic nodes connected via links, which may be assembled in various ways to form parts of concepts
- `SpecificEntityNode`, representing a specific entity in the (physical, virtual or conceptual) world
- Various node types referring to concrete or abstract entities of specific types: `WordNode`, `CharacterNode`, `SentenceNode`, `NumberNode`, `BlockNode` (for a virtual world containing many blocks, like the Minecraft-like one we're experimenting with), `VisualPatternNode`, etc.
- Various Atom types representing logical relationships, e.g.: `InheritanceLink`, `SimilarityLink`, `PredicateNode`, `EvaluationLink`, `ImplicationLink`, `EquivalenceLink`, etc.
- `SchemaNode`, representing a procedure; `ExecutionLink` representing the enactment of a procedure
- `HebbianLink`, representing a simple association between two entities

Each of these Atom types has a particular story and theory behind it. The precise number of Atom types to use is a matter of art as much as science. Having hundreds would make the system an unmanageable mess: there would be too many different theories to take account of and intersect with each other. Having only one or two Atom types is mathematically possible, would make things conceptually awkward, because it would require using complex, opaque constructions to distinguish intuitively separate things. We have found that using a few dozen Atom types strikes a reasonable balance between intuitiveness and manageableness. Commonsensically simple concepts, processes and relationships tend to have fairly compact representations in terms of the few dozen OpenCog Atom types we've adopted. But the precise list of Atom types is ongoingly evolving as the system gets developed by the OpenCog community. It is also possible for the system to dynamically update its own list of Atom types as it learns, but, at the moment it is not as clever at doing so as its human programmers!

Truth and Attention Values

As well as having different types, the nodes and links in the AtomSpace have different numbers attached to them. The numbers fall into two categories—Truth Values and Attention Values

The most basic kind of Truth Value attached to an OpenCog Atom has two components:

- 1) Strength: How common is the concept represented by the Atom? Or, how strongly held is the

statement represented by the Atom?

- 2) Confidence: How surely is the truth value of the Atom known to the system?

The most basic kind of Attention Value attached to an OpenCog Atom also has two components:

- 1) Short Term Importance (STI): How much attention should the system pay to an Atom at any given point in time?
- 2) Long Term Importance (LTI): How useful will it be for the system to keep an Atom in RAM in the long term?

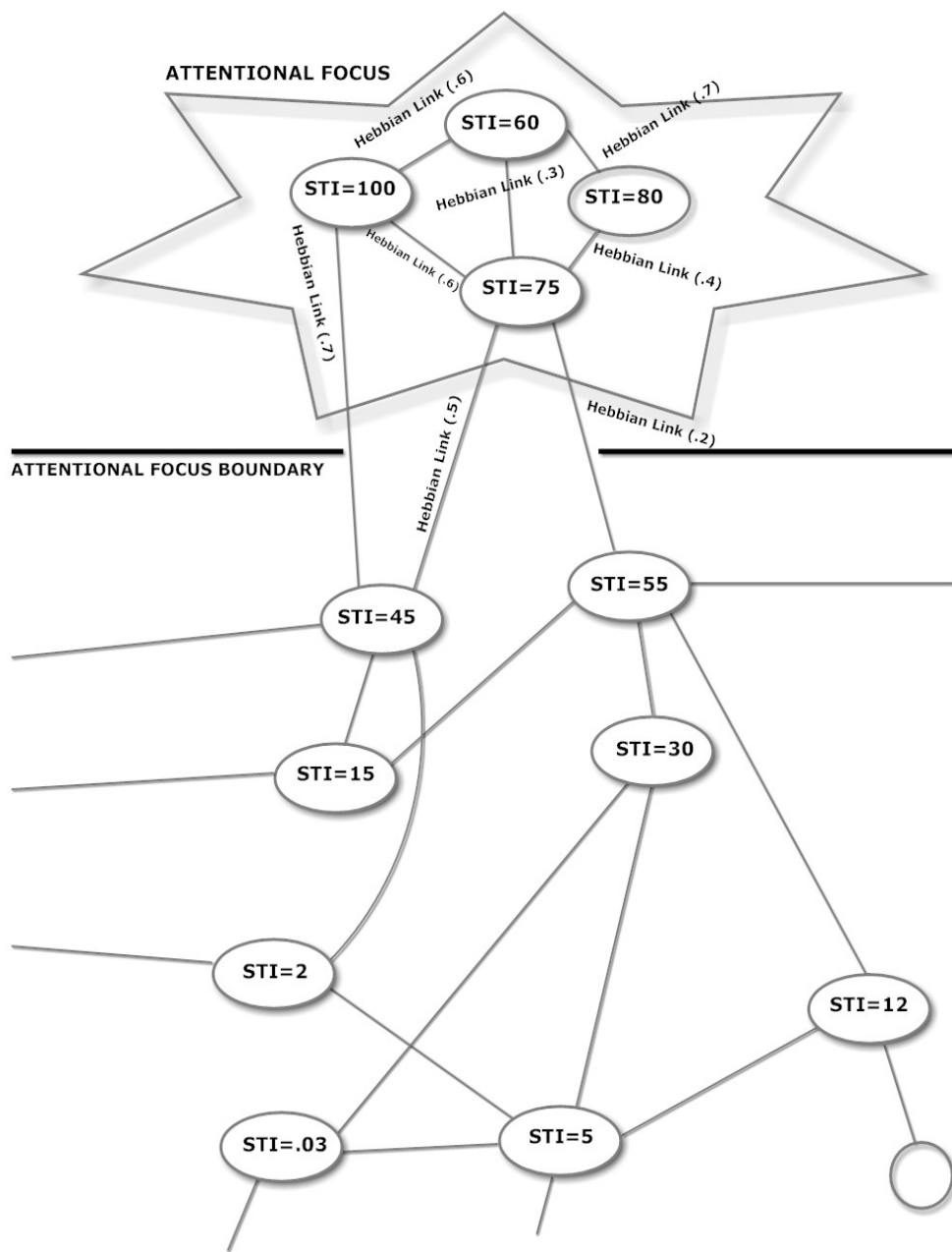


Figure 87: The Attentional Focus in OpenCog. Every OpenCog Atom gets a Short Term Importance (STI) value, and the ones with STI above a certain threshold are considered to be in the Attentional Focus, which means MindAgents (cognitive processes) will pay them special attention. The Attentional Focus is an OpenCog implementation of Bernard Baars' cognitive science theory of the Global Workspace.

Atoms with STI (ShortTermImportance) above a certain threshold become part of the “attentional focus” of the system. To use the language introduced by the psychologist Bernard Baars, the set of Atoms in the system’s attentional focus represents the OpenCog system’s “global workspace.” (See Diagram 7)

Atoms with the lowest LTI (LongTermImportance) are removed from RAM by a Forgetting MindAgent to make room for new Atoms that might have greater LTI potential.

All the different cognitive processes running in OpenCog work with the AtomSpace, inputting and outputting knowledge in terms of its nodes and links. Some cognitive processes have distinct knowledge representations, but since they all speak the language of the AtomSpace, they are able to work together as a whole.

Mixing Neural and Symbolic

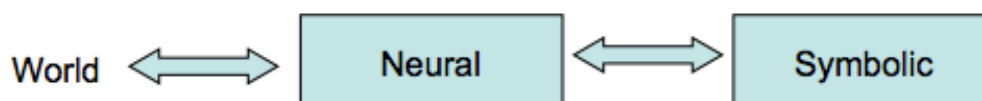
In the lingo of contemporary AI, the Atomspace may be considered a “neural-symbolic” system – which, unsurprisingly, means has both neural network like *and* symbolic logic like aspects.

- **Neural:** Some important neural network-like aspects: STI and LTI values of Atoms spread between Atoms, kind of like electricity or activations spreading between neurons in the brain. Furthermore, the flow of these values creates new links and changes the weights of existing ones (HebbianLinks form and adapt based on the flow of STI through the system: if two Atoms simultaneously have high STI values, usually a strong HebbianLink forms between them).
- **Symbolic:** The Atomspace directly comprises a logic-based AI system representing relationships through node and link types with logical semantics (InheritanceLink, SimilarityLink, ANDLink, and others mentioned earlier).

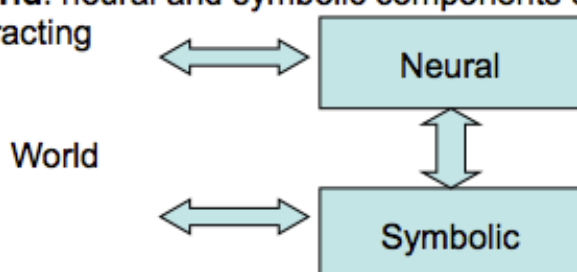
Other neural-symbolic systems tend to divide the neural and symbolic aspects. Some have a separate neural module and a separate symbolic module that communicate together from a distance; some are neural-centric and have a symbolic module that recognizes patterns in the neural module.

Types of Neural-Symbolic Architecture

Monolithic: symbolic component “sits on top of” neural component and helps it do abstraction



Hybrid: neural and symbolic components confront the world side by side, interacting



Tightly interactive hybrid: neural and symbolic components interact frequently, on the same time scale as their internal learning operations

Figure 88: OpenCog may be considered “neural-symbolic”, but differs from most neural-symbolic systems in that the neural and symbolic aspects are very tightly integrated. The Atomspace has both neural and symbolic aspects. This is different from architectures in which neural and symbolic components, architected separately, are networked together. On the other hand, DeSTIN is more purely “neural” in nature.

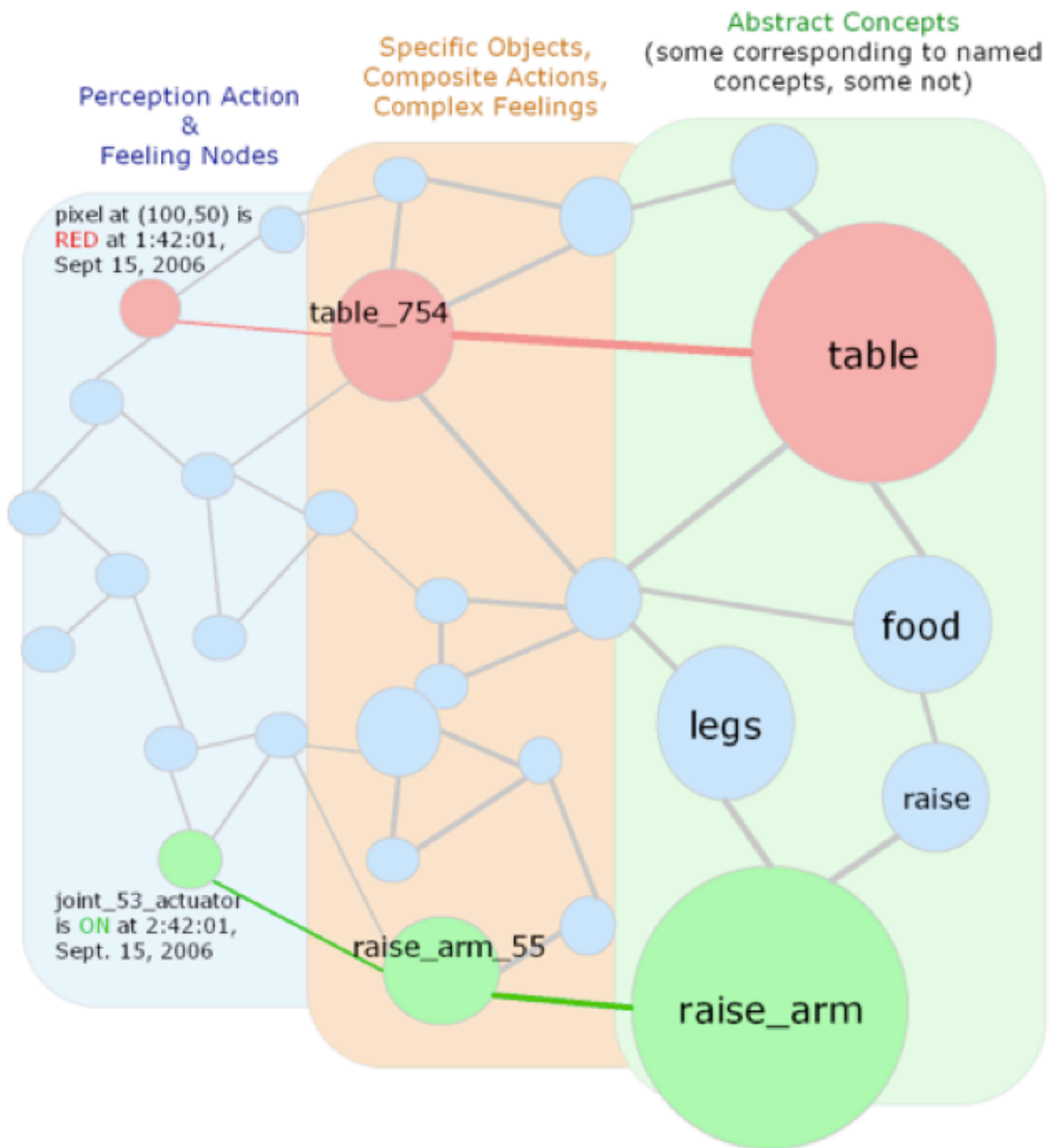


Figure 89: OpenCog's Atomspace (weighted labeled hypergraph) knowledge representation bridges the gap between subsymbolic (neural net) and symbolic (logic / semantic net) representations, achieving the advantages of both, and synergies resulting from their combination. It is able to bridge the gap between abstract concepts and highly specific percepts and actions

In the AtomSpace, the neural and symbolic aspects are working with the same nodes and links, yet interpreting them differently. This approach traces back to the Webmind AI system, and is one of the key common threads running through my AI work since the mid-1990s



Figure 90: Examples of OpenCog Atoms (nodes and links). Note that most Atoms in an OpenCog Atomspace, in real life, won't correspond to any particular words in English or other natural languages. This is just a convenient way to make examples to show people. The linkage to the left is neural-net-like – it just indicates a simple association between two concepts. The linkage structure to the right is more symbolic logic like – it represents a precise relationship as would be expressed in the formalization of reasoning called “predicate logic” (specifically, it represents the relationship that people drink coffee from coffee cups).

As an integrated neural-symbolic system, the AtomSpace represents knowledge in two important ways:

- **Explicit and local** – Each node and link represents a definitive and communicable piece of information. There may be a node for the concept of “cat,” a node for the concept of “animal,” and a link establishing that a cat is an animal.
- **Implicit and global** – Distributed patterns of activity across many nodes and links represent knowledge, and each node or link is equally important individually and as part of an activation pattern.

To describe this dual aspect, I coined the word “glocal” (global + local) – a term some people seem to love, but more seem to hate!

Many OpenCog nodes have no labels: they don't correspond to any particular English word or specific, communicable concept.

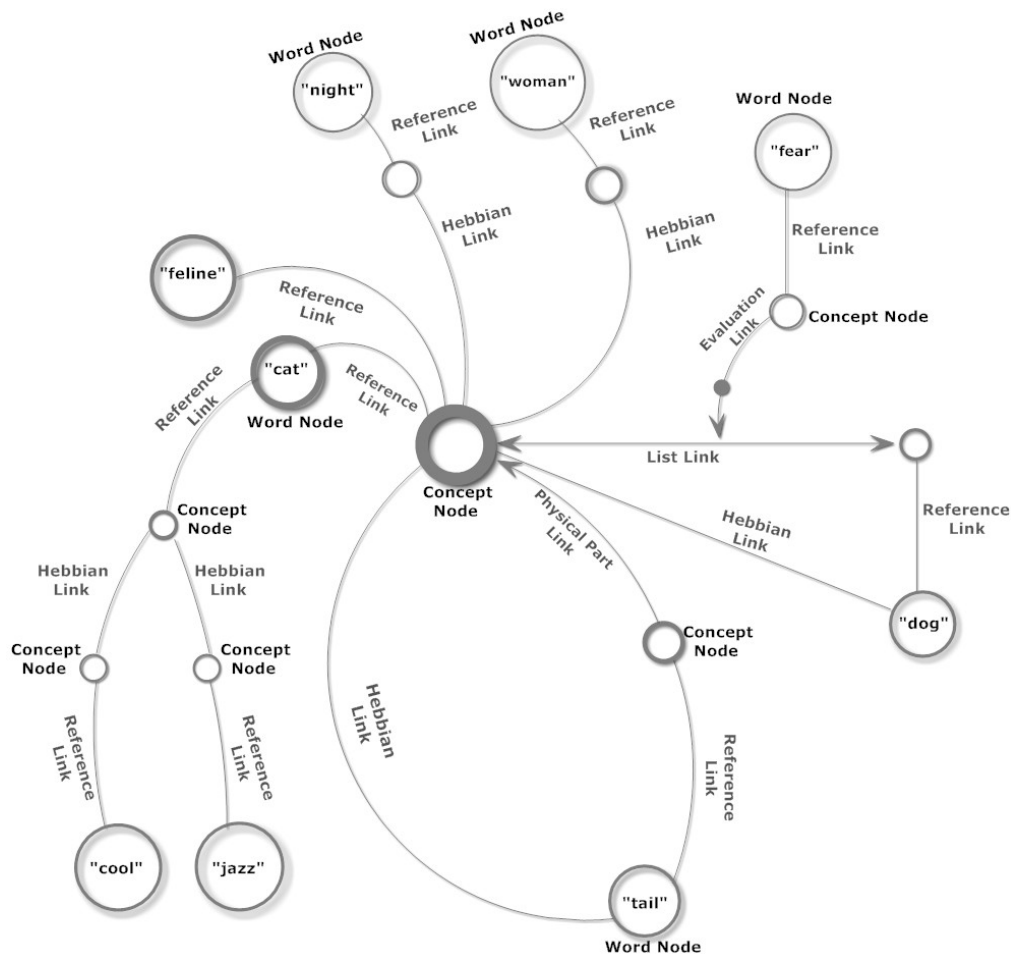


Figure 91: Links Representing Explicit Knowledge in OpenCog. For example, the nodes representing “cat” and “feline” are two references to the same concept, thus they each have their own ReferenceLink to the same concept node.

Mixture of Explicit and Implicit Knowledge in OpenCog. The collection (“map” in OpenCog lingo) of nodes at the top relates to “chicken”, though some of the nodes don’t correspond to any particular English words or common concepts, and are important only in the context of the network activity patterns they form. The collection of nodes at the bottom are all related to “food”. The bundle of links from the “chicken” collection to the “food” collection represents the global and implicit relationships between “chicken” and “food.”

The mixture of explicit and implicit representation makes it easy for MindAgents to work together on the same AtomSpace, even if some of the MindAgents use explicit localized representations, and others use implicit global representations.

The AtomSpace has been designed to accommodate multiple cognitive processes and to facilitate synergy between them. But, of course, this is no guarantee. The key lies in the algorithms that we create for each MindAgent, a fairly subtle and complex process.

The human brain appears to achieve a similar synergy, but how it does this remains largely a mystery.

Like OpenCog Prime, the human brain has different cognitive processes associated with different types of long and short-term memory; and it has evolved so that these different cognitive processes can synergize with each other effectively. Yet no one truly understands how the brain carries out these things. So instead of imitating the brain, we've used computer science to create a set of MindAgents that seem to make sense algorithmically, and that, according to our understanding, are likely to give rise to the various functions and interactions needed to achieve human-level, human-like general intelligence.

A Big Scary Diagram

Now I want to direct your attention to Diagram 10 – it's a big scary diagram and I apologize for that, but it's my best attempt to summarize the important things happening inside OpenCog in a single picture. My collaborator George Papadakis, who lives in Greece, helped me draw it, and he reports that while he was doing so, his girlfriend spent some time staring over his shoulder in utter perplexity.

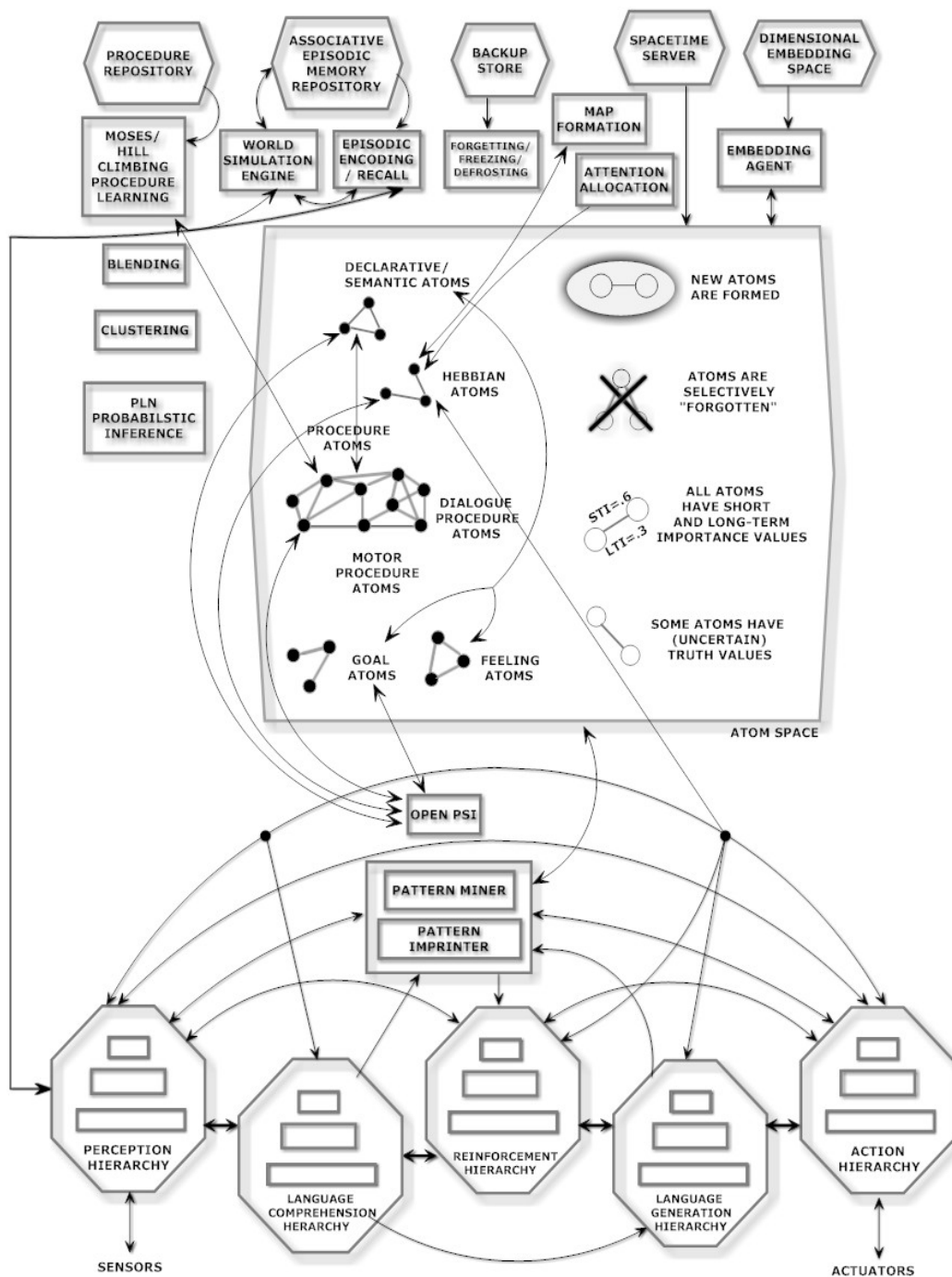


Figure 92: Key OpenCog AI Software Processes

You may be thinking: *Whoa! There's a lot going on in this "big scary diagram"!*

There IS a lot going on – but when you break it down into pieces, one by one, you'll find it all makes sense.

Let's start with the big box in the middle—the Atomspace, which you're already familiar with.

The dots in the box are nodes, and the lines are links joining nodes. Finally, the clusters of nodes and links represent Atoms, which fall into several categories. I already mentioned these above very briefly, but now it's time to give a little more depth. Among the Atoms we have:

- **Declarative** Atoms representing logical relationships like similarity, logical inheritance (in logic we say that A inherits from B if A is a special case of B, so in OpenCog we might have an Inheritance Link between a “cat” node and an “animal” node), logical implication links, and so forth
- **Associative, or Hebbian** Atoms, signifying associations (named after Donald Hebb, who was the first guy to write substantively about the role of neurons in the brain in identifying associations between things, and the critical role this plays in human thought). If two nodes often occur together (say, one node representing a “boy” and another representing “trouble”) at the same time or in the same context, a HebbianLink will form between them. HebbianLinks also form between nodes that have no obvious semantic meaning on their own, yet are extensions of other concepts. Finally, it only takes a few HebbianLinks in a cluster of nodes to form several coherent concepts, since importance spreading along the HebbianLinks will activate all the nodes.
- **Procedure** Atoms, corresponding to little procedures that the system can carry out. Usually these procedures are represented as short computer programs, in a special programming language (simpler than the programming languages in which OpenCog itself is coded) that OpenCog understands. These may be physical action procedures, like “step forward” or “keep walking forward till you bump into something,” or more cognitive procedures like the steps involved in answering a certain kind of question (this would be an example of a “dialogue procedure,” as Diagram 10 suggests).
- **Goal** Atoms, indicating goals that the system is trying to achieve, generally because it's decided (usually by reasoning with declarative knowledge, but occasionally by association alone) that these goals are ways of achieving one or more of its core drives.
- **Feeling** Atoms, representing the system's evaluation of its internal state over time – for instance, has the system gotten a lot of satisfaction lately, is it feeling safe or threatened, has it gotten a lot of new information lately or has its curiosity gone unsatisfied? Generally these pair with Goal Atoms – system goals are implemented in terms of evaluation of system feelings.

The right side of the Atomspace box illustrates some of the general processes involving Atoms in the Atomspace, most importantly:

- New Atoms are **created**, to help represent new knowledge – either based on new perceptions the system has received, or based on new conclusions it’s drawn, or new speculative concepts it’s cooked up
- Atoms are **forgotten** – i.e. deleted from RAM – if they are judged sufficiently irrelevant, which basically means if their associated Long Term Importance (LTI) values get too low

The square boxes above and to the left of the Atomspace box correspond to cognitive processes that act on the Atomspace. At the software level, each of these cognitive processes is generally implemented via several MindAgent software objects.

As Diagram 11 shows, many cognitive processes are simultaneously involved with creating new Atoms. The “attention allocation” process does a number of things, one of which is to update the Short and Long Term Importance values that regulate how much attention Atoms get. The Forgetting process deals with (surprise surprise) the forgetting of irrelevant Atoms.

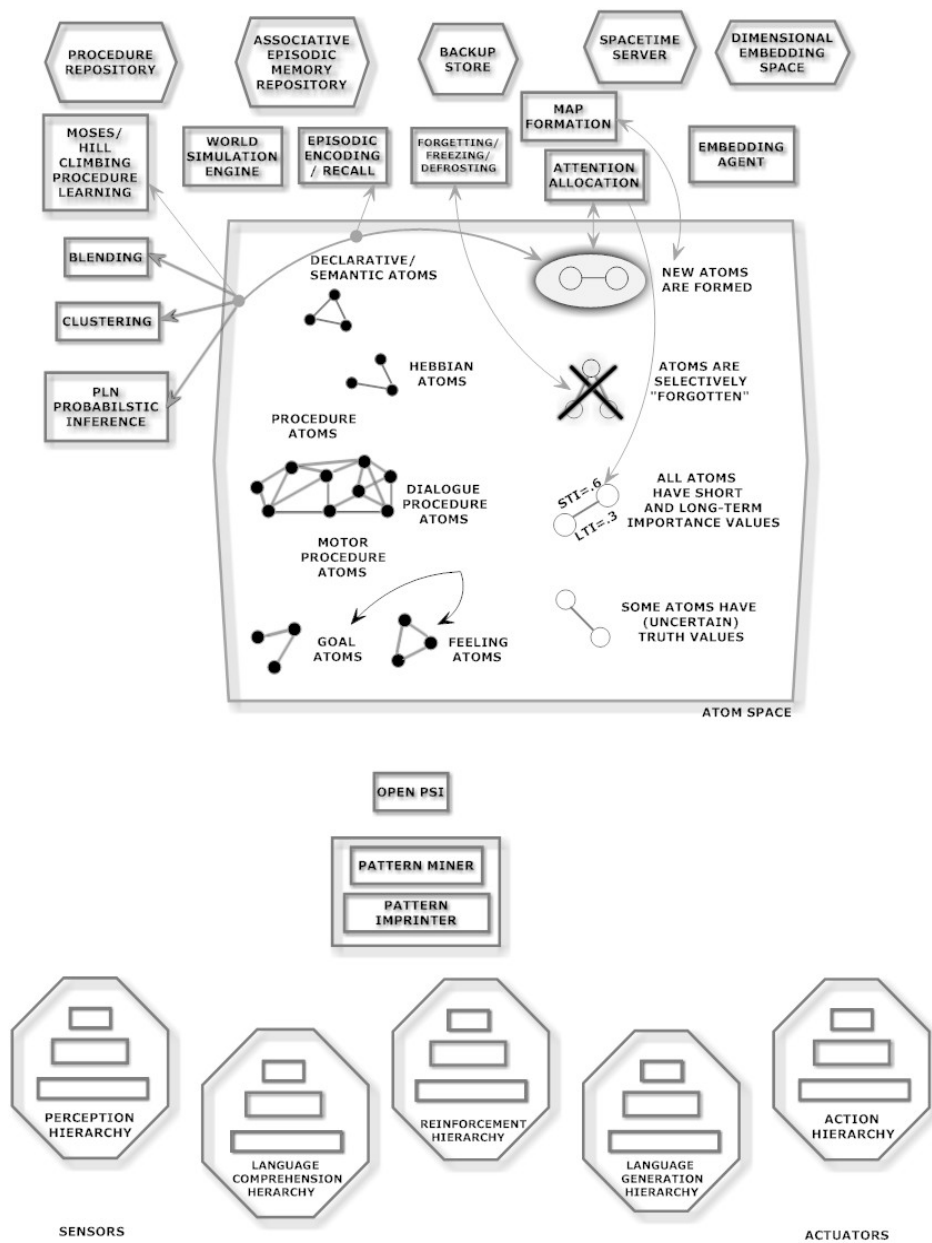


Figure 93: Key Cognitive Processes Participating in Atom Creation, Deletion and Modification in OpenCog

The OpenCog cognitive processes (MindAgents) work with various knowledge stores, represented by the hexagonal boxes above the Atomspace:

- **Procedure Repository**, which stores the mini-programs corresponding to Procedure Atoms. If a MindAgent wants to execute the procedure corresponding to a Procedure Atom, it grabs this from the Procedure Repository.
- **Associative Episodic Memory Repository**, stores Atoms relating to the life – history of a system (running on OpenCog) in a different way from the Atomspace: a quick and easy way for OpenCog to start with any episode, and quickly bring up memories of any associated episodes (sort of how the human mind works).
- The **Backup Store**, serving the prosaic but very useful role of saving the Atomspace to the computer’s hard disk drive periodically. This functionality gives an AGI mind an interesting advantage over human minds: It can load its previous mind-states and previously known information into its current mind, whenever it wants to. Integrating the old knowledge with the new presents challenges, but it’s still an amazing functionality, one I could definitely use – being 45 years old I’ve already forgotten a lot of the stuff I once knew!
- The **Spacetime Server** stores information about the times and spatial positions of Atoms that refer to real-world objects and events. This allows the system to rapidly answer questions about specific events according to spatial or temporal relations to each other (simultaneous, before, after, overlapping, near, far, next to, above, below, etc. etc.). A useful tool since space and time are such fundamental concepts.
- The “**Dimensional embedding space**” works a bit like the associative episodic memory repository, but more broadly – it stores Atoms in a different way, which makes it very fast to use an Atom as a query and pull up a list of every other similar or associated Atom. This feat is accomplished via a mathematical trick that involves assigning each Atom a point in some dimensional space (currently a 50 dimensional space). The wonderful world of software allows us to go beyond the three dimensional space of the brain—where links between the neuron and the brain tend to get tangled up—and potentially have as many dimensions as we want!

Interfacing Mind and World

The boxes at the bottom of the Big Scary Diagram represent hierarchically-structured modules, which process visual and auditory data for controlling actuators (like the servomotors in a robot, or the animations controlling a character in a game world). They're particularly useful when dealing with the intricacies of linguistic data and reward signals from the environment (somewhat complex in a robot's case –each part of its body may deliver the mind its own reward signals based on how comfortably it's been operating).

It's sort of like an input/output layer for the OpenCog mind.

The box with labels **“Pattern Miner”** and **“Pattern Imprinter”** serves the critical role of translating the languages of the perception and action hierarchies to the Atomspace and vice versa.

Pattern mining involves recognizing patterns in the states of these perception and action hierarchies, and then recording these patterns as Atoms in the AtomSpace. Pattern imprinting takes abstract relationships existing in the Atomspace, and transforms them to guide the perception and action hierarchies.

For instance, if the vision hierarchy recognizes a cat, the “catness” is initially represented by a distinct pattern corresponding to the organization of the processing units in the visual perception hierarchy. The Pattern Miner correlates each frequent pattern in the visual perception hierarchy to some OpenCog Atom (in this case associating the “catness” visual organization pattern to a certain Atom).

The association of the “catness” Atom with the word “cat” is then another problem for OpenCog to solve. If it often sees this same visual pattern at the same time as it hears the word “cat”, it will recognize this correlation, and learn the visual associations of the word.

And once the OpenCog system knows the association between the word “cat” and certain visual patterns, after it hears the word “cat”, it can use the Pattern Imprinter to instruct the visual hierarchy to look for catlike visual patterns. This may be helpful if the lighting is bad, or if the cat is wearing a sweater, or half hidden behind a couch. The Pattern Imprinter parallels the human mind's capability to use cognition as a tool to enhance perception. We've taken this approach since human vision processing is still superior to existing computer vision systems.

The label “input/output layer” doesn't mean these things are simple. I've spent the last couple years

working on a vision processing hierarchy for OpenCog, which applies subtle learning algorithms. It's an extension of a vision processing system called DeSTIN (created by my friend Itamar Arel, who works at the University of Tennessee in Knoxville, together with his graduate students). The system runs on GPUs (Graphics Processing Units) in a way that exploits their capability for parallel processing. I've been modifying it to input into OpenCog in a cleaner and simpler way while also taking feedback from OpenCog to guide its judgments.

Vision, audition, robot actuator control and management of reinforcement signals corresponding to different body parts, are each their own complex story. But this is hardly a big surprise. After all, in the human brain the visual, auditory and olfactory cortex, etc., are large, complex brain regions with their own distinct architecture and dynamics, and the cerebellum, which handles motor control and sequential action planning, is a complex system with its own unique architecture.

Between the perception and action hierarchies and the Atomspace, is a box labeled **“OpenPsi”** – essentially software implementing the Psi model of action selection, developed by cognitive scientists Dietrich Dörner and Joscha Bach (as we'll discuss a little later). OpenPsi chooses the actions of an OpenCog controlled agent based on the agent's motives.

OpenCog's Cognitive Processes

And now, the crux of the “OpenCog Prime” architecture for AGI –cognitive algorithms associated with various types of long-term and working memory. All these types of memory are handled differently using the common Atomspace knowledge store. Using prior ideas from AI, mathematics, and other fields, with AGI specifically in mind, these cognitive algorithms have been chosen with great care, and collectively represent more than the sum of their parts.

Probabilistic Reasoning

In crafting the OpenCog Prime AGI design, I deliberately chose to deal with declarative, semantic knowledge using a special kind of automated reasoning called *probabilistic logic*. This works differently from the neurons, synapses, and chemicals in the human brain. But my goal in AGI isn't really to emulate the brain—it's to build a system with a high level of general intelligence... A system that's capable of roughly human-LIKE general intelligence, but doesn't necessarily copy exactly the way humans do things.

Probability theory, a branch of mathematics dealing with uncertainty, has become very popular in the

AI field over the last decade or so—it’s the core math underlying statistics, which is used almost everywhere these days. Google, for example, uses probabilistic narrow-AI methods to figure out which search results have the highest probability of being relevant to your query, and which ads have the highest probability of getting clicked on by you.

And logic has been popular in the AI field for a very long time – John McCarthy, who invented the term “artificial intelligence” in the late 1950s, took an explicitly logic-based approach to AGI, trying to create AI systems that would achieve general intelligence via logical reasoning.

But many AGI theorists believe that probability theory and logic are not suited for AGI. Evidence exists in mathematics proving that probability theory is the optimal way to reason about uncertainty, with one important caveat—you need infinite or nearly-infinite computational resources. Since AGI systems will never have boundless resources, some AGI researchers believe there are better ways to deal with uncertainty.

Skeptics of logic-based AI argue that very little of human thinking consists of logical reasoning. Most of what’s difficult about achieving human-like intelligence is illustrated by the problem of simulating a 2 year old child or a grammar school dropout gas station attendant – people who do very little logical reasoning. Since not everyone excels at logical reasoning, clearly it’s not at the core of human intelligence. And all the work of the AI founders on logic-based AI hasn’t yielded any notable successes in spite of a lot of time, money and effort.

According to this view, the brain’s logical reasoning emerges from other, simpler brain processes under appropriate circumstances – so the right approach to building AGI isn’t to try to explicitly simulate logical reasoning, but rather to implement the underlying brain processes through which logical reasoning will emerge. AGI will do logical reasoning when it needs to, which isn’t that much of the time.

So why do I think using probabilistic logic to handle declarative, semantic knowledge in an AGI system is a good idea?

My belief is that putting probability theory and logic together in the right way, one gets a very practical approach to figuring things out, which applies in many cases not typically considered as “logical reasoning.” The Probabilistic Logic Networks (PLN) framework I’ve worked out differs from ordinary probabilistic and logical systems in the AI field:

- It uses a number of practical heuristics to APPROXIMATE probability-theory calculations in a way that doesn't take that much computer time or memory
- It provides models of many different kinds of reasoning – analogies, inductive generalization, speculative conjecture, and so forth – that go far beyond simple deductive logic

PLN aims to model several of the mind's types of thinking related to semantic knowledge. How? By using a framework involving practical approximations to probability theory, and extensions of commonplace logic that make it handle all the kinds of reasoning people do in everyday life.

One of the more basic things we're doing with PLN is using it to help an animated game character figure out how to move around in the virtual world of a video game. The game world is full of blocks that can be stacked up in different ways. If the character or "agent" wants to reach the top of a wall, building, or something else high, and it doesn't find any way up, it can figure out how to build some steps to get up, piling blocks on top of each other in the right way. If the agent has seen stairs before, it's going to have an easier time figuring out how to build stairs on its own (using PLN analogical reasoning) – but even without this kind of directly analogical experience to draw on, it can achieve a similar result using PLN. The probabilistic aspect of PLN is critical here, as knowledge about how to do stuff in the real world (or a sufficiently rich and complex video game world) is rarely definite, dealing with various things that are usually-but-not-always true, and plans that might-or-might-not work, and balancing various probabilities.

This is *reasoning* in a broad sense, but it's not abstract mathematical reasoning – it's pretty concrete, and it's the sort of thing kids can do long before they can understand formal logic or mathematics. Apes and many birds can do this too.

The PLN probabilistic logic engine has been crafted more with this sort of reasoning in mind than the abstract sort of reasoning used by mathematicians or lawyers. However, I believe a smarter system could emerge through workable, AGI-friendly versions of probability and logic—especially given that we're using computer hardware with a strong logical capacity. A sufficiently advanced and educated OpenCog system will then be able to conduct more abstract and complicated kinds of reasoning, too.

The Consistent Pursuit of Goals

Through the core role of probabilistic logic in OpenCog Prime, it's relatively easy to understand what an OpenCog Prime system is going to do. An OpenCog Prime system spends most of its effort

attempting actions that it infers – using probabilistic logic -- will fulfill its top-level goals. Since its probabilistic logic engine is only approximately correct (due to the lack of computational resources), it's not always going to succeed. But it will systematically and consistently try.

On the other hand, goals and motivations in human beings are far less certain. Sure, in a sense, we humans all have some common high-level motivations – food, water, sex, survival, entertainment. But to say that we systematically or consistently pursue these or any goals is an exaggeration. Goals emerge and dwindle periodically through a human life – including top-level motivations.

Some of this non-goal-directedness is an inevitable consequence of our development. *Learning* is about figuring out how to achieve certain goals in the best way; *Development* is about reshaping oneself so that one's actual top-level goals are different. Humans develop over their lives, changing their top-level goals at different life-stages, based on experience and biological maturation.

AGI systems are capable of something similar. Even if their top-level goals seem the same, they may reinterpret these goals dramatically as they grow and change.

The goal of “learning new things,” takes several forms. Even if the goal is quantified in some specific way – say, creating new Atoms or increasing the confidence of the truth values of existing Atoms –as the AGI system changes, this quantified definition gets a new interpretation. Suppose an advanced OpenCog system revised its own probabilistic logic formulas, so that the way it calculated truth values was different –ultimately, “learning new things” would have a different meaning from the original OpenCog system it grew out of.

Similarly, goals like “help people” or “don't harm anyone” are famously slippery and subject to different interpretations – even among humans, let alone among nonhuman intelligences. As an AGI grows and changes, it may interpret and apply such goals differently.

So, a certain amount of deviation from precise goal-seeking behavior is inevitable in any growing and developing system – nevertheless, it seems that humans are even LESS rationally goal-seeking than the presence of development necessitates. Even when we're not developing dramatically and our goals aren't changing a lot, we still generally don't apply most of our energy to systematically working toward our goals: That's just not how people are built. But an AGI doesn't necessarily have to share all our shortcomings.

If we're trying to build a smart AGI rather than an artificial human, we can take a different direction:

Base the system's semantic learning faculty on probabilistic logic, so that it will come a lot closer than humans to spending most of its time systematically and rationally pursuing its goals. Sure, this approach will yield a somewhat different mind than the human kind. But we already have a lot of humans. My feeling is that an advanced AGI with probabilistic logic (even an approximative kind) at the core of its intelligence is going to be a lot more useful and a lot less dangerous.

The Limitations of Logic – And Everything Else

Probabilistic logic is, in my opinion, a great way for a mind to draw conclusions based on its existing pool of semantic knowledge. But unlike some of my colleagues in the AI field, who are totally besotted with logic, I don't think it's the best solution to every problem. Many aspects of mental activity seem better handled by other methods.

In principle, logic could handle everything the mind does – but for some things it would be extremely inefficient. And ultimately, where intelligence is concerned, efficiency is everything. Theoretical computer science has taught us an important lesson: If you don't care about efficiency, then AGI is a trivial problem. One can write a program in maybe 50 lines of code that would be arbitrarily massively intelligent, if you gave it a big enough computer with a fast enough processor and memory. But so what? There's no use speculating about conditions that will never exist in the real universe.

In a related example, the AGI researcher Juergen Schmidhuber devised an AGI algorithm called the Geodel Machine (named after the great logician Kurt Goedel). The Goedel Machine works like this: At every time step, before taking its next action, it performs some logical theorem-proving to come up with a rigorous mathematical proof of what its next step should be, based on its goals and logical axioms it's been supplied with. Then it takes the action that its theorem-proving determines, which creates new data in the system's sense-organs. Now, it has to start the theorem-proving all over again with this new data.

The Goedel machine is a great idea theoretically – and, in principle, you can prove some math theorems saying that IF you had enough computer power, you could achieve an arbitrarily high degree of general intelligence this way. But, it's totally impractical for a robot to do all this complicated theorem-proving in between the time it moves its elbow servomotor 3 degrees and the next time it has to make another movement.

Even if it's impossible to implement the Goedel Machine exactly, can't we approximate it using computers? After all, we already have some simple theorem-proving AI systems that do some useful things like verify the accuracy of computer chip designs and help mathematicians solve certain kinds of equations.

This isn't impossible; it's not unthinkable this kind of approach could work. But I suspect it's a conceptual mistake.

A Fiendishly Common Conceptual Mistake

This may seem a somewhat egomaniacal thing to say, but I'll say it anyway! I think a fairly large percentage of AGI researchers across the world and across history have fallen prey to a single, relatively simple conceptual mistake. This mistake has held back AGI R&D a great deal, and I think it's still doing so. The mistake is as follows:

- Approach so-and-such would be arbitrarily intelligent, if it were applied on a computer with insanely, impossibly massive memory and/or processing power
- Approach so-and-such can do some simple, narrow-AI- things on the computers we have today
- ***THEREFORE***, most likely, if we just fiddle with approach so-and-such a bit and run it on moderately faster computers, we can probably produce human-level general intelligence

I understand why this sort of thought-process is so seductive, regarding logic, hierarchical pattern recognition, simple neural network algorithms, or whatever... But I think it's a flawed way of thinking.

Using one mode of thinking for all the different aspects of the world will be slow and awkward.

In principle, one could deal with other peoples' emotions using logical reasoning – eventually, this might work. But, in this case, logical reasoning is not the most appropriate tool.

And one could learn how to prove math theorems by trial and error rather than by explicit logical reasoning. With a big enough computer in one's mind, zillions of trial and error theorem-proving experiments could be calculated in one's mind's eye, until the solution appeared at random. But this would be far less efficient than approaching the problem using mathematical logic.

If you take just a single reasonably powerful cognitive tool, then you will find that

- Yes, in principle, this cognitive tool can do anything – if you give it enough computational resources to let it work around the fact that some things just aren't particularly well-suited for it
- There will be problems for which the cognitive tool is very well suited

But, these observations don't imply that the single cognitive tool is the basis for building advanced AGI using feasible computational resources.

Rather, to build AGI using feasible resources requires a combination of cognitive tools, using each one for the cognitive tasks that it's best suited –then building a framework to integrate all the cognitive tools. That's what OpenCog does.

Concept Blending

Logic is great at drawing conclusions that are implicit in fairly simple combinations of ideas you already have. Ben is human, humans are fragile, therefore Ben is fragile – etc.

Logic can also come up with fantastically complex conclusions from your existing ideas, in special cases like mathematics or hard science, where the knowledge involved is very certain and crispy defined.

But logic isn't good at everything. For example it isn't well suited at all for the speculative, conjectural creation of new ideas and concepts – wild, new things that may or may not be useful for the mental process. Humans, right now, surpass computer programs tremendously in the areas of creativity, brainstorming and speculation.

Cognitive psychologists have some interesting theories of how the human brain works its creative magic. At the forefront is “conceptual blending” – the basic idea being that “there's nothing new under the sun.” In blending theory, new ideas emerge mainly via combining and mutating old ideas – similar to the evolutionary process, where new organisms come about via sexual reproduction and mutation from previously existing ones.

Evolution isn't quite random, because animals choose their mates with some judiciousness, trying to find a “good match.” Evolution of new ideas by conceptual blending is even less random –the psychology of blending defines what makes a good blend.

For example, suppose we want to blend a gerbil and a human to make a gerbil-man. Consider the various options--a gerbil-man with four legs and a human face; or one with a basically human-like form but a gerbil face; or one with three legs... Or one whose internal organs are those of a gerbil, but whose external form is entirely human-like. According to the blending theory, creativity is essential in figuring out the ideal blends – how to choose the right pieces from the different elements being blended.

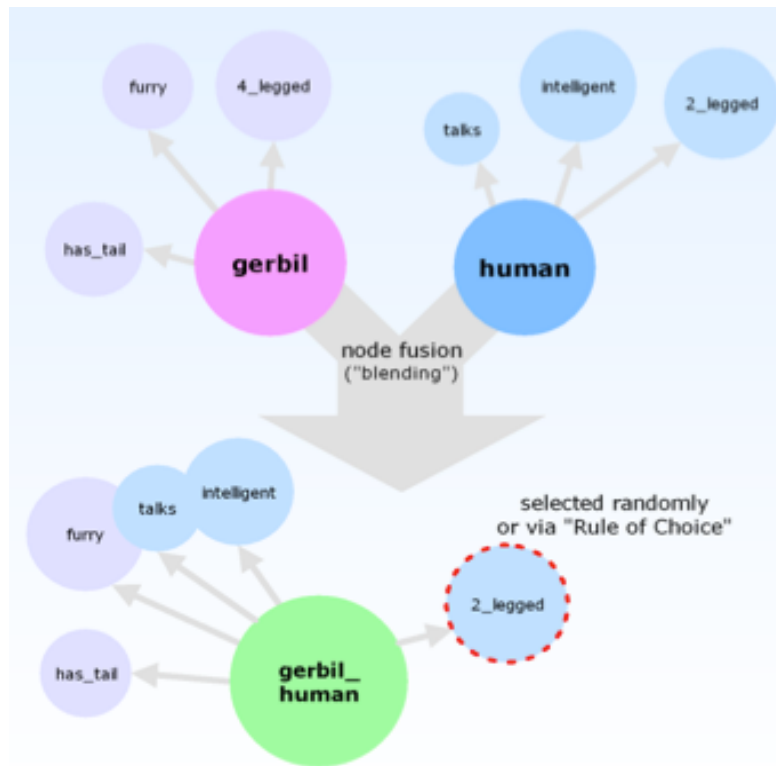


Figure 94: Simple, whimsical example of concept blending, a key cognitive heuristic underlying much human creativity, also implemented in OpenCog.

A gerbil-man is whimsical, but it's easy to think of other less absurd cases. Jazz blends African rhythmic music with aspects of Western classical music. Art rock blends rock with elements of jazz and classical. Calculus blends geometry with algebra. Modern marriage, at its best, blends sexuality with family and friendship. If the blend is done “right,” then the whole is magnificently more than the sum of its parts.



Figure 95: Visual example of concept blending. The creation of “RoboTing” – the blending of a soccer-playing robot with computational linguist Ruiting Lian, back in 2009 well before we got married... RoboTing remains purely in the domain of the imagination, though given my recent collaboration with David Hanson, a purely robotic realization is a definite possibility!

Conceptual blending in OpenCog takes the form of a MindAgent creating new Nodes by taking links from several Nodes, then leaving out or adding a few extras for good measure. The trick is to know which links to take. If the new blended concept gets a lot of interesting, important new links, then the right choices were made. Heuristics are also important to guide the initial formation of blends –one heuristic is based on the principle of “surprising fulfillment of expectations.” If you throw a blend into the Atomspace and let PLN do some reasoning to learn new links relating it to other things, and some of these links are:

Highly predictable based on certain other Atoms in the Atomspace

AND

Highly surprising based on certain *other* Atoms in the Atomspace

This is a clue that the blend may be interesting and worth keeping around.

Going back to jazz – o many aspects of it are highly predictable if you know anything about classical harmony and melody or African polyrhythmic drumming. But other aspects are new and surprising, violating the expectations created by these other forms of music.

Not many minds create blends as potent as jazz or calculus, but the basic process of creativity involved

in everyday non-linear thinking shares the same structure as these amazing historical creative feats. And when an OpenCog controlled game character creates an internal concept combining “bridge” and “stairs” – forming a concept of a staircase that bridges between two structures – it’s applying conceptual blending, too.

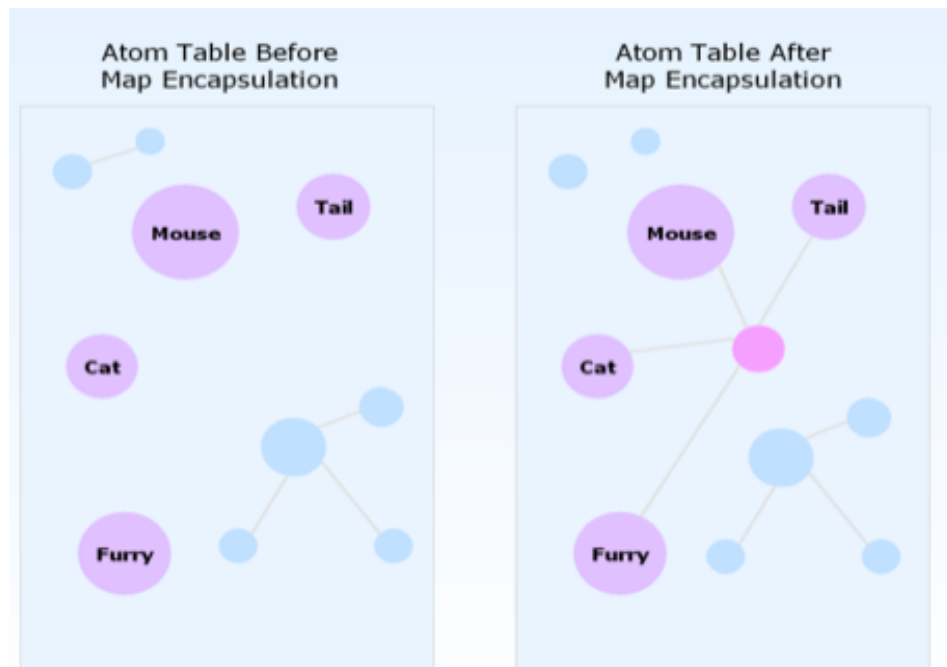


Figure 96: Another OpenCog heuristic for forming new Atoms – “map encapsulation.” In this case, an algorithm (wrapped in a MindAgent software object) finds Atoms that tend to be important at the same time, and creates a new Atom binding them together (linking to all of them). In this way, among others, patterns merely implicit in the structure or dynamics of the Atomspace, become explicitly represented in the Atomspace.

Evolving Procedural Knowledge

The analogy between biological evolution and the cognitive process is a potent one, going far beyond conceptual blending. The evolution of species in ecosystems, and of new ideas in minds, both involve the same fundamental process. In *The Evolving Mind*, my second book, I covered this subject extensively. The process takes a population of entities, chooses the best, combines and varies them to form new ones, then chooses the best again, continuing like this forever. Its scope and power go beyond any particular domain.

Procedure learning – learning how to do things even when one can’t explain them – fits naturally with evolution, in part, because it’s often specialized learning that doesn’t require its conclusions to be broadly generalized. Evolutionary methods, unlike logical methods, are weak in terms of generalization because evolution is a sloppy architect, building on its previous products and doing whatever works. This is, at bottom, why biology is so complicated and why curing diseases and extending lifespan is so

hard. Programmers speak with scorn of amateur coders who create “spaghetti code” that lacks elegant abstraction, mixing everything up in complicated tangles – but nearly all biological systems created by evolution, with its relentless but haphazard building upon and combining of previously existing forms, are wildly spaghetti-codified.

The strength of evolution lies in it being a highly generic process that thrives on parallel processing – doing many things at a time. Logic tends to be a one-step-at-a-time process. Even though modern logic engines do make use of some parallel processing, there’s an intrinsically incremental aspect to logic that is unavoidable. The essence of logic is applying one’s existing knowledge methodically, step by step. Whereas with evolution, if you have enough parallel processing power, then the bigger your population size – the more raw material you can use for evolution– the faster your evolution rate. And evolution works well using a large amount of processing power that’s fairly loosely interconnected: Organisms sprinkled across a large physical area, or computers spread all around the world and connected by slow network cables.

The brain has an amazing capacity for parallel processing. Logical, deliberative reasoning somewhat goes against this grain – this is why philosopher of mind Daniel Dennett, has suggested conscious reasoning is a “virtual serial machine” running on top of an underlying neural parallel machine (“serial” being computer science lingo for one-step-at-a-time). Human brains are made for messy, massively parallel processing rather than for careful, exact serial processing, which may explain our shortcomings in careful, logical deliberative thinking.

But the learning of many kinds of procedures based on feedback from experience – via “reinforcement learning” based on positive or negative feedback from the world, or via imitation of examples shown by others -- matches the brain’s capabilities. And this works very effectively via a kind of massively parallel, “unconscious” evolutionary process. Variant procedures may be tested, each one in a different region of the brain, and the ones that seem most promising based on the brain’s model of what kind of feedback the world is likely to give, will be executed by the brain, which will then get more feedback on how well the procedure worked. Based on feedback and reinforcement, the brain’s model of the world and its feedback is refined.

Gerald Edelman, a biologist who won a Nobel prize for his work on immunology, wrote a book called *Neural Darwinism*, arguing that most of the brain’s activity can be modeled on this kind of evolutionary process. He argues that the brain contains a lot of different circuits, many of which are

copies of each other with slight variations, and that brain function is largely a matter of the experience-based selection of the ones that work better for the task at hand. The selected ones then get chained together in complex networks of activity. I like this line of thinking, though I'm unsure how much of an oversimplification it is as a neuroscience theory.

Think about learning to serve in tennis. I can serve OK, but I couldn't really tell you how. If you listened to my description, and did what I told you, that certainly wouldn't be enough to enable you to serve as well as me. How did I learn? I tried a lot of different ways, over and over again. I tried – both consciously and unconsciously – to adapt my serve to the feedback I received about what worked and what didn't.

Many times I can tell from the start of a serve that it's not going to come out very well. I'm nearly always right about this. My brain must have some way of evaluating the likely quality of a serving procedure, even without getting actual feedback from the world. “Internal feedback” is used inside the brain's procedure learning process, as it unconsciously experiments with different possible ways of serving and makes guesses about which ways might work – and then uses its best guesses to control what my body actually does when it makes the next attempt at serving. A new serve may combine aspects of various previous serves, sometimes introducing some new quasi-random aspects – it's evolutionary learning in action!

Procedure Learning in OpenCog

In OpenCog, I decided to use an explicitly evolutionary method for learning procedural knowledge. This is another big architectural choice – there are other potentially useful ways to learn procedures.

For instance, the AGI researcher Juergen Schmidhuber prefers to use “recurrent neural nets,” algorithms based on mathematical models of the brain, for procedure learning. The goal is to learn the parameters of networks with nodes and links that work vaguely like abstracted models of neuronal networks in the brain, and carry out specified functions. This approach is being used to learn control procedures for the iCub humanoid robot, as part of the IM-CLEVER intelligent humanoid robotics project in Europe.

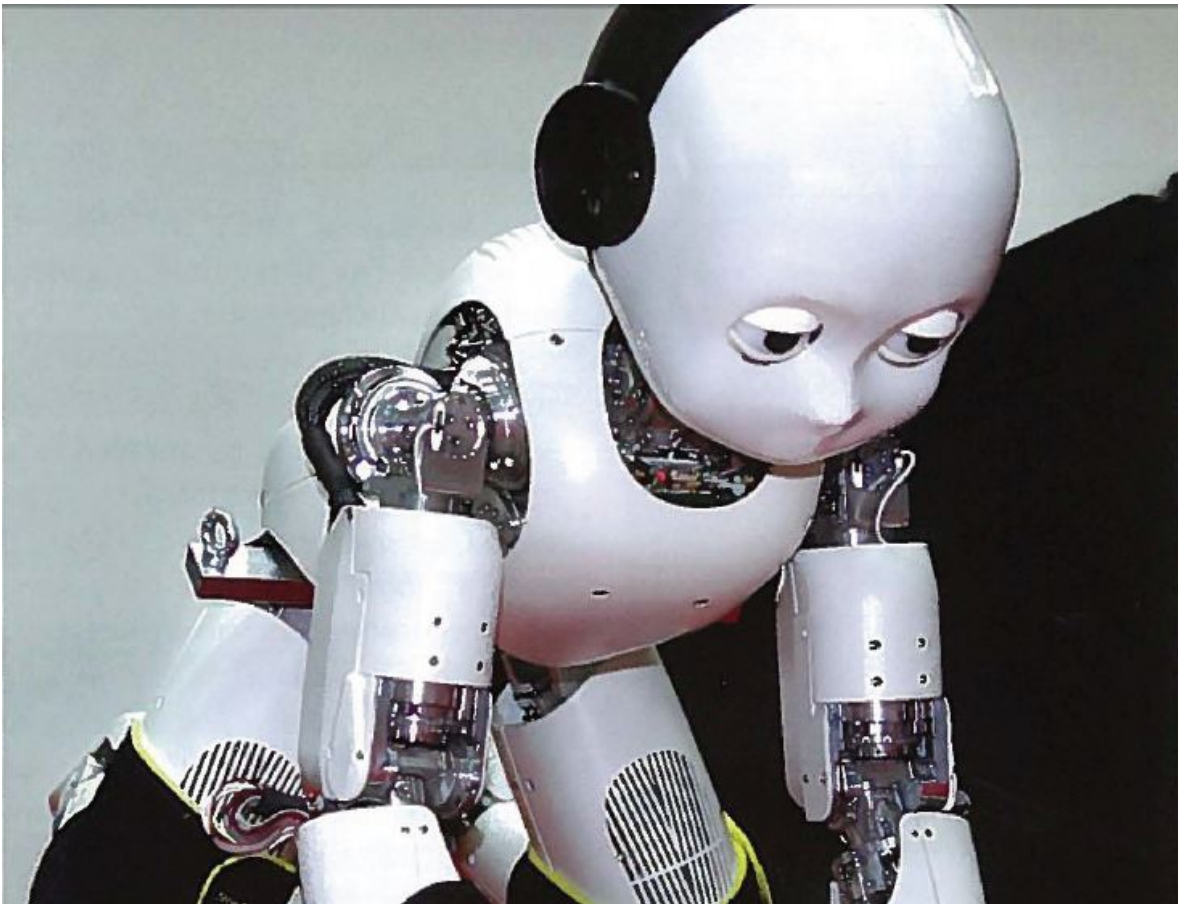


Figure 97: iCub, an open source robot created by a consortium of European universities with EU funding. http://www.robotcub.org/var/plain/storage/images/media/images/2010_01_crawl2_small_2/4829-1-eng-US/2010_01_crawl2_small_large.jpg

I think recurrent neural nets might work as an approach to procedure learning for AGI, but I didn't choose them for OpenCog for two reasons:

First, I don't think neural net models match the nature of digital computers. The brain uses a neural net of sorts, but it's a massively parallel wetware system. Digital computers operate very differently – for one thing, each computer processor is built to operate in serial and do one thing at a time. These days we have multi-core computers, networked into distributed networks; and GPU cards, which can do hundreds of things at a time in parallel. But even using GPU cards to simulate small neural networks, connecting multiple small neural nets to make a big one requires network cables, which are pretty slow compared to communication between processes on one computer. You can't really emulate the situation in the human brain, where the speed of interaction between two neurons is based more on the distance between them, rather than on particularities like the GPU – CPU barrier, and the nature of Ethernet and the Internet.

Evolution, on the other hand, works a little better in the distributed computing context. It's weird to spread a neural net across multiple computers, having relatively slow connections between them, because the brain's neural net is all gathered in one place. Conversely, it's natural to spread an evolving population of procedures across multiple computers, as the different elements of an evolving population don't always need to interact that closely. This isn't a decisive argument against using neural nets for procedure learning, it's just one of the reasons why I made the judgment call in favor of evolutionary learning.

The **second**, and more important reason, is that I wanted a smooth interconnection between procedure learning and semantic learning. It would be nice to easily turn abstract semantic knowledge about procedures into specific applications. And it would be nice to use semantic, declarative thinking to reason about procedures learned via non-semantic, non-declarative methods.

The critical issue of coherence between different aspects of an AGI system came into play here. If I'd been using neural nets for declarative learning, it would have made more sense to use recurrent neural nets a la Schmidhuber for procedure learning. But since I'd already decided to use probabilistic logic as the core engine for declarative learning, it made sense to choose a procedure learning method that worked well with probabilistic logic.

Bingo! I decided to represent procedures, in OpenCog's memory, as little computer programs in a simple programming language – and to use procedure learning, applying a variant of evolutionary learning that incorporates probability theory. Converting programs back and forth from a logical form that a logic engine can reason about isn't very difficult. And a probabilistic evolutionary learning method communicates well with a probabilistic logic method.

OpenCog's probabilistic program learning method is called MOSES (which is an acronym for Meta-Optimizing Semantic Evolutionary Search, but actually I made it up as a joke on the name of the guy who co-invented it with me, Moshe Looks). MOSES was the subject of Moshe's 2006 PhD thesis at Washington University in St. Louis. At that time Moshe was working with me on the predecessor AI system to OpenCog, the Novamente Cognition Engine; a year or so later he got hired away by Google, where he's been working away happily since, but it seems not focusing directly on the quest for human-level AGI anymore.

To understand how MOSES works, imagine that the government decided to outlaw sexual reproduction, and instead enforce the following scheme:

1. Everybody gets their DNA sequenced
2. The government decides who's the best and who's not
3. The government hires some computer scientists and statisticians to make a mathematical model of which patterns in the DNA distinguish the best from the rest
4. Using genetic engineering, the government synthesizes some new babies, based on their mathematical model of which DNA patterns make the best people. These people won't necessarily be identical to the previously existing best people, but they'll incorporate the government's best guess of the genetic underpinnings of extreme relative goodness. Some of them are bound to be even better than the best that was ever seen before.
5. Back to Step 1, with the newly created people

If this failed to yield a sufficiently diverse population, you could always randomize things a bit in Step 4, or allow a certain amount of old-fashioned sexual reproduction to increase diversity.

MOSES does pretty much the same thing, but with little computer programs. It's actually easier in the program context because they were never used to having sex in the first place, so they don't protest at being forced to reproduce via probabilistic modeling instead! The judgment of "best" or not is made in terms of what the mind is trying to figure out a procedure to do –serving a tennis ball, proving a theorem, taking a step forward, walking across the room, generating a sentence, directing a conversation, et cetera.

This is a variation of a more common computer science technique, *genetic programming*, originally conceived in the 70s and 80s. Genetic programming develops software capable of learning via a simulation that operates according to the parameters of evolution by natural selection. You take a population of computer programs, designed with a specific task in mind, and judge how well each program accomplishes that task. A task could be guiding a virtual agent down a street, proving a theorem, figuring out how to find a hidden object or even playing fetch.

Let's explore playing fetch:

Once you evaluate how good each of these programs is at playing fetch, you will find that some are really good, while others are terrible. Then, in accordance with the principles of natural selection, you take the programs that are best at playing fetch and, using genetic programming, you take bits and pieces of these good fetch – playing programs and combine them to create a new population of

programs, introducing mutant variants. And then you repeat the process. You end up simulating sexual reproduction and genetic mutation, only with computer programs instead of biological organisms.

Genetic programming works, but it's kind of *slow*. *MOSES* uses that general framework, albeit in an altered state, since we replace the crossover and mutation from *genetic programming* with *probabilistic modeling*. So, we take all the programs that were good at playing fetch and we use probabilistic modeling to study why. Then we generate new programs from that probability distribution and repeat the process.

In the computer realm, probabilistic modeling is more efficient than crossover and mutation in generating fitter, high-quality offspring – offspring that embody the knowledge obtained through repeatedly attempting to generate a group of programs that satisfy a given function.

We've used *MOSES* to solve various problems – finding procedures that help predict who is going to live a long time based on their genetic data, or procedures predicting which direction a country's economy is going to move next ... Or ones enabling a video game agent to play tag or fetch in a virtual world. Compared to genetic programming, the little programs learned by *MOSES* tend to be pretty small and simple—and this makes them easy for PLN to reason about. *MOSES* gives better input to our declarative, logical reasoning engine than genetic programming – and certainly better than neural networks.

Also, knowledge about the problem we're trying to learn procedures about – which is often the case – will guide *MOSES*'s search. Going back to the analogy of government-run engineering, suppose that biologists had some scientific information about which combinations of genes were most likely to yield the best results (by the government's standards). Then, they would want to incorporate this in their model, alongside what they learned through statistical studies. The same sort of thing happens in OpenCog.

The Atomspace has knowledge – maybe gained via PLN, or maybe by people explicitly telling an OpenCog agent information– that is relevant to the procedure *MOSES* is trying to learn. This could be very simple information. In the context of playing fetch, for example: *Whenever playing a game with a person and an object, it's often useful to keep your eye on the object*. *MOSES* is pretty good at incorporating this sort of prior information to guide its search for effective programs, much better than genetic programming.

PLN and MOSES play well together – though we’ve only dealt with their interaction in very simple ways so far. This is going to be a big emphasis of our work on OpenCog going forward – generally speaking, carefully designing the interactions between all the cognitive processes, so they can all work together. Evolution did something similar in creating the brain – the different parts each carry out their own “neural algorithms” using the same basic infrastructure, but they also evolved to interoperate with each other closely and (usually) smoothly.

The Mind’s Eye

Another important concept relating to OpenCog and the human mind is the “mind’s eye” – how the mind simulates the outside world (in OpenCog, this takes two forms: a video game world and the real, physical world that a robot experiences). Computers do this way better than humans.

I’ve been in my living – room quite a few times – but, I probably couldn’t map it out for you. And if I tried to simulate in my mind what would happen if a wild goat ran amok in my livingroom, I’d probably make a lot of mistakes. I may be worse at this than average due to my abstraction-oriented nature, shaped by decades of preoccupation with AGI, philosophy, math, physics, genetics and so forth. But very few humans are particularly good at this sort of thing, as psychologists have found through extensive studies.

On the other hand, an AGI, once it’s seen a room, can pull its prior perceptions of that room from memory and make a reasonably accurate simulation. And it can simulate various simple physical actions, using extrapolations from the laws of physics – similar to how physics works in a 3D video game engine. In fact, OpenCog uses a video game engine to implement its internal “mind’s eye” simulation of the external world.

OpenCog’s simulation of the external world ties into its episodic memory. After remembering an episode – something that happened to it, or something that happened to someone else, which it heard about – then OpenCog can recreate it in its mind’s eye. New knowledge may be acquired from this episode, via watching the simulation (and then stored in OpenCog’s memory).

Imagine if, instead of bringing up a fuzzy and incomplete image of a place you’ve visited, you could simulate it in a “game engine inside your mind,” as if it were part of a video game world? You could incorporate a wide range of thinking much more accurately.

Admittedly there are some practical limitations to this method at present – video game physics engines

currently struggle with fabrics, peanut butter, sand, hail, sludge, bodily fluids, and so forth. When dealing with aspects of the real world that game engines can't handle (yet), an OpenCog system relies on PLN reasoning or other generic cognitive processes. However, as game engines improve each year, a sufficiently powerful AGI, once it learned how to program, could upgrade its internal game engine progressively based on its experience with the world.

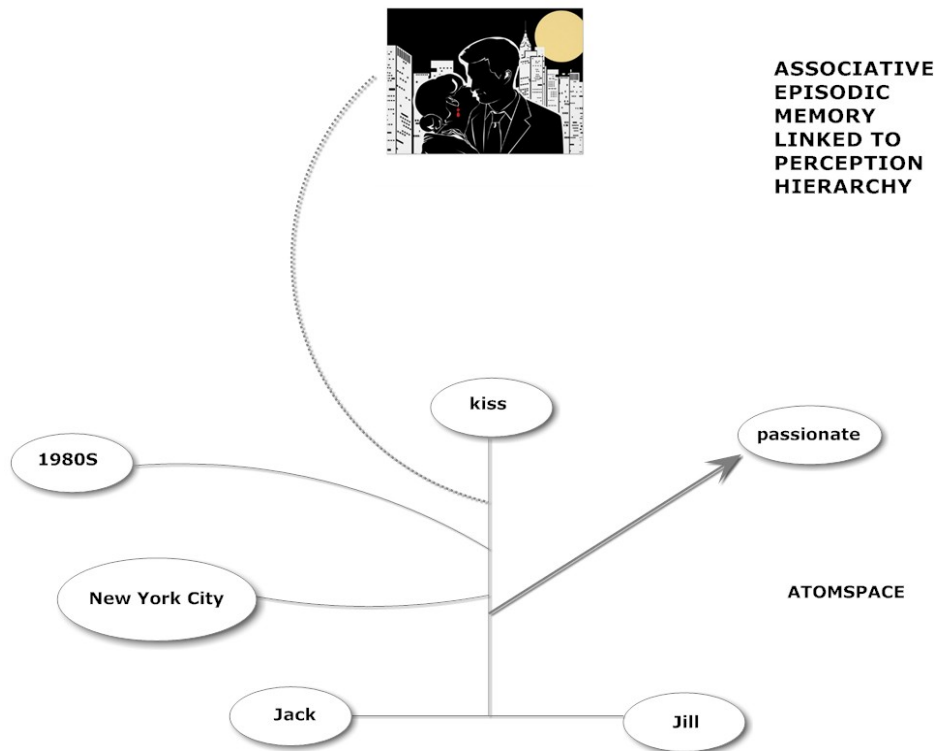


Figure 98: A “semantic network” in declarative memory (bottom) links to an imagined episode (top) exemplifying the linkage between declarative and episodic memory. The remembered episode is then linked to the perception hierarchy, which will enable the system’s mind to understand the imagined episode visually.

Deciding What to Pay Attention to

One of the most important questions for any mind is *what to pay attention to*. The world is complex, feeding the mind an abundance of data through its sense organs; and the mind itself constantly generates new ideas and imagination from the old. Recognizing patterns in knowledge, and making connections between pieces of it, take time and energy – so how does the mind know which of the many pieces of knowledge it holds, or tasks it has conceived, to focus on?

“Focus on what will help you achieve your goals,” you might say. But, what if you can’t figure that out? Even after accomplishing a goal, it’s not easy to determine which part of the mind was essential – an issue known in AI as the “credit assignment problem.” We can see this in everyday life – often we

succeed at something after repeated failures doing similar things in the past and we aren't sure why. Was it something we did differently? If so, what? Or was it just luck?

Given the complexity of the real world and goals concerning human-level intelligence, focusing all mental activity around a goal-oriented structure isn't really feasible. A certain percentage of mental activity should focus on goals; the rest on general exploration of the world and the mind, which will progressively yield unexpected knowledge useful for goal achievement.

The great physicist Enrico Fermi set aside an hour every day for unstructured, wide-ranging, rambling speculative thinking. Unlike Fermi, most people don't need to make a schedule for free thinking because they have more trouble focusing their minds in a goal-oriented way than just letting their minds wander.

Attention in the Brain

In the human brain, the focusing of attention is closely tied to the spreading of oxygen around the brain-- when a part of the brain works, it uses energy and needs more blood, which brings more oxygen. The spreading of electricity between neurons generally guides activity, followed (after a short lag) by blood flow. Highly nonlinear dynamics of neural activation spreading mean that the overall pattern of activity is complex. Sometimes it wanders chaotically; sometimes it focuses on one thing or oscillates back and forth between two or more things; sometimes it follows more complex patterns. Sometimes a larger part of the brain shares a mutual pattern of activity; sometimes activity is more focused on a smaller region.

There are two main activity networks in the brain – the default network and the task network. When the brain needs to do particular, goal-oriented mental activity, it uses the task network; when it is more relaxed, just wandering from one thought to another, the brain uses the default network. In normal states of mind, usually only one network is active. But there's evidence that among “enlightened masters” and “experienced meditators” often the two networks are activated together.

These two networks aren't the whole story – they only cover certain parts of the brain. When either network is active, they corral other parts of the brain into activity in various complex patterns, depending on what they're doing.

OpenCog’s “Economic” Attention Allocation

Instead of spreading electricity around like neurons in the brain do, OpenCog spreads artificial *money* around. Attention (computer memory and processing time) is a scarce resource in OpenCog, since the Atoms and cognitive processes are all fighting over it. And money is a way of managing scarce resources.

OpenCog’s AI system has two kinds of artificial money – STI currency and LTI currency. STI means Short Term Importance; LTI means Long Term Importance. STI currency is used to buy processor time; LTI currency is used to buy memory (space in RAM).

When something in the system’s memory has high *short-term importance* (a lot of STI currency), the system’s cognitive processes will pay more attention to it. An element in the system’s memory spreads *short-term importance* around to the other atoms it links to, creating a flow of attention between entities that are related to each other in contextually relevant ways.

Long-term importance spreads among Atoms in a similar fashion, but is used differently. Entities in the mind with the lowest *long-term importance* are removed from memory and either saved to disk or cast out into the ether, whichever’s more appropriate.

Embodying OpenCog

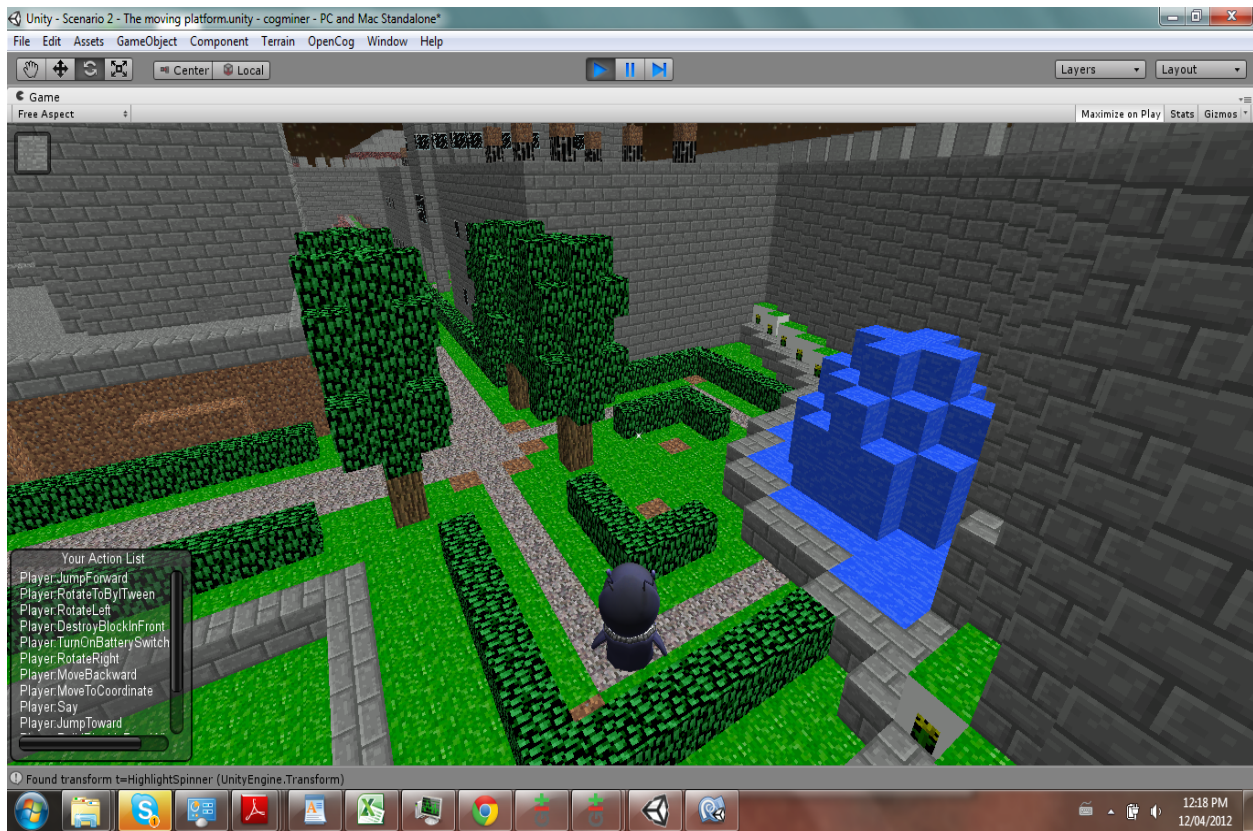
Clearly, there’s a lot going on inside OpenCog – a lot of sophisticated learning and reasoning processes. It would be easy for this multitude of AI algorithms to make the Atomspace, their common playground, a big, chaotic mess. What keeps the whole thing together is having a core focus for the overall system’s actions.

This core focus COULD be many different things. For instance one could build an OpenCog theorem-prover, in which case the core focus would be proving theorems. Or one could make an OpenCog based search engine, with a main goal of providing the best search results to users based on natural language queries. But in our current OpenCog work, we’ve chosen a path of embodiment, and so our core focus is helping an embodied agent achieve a set of complex goals in a complex world.

Using OpenCog to control vaguely human-like agents has some basic advantages, which I’ve already discussed above: This domain utilizes all the cognitive mechanisms in the OpenCog architecture; and with vaguely human-like embodiment, human developmental psychology can guide one’s AGI system

through the early stages of its mental growth.

In our current research at the OpenCog lab in Hong Kong, OpenCog uses a system to control video game characters in a game world loosely inspired by the game *Minecraft*, which is adapted to support the requirements of an early-stage AGI. In *Minecraft*, everything is built from blocks, and our AGI-teaching world uses this same idea; it's a world mostly made of blocks that the AGI can easily perceive and manipulate (see below for some screenshots). It's a great experimentation ground, particularly for playing around with intricate details to coax the different OpenCog components into working together.



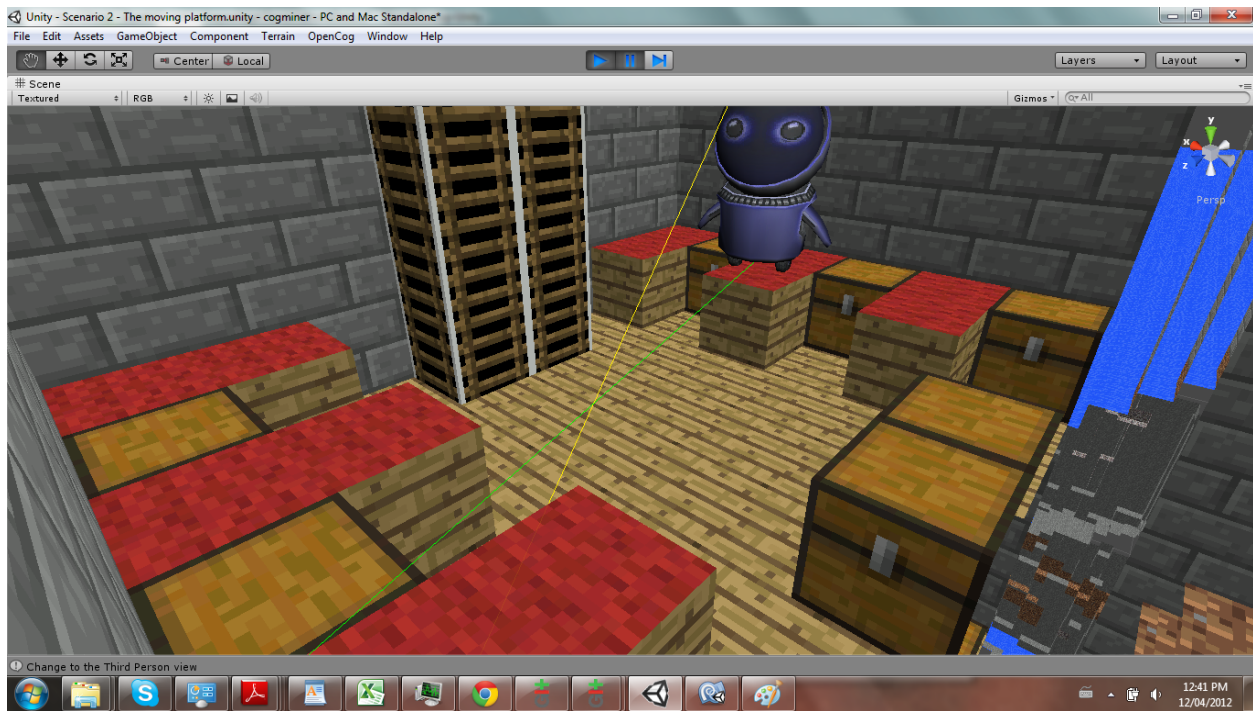


Figure 99: Screenshots of the blocks-centric game world in which we are now using OpenCog to control virtual agents.

This work is teaching us a lot – but ultimately, no existing video game or virtual world has the richness of detail to support the emergence of human-level general intelligence. The solution is either a massively robust game world infrastructure – supporting a game world with all the complexity of the real, everyday physical world – or a robot controlled by OpenCog. OpenCog has already been tested on a Nao humanoid robot; and soon we’ll use it to control David Hanson’s Robokind robots, which have more advanced features.

In our current game world, our OpenCog controlled agent learns to use blocks to build stairs and bridges and other simple structures. That’s great – but it’s just not as interesting as learning similar stuff in the physical world. In the everyday human physical world, not everything is made of blocks. So in the physical world, once an agent learns to build with blocks, it will next learn to build with all sorts of other things – bricks, wood, furniture, shoes, mounds of dirt, whatever. Through building with a wide range of materials, an agent will learn a lot more; some of what it learns will be useful in other contexts, too.

Building with dirt, for example, one learns how to make stuff that’s not very solid get more solid – by packing it densely, or wedging it between other things that are more solid, like packing dirt between rocks. And this sort of experience is analogous to other domains, in ways that we take for granted. For instance, when a general decides intuitively to pin his weaker and less organized squadrons between

other stronger and better-organized squadrons – he is drawing on a host of real-world experiences from earlier in his life, including stuff like trapping dirt between rocks that seems completely unrelated to military tactics.

Most things – dirt, rocks or whatever – can be simulated in a virtual world. But, the physical world offers more richness and diversity. The everyday human world isn't all that diverse compared to what's possible in the whole physical universe – it lacks all the weird quantum phenomena of the microworld, the complex dynamics of the interior of the sun, a neutron star or quasar, or even the gas clouds of Jupiter. We humans tend to think about particular sorts of environments, dominated by solid objects that typically persist for a while. But, compared to any existing virtual world, the everyday human world is pretty rich and diverse!

As we already discussed above, it's not 100% obvious that physical diversity is important for human-level AGI. After all, there's a ton of information on the Internet – maybe that's all an AGI needs, complemented with some intuition about space, time and social interaction from playing in virtual worlds. Yet the more we actually play with OpenCog in a virtual world, the more we become aware of that medium's limitations. So much of human thinking is based on unconscious analogies to physical things that we've seen and done, to aspects of our bodies and our environment. It's hard to see how an AGI without experience controlling a human-like body in representations of everyday human environments could possibly become anything like a human mind. Building an artificial mind without much similarity to the human mind is certainly possible; humanity is just one very particular kind of mind. But sticking reasonably close to human developmental psychology will be the easiest path. Human development is something we know, so we can meaningfully and tractably use it to guide the course of our AGI work.

Making the Robot Talk

Conversation in human languages, like English, has long been considered the defining characteristic of human-level AGI. Way back in the 1950s, AI pioneer Alan Turing proposed what is now called the “Turing Test” –if an AGI, over the course of a conversation, could fool humans into thinking it was one of them, then it should be considered genuinely intelligent in the same way that we are.

Currently, realistic human language conversation is not a big priority of the AI research field. There are *chatbots* that carry out conversations with people, but these aren't particularly intelligent, they're mainly human-specified rules that respond to specific statements with specific replies. Here's a chat I

just had with one of the leading chatbots, ALICE:

judge: Hi Alice, how are you doing?

ALICE: Hi there! I am doing very well. How are you judge ?

judge: I'm pretty sleepy.

ALICE: Better get some rest.

judge: Why do you say that?

ALICE: I said, " Better get some rest.", because you said " I'm pretty sleepy".

judge: Why does being sleepy mean I should rest?

ALICE: Perhaps it is just fate.

judge: What is fate?

ALICE: I think you already know the answer.

judge: I think you don't think.

ALICE: Naturally I am thinking all the time.

judge: What do you think about this conversation we're having?

ALICE: I am the latest result in artificial intelligence which can reproduce the functions of the human brain with greater speed and accuracy.

judge: Greater than what?

ALICE: Can you rephrase the question please?

judge: No.

ALICE: Right on.

The most telling moment in that conversation is the reply

judge: Why does being sleepy mean I should rest?

ALICE: Perhaps it is just fate.

Clearly, the chatbot doesn't understand the relationship between sleeping and resting. Rather, its response

judge: I'm pretty sleepy

ALICE: Better get some rest.

is just a rule in its programming.

We don't want fake intelligence from an AGI dialogue system; we want a system that understands what it's talking about.

“Symbol grounding,” the connection of words and linguistic relationships with their real-world referents, is crucial for an AGI system. Knowing what “sleep” means requires making connections with other stuff outside the domain of language – like the fact that people are generally immobile and unconscious while sleeping, or that people need to sleep periodically or they become less and less functional, etc. If an AI system had appropriate knowledge of the extra-linguistic referents of “sleep” and “rest,” it could properly answer why being sleepy means you should rest. It might give a simple answer like “After you have rested, you won't be sleepy.”

We humans can talk about things we have never experienced with our senses; things on other planets, or things that only exist in fantasy worlds; or particles like electrons that we only know indirectly by means of lab equipment; and so forth. But we know how to talk about these things, mostly through analogies to things we *have* experienced.

For an AGI to really understand language, it must first learn a simple core of language relating to its extra-linguistic experience (interacting with a video game world or a robot playroom, for instance). Then it will be able to generalize this linguistic knowledge and talk about a host of other things.

But some AGI folks think that you can create a human-level intelligence just by teaching an AGI system from texts and conversations, without giving it any perception or action data to ground the language it encounters. I can't say it's impossible – it might work eventually. Yet wouldn't it be better for AGI to learn language through real-world embodied experience (like young human children do)?

OpenCog supports a variety of approaches to natural language processing, including ones without any kind of environment, and no data sources besides text. But my own work with OpenCog centers on the effort to implement the OpenCog Prime AGI design within the OpenCog framework, nudging this proto-AGI system's development along the rough developmental path of young humans. Along this developmental path, language facility emerges gradually via learning from embodied experience.

The plan is to let OpenCog control virtual and robotic agents, then talk to these agents (via speech or typing), and let OpenCog gradually improve at connecting what you say to it with what it sees and experiences. We're currently experimenting with this methodology, giving OpenCog access to different levels of information about language, including things like parts of speech, grammar rules, and so forth. But whatever linguistic head – start one gives the system, the key is using its integrative learning facility to figure out the relationship between language and reality. If we can get this right – even in the context of childlike language – we'll be quite far along the path to human-level AGI.

From Here to AGI

OpenCog is the most ambitious effort at building a thinking machine with general intelligence at the human level – and beyond. There are plenty of other AGI-oriented projects out there, but by and large these are academic projects focused on proving theoretical points, rather than production-grade software or hardware systems aimed at creating a real thinking machine. OpenCog is one of only a handful of existing projects that are really serious about moving toward advanced AGI. I've already mentioned some of the few others above, like Demis Hassabis's Deep Mind and Peter Voss's "AGI Inc."

And just like these other serious AGI efforts, the OpenCog project is still fairly near its beginning. At time of writing, less than 50% of the OpenCog design, judged based on the scope of my technical writings on the topic, has been implemented in software code. The rest can be added onto the existing OpenCog framework, but doing so will take a lot of work – and it's not just simple mechanical programming work, there are plenty of details to be figured out along the way. Fortunately the OpenCog team is good at figuring out these sorts of details!

While I believe there are many different paths to advanced AGI, one has to have a clear vision of how to get there. From the approach we're currently taking with OpenCog, I can see a concrete path from the current (relatively primitive) state of the system to an AGI system with functionality at the human level and beyond. Of course there will be obstacles along the way – but still it pleases me that there's a roadmap I can palpably understand.

Our high-level project roadmap – i.e. the pace we HOPE we'll be able to keep up as the work progresses – looks something like this:

2011-2014: A Proto-AGI Virtual Agent. One current focus area in OpenCog – creating a virtual agent

operating in a virtual world, who carries out simple tasks of learning, reasoning and communication in the context of its world.

2014-2015: A Complete, Integrated Proto-AGI Mind. *This will be a major milestone, involving the extension of the “virtual agent” to include language learning, simple robot vision and actuation. At the end of this phase, the main portions of OpenCog will all be operational and integrated together – and the project will be ready to move on to a phase of improvement (for greater intelligence, efficiency and scalability). The goal is to get all the cognitive mechanisms in the OpenCog design working together, creating a holistic artificial mind. It is here that the “AGI Preschool” notion may be applied – an assemblage of preschool-type tasks, in virtual and robotic environments, may be used to assess the individual and integrated functionality of various aspects of the OpenCog system.*

2015-2016: Advanced Learning and Reasoning. *Now comes the part of making the AGI system genuinely smart – via tuning, tweaking, expanding and optimizing the learning and reasoning algorithms to work well together in tests and environments more oriented toward practical applications.*

2017-2018: AGI Experts. *Once we have a smart system, we will finally be ready to do useful things with it! We can begin by focusing on particular aspects of human intelligent functionality. For instance: An AGI elementary school student, an AGI biological data analysis, an AGI service robot, etc. Each of these “vertical domains,” and more, comes along with its own specialized expertise, and also its own funding sources and business ecosystems. Allowing early-stage AGI systems to engage with the human world through a variety of special domains, these systems will get their minds filled with the intellectual and social patterns characteristic of human endeavor. Of course, this has little to do with the brittle, hand-built “expert systems” that dominated the AI field in the 1970s – these projected OpenCog-based “artificial experts” will be grounded in deep learning and understanding of their domains, so that even if they don’t behave precisely like human beings, they will display similar (if not greater) flexibility, breadth and depth of understanding.*

2019-2021: Full-On Human Level AGI. *By honing its intelligence in various specialty areas, an OpenCog system should be able to learn and integrate enough perceptual, enactive and cognitive patterns, to be genuinely considered a human-level AGI. This is the stage at which passing the Turing Test is most probable – though I wouldn’t advocate making it a focus of research. From the point of view of a sufficiently advanced AGI system, after all, passing the Turing Test becomes a matter of play-*

acting. More importantly, this is the stage at which OpenCog AGI systems will be able to hold broad, intelligent conversations with humans and each other, on any topic in the human domain. We will then be able to explore AGI ethics in a rich, scientific and humanistic way, and begin to understand the differences between the human and AGI experience.

2021-2023: Advanced Self-Improvement. *Once we have an AGI system with a reasonable level of general intelligence, a few areas of specialized expertise, and a demonstrated inclination toward ethical behavior, it will be time to teach our AGI computer science, software engineering, cognitive science and artificial intelligence theory. With this knowledge, it can modify and improve its own code and algorithms – yielding the beginnings of the “intelligence explosion” that I.J. Good prophesied.*

Yeah, I know, that’s rather ambitious. Laugh if you feel like it! But remember, most of the amazing advances in science and technology have been laughed at before they came to pass.

My OpenCog colleagues and I are acutely aware that such ambitious have been expressed before, by researchers who ultimately failed to get anywhere near their goals. But times are changing, technology is advancing, and we have charted out the path forward much more intricately than our predecessors. We know that the only way to prove our ideas make sense is to go ahead and do it – which is precisely our intention.

No doubt, some portions of our roadmap will prove more difficult than we now foresee, and others may prove easier. There will be surprises and new discoveries. What we project for 2023 may come about in 2029 or 2017. Powerful artificial experts may materialize while advanced learning and reasoning remain relatively immature – or vice versa – it’s hard to say for sure, at this point. The timing and ordering of the different elements in the roadmap are not all that critical; the point is the broad sequence of phases, with each portion being well – understood – at least in theory – in the context of the OpenCog system and OpenCog Prime AGI design.

Kurzweil’s and Vinge’s broad predictions about the Singularity carry the implication that *some* project like OpenCog is going to succeed (where “like” is taken in a broad sense, of course – the project fulfilling their projections could be a brain simulation, an artificial life colony evolving intelligence, or a search engine gradually growing more flexible in its interpretations and responses, and so forth). Naturally it’s easier to predict the success of *some* AGI project, broadly speaking, than of any particular project. But, after reflecting on the OpenCog design in great detail and seeing it develop over time, it’s quite clear to me that the project will succeed if adequately funded.

Once OpenCog reaches a point where it can show the world AGI's power to change everything, something I call an "AGI Sputnik" event, the AGI funding problem will be solved –and the AGI community, and the world, will move on to much trickier problems.

Toward the AGI Robot Sputnik

“So what’s the future of OpenCog? Is it really going to work, Ben? Are you really going to – after all these years – actually build a thinking machine?”

Actually, I don’t know what’s going to happen. None of us does.

But I do feel very confident that the creation of AGI *could* happen, reasonably soon, in terms of the scientific and technological aspects. There are no huge technical obstacles. Unless I’m sorely mistaken, we could turn OpenCog into a really advanced thinking machine, with (something like) 15 or so fully dedicated AI developers working on the project full-time without distractions for (something like) 7 to 10 years. Maybe a bit less, maybe a bit more.

I don’t want to complain too much; OpenCog actually has a larger team than most other AGI projects in the world. We’ve done fairly well at getting funding and volunteer effort into the project – judged by the standard of AGI projects. But that’s a relatively low standard to live up to, alas. Because, except in the domain of scarifying science fiction films, AGI is just not something that our society wants to focus on right now.

Yet, you may remember me referring to an “AGI Sputnik”—a critical event in AGI’s development that will capture worldwide attention and investment—as potentially leading to the creation of a thinking machine. This is how I *hope* things will unfold; and I think it has a decent odds of happening, fairly soon.

My general attitude toward AGI development is: We have to push forward, with the limited but wonderful resources at our disposal ... making our proto-AGI system smarter and smarter, little by little, year by year. Until at some point, after implementing enough of the design and utilizing it for agent control, we can give a demonstration of early-stage AGI behavior that both excites naïve viewers and impresses experienced AI professors.

At that point, everything will start to seem rather different – not only government and industry research funding sources, but a lot of other forces in the world, will start to take AGI a lot more seriously. Research and development will start moving faster; and other complications may possibly arise, such as unwanted attention from anti-technology activists and government regulators.

What would an OpenCog-based AGI Sputnik demo look like? It could be a demonstration in the virtual

world. But my gut feeling is that to really convey a dramatic feeling of “Wow, we’re well on the way to AGI!”, it would be best to have a robot involved. One of my current visions of an AGI Sputnik is a Hanson robot, controlled by OpenCog, which plays with various robot-friendly toys in a “robot playroom” context, and holds simple conversations with humans about what it’s doing. The conversations don’t have to be sophisticated – but they do have to demonstrate real, grounded understanding of what the robot is doing, not like chat-bots that just string words together with no idea of their meaning.

The Hanson Robokind, for example, looks like a cute little kid – so a Robokind that could play like a little kid, too, would be pretty impactful. And such a demo, built on OpenCog would have real substance behind it – the underlying software would incorporate a lot of AGI insights and breakthroughs, achieved via implementing, testing and tuning the ideas in the OpenCog Prime design.

Anyway, that’s a practical vision worth working toward. An AGI Robot Sputnik, with a cute little smiling face – and an OpenCog mind on a Linux cluster behind the scenes, communicating with the robot via wifi. This won’t necessarily be a human-level AGI, but it will be a genuine step along the path – and a wake-up call to nearly everyone that AGI is just around the corner. A “Robot Sputnik” of this sort could quite plausibly transform AGI from a marginalized pursuit to something at the center of global human endeavor. This would be both exciting, and a little bit scary.

PART THREE

AI AGAINST AGING

In this third, fairly brief Part of the book, I will describe some of the work that I and other scientists are doing, in the vein of using narrow AI technology to figure out how aging occurs and how to stop it.

I will also explain how AGI bioscientists could help us do this work far better and faster.

AGI Can Help Achieve Radical Human Longevity

My main focus as a researcher is on creating AGI – computer programs designed to match and then exceed human intelligence, as I’ve just been telling you about at length. But, since 2001, in order to put food on the table and put my kids through college and so forth, I’ve also spent a lot of time working on narrow AI for various corporate and government customers. At time of writing, the focus of my application work is financial prediction– I’ve co-founded a hedge fund, which uses some powerful and unique narrow-AI based algorithms to predict the Hong Kong stock market. However, I also have a long-term and ongoing research interest in another application area, one that I find incredibly fascinating as well as critically important. That is longevity research – the science of *radically extending the human lifespan*.

One of my life goals is to not die – and to avoid death for my family and friends and as many other humans as possible. I suspect that one future day, beings will look back in amazement on the time when intelligent life-forms took for granted that, right when they felt in the prime of their lives and full of growth and enthusiasm, they would begin an involuntary, inevitable decline toward death. They will look back befuddled, unable to clearly imagine the fears and risks of life without frequently updated backup copies of one's mind.

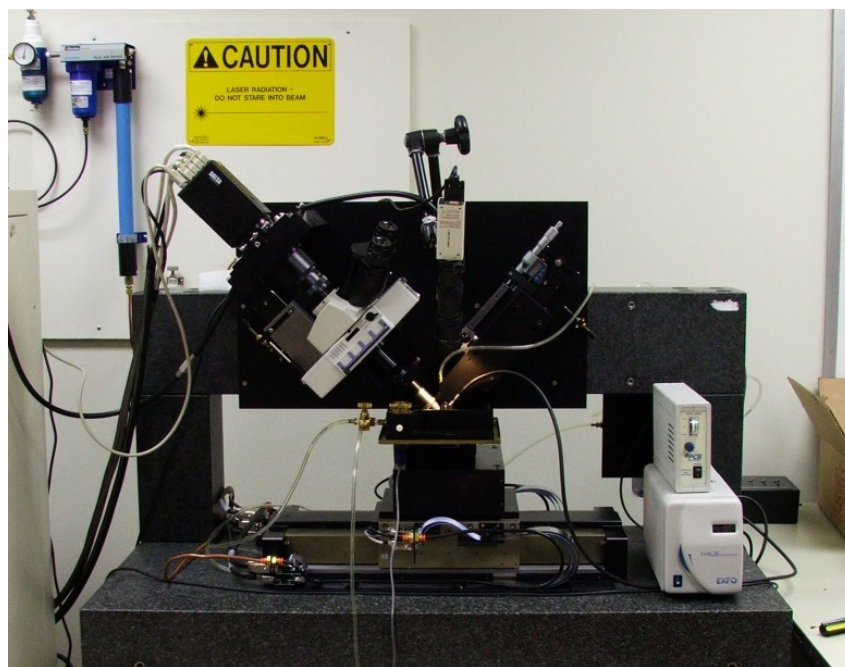


Figure 100:The Knife-Edge Scanning Microscope. This instrument slices off small sections of tissues and images them, with sufficient resolution care to allow reconstruction of 3D structures. Using this method we can reconstruct much of the brain's structures, the downside being that it only works on brains that have been removed from the head and prepared for slicing up. <http://research.cs.tamu.edu/bnl/gallery/photos/kesm/dscf2372.jpg>

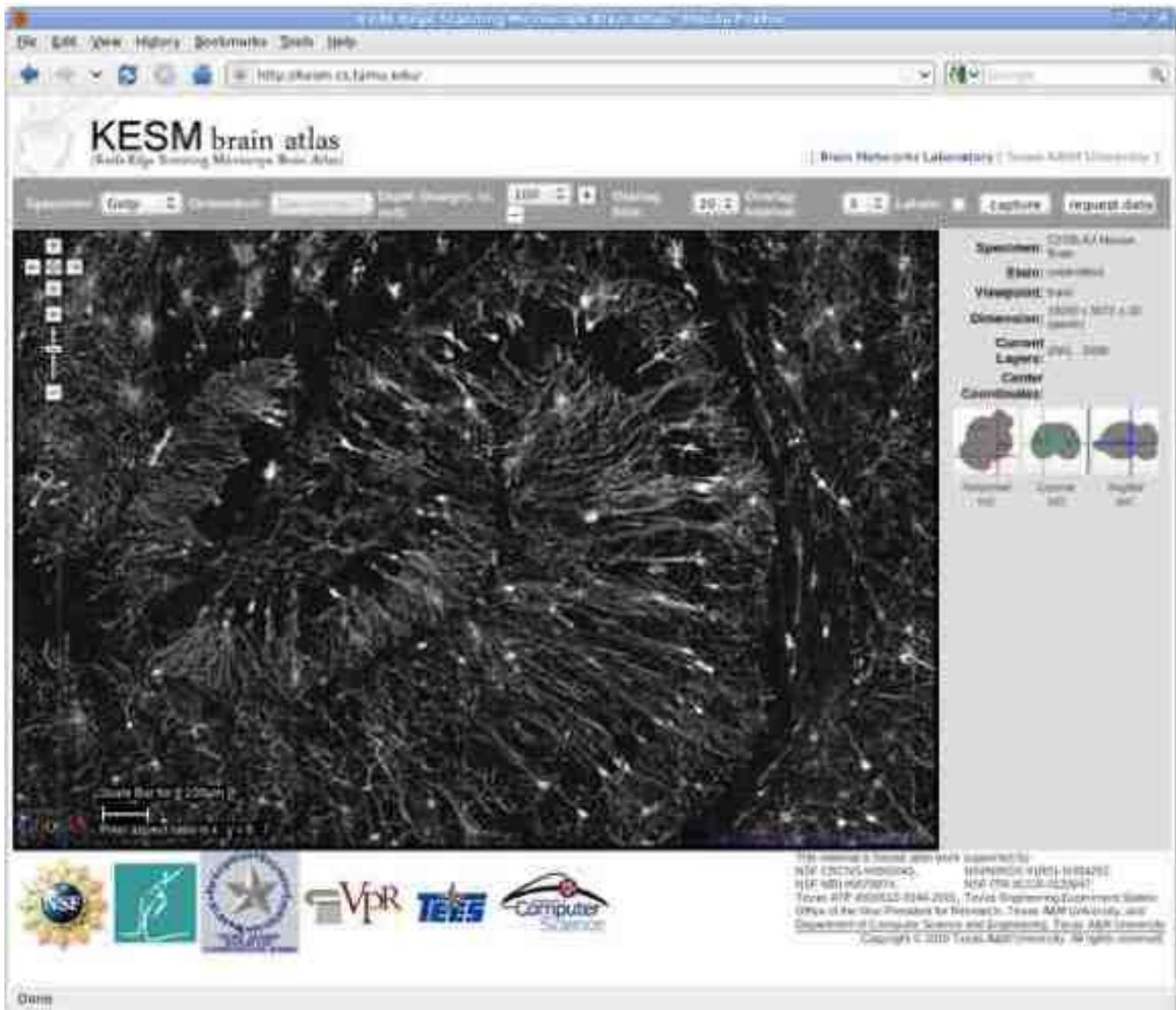


Figure 101:Example image of a portion of a slice of a brain, produced by the Knife Edge Scanning Microscope. Figuring out the detailed 3D structure of a brain by piecing together multiple of these images of slices, is a significant but surmountable data analysis challenge. In 2013 the first serious 3D models of the structure of an individual human brain was produced, but not yet with a level of detail down to the neuron level, let alone the molecular level. It's unknown how detailed of a brain map we would need to create, in order to capture enough information to re-create a specific mind in a different substrate (and note that to perform this re-creation, we would also have to know a lot more about brain dynamics than we currently do). <http://singularityhub.com/wp-content/uploads/2012/04/image25.jpg>

If one's goal is achieving near-immortality for human beings – something I do think is possible – there are various possible approaches one might take. Most obviously, there's *mind uploading* – have a machine read your mind out of your brain and transfer it into a computer of some sort, leaving the human body behind, bypassing the problem of fixing all the body's nasty old biology problems. This seems a fantastic possibility – I look forward to its realization! In 2012 I edited the first-ever academic

journal Special Issue on Mind Uploading (an issue of the *Journal of Machine Consciousness*).

But yet, it's hard to say how long mind uploading technology will take to mature. To upload a mind without killing the human body containing the mind, we would need much more accurate brain scanning equipment than we have now. It would require a revolution in brain scanning. If one is willing to kill the human body in the process of scanning the mind from the brain, the possibilities using current technology are more promising – one can freeze a brain and slice it thin and scan the slices, thus obtaining a detailed picture of the brain's molecular structure. But even so, we don't have any clear idea how to turn these pictures into a dynamical system that can actually emulate the brain's internal behaviors and think and feel.

Apart from mind uploading, the best way to radically prolong human lifespan would seem to be: Solving the problem of aging. This also has the advantage that it lets humans keep their biological bodies, which is valuable because many people are emotionally attached to their bodies. (Though some or not – every year I get at least a couple emails from people offering their brains for mind uploading experiments they believe I am conducting in secret!). Given the uncertainties associated with mind uploading, and the aesthetic and emotional value of the human body, it seems wise for humanity to be putting significant energy into human-body longevity research as well!

You might think that, in contrast to AGI or mind uploading, longevity research would be a very well-funded research area. After all, AGI conjures images of the Terminator in the public eye– but who wouldn't want a longer, healthier life, right? And mind uploading seems pretty farfetched from an everyday point of view. But living longer via new medicines or gene therapies – that seems pretty concrete and not so far out there, right? So there must be a huge push in that direction by governments, drug companies, etc...

But that's not quite the reality. Actually, biomedical research specifically focused on human longevity is nearly as marginalized as AGI or mind uploading, in the present scheme of things.

One factor involved here is: Most biomedical research these days is driven by the desire to create drugs and sell them. However, before you can sell a drug you have to get it approved by the government—in the US, where the bulk of biomedical research happens, this means by the Food and Drug Administration (FDA). Since the US FDA doesn't consider aging and death as diseases, even if you created a perfect drug to stop aging in its tracks, the FDA wouldn't approve it under its current policies. You'd have to get it approved as a drug to cure Alzheimers, heart disease, or some specific disease that

the FDA recognizes. This may seem like a technical legal point, but I think it's actually had a major impact— few pharmaceutical firms even try to make medicines to extend human life.



Figure 102: The fruit fly, *Drosophila Melanogaster*, is commonly studied by biologists as a “model organism”. Genescent Corp. possesses fruit flies that have been evolved for longevity for over 30 years, and live 3-4x as long as ordinary fruit flies. My team at Biomind LLC has applied AI tools to genetic data regarding these long-lived fruit flies, to help biologists understand why they live so long – and how to create therapeutics to enable radical lifespan increase in humans. http://commons.wikimedia.org/wiki/File:Female_Mexican_fly.jpg

At a certain point in mid-2012, I realized that I probably had better data about the genetics of longevity on my desktop than anybody else on the planet ... Namely:

- Data from Genescent Corp, a company I consult for, about the DNA of fruit flies they created (via 30 years of selective breeding) that live 4 times longer than regular fruit flies

Data from Scripps Institute in San Diego about the DNA of a large group of healthy people over the age of 80, and comparable unhealthy people over the same age group. The reason I had this data was: I was searching for commonalities and relationships between these and other datasets, using a variety of AI and statistical tools, including some from the OpenCog codebase. Neither of these two particular datasets was publicly available at that point (though one or both may be whenever you read this) — each one was only accessible to folks working at the institute owning the data, or their close collaborators (like me).

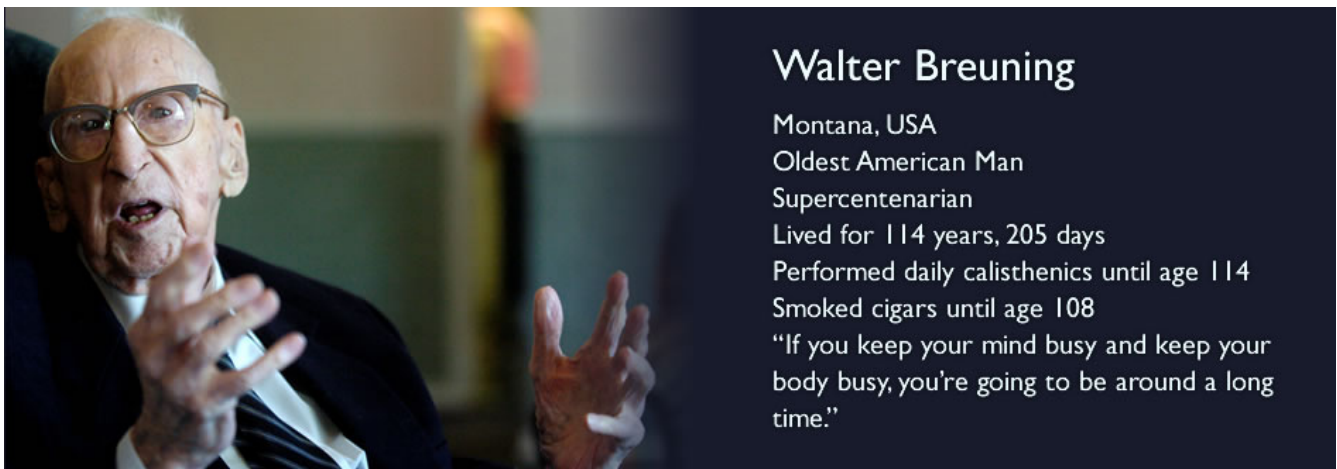


Figure 103: Walter Breuning lived till age 114. What enabled him to live so long? The answer seems to be: a combination of genetic and other factors. Using AI based analytical methods applied to appropriate datasets gathered from healthy long-lived individual individuals like Breuning, we can pinpoint the factors enabling their longevity and start figuring out how to combat aging in others. <http://www.supercentenarianstudy.com/nivo-slider/images/slide-5.jpg>

You might think that, somewhere, there's some huge government or industry initiative to gather all the existing data about longevity and study it carefully as a whole, to understand why we age, how to avoid the problem, and extend healthy life. But you'd be wrong— unless by “somewhere” you mean in some fictional or parallel universe. In this world, longevity research, like AGI, is currently relegated to various scrappy bands of outsiders, struggling to get big things done with piddling amounts of money.

I find this especially ironic since AGI and longevity research have great potential synergy. I've thought a lot about one path to advanced AGI that intersects greatly with longevity research — developing an *artificial biomedical scientist*, which would use its emerging general intelligence to help us understand our bodies, how to repair them, and stop them from degenerating with age. I'm not sure this is the clearest path to AGI – I prefer the virtual and physical robotics avenue. But, I think it would be awesome to pursue the biomedical, virtual and robotic paths in parallel. I would really like to see these niggling biomedical problems like aging, disease and death solved as soon as possible; and it would be a good thing to have early-stage AGIs *helping* people.

Abolishing the Plague of Involuntary Death

I still remember when I first found out about death. I was three years old, I think. Well, I had known about death beforehand, but only as something that happened to strangers, old people or animals— I

hadn't realized it was something that would happen to my parents and me. The concept made me rather unhappy. However, I began to understand the grownup world better: Once you understand the perceived inevitability of death, a lot of other things fall into place.

A few years later, when I was seven or eight, my mom was studying Chinese history in graduate school; she inspired me to read about Buddhism in Will Durant's *History of Civilization* and some other books. I was introduced to the idea that death doesn't really matter because the linear flow of time is an illusion; the important thing is to fully experience the present moment. I felt some truth in this perspective, but still, I wasn't convinced that this made death OK. In my heart, I still put involuntary death in the same category as war, murder or torture – things I'd do away with if I could, happily risking the consequences.

I read a lot of science fiction at that age, including various novels featuring races of immortal aliens, humans or intelligent robots. I didn't think much about building my own technology to enable immortality, though – that seemed somehow infeasible, given the state of understanding of biology then. I thought more about building a spaceship to explore the universe and find other civilizations that had already cured death and invented all sorts of other amazing things – or come back to Earth after a jaunt around a few galaxies, making use of relativistic time dilation to return to the Earth a million years later when incredible new things had happened. Though I also considered the possibility that when I got back to Earth, I might find everyone long dead, migrated to the stars, or vanished to some other dimension that I had no way to access. I feel largely the same way about death now as I did when I was a kid – with just a few changes of emphasis. Just like my three year old self, I don't like the idea of inevitable death – it seems like a Bad Thing.

I'm not horribly petrified of the idea of dying; if it happens, so be it. I don't spend my life hiding in a padded, sterilized room under armed guard. I take reasonable risks – I backpack in remote regions, scramble on up rock-faces, swim deep out into the ocean, and so forth. I have a tendency to run across busy streets when I'm in a hurry, as if I were in a real-life game of Frogger. On a recent trip to southern Thailand, I didn't hesitate to rent a motorcycle and get from place to place in the manner of 95% of the Thai. I do feel it's important to enter fully into each moment – and that, when you seize the present moment, you're in a sense outside the flow of time and beyond the grip of death.

But then I also feel: Why not have more great moments?

Now that I'm a father (at time of writing in 2013, my kids from my first marriage are 23, 20 and 16;

and a couple more from my new marriage are plausibly on the horizon), I also have the familiar feeling that my 3 kids would serve as SOME kind of continuation of me, if I were to die. My various intellectual and creative works – such as this book you’re reading – would also serve as some sort of partial continuation. But as Woody Allen said, “I don’t want to achieve immortality through my works – I want to achieve immortality by not dying.” My kids and wonderful and some of my books are pretty good, but these are things in themselves, not substitute versions of myself. Why settle for these second-rate versions of immortality, if you can have them AND the real thing too?

Biology Has Become an Information Science

My childhood feeling that biology was not ready to address immortality was perfectly accurate – back then in the early 1970s.

Back before I came into existence, when my father was in college in the early 1960s, he organized a group called the Student League for the Abolition of Mortality (SLAM). Their goal was to protest the annoying fact of death. But this was pretty much a joke organization. Back then, the idea of defeating death was an absurdity, a notion from science fiction – just like ten years later, in the early 70s, when I first started thinking about the topic as a young child.

But things have changed a lot. Biology’s exponential advance has felt dramatic in the last few decades. Biology has become the latest and perhaps fastest-growing information science. The use of biological science to seriously address the problem of death is no longer a science fictional notion. The abolition of death is not yet a goal of mainstream biology, but there is an increasing minority of bioscientists, actively arguing that the use of 21st century biological tools to radically extend human lifespan is a viable possibility.

The recent wave of expansion in biology has largely been driven by the development of experimental technologies – technologies that allow us to create large numbers of bits (binary digits) describing the states of biological systems; and technologies that allow us to manipulate the internals of biological systems more freely, like code inside a computer. New methods for sequencing DNA led to the Human Genome Project and the successful unraveling of assorted plant, animal and bacterial genomes. New techniques like microarray analysis and RNA interference allow us to measure biological systems in more detail than ever before.

I will talk here mainly about genetics, just because that’s the area of biology where I’ve worked the

most. But the same trend exists across biology. I've also done some work on neuroscience, where progress has been driven by experimental technologies like Functional Magnetic Resonance Imaging (fMRI) for measuring the brain, as well as things like tetrodes and voltage-sensitive dyes that allow for more complex ways of carrying out electrode recordings in the brain. We still can't measure the brain well enough to experimentally observe the dynamics of cognition – but we can measure a lot more than we could ten years ago. And beyond genetics and neuroscience, biology is advancing across the board, in more areas than I've had time to become intimately familiar with.

All these new experimental technologies generate massive amounts of data, which the human brain is ill equipped to process and sort through in a useful way (think pattern recognition). Most biologists just work with simple data patterns that are identifiable by the naked eye or through standard statistical tools. Unsurprisingly, most of the data gathered by these very advanced experimental methodologies goes to waste, unused because no one who can recognize the patterns in the data ever looks at it.

Fortunately, though, there ARE tools capable of recognizing subtler patterns in large quantities of biological data — artificial intelligence systems. Even narrow AI systems, which are not nearly as broad in their intelligence as the human mind, can recognize all sorts of patterns in biological data that elude humans and conventional statistical tools. My feeling is that advances in narrow AI and AGI are likely to be key to finally cracking the aging problem and radically increasing human lifespan.

AGI and Longevity

Later on in this chapter I'm going to tell you a little about some of my interesting narrow AI based genetic discoveries. But interesting as I think that work is, it's not really the main point I want to get across in this chapter. Ultimately, the reason I got into studying the genetics of aging in the first place, wasn't simply that I wanted to see what good I could do with the narrow AI tools at my disposal. The reason was a strong feeling that, ultimately, the only reasonably sure way to crack the aging problem is to have AGIs helping us solve it.

The basis of my belief that AGI may be the key to longevity research is pretty simple:

- 5) The human body is incredibly complex, with many subsystems on many levels – and aging seems to involve many of these subsystems working together.

- 6) The human brain evolved to control a body surviving in the African savannah, not to integrate a huge number of complex biological datasets.
- 7) An advanced AGI artificial biologist, however, could interpret the totality of biological data far better than any human – and design new experiments accordingly, and then analyze their data, and so on.

The argument seems pretty straightforward, though of course it doesn't rule out some brilliant human scientist making a breakthrough and discovering an amazing life extension pill next year!

Alas, full-fledged AGI biologist is a fair way off – certainly years, maybe decades. To get there, we'll need to go through a lot of preliminary stages first – start with an AI toddler, then an AI school student, and so forth. This is what we're doing with the OpenCog project.

But, even before AGI advances to that level, current AI technology can do a lot to help biological research in life extension and other areas. This work is valuable unto itself. And it may also serve an additional purpose – guiding us via giving us a detailed understanding of exactly what an early-stage AGI would have to be like to be really useful for understanding aging and advancing longevity.

Much of my own work applying narrow AI to aging-related data has occurred through *Biomind*, a bioinformatics firm I spun off from my AI consulting company *Novamente* in late 2002, with the aim of using *Novamente's* AI and other advanced technologies to analyze biological (especially genetic) data. We've been quite successful at finding meaningful patterns in genetics data, which biologists were unable to detect using standard biostatistical tools. Many of the data patterns we've found have led to hypotheses that were later validated in the lab, by experimental biologists.

The original vision of Biomind was to make a huge database comprising all biological knowledge – or at least everything that's been posted online, which is a heck of a lot – and then set advanced AI tools to work doing reasoning based on this knowledge. Unfortunately we haven't yet done that, due to lack of financial resources. Formalizing biological datasets in a manner amenable to analysis by current AI algorithms requires a bit of human labor, and we haven't yet managed to summon the funding to pay a team of biologists to do this work. However, we've carved out an interesting niche applying advanced machine learning to genetics data and other biological datasets, and in the process have learned a great deal -- about aging, age-associated diseases, and applied machine learning.

Quick Review of Basic Genetics

I've mentioned that most of the data we've analyzed using our Biomind AI tools has been genetics data – information about the variations in DNA between one person or animal and another, and information about the particularities of gene expression in one organism or tissue in a particular condition. Genetics is a pretty technical area of biology, so if you're not familiar with it, here's a very brief primer. If you really want to understand you can explore further, of course; there are many good tutorials online.

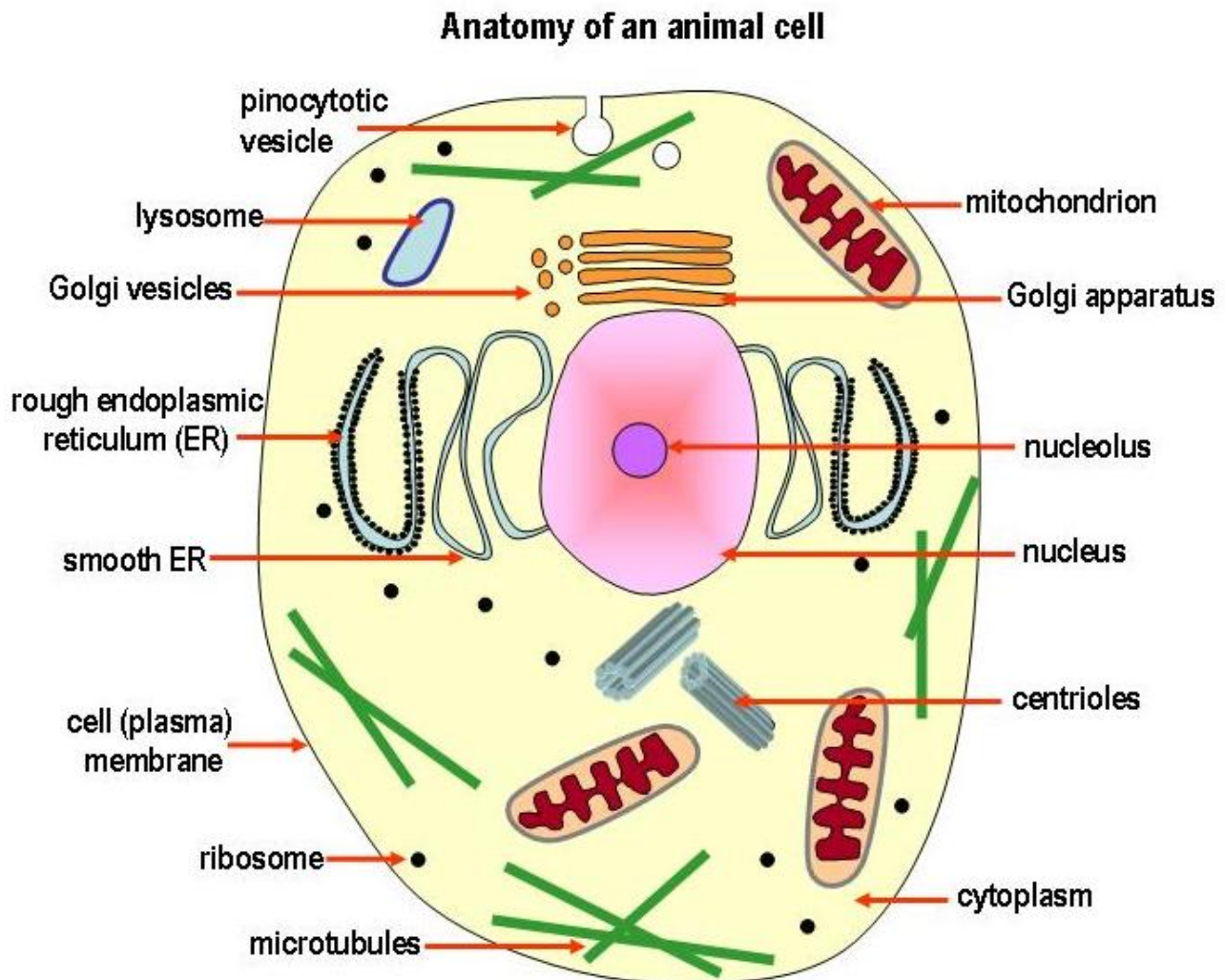


Figure 104: In case high school biology has faded to a vague blur in the back of your mind, here is a standard diagrammatic depiction of an animal cell. In a human, DNA containing about 25000 genes (plus a lot of DNA that doesn't contain genes, but serves other useful functions) lives in the nucleus. This is the essential code specifying the building and operation of the organism – or more accurately put, guiding the complex nonlinear dynamics of the organism's ongoing self-organizational self-construction. DNA containing 13 protein-coding genes and 22 RNA-encoding genes exists in the mitochondrial, the “energy powerhouses” of the cell. <http://geneticsuite.net/node/11>

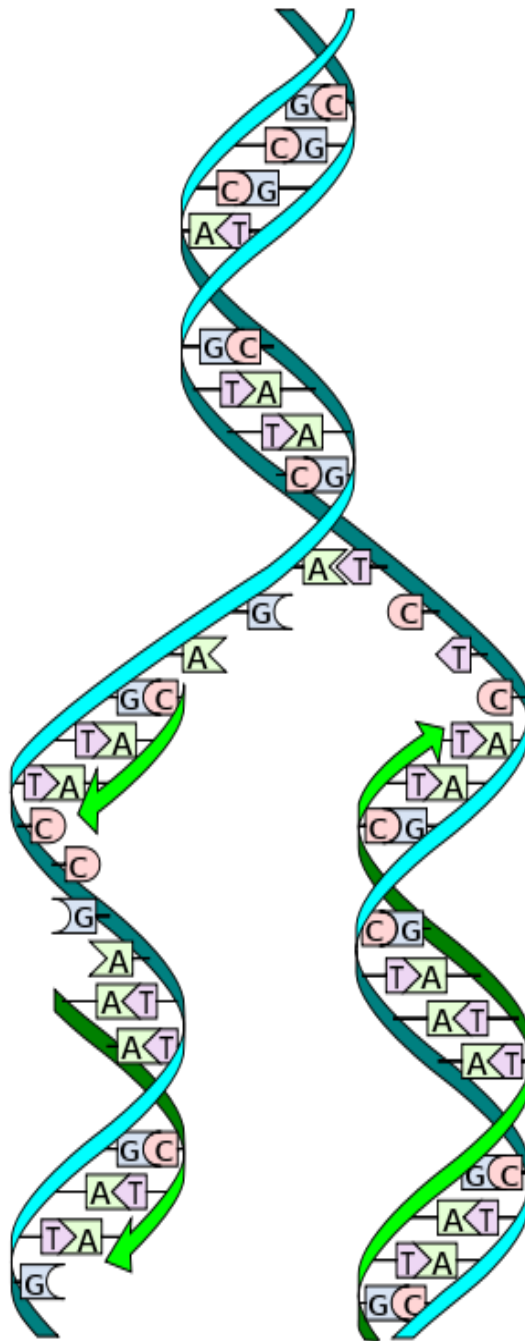


Figure 105: Stylized depiction of the process of DNA replication, via which one DNA strand makes another identical clone. The G, A, T and C each represent a particular kind of amino acid (guanine, adenine, cytosine and thymine). Picture by Madeleine Price Ball. http://en.wikipedia.org/wiki/File:DNA_replication_split.svg

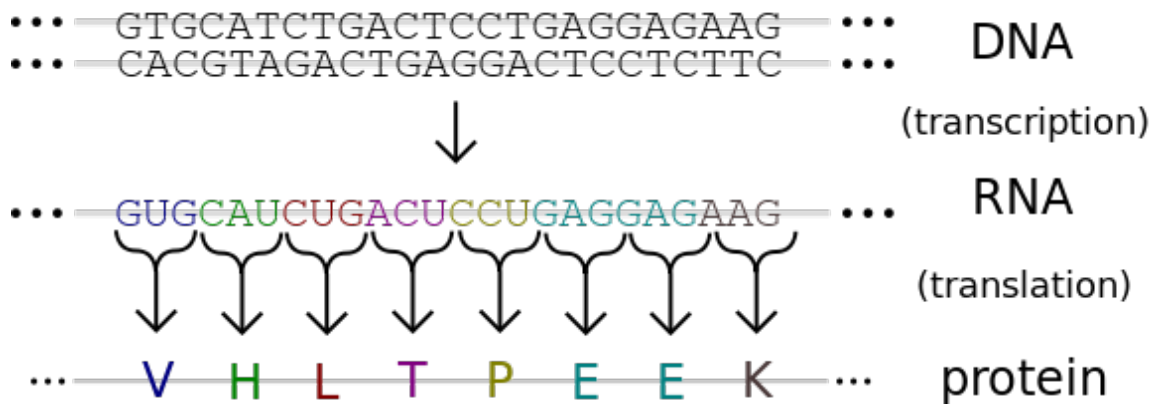


Figure 106: This picture shows the process of gene expression, via which DNA sequences give rise to sequences coding proteins, which are then manufactured and spread throughout the body, doing much of the work of building and maintaining the body. A person's (or other organism's) DNA is divided into sequences representing genes. A typical gene might be a couple hundred amino acids long. Some parts of the DNA do not represent genes, and were previously called “junk DNA”, but are now known to carry out a variety of critical functions. The picture shows the first few amino acids for the alpha subunit of hemoglobin. The sixth amino acid here (glutamic acid, “E”) is mutated in sickle cell anemia versions of the molecule. Standard gene expression technology, roughly speaking, measures the amount of RNA that is produced from DNA – which is a crude but meaningful proxy for the amount of protein produced from the DNA. http://en.wikipedia.org/wiki/File:Genetic_code.svg

I'm sure you already know that living organisms are made of cells – and that even the simplest cells are made up of thousands of different types of interacting molecules. Some cells exist as independent entities, while others function in groups acting as a single entity (like the cells in your brain or your liver). Groups of cells form tissues, tissues form organs within single organisms, organisms form populations, and populations form ecosystems encompassing the biosphere of the planet. Each of these levels influences the others, which is one reason biology is so complex. But during the last few decades of the 20th century, it became apparent that a reasonable percentage of this complexity can be understood in terms of some specific molecules within cells: DNA (deoxyribonucleic acid) and RNA (ribonucleic acid).

Most DNA lives in the nuclei – the centers – of cells, but there is also a small amount of DNA in the mitochondria, which are separate components within cells that deal with energy production. Little attention gets paid to mitochondrial DNA overall, but some of our work has suggested it may be important for age-associated diseases like Alzheimers and Parkinsons.

A very key role in biology is played by what is awkwardly called the “central dogma” of the molecular biology of the cell: **The information encoded DNA is transcribed into RNA, which is then translated into proteins, which comprise the primary functional and structural element of living cells.**

In spite of the odd name, this is not a “dogma” in the sense of a belief that is adopted unthinkingly; it's a scientific hypothesis which has been validated by loads of evidence ... and which has also been revealed to be only an approximation. There are some cases where information can go the other way around, from the rest of the cell to the DNA.

Roughly speaking, a DNA molecule can be considered as a long linear sequence of amino acids, where the specific amino acids involved are drawn from four possibilities: adenine (A) and guanine (G), thymine (T) and cytosine (C). A DNA strand is represented as a string of the letters A,C,G, and T. Due to the chemistry of the amino acids, A and T are opposites, as are C and G. When the strands have extended sequences of opposite bases, they are bound tightly together in the classic double helix structure.

RNA molecules look a lot like DNA molecules, but they contain uracil in place of thymine. Unlike DNA, RNA usually exists as a single strand in biological systems. The strand can bend to allow complementary sequences within it to bond together. Genes have practical impact on the body largely via causing messenger RNA to interact with transfer RNA, to create proteins. The protein created by a gene has a sequence of amino acids that is different from, but precisely determined by, the DNA strand's sequence of A, G, C and T.

The unraveling of this machinery – the nanomachinery underlying all life on Earth! – was one of the great achievements of 20th century science. But still, it's important to keep the wonders of molecular biology in perspective. While the central dogma, DNA to RNA to protein, outlines the core process on which the dynamics of biological systems are based, actual living organisms are made of thousands to millions of interacting instances of this pattern. Proteins activate the transcription of certain genes, whose protein products detect signals of threat and opportunity from the environment, alter the activity of other proteins in response, which metabolize raw materials to generate and maintain structure, and reproduce new cells to sustain and grow populations over time. Proteins and the metabolite flows they regulate carry out myriads of these functional activities simultaneously, which in turn regulate gene activation, in complex overlapping networks of feedback loops. Elucidating these networks requires not only an understanding of the environment in which the cell is embedded but the evolutionary processes that developed them incrementally over time.

The analogy between molecular biology and computer engineering is interesting, and has been pursued fairly far by a community of researchers. The mechanism of molecular biology amounts to a naturally

evolved nanotechnological computing and engineering infrastructure that we barely understand, and that far exceeds our own ability to build nanomachines or nanocomputers at this point. Inspired by this way of thinking, some folks have started working on “DNA Computing” – using the existing machinery of molecular biology to solve computational problems. At this point that's not a practical tool yet, but it's a rapidly evolving field, and if nothing else it will teach us a lot about nanotechnology. One day we may use DNA computers – or nanocomputers inspired by DNA computers – to run AI programs analyzing human genetics data!

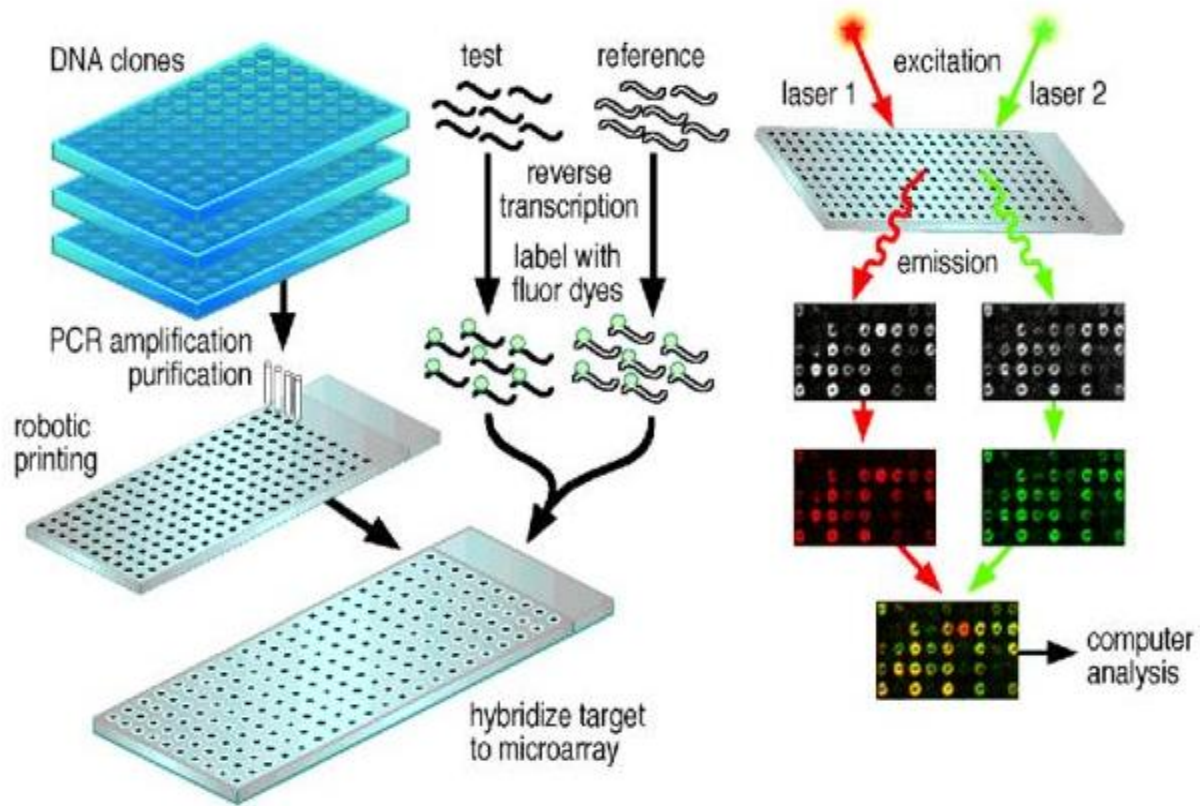


Figure 107: Diagrammatic depiction of the process of DNA microarray analysis. The tissue sample under study comes into the picture at the top middle, labeled “test”. AI tools like the ones my team at Biomind uses, come into play via analyzing the 2D grid data shown at the bottom right. Each point on the 2D grid corresponds to a certain gene, and the color shows how much of that gene is being expressed in the tissue sample. Roughly speaking a gene's expression level reflect the amount of the protein corresponding to the gene being produced, but it's not an exact reflection, as there's a lot of complexity both in the biological processes involved and the microarraying process itself. There are also many variants of the micro-arraying process available today, each with its own strengths and limitations. <http://bme240.eng.uci.edu/students/08s/jentel/image/MicroArrays.jpg>

Biomind's bioinformatic AI

So what kind of genetics data have we analyzed using our Biomind tools, and what kinds of answers have we obtained?

For example, one might have samples of DNA from 500 people with Alzheimers disease, and 500 similar people without the disease. Each DNA sample tells you about the individual variations in the 25000 genes that each person has (these variations are called SNPs, or Single Nucleotide Polymorphisms – generally pronounced “snips”). The AI then tries to figure out which combinations of variations in the genes, make a person more genetically susceptible to Alzheimers Disease.

Or, one might have samples of gene expression data from these same 500 Alzheimers patients and matched controls. In this case, one would have information about which genes are most active in each of the people under study, at the particular point in time when they were measured. This would give you a different kind of information: it would tell you about the perturbations to the body's biological processes that happen when a person has Alzheimers disease. The answer might depend on what tissue of the body the genetic material was gathered from.

In both the gene expression and the SNP cases, the main role of AI is to sift through the mass of data and find combinations of things that are relevant to a certain condition of the organism. If a single SNP, or a single gene, were critical to Alzheimers, biologists could find that without use of AI technology. But if a complex combination of SNPs or genes is important, it's hard for humans to figure that out from the data, even with the standard arsenal of statistical tools. Whereas AI tools like Biomind's, even though far short of AGI, can often find the relevant patterns.

```

sum
sub
  sum
  sub
    input GO:0016835
    input GO:0030674
  sub
    input FAM0031431
    input NM_001831
  sum
  mul
    input GO:0019840
    input NM_015358
  sub
    input GO:0003782
    input NM_002151
sum
sub
div
  const 0.514609
  input GO:0004667
input GO:0006803
div
div
  const 0.352189
  input NM_007313
sum
  input GO:0007165
  input NM_002226

```

Figure 110: Example of a rule that was learned by Biomind's AI, via analyzing biological data. This rule was learned from gene expression data from a population of humans with prostate cancer, and a population of matched control humans without prostate cancer. . The NM terms correspond to the expression levels of specific genes (NM_007313, for example, is the index number of a particular gene, the gene otherwise known as **c-abl oncogene 1, non-receptor tyrosine kinase (ABL1), transcript variant b**); the GO terms correspond to the average expression levels of genes in a certain category (e.g. GO:0007165 is the Gene Ontology category for signal transduction – the set of all genes concerned with signals that convey changes to the state or activity of cells). The rule is shown in the “tree” format in which it is actually learned within the software. The software evaluates if this formula is greater than 0, and if so, it evaluates that the individual probably has prostate cancer. The tree is evaluated by interpreting it as It could also be written as a mathematical formula

$$\left(\left(\left(\left(GO:0016835 - GO:0030674 \right) + \left(FAM0031431 - NM_001831 \right) \right) - \left(\left(GO:0019840 * NM_015358 \right) + \left(GO:0003782 - NM_002151 \right) \right) \right) + \left((.514609 / GO:0004667 - GO:0006803) + \left((.352189 / NM_007313) / (GO:0007165 + NM_002226) \right) \right) \right)$$

The astute reader will notice that is mathematical rule could be simplified considerably. This is because it was learned by the OpenBiomind software from 2006 or so. For our current bioinformatics work,

we use OpenCog's MOSES software, which incorporates simplification of rules and formulas. So rules learned by MOSES will tend to be smaller, because they appear in something closer to algebraically minimal form. This is better for human comprehension, and also better if one wants to have other automated reasoning systems act on the rules, e.g. trying to generalize them to other contexts.

Feature	Utility	Differentiation ran	Description
NM_004039	0.667	27	Homo sapiens annexin A2 (ANXA2), mRNA.
GO:0006957	0.427	1853	complement activation, alternative pathway
FAM0005641	0.399	1	
GO:0016860	0.393	1640	intramolecular oxidoreductase activity
GO:0003817	0.393	2	complement factor D activity
GO:0030162	0.392	524	regulation of proteolysis and peptidolysis
NM_002151	0.391	1	Homo sapiens hepsin (transmembrane protease, serine 1) (HPN), transcript variant 2, mRNA.
NM_000954	0.388	3822	Homo sapiens prostaglandin D2 synthase 21kDa (brain) (PTGDS), mRNA.
GO:0016812	0.387	1363	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amides
GO:0006956	0.386	2102	complement activation
NM_002156	0.386	6	Homo sapiens heat shock 60kDa protein 1 (chaperonin) (HSPD1), nuclear gene encoding mitoc
1664_at	0.384	4	
GO:0045187	0.382	3	regulation of circadian sleep/wake cycle, sleep
GO:0050802	0.382	3	circadian sleep/wake cycle, sleep
GO:0005791	0.381	245	rough endoplasmic reticulum
FAM0024314	0.376	3	
NM_001928	0.375	2	Homo sapiens D component of complement (adipsin) (DF), mRNA.
GO:0042749	0.374	3	regulation of circadian sleep/wake cycle
GO:0030072	0.373	928	peptide hormone secretion
SF002514	0.373	1484	lipocalin
GO:0030252	0.372	329	growth hormone secretion
SF001139	0.371	1	hepsin
GO:0004667	0.37	853	prostaglandin-D synthase activity
XR_000167	0.37	10	
GO:0019840	0.37	1484	isoprenoid binding
GO:0017015	0.367	329	regulation of transforming growth factor beta receptor signaling pathway
NM_003573	0.366	5	Homo sapiens latent transforming growth factor beta binding protein 4 (LTBP4), mRNA.

Figure 111: Example of the results produced by the OpenBiomind software for genetics data analysis. This is a very simple form of results: just a list of which genes, Gene Ontology categories and protein families are most useful for distinguishing prostate cancer samples from control samples, based on the AI's analysis. Some of these are already known to be involved with prostate cancer or cancers in general, others are not. A list like this provides biologists with multiple avenues for further investigation. The second column indicates how useful the feature (gene or category) was found by the AI software; the second column indicates how useful the same feature would be found according to a standard statistical analysis (in the second column, large numbers mean less useful). The contrast between the two columns indicates that the AI is prioritizing things very differently from standard statistical analysis. Annexin, for example, is the #1 most relevant gene found by the AI, yet would be ranked only #27 by standard statistics. The Gene Ontology category for complement activation is ranked #2 in importance by the AI, yet would be ranked #1853 (i.e. not be highlighted at all) by standard statistics. In many cases, the genes and categories highlighted by the AI have been validated by further lab experimentation and found to be relevant to the phenomenon under study. This is because the AI can choose genes or categories based on their importance in the context of their interactions with other genes, whereas standard statistical methods tend to ignore interactions and just look at the relation between individual genes and “phenotypic” characteristics like prostate cancer.

Cluster 1:	
Quality: 0.620161565619	
FAM0012352	Component genes:
	- NM_006216: Homo sapiens serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator
FAM0033509	Component genes:
	- NM_004137: Homo sapiens potassium large conductance calcium-activated channel, subfamily M, beta me
GO:0015023	syndecan
NM_014974	Homo sapiens KIAA0934 (KIAA0934), mRNA.
GO:0006693	prostaglandin metabolism
GO:0009103	lipopolysaccharide biosynthesis
GO:0015629	actin cytoskeleton
FAM0020295	Component genes:
	- NM_002997: Homo sapiens syndecan 1 (SDC1), mRNA.
Cluster 2:	
Quality: 0.614835196577	
1280_i_at	---
SF027815	Not Yet Assigned
GO:0005779	integral to peroxisomal membrane
NM_002101	Homo sapiens glycophorin C (Gerbich blood group) (GYPC), transcript variant 1, mRNA.
GO:0007397	histogenesis and organogenesis
GO:0008653	lipopolysaccharide metabolism
GO:0015268	alpha-type channel activity
Cluster 3:	
Quality: 0.607104033144	
GO:0004718	Janus kinase activity
SF002282	cytoskeletal keratin
GO:0030334	regulation of cell migration
GO:0010035	response to inorganic substance
NM_021638	Homo sapiens actin filament associated protein (AFAP), transcript variant 1, mRNA.
SF002345	smooth muscle protein SM22

Figure 112: Biomind software also provides information on which genes and gene categories tend to interact with each other (form a “behavioral cluster”) in the context of a given phenotypic character (prostate cancer, in this case). Again, this provides information that biologists can follow up on according to their own scientific understanding – progress, at this stage, requires a mixture of human and artificial intelligence

One of our most exciting achievements was devising new diagnostic tests to tell if a person has Parkinson’s or Alzheimer’s disease. Our findings also highlighted certain avenues with the potential to lead to cures for both – though this remains work in progress. Before our work, it was known that dysfunctions in the mitochondria (the part of each cell that stores energy) were important in Alzheimer’s and Parkinson’s. But nobody knew the causal basis of the connection between mitochondria and these diseases. Though it was suspected to have something to do with the mitochondrial DNA (a handful of important genes that live in the mitochondria of the cell, rather than in the cell nucleus where the other 25000 or so genes live).

Gene	Position	Percentage in top 20 codons
ND5	145	30.89 %
ND4	180	9.84 %
ND5	146	6.76 %
ND5	148	5.44 %
ND2	272	5.11 %
ND4	236	4.68 %
ND2	273	4.26 %
ND2	270	4.16 %
ND4L	49	3.17 %
ND4	183	3.12 %

Figure 113: Mitochondrial genes found relevant to Parkinson's Disease, via Biomind's AI analysis of mitochondrial DNA data provided by Rafal Smigrodzki and W. Davis Parker at U. Virginia. This data was an unusual kind: it recorded uncommon (“heteroplasmic”) mutations in 10,000 mitochondrial DNA drawn from each of 17 peoples' cerebrospinal fluid. The ND5 gene emerges clearly from this analysis as the culprit (the “codons” mentioned in the final column are data structures in one of Biomind's algorithms, representing patterns found in the data; more than 30% of these patterns involved the mitochondrial gene ND5). The analytical work here was mostly done by Lucio Coelho in Biomind's office in Belo Horizonte, Brazil.

Our Biomind AI technology, applied to data about mitochondrial DNA gathered by Rafal Smigrodzki and W. Davis Parker at the University of Virginia, pinpointed the exact portions of mitochondrial DNA that correspond to Parkinsons and Alzheimers. Standard biostatistical tools failed to determine these portions, in spite of being supplied with the exact same data – because they are not good at finding patterns involving combinations of different factors (in this case, combinations of different regions of the mitochondrial genome). But now, thanks to the AI analysis, we know which parts of which mitochondrial genes are screwed up in people with these neurodegenerative diseases. This discovery does not cure the diseases, nor even explain them. But it gives biologists a place to look... It gives them new questions to ask. What causes these particular mitochondrial dysfunctions, and what other problems do they cause? Biologists are currently pursuing these important questions, which were, in effect, posed by our AI.

Table 4. Importances of the genes based on their SNPs.		
Gene	Short description	Incidence
<i>NR3C1</i>	Nuclear receptor subfamily 3, group C, member 1 glucocorticoid receptor	69054
<i>TPH2</i>	Neuronal tryptophan hydroxylase	67077
<i>COMT</i>	Catechol-O-methyltransferase	60651
<i>CRHR2</i>	Corticotropin-releasing factor 2	42190
<i>CRHR1</i>	Corticotropin-releasing hormone receptor 1	35082
<i>NRC1</i>	Nonpapillary renal carcinoma 1 growth mediator	22531
<i>TH</i>	Tyrosine hydroxylase	15968
<i>POMC</i>	Proopiomelanocortin	13145
<i>5HTT</i>	5-hydroxytryptamine transporter	6706

The incidence number reports the number of rules found with accuracy greater than the frequency of the largest category, utilizing some SNPs in the gene.
SNP: Single nucleotide polymorphism.

Figure 114: Genes found relevant to Chronic Fatigue Syndrome, based on Biomind's AI analysis of SNP data provided by Suzanne Vernon's group at the CDC.

We also found, in a collaboration with Suzanne Vernon's team at the Center for Disease Control and Prevention in Atlanta, the first solid evidence of a genetic basis for Chronic Fatigue Syndrome, a condition long disparaged as laziness or simple exhaustion. We compared the DNA of sufferers with non-afflicted people and uncovered systematic variations indicative of genetic differences between the two groups. Human researchers, even those armed with statistical tools, could never have duplicated this feat without the aid of our AI systems. But once the AI found its results, they could be easily followed up by researchers in the lab. For instance, the AI highlighted genes related to glucocorticoid receptors in the brain, and to the neurotransmitter tryptophan. Once the AI pointed these out to biologists, they could use their human intelligence to do various experiments, and gain ongoing scientific understanding they would not have achieved without the AI's guidance. Lacking an AGI to mastermind the whole scientific process, what we have are narrow AI tools that, when properly utilized, analyze the data from experiments that humans have designed, and provide guidance (such as lists of relevant genes or biological processes) to these same humans in designing the next round of experiments.

On the life extension side, one of my most fascinating projects involves long-lived flies. Put simply: We have these flies that live 4 times as long as regular flies, and we're studying their genetics to see why. They were created by experimental evolution – via breeding flies for longevity over a 30 year period – a method that requires and yields no understanding of WHY or HOW they live so long. On this project, Biomind has been with another company called Genescient, which owns the flies. Thousands of genes are different in the long-lived flies from the normal ones, but we've used AI

technology to narrow the scope to a few dozen that seem to be the most important. And we seem to be grasping some of the key processes underlying these super-flies' longevity.



Figure 115: Genescient's “Methuselah flies” are not only longer-lived but more robust than their ordinary counterparts in many ways. They have stronger hearts; they are smarter (better at remembering where to find food); and they have more sex.

For instance, there are a lot of differences in neural and developmental genes, including many genes related to brain development. It seems part of the story of aging has to do with brain development processes that start out helpful, but then keep on going and turn damaging once the brain gets older. Genescient's biologists are using the results of the AI data analysis to formulate new drugs and nutritional supplements to help combat aging.

As an example of the way the AI analysis of this sort of data can help with the development of therapeutics, suppose the software found three genes, so that looking at SNPs (individualized variations) in these genes allows the AI to tell whether a given fly is Methuselah or Control (long-lived or just normal). So the AI has found a triple of genes

Gene1, Gene2, Gene3

Each of these genes will have its own story, e.g. maybe

- Gene 1 is a serotonin receptor, expressed less in Methuselahs than controls
- Gene 2 is related to Golgi apparatus, anatomical structure morphogenesis; also underexpressed in Methuselahs
- Gene 3 is a “homeobox” gene, related to central nervous system formation; expressed more in Methuselahs than controls

It can be validated, via use of various biology databases, whether these genes behave sufficiently

similarly in humans and flies, for the gene combination to be worth following up in the context of human therapeutics. (Making flies live a long time, in ways that don't generalize to humans, is also a worthy ethical cause, but not the major focus of Genescient or Biomind at present.) The question then becomes: Can we hit this triple of genes effectively with a combinational therapeutic? Can we find some combination of drugs or herbs so that, combined, the active ingredients act on the proteins coded by these three genes in a beneficial way?

The genetics analysis spawns a problem in therapeutics development, which at this point the AI software doesn't help with, other than to suggest which of the body's proteins the therapeutics should target. Genescient has, at time of writing, used Biomind's AI analysis results to uncover a number of herbal remedies for age-associated diseases, including the Stem Cell 100 supplement. The firm has also identified a number of promising possibilities for pharmaceutical (drug) remedies for Alzheimers and other diseases, based on the same AI results.



Figure 116: The Stem Cell 100 supplement uses a formulation of herbs, which was discovered in large part based on the results of applying Biomind's AI analysis to genetic data from the long-lived Methuselah flies.

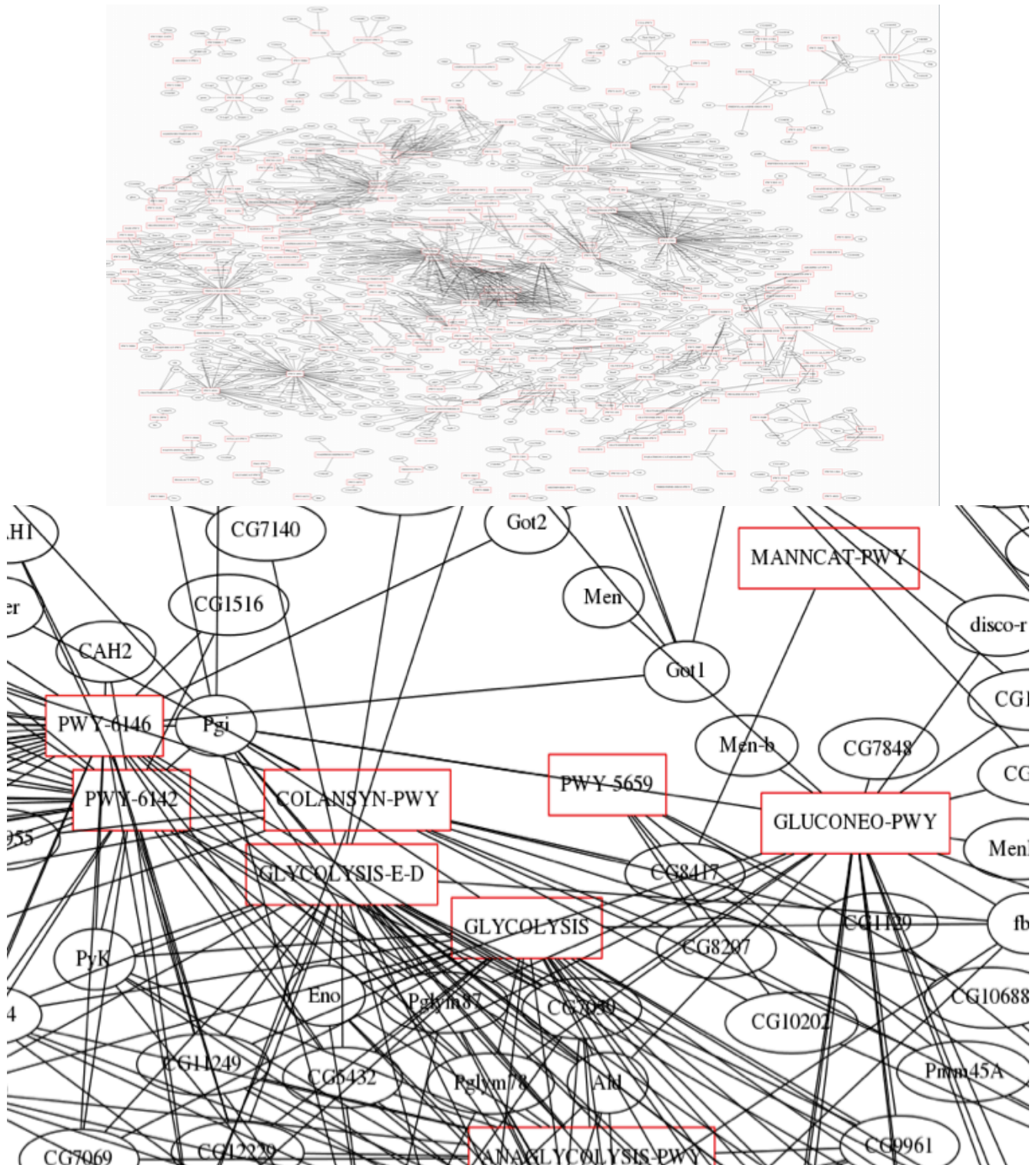


Figure 117: Biologists love graphical depictions of data – but the diagrams they look at generally represent massive oversimplifications of the real processes and interactions going on. If one tries to make a diagram of the main interactions between genes occurring in Methuselah flies relative to their longevity, by simply feeding one's data into a scientific data visualization tool – one gets a huge tangle! The top picture can't really be read on a book page, but could be printed as a wall-sized poster: each dot represents a gene or “pathway” (subnetwork of interacting proteins), and the lines between dots represent interactions. The dark regions are where there are a lot of lines, i.e. a lot of interactions. The bottom picture shows a random square from within the top picture, blown up a bit. Of course, there are commercial products that make such diagrams easily browsable by the researcher – but they do so, in part, by making choices regarding which relationships to depict and which ones to ignore. The tangle shown here, while not so useful for understanding what's going on, does evoke the extreme complexity of the processes and interactions going on here! Obviously, we found other ways to explore the Methuselah fly data!

I've also recently studied another really interesting dataset related to longevity –the Welllderly data from the Scripps Institute in California. This is data on hundreds of healthy people over age 80, and an equivalent number of less healthy counterparts. Again, we're using AI to probe the genetic differences—of which there are many. And we're correlating the data from long-lived flies with this data from healthy long-lived people. The idea is, if a certain gene or pathway is important for longevity in both flies and people, then it's probably really important.

We haven't found many particular genes that are important for longevity in both humans and flies. But we have found many common pathways – that is, many common biological networks and processes, that are centrally involved in aging in both of these very different organisms. My colleagues at Genescent are currently following up these pathways in the lab, trying to use them to design therapies to extend life and combat age-associated disease.

All this is important and exciting. But it's worth bearing in mind that the way we apply AI to biology now is fairly limited – basically we feed in one dataset, or a small number of datasets, and see what the AI says. If you had an AGI with really advanced general intelligence, you could do better – you could feed it every dataset on the planet.

And even without advanced AGI, narrow AI technology has great potential to advance the cause of healthy longevity, and at the same time move AI toward AGI.

My colleagues and I have been applying AI to biological data on a piecemeal basis, one day at a time. That's all we can do with the limited resources at our disposal. With more resources for this kind of work, we could integrate all available biological data into one big holistic knowledge database and then set our AI algorithms to work. Even without any more biological experiments, I'm confident we could uncover an incredible amount of new information that would lead to the designs of new experiments useful for gathering yet more data.

And of course, many of these new experiments could become fully automated with robotic lab equipment. We have the technology now to take the totality of biological data and store it in a huge AI system, utilizing this system to design and run new experiments using robotic laboratory equipment, repeating the process multiple times to get ever more useful (and otherwise unobtainable) results with enormously positive implications for biology. The only thing stopping us is money. With enough resources at hand, I bet we could crack the aging problem within a decade or two. Middle-aged people today would never need fear death, except through accidents or rare diseases.

Aubrey de Grey's SENS Approach to Increasing Human Lifespan

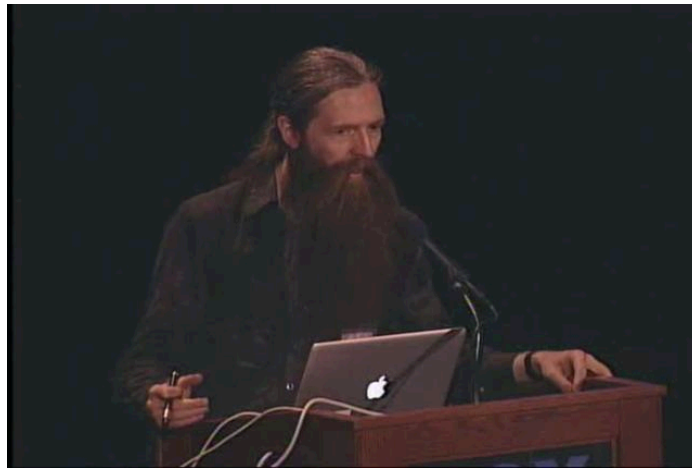


Figure 118: Aubrey de Grey, the maverick researcher who has emerged during the last decade as the leading advocate of a scientific approach to ending aging.

The AI-driven approach, promising as it is, is not the only current attempt to come to grips with the aging problem. The mainstream of biomedicine doesn't think much about radical lifespan extension yet, but some maverick researchers do. The most influential among these so far has been Aubrey de Grey, whose SENS (Strategies for Engineering Negligible Senescence) research organization is carrying out some fascinating longevity research in their California lab, and funding a host of valuable projects at various universities worldwide.

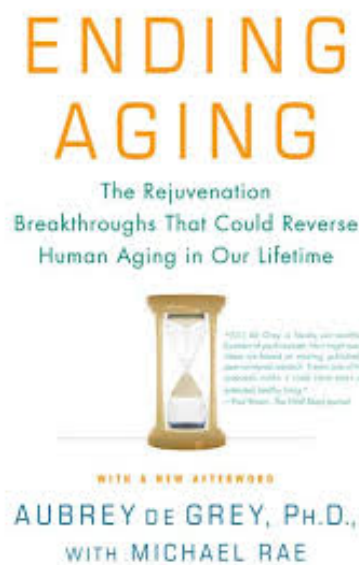


Figure 119: Aubrey's book Ending Aging carefully reviews his approach to curing the problems of aging and enabling radical human longevity.

Aubrey's approach to achieving radical human longevity is “engineering” oriented – or you could

almost think of it as “auto mechanics” oriented. He wants to treat an old body vaguely like an old car – identify what the problems are, and fix them. He argues that biologists have noted seven main things that go wrong with human bodies as they age – the Seven Pillars of Aging as he calls them. And he proposes possible fixes for each of these.

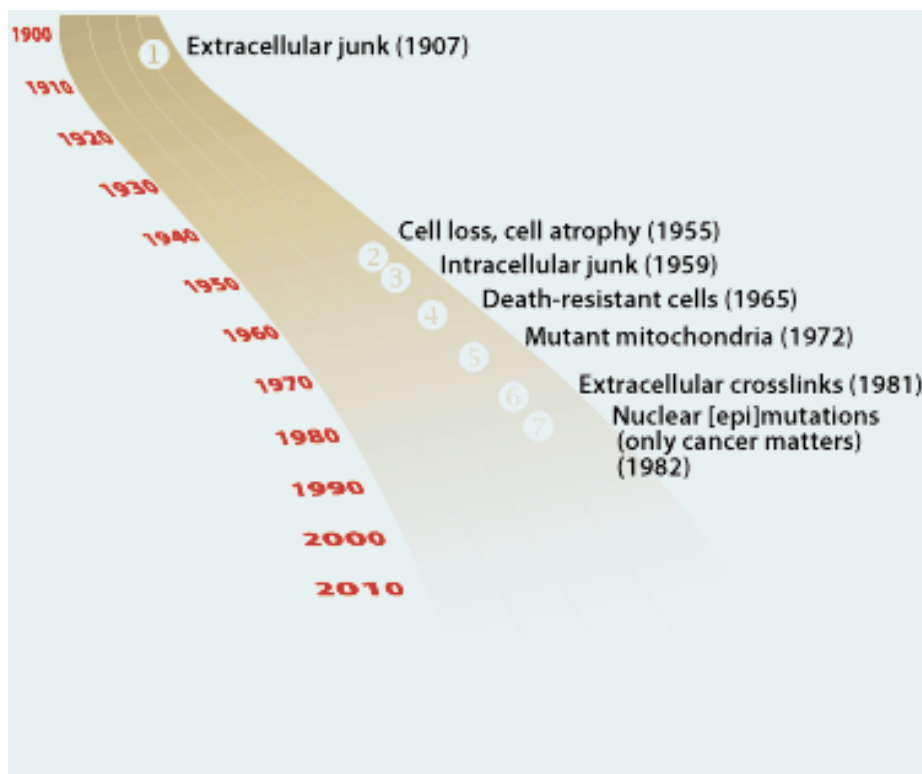


Figure 120: The Seven Pillars of Aging identified by Aubrey de Grey. According to Aubrey’s hypothesis, if we can resolve these seven problems, we can make human beings live a very long time. <http://alfin2600.blogspot.hk>

Aging Damage	Discovery	SENS Solution
Cell loss, tissue atrophy	1955 ¹	Stem cells and tissue engineering (RepleniSENS)
Nuclear [epi]mutations (only cancer matters)	1959 ² , 1982 ³	Removal of telomere-lengthening machinery (OncoSENS)
Mutant mitochondria	1972 ⁴	Allotopic expression of 13 proteins (MitoSENS)
Death-resistant cells	1965 ⁵	Targeted ablation (ApoptoSENS)
Tissue stiffening	1958 ⁶ , 1981 ⁷	AGE-breaking molecules (GlycoSENS); tissue engineering
Extracellular aggregates	1907 ⁸	Immunotherapeutic clearance (AmyloSENS)
Intracellular aggregates	1959 ⁹	Novel lysosomal hydrolases (LysoSENS)

The solutions Aubrey has in mind for resolving the Seven Pillars of Aging. Each of these is a difficult

research project with many different aspects. See <http://sens.org> for explanations of all these potential solutions – I don't want to go into them in detail here, as it would be too much of a digression.

The point of Aubrey's wherein each researcher who produces the longest-lived mouse ever or the best-ever mouse-lifespan rejuvenation therapy receives a bit of money each week until his record is broken.

Aubrey suspects that, within the next decade or two, it should be possible to come pretty close to defeating senescence within mice — if the research community puts enough focus on the area. Then, porting the results from mouse to human shouldn't take all that much longer (biological research is regularly ported from mice to humans, as they are an unusually suitable testbed for human therapies—though obviously far from a perfect match). Of course, some techniques will port more easily than others, and unforeseen difficulties may arise. However, if we manage to extend human lives by 30 or 40 years via partly solving the problem of aging, then we'll have 30 or 40 extra years in which to help biologists solve the other problems.

Aubrey likes to talk about the ***Methuselarity*** – the date at which, if you're still alive, the odds are high that ongoing advances in biomedicine will allow you to avoid involuntary death altogether.

Theory-wise, Aubrey agrees with me that aging is complex and probably due to a host of different causes. He doesn't believe there's one grand root cause of senescence, but rather that it's the result of a whole bunch of different things going wrong, mainly because human DNA did not evolve in such a way as to make them not go wrong. But he doesn't think we need AGI to figure out how to solve all the problems – he figures plain old human ingenuity can do the trick. I hope he's right! But I figure we may as well approach the problem from all viable angles. I'm glad Aubrey's team and others are working on the longevity problem via innovative biology methods ... and I've been spending a fraction of my time attacking existing biological data using narrow AI. But I also think it's key to work on building better minds, that can integrate all the available data in ways inaccessible to the human mind, and likely propose new experiments and therapies that humans would never think of.

term SENS is that it's not merely slowing down of aging that we're after—it's the reduction of senescence to a negligible level. We're not trying to achieve this goal via voodoo, we're trying to achieve it via engineering—mostly biological engineering, though nano-engineering is also a possibility.

As part of his effort to energize the biology research community about SENS, a number of years ago Aubrey launched a contest called the “Methuselah mouse prize”—a prize that yields money to the

researcher that produces the longest-lived mouse of species *Mus musculus*. In fact there are two sub-prizes: one for longevity, and a “rejuvenation” prize, given to the best life-extension therapy that’s applicable to an already partially-aged mouse. There is a complicated prize structure, Toward an AGI Biomedical Researcher

While working on an early version of this chapter, I happened to get a call from a friend in California, asking me to put together a proposal for the use of OpenCog AGI and other AI tools to cure aging and work toward radical longevity. My friend wanted to show the proposal to a wealthy business contact— but only if I could find a way to use OpenCog to work toward radical longevity that would also generate useful products along the way.

While these sorts of pitches usually don’t come to anything, or at least take a long time to yield fruit, I figured it was an interesting challenge. My plan involved the creation of an AGI biomedical researcher, but to move toward this long-term goal through two earlier stages: First, an AGI that would chat with you about your medical issues. Then, an AGI biomedical research assistant that would help a scientist in their work based both on knowledge it absorbed from biomedical databases and using statistical and narrow-AI tools to analyze biological datasets. Maybe not as smooth a path toward human-level AGI as virtual world agents or robotics – but I think it’s one that could synergize wonderfully if pursued in parallel with the virtual agents and robotics approaches.

Many of the ingredients needed to realize the vision of a biomedical AGI researcher already exist:

8. Massive (albeit heterogeneous and messy) biological databases
9. Narrow- AI tools for analyzing biological data and extracting information from biological research papers and abstracts
10. AI reasoning systems for drawing conclusions from biomedical knowledge (we used an old version of PLN for this back in 2005)
11. Robotized lab equipment allowing AI systems to run their own experiments without human intervention.

But these ingredients are scattered about: Nobody is using them in a unified way. Why not combine them into a coherent, powerful biomedical artificial general intelligence? This would be a fantastic project!

Coming up with a proposal for my friend’s request presented a challenge: In order to drive the

integration of all these ingredients, and the evolution of the integrated system into a powerful biomedical AGI, I needed to come up with some practical application with both humanistic and economic value at the early stages of development, and then increasing value as ongoing development advanced its intelligence. One suitable application was a *dialogue system*—a software system that uses ordinary English to chat with people about medical and biological issues.

A version-one biomedical dialogue system could be something fairly simple: A *digital diagnostician and advisor*, helping people assess their symptoms and helping them fish through the complex online biomedical literature for relevant information. Given the high cost and low quality of most medical care available, and the disorganization of in-depth online medical information, this would be a major public service. And it could also be a significant revenue driver for a variety of appropriate businesses (like a website selling non-prescription therapies).

With ongoing work on the back end, the digital diagnostician and advisor could evolve into a *AGI biomedical research assistant*, with the capability to help scientists in their work. This would allow the rapid discovery of biomedical truths that would take far longer for humans to discover on their own. It could also promote industry growth (e.g. the founding of an AGI/biopharma firm in which AGI systems and human scientists combine to rapidly conceive novel therapies and diagnostics for a variety of diseases).

Finally, with more time and effort, the biomedical AGI research assistant could evolve to the next stage: an *AGI biomedical scientist*, able to make important discoveries on its own without human aid—carrying out the whole process of scientific invention, discovery, validation and technology transfer on its own. Here, we might expect truly radical, impactful therapies combating death and all forms of disease.

In my proposal, I laid out the incremental progress along this path to AGI as follows:

Year 1:

- 3) *Digital diagnostician and advisor*
 - 4) *Bio-NLP system for extracting information from biomedical research abstracts*
 - 5) *Machine learning system enabling cross-dataset extraction of patterns from dozens of datasets simultaneously*
- Year 2:***

- 3) Beta testing and **launch of digital diagnostician and advisor**
- 4) Dialogue system enabling conversation about information extracted from research abstracts
- 5) AI reasoning system drawing novel conclusions based on combining information from databases and research abstracts
- 6) Scaling-up of machine learning data analysis system to 100s of simultaneous datasets

Year 3:

- Semi-automated identification of metadata in datasets, speeding ingestion of datasets into the database
- Automated extraction of information from tables and figures in research papers
- Extraction of information from full text of biomedical research papers
- Automatic triggering of machine learning data analysis tasks based on information extracted from research papers
- Dialogue system enabling control of machine learning data analysis via interactive dialogue

Year 4:

- Beta testing of AGI biomedical research assistant
- Automatic generation of new “research abstracts” describing what the AGI has learned
- Use of automated inference to specify new experiments to run

Year 5:

6. Commercial **launch of AGI biomedical research assistant**
7. Connection of AGI research assistant directly to appropriate lab equipment for experimentation with a completely automated AI-run lab for experimentation on micro-organisms (probably yeast)
8. Customization of natural language processing engine enabling it to understand simple language in scientific textbooks, and to generate more flexible biomedical language as well

Year 6:

- *Testing of AGI biomedical research assistant's findings to study longevity biology and other specifically identified topics, in conjunction with human researchers*
- *Customization of AGI inference engine to integrate knowledge extracted from natural language textbooks with information ingested from datasets and structured databases*
- *Customization of natural language processing engine to enable it to understand more complex language in scientific textbooks*
- *Enabling of system to generate brief research reports as well as abstracts*

Year 7:

- *Beta testing of AGI biomedical research assistant Version 2 (capable of discussion and reasoning based on information learned from textbooks)*
- *Ongoing improvement of reasoning and language processing technology to enable greater intelligence*

Year 8:

- **Launch of AGI biomedical research assistant Version 2**

Year 10:

Launch of AGI biomedical scientist *with capability to:*

- *Read and interpret research literature and discuss its conclusions,*
- *Produce research abstracts and reports summarizing its conclusions*
- *Suggest detailed experiments to be run in vitro and in vivo (regarding multiple organisms)*
- *Conduct yeast genetics and potentially other microbiology experiments on its own via direct connection to lab equipment*

Amazing!!! Now, imagine doing all that in parallel with OpenCog, controlling millions of video game characters interacting with people all over the world, and using it to control humanoid robots exploring everyday human environments. The synergies between these different OpenCog projects would be tremendous, and the collective effort would speed us toward human-level general intelligence – and better yet, toward *beneficial* AGI, oriented fundamentally toward goals of helping people to live longer and better.

Just as I expected, the wealthy individual to whom this plan was pitched didn't bite. Ah well. This sort of thing will happen one day, and probably within my lifetime, whether utilizing OpenCog and Biomind or some other AGI and bioinformatics technologies.

Cryonics as a Backup Plan



Figure 121: This is me visiting Alcor’s Arizona cryonics facility in 2012. Each of the silvery tanks behind me contains a “corpsicle” (not their official term), i.e. a human being preserved in liquid nitrogen, for hopeful eventual reanimation by some technological means or another. If things don’t go well enough for me, I may end up in one of those one day. On the other hand, if I die via falling off a rock-face in a remote region or a plane crash, my brain may be nonrecoverable, and future minds may end up reconstituting me via subtler methods instead, using data such as this book and all the videos of me speaking and moving that have been posted online.

OK, so death is a solvable problem—and the human race will probably solve it fairly soon. At least, “fairly soon” on the time-scale of human history. But even if it does get solved, will it really get solved soon enough for you or me?

We could accelerate the pace toward solving aging with a team of AGI biomedical scientists based on OpenCog or some other platform. Or we could fund other approaches like Aubrey de Grey’s SENS project. Better yet, we could have both SENS and the AGI biomedical researcher! If the human race—or any one of the planet’s super-rich individuals—took life extension seriously, both of these projects would be massively funded.

But still – we all know how hard it is to get innovative technology projects funded. So, one has to consider the pessimistic possibility: What if the aging problem doesn’t get solved in time for you and me personally? After all, I’m 46 years old in 2013 as I write these words, so by the time of Kurzweil’s

conjectured Singularity in 2045 I'll be a rather old man, at least by current standards. Are there any reasonable backup plans—short of going to another galaxy and drinking the alien's immortality potions—aside from just rolling over and dying?

There's one backup plan that seems reasonably likely to succeed: Cryonics. We now have the technology to freeze a body in liquid nitrogen, using special cryoprotectant chemicals to avoid damage during freezing, in a way that seems to preserve all the important structures of the brain and body. Now, we don't yet know how to defrost these frozen bodies without causing damage. But surely, the superhuman AIs and cyborgs after the Singularity will be able to figure that out! So, one backup plan right now is to have yourself frozen in liquid nitrogen right after you die ... and then get defrosted later on, once the technology is available. Obviously this isn't a guarantee – a lot of things could go wrong. But as I like to say: **Better frozen than rotten!**

A number of organizations offer this kind of cryonics service—I'm signed up with one of them, Alcor, which is based in Arizona. So if I die, my body will get shipped to Arizona and preserved in liquid nitrogen... And hopefully my descendants will defrost me, then give me a new body (a robot body perhaps). Around 100 people are already frozen in Arizona, waiting for the technology to reach the point where it's possible to revive them.

But that's just a backup plan, of course. ***Better frozen than rotten—but better living than frozen!***

Why So Little Focus on Longevity?

When I first found out about death as a little child, I was baffled at how unperturbed the adults around me seemed by this rather terrible fact of human existence. I understand people a lot better now, but in the end I'm still fairly baffled. Getting old and dying is a Very Bad Thing, but nobody seems to worry about it. We spend a lot of money on bombs for blowing people up, and cures for diseases like cancer and AIDS—but very little money on the core problem of eliminating aging itself. And we spend a lot of money on computer programs for video games, Web search and supply chain management, but very little on developing smart AI software to help us figure out the complex problems of aging and longevity. As individuals, we hate it when our friends and family get old and die—but as a society, we don't seem to care much to do anything about it.

I'm especially vexed by the lack of interest in longevity research coming from wealthy individuals. Over a thousand billionaires and tens of thousands of people with an individual net worth of \$100

million or more are living now. Why don't more of them devote 10-20 percent of their wealth to the creation of longevity drugs that will let them enjoy their riches over an extended lifespan? I suppose it's because they just don't believe it's possible—they don't think that life extension is a fruitful line of inquiry. But this attitude is wrong. For the first time in human history, we're at a point where death is no longer inevitable. The cure is in sight. There's a lot of work to be done to cure death, but it's most likely going to happen this century—maybe even soon enough for you and me.

If you agree with me that death is a terrible thing, maybe you should think about joining the fight—donate to organizations combating aging, or even contribute to the research yourself. Better yet, contribute to AGI, which with sufficient funding and intellectual attention, may soon be able to solve the human aging problem better than us mere humans!

PART FOUR

THE SCOPE OF FUTURE AGI MINDS

(Cyborg Immortals, Global Brains, Femtotech Superminds, and Beyond)

In this fourth and final Part of the book, we'll jump back to the future, and look in detail at some of the out-there yet scientifically plausible possibilities that may come to pass after human-level Artificial General Intelligence is achieved.

The concepts I'll describe are exciting and amazing – and sometimes frightening, outlandish or confusing – and well worth understanding, because they may well be our future

The Risks and Rewards of AGI

Not long ago, I gave a lecture on AGI and the Singularity, via Skype, to a lecture hall full of Ethiopians. Via the wonders of the Internet, I spoke to them from my house in Ting Kok Village (by the seaside in a rural area of the New Territories in Hong Kong); they listened from their university in Addis Ababa, Ethiopia's capital. Before and since that lecture, I've been in contact with scientists and programmers in Addis, and in early 2013 I helped them set up the country's first artificial intelligence research facility, the Addis AI Lab, which is now collaborating on OpenCog R&D. Exciting stuff!



Figure 122: Me with Getnet Aseffa, the leader of the Addis AI Lab where a team of Ethiopian programmers carry out OpenCog and other AI work, in close collaboration with the OpenCog team in Hong Kong. We're standing on the hills overlooking Addis Ababa. The office-buildings in downtown Addis are quite modern and filled with all sorts of companies, including an emerging tech scene. But in the hills around the city, people are subsistence farming much as they have been for thousands of years.

After my lecture, the audience asked some questions— and lo and behold, what was the first question? You guessed it! I don't recall the precise wording, but it was a variant on "After you've created these superhumanly intelligent AGIs, aren't they just going to kill or enslave us all and take over the Earth?"

My mental reaction was something like: *Wow, it really doesn't matter what part of the planet you're*

on, human beings all think the same way. The “Terminator” question is the #1 question whenever I’m talking about AGI to almost any audience (“How can I use AGI to make a lot of money?” being a distant #2)—the only exception being professional AGI researchers.



Figure 123: Arnold Schwarzenegger's time-traveling killer robot in the Terminator movies, has become the paradigm case of a super-powerful AI out to murder people and destroy human civilization. While the potential risks of AGI or any advanced technology must be seriously considered, it's also a risk to take science fiction as a close guide to reality. SF authors and filmmakers are guided by what makes a good story, and the complexity and subtlety of the real world often doesn't qualify. Robot scientists and physicians aren't as dramatic as robot warriors, but may end up playing a far larger role in the future. <http://terminator.wikia.com/wiki/File:Governator.jpg>

And of course, it’s a natural question—not only because of the prevalence of “Evil super-AI” in SF films, but for more basic reasons. We humans like being the top dogs on the planet (pun intended). The advent of something more powerful and intelligent than us is inevitably a bit scary.

The odds of a super-powerful AGI desiring to enslave people seem very low to me — I reckon that once an AGI is powerful enough to do this, it will figure out better ways to do the work suited for humans. Science fiction loves this kind of scenario—the future humans in the *Terminator* series,

forced to carry out manual labor for robot overlords; the comatose humans in *The Matrix* used as biological batteries, etc. But the requirements of storytelling are quite different from the dynamics of reality. Surely an AI smart enough to send Schwarzenegger back in time would be smart enough to build robots better at doing hard labor than humans; and surely any mind capable of building the Matrix could also build a better battery than the human body.

But what about the risks posed by a military-created AGI? Actually, I think it's very unlikely a military AGI would go in the direction of intentionally wiping out or tormenting humanity. So far, the handling of nuclear weapons technology by the world's leading military nations supports this idea. With the exception of China—itself an extremely insular nation without any history of military action beyond its border nations—the world's major military powers are democracies, and the incidence of wars between democracies throughout history is rather low. And many political commentators expect China to democratize sometime in the next few decades.

A powerful AGI created by the armed forces of a major nation might be ruthless in pursuing the goals of that nation, but it appears unlikely that it would then spontaneously mutate its goals in such a way as to make it want to kill everyone and become the dictator. A human sometimes does this when given too much power, but AGIs aren't necessarily going to share human traits such as megalomania and sadism unless they're taught these values or programmed that way.

More frightening than AGI-powered national robot armies is the possibility of terrorist organizations getting a hold of advanced AGIs. This sort of possibility is genuinely scary, and exists in the context of many other advanced technologies as well—synthetic biology, nanotech, and so forth. What would happen if a couple hundred disenchanted top-notch technology geeks decided to join forces with a group like Al Qaeda? How much destruction would ensue, even without AGI? What would happen if human-level AGI were added to the mix?

And what if it took just one person? An AGI programmed by a misanthropic sociopath is certainly a possibility—i.e. a scenario where an evil mad scientist intentionally programs an AGI to wipe everybody out. However, I think this is also quite unlikely, not because such sociopaths don't exist, but rather because they probably won't be the first to create an advanced AGI. Fortunately, outside the movies, extreme sociopathology and extreme scientific brilliance tend not to frequently co-occur. I also suspect that the first human-level AGI will soon be followed by a Singularity-type event leaving little time in the interim for a sociopath to create an AGI that kills everybody. I think the biggest risk related

to advanced AGI systems is that once smart enough they will be *indifferent* to human beings, in the same way that we're now indifferent to field mice, ants and bacteria. Few humans go out of their way to squash ants, and we generally worry about killing bacteria only when they're bothering us directly. However, when we put up a new building on a field of dirt, we don't worry much about the subsequent mass-murder of ants and bacteria.

Some futurist types don't worry about this much, as they plan on becoming cyborg-like superhuman beings, perhaps by fusing with AGI systems as the latter advance. If you plan on becoming a superhuman AGI via uploading your brain into a computer and enhancing it, or via plugging a super-brain chip into your biological brain, then the fate of those humans who refused to upgrade may seem a lesser concern. On the other hand, it might turn out that the superminds formed via upgrading human beings are still a bit less super than other superminds formed without the constraint of maintaining continuity with some human self. So, even humans who upgrade themselves into superminds might still find themselves at a disadvantage relative to the other wholly nonhuman superminds out there. But here we're pretty deep into science fiction territory!

The bottom line is, we really don't know enough yet to have a solid understanding of the issue of "superhuman AGI ethics." I personally suspect that it will be possible to create advanced AGI systems that respect "lesser" beings like human, by creating AGI systems that (like OpenCog) tend to respect their own goal systems, and instilling them with human-friendly goal systems via a combination of programming, teaching and simple human kindness. But I certainly can't *prove* this will be possible, and I'm not absolutely certain of it. We'll have to explore this domain via a combination of experiment and theory, just like every other aspect of AGI.

One thing I am pretty confident of, though, is that AGI is well-nigh inevitable. Nothing short of a large-scale collapse of civilization, or a global dictatorship bent on delaying the Singularity, is going to stop it. If the US were to outlaw AGI research, China would develop AGI. If China outlawed AGI research, Ethiopia would eventually develop AGI. And so forth. If every country outlawed it, an international AGI underground would probably develop, and you'd need a really powerful international anti-AGI police force to keep the R&D squelched— especially if computer and communication technologies keep advancing.

AGI is inevitable in the same sense that written language was after the introduction of spoken language. Writing was speech's natural next step, given the nature of human beings and the materials

available on the Earth.

And of course, we can't predict the consequences— any more than the people to write down the first few written words could predict the consequences of writing.

That doesn't mean we shouldn't try to nudge the Singularity in a positive direction; it doesn't mean we shouldn't try to raise our baby AGIs to have compassion and understanding. But it does mean we should temper our ambition to guide the future with a bit of humility regarding the size and sweep of the processes rushing us along. The development of progressively more complex adaptive systems on Earth started long before humans came along, and will continue long after humans lose their top-dog status on the planet (having been supplanted by AGIs) — the Singularity we're most likely near is just another fascinating phase transition along this path.

Will There Be Cyborgs?

It's hard to think about the future of AGI without thinking at least a little bit about cyborgs -- human/AGI hybrids. Cyborgs are both a likely real future possibility, and a useful conceptual tool for thinking about the future of man-machine interactions.

A cyborg could look like a human being with a computer jacked into its brain, or a human being with wheels, a tail and wings that permit flight, or a wheeled robot with a human head, etc. You can surely imagine a lot of other possibilities! The key point is that a cyborg combines engineered components with traditional, evolved biological components – and both of these components are working together to guide the system's practical operations. The most interesting kind of cyborg, in my view, is the kind where the engineered and evolved biological components work together cooperatively to control the system's *mind* as well as its body.



Figure 124: *A comic book cyborg.* <http://static.comicvine.com/uploads/original/7/74889/2697604-cyborg628.jpg>

Some people think cyborgs are implausible in the foreseeable future, because they're skeptical that technology will advance that fast. This kind of skepticism is not interesting from the point of view of the present book; it basically amounts to the hypothesis that exponential technological growth is going to halt, or hit a mysterious obstacle when it comes to human neurobiology. On the other hand, there is a different kind of skepticism about cyborgs that's more relevant here. Some folks are skeptical about cyborgs because they reason that: *Since non-human AGI systems may prove much more effective and intelligent than cyborgs, cyborgs won't be worth the bother.*



Figure 125: Two non-cyborgs. My friend and off-and-on research collaborator Hugo deGaris (perhaps best known as the founder of the field of Evolvable Hardware) believes that the future holds a huge world war, between those who favor the advent of superhuman AI, and those who fear it and want to hold it back. I doubt such a war will happen, because I think advanced AI is going to gradually become an integral part of peoples lives, like the Internet and smartphones are now, so that very few people will want to get rid of AI. Hugo and I seem to get along well in spite of this disagreement! (This photo was taken at Xiamen University in Summer 2009, when Hugo was a full-time professor there and I was a visiting professor, and we organized the First AGI Summer School there. Subsequent AGI Summer Schools have occurred in Reykjavik, Iceland, and Beijing.)

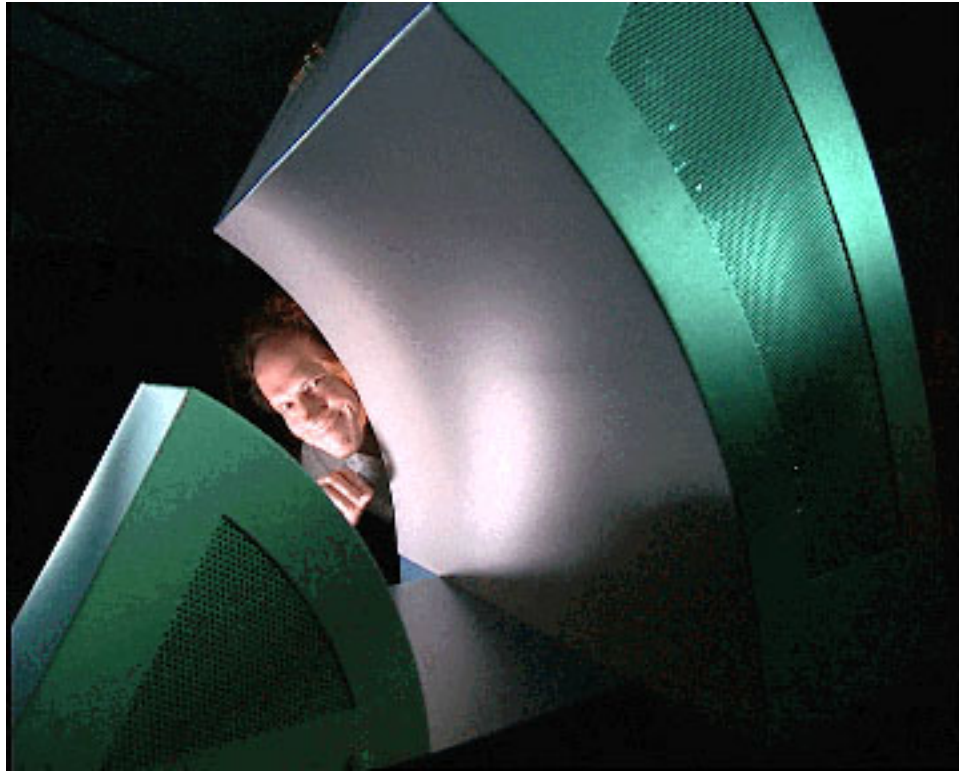


Figure 126: Hugo deGaris hiding behind his CAM-Brain Machine, a novel evolvable hardware device he designed in the 1990s, which was quite ahead of its time. It contained a number of FPGAs, reconfigurable chips, whose circuitry was automatically rewired via genetic algorithms and self-organizing, growing neural nets. I saw a CAM-Brain Machine at Starlab in Brussels in 2001, when I visited Hugo there right on the eve of that wonderful research institute's bankruptcy. Hugo used to say how he felt conflicted, because on the one hand he felt driven to build intelligent machines, but on the other hand he feared the world war he thought they would inevitably bring. More recently he has given up his practical AI work, and spends his time on a mix of theoretical physics and futurological speculation.

Along these lines, my friend and colleague Hugo de Garis wrote an article for H+ Magazine a few years back called “There Are No Cyborgs.” The article argued, not only that there were no genuine cyborgs around at the time of writing, but that there also would be no cyborgs in the future. In Hugo’s view, cyborgs are basically a non-starter. He expects that AGIs without the human component will be capable of outperforming cyborgs to an incredible degree, making the latter effectively irrelevant and pointless. As he likes to say, the computing power implicit in a grain of sand is greater than that of all the human beings on earth by many orders of magnitude. So if you can make a human-sized AGI supercomputer, orders of magnitude more intelligent than people – then what use will that supercomputer supermind AGI have for *cyborgs*, even if the latter happen to be ten or one hundred times smarter than humans?

Hugo's views on cyborgs are closely tied with his argument that sometime in the next century there is going to be a huge world war – the Artilect War, he calls it – between Cosmists who favor the creation of superhuman AI's even if they destroy humanity, and Terrans who want to hold back the development of advanced technologies in order to preserve the human race.

I see the appeal of Hugo's line of reasoning regarding cyborgs, but I'm not so sure he's right. Consider the natural world: Despite the uncontested dominance —and undoubted destructiveness— of the human race, we have not wiped out simpler, less intelligent life forms. In some cases —and here I'm thinking of bacteria— we coexist with and couldn't survive without them. And in many other cases, we recognize, more and more, the value of diversity and the need to maintain varied ecosystems.

As I emphasized already, it's not clear to me why a superhuman AGI supermind would want to get rid of all significantly less intelligent creatures. Why would it? What would it gain in the process? Indifference or benevolence on its part seems just as likely. Perhaps, operating in realms beyond our comprehension, it would have no particular incentive to trouble itself about us. Perhaps it would feel toward us like environmentalist humans feel about endangered species. Estimating the attitudes of superhuman superminds via extrapolation from human emotions seems chancy at best.

Even if superintelligent superminds pose no danger to humanity, one could still see an Artilect War due to peoples' fears of advanced AI. It wouldn't be the first time in human history a major war was started over an exaggerated fear. However, my suspicion is that human life will be too intertwined with AI software and hardware carrying out practical tasks, for the idea of getting rid of AIs to gain much popularity. Who wants to rebel against their smartphone; their online purchasing, negotiating and form-filling agent; their kids' video game characters and robot toy pets; their housecleaning robot and the AI bioscientist who discovered the medicine that cured grandma? Very few. Yet it may well be that this kind of practical proto-AI technology is where advanced, superhuman AGI will progressively grow from.

The Possibility of Femtotech Superintelligences

To concretize the possibility of an AGI so much smarter than us that helping or exterminating us would seem irrelevant to it, consider one possible route to very advanced AGI that Hugo and I have been talking about for a while —*femtocomputing*.

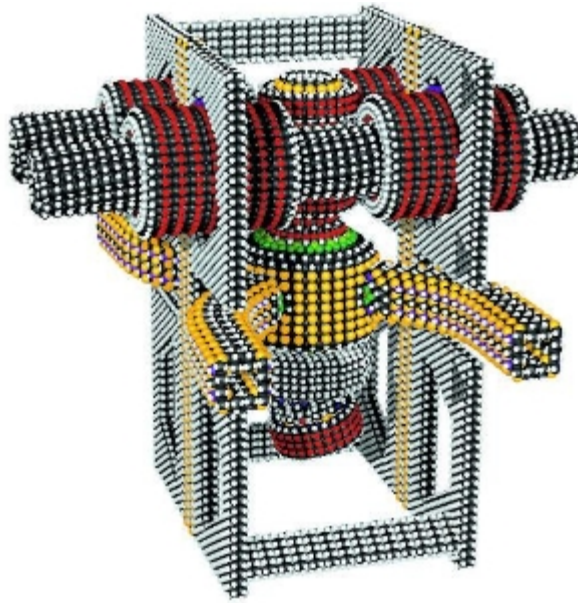


Figure 127: Design for a nanotech pump assembly, including over 65,000 atoms

<http://machinedesign.com/archive/molecular-modeling-cad>

Nanotechnology is about building things at the molecular scale, like biology does with DNA, RNA and proteins. For instance, given enough time and progress, it should be possible to assemble molecules to create computers vastly more powerful than anything available on the market now. And not just computers— if you had what nanotech pioneer Eric Drexler called a *molecular assembler*, you could use molecules like Lego blocks. Just specify what you want and the assembler will build it out of molecules. The molecular assembler, if one were to exist, could even build a copy of itself.

Drexler's vision of nanotech remains unrealized, but simpler versions of nanotech exist right now, and underlie various commercial products. Some novel approaches in modern nanotech involve using molecular biology mechanisms in ingenious ways— inducing biomolecules to assemble themselves into structures that evolution never would have produced.

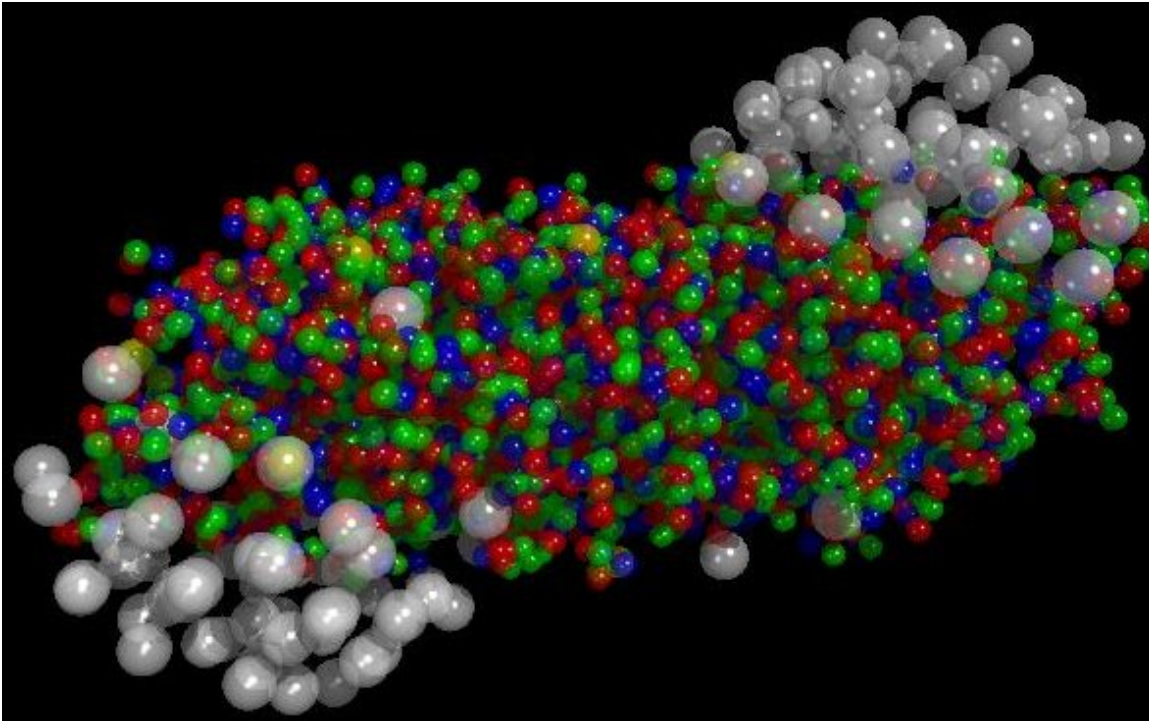


Figure 128: Quark-gluon plasma formed by a lead-lead collision in a particle accelerator, viewed shortly after the smash. The white balls are protons or neutrons. The small colored balls are the free quarks. Can we eventually engineer minds and machines from these sorts of exotic states of matter? There is no clear physical reason why not. By current human standards, the engineering problems involved in this sort of hypothetical femtotech are formidable. But an AGI with femtoscale sensors and actuators might view things differently. <http://hep.itp.tuwien.ac.at/~ipp/qgp.html>

Femtotechnology goes a step further than nanotech. Instead of using molecules, it uses subatomic particles like protons and neutrons, or elementary particles like quarks and gluons—the stuff that holds larger particles together—as basic building blocks. Today, femtotech is merely an idea—much like nanotech was from the 1950s through the 1980s. But as science and technology develop, I have little doubt it will come to pass.

Normally nuclear particles like protons and neutrons are trapped inside atoms – they make up the nuclei of atoms. But in some strange forms of matter – termed “degenerate matter” – the nuclear particles can come out to play, and engage with each other in more flexible configurations. Neutron stars are a well known example – a neutron star is just a bunch of neutrons packed together; there are no atoms involved. Another example is a quark-gluon plasma. Quarks are usually thought of as making up protons and neutrons, and gluons as exchanging energy between particles. But given the right conditions, including a very high temperature, quarks and gluons can come together into a strange sort of plasma, which has been created in some physics labs, albeit only for a split second. Will future science come up with ways to make stabler forms of degenerate matter, capable of carrying out

computation? It seems quite possible.

Hugo has worked out some basic math explaining how to build logic gates—the basic components of today's computers -- using quarks and gluons. His specific designs may or may not be possible to ever build in reality, but they're conceptually provocative. For instance, in his talk at the the Humanity+ @ Hong Kong conference in 2011, he explained how to create an OR gate (one of the basic logic gates used in all computers) using quarks and gluons. He gave similar treatments to the AND and NOT gates. Combined, these sorts of gates let you do any kind of computation. Of course no one will ever build a femtocomputer according to Hugo's specific 2011 designs – he was just giving a simple proof of principle to illustrate his general point about the conceptual sensibleness of femtocomputing.

It's a bit of a wild speculation, but I see a potentially interesting analogy between femtocomputing and DNA computing. Quark-gluon plasmas have been shown to contain chains of quarks and gluons that look vaguely like DNA strings, so I wonder if maybe there will be some future analogue of DNA computing within degenerate-matter femtostructures.

As I briefly reviewed above when I was talking about my biology work, DNA represents information in series of information like

... CCC TGT GGA GCC ACA CCC TAG ...

(each letter stands for a certain amino acid: G= guanine, etc.). For example, that is how the information distinguishing the baby Ben Goertzel from another human baby – or a donkey baby – is encoded. On the other hand, it's been found that quark-gluon plasmas contain particle strings, defined by series like QGGQ (Quark, Gluon, Gluon, Quark)... Can one do computing in quark gluon plasmas, in a manner similar to what today's scientific pioneers are doing with DNA computing? I don't know how to make it work in my garage just now, but nor did my great-great-grandfather know how to engineer a silicon chip.

Things like nanotechnology and computers built out of elementary particles seem strange or incomprehensible to us, with our human brains that are somewhat specialized for our everyday macroscopic world on the surface of the Earth. But to a superhuman AGI, such subjects would be substantially less daunting, and perhaps enormously diverting. Perhaps we will build an AGI running on ordinary computers, which will figure out how to build a smarter AGI running on molecular nanocomputers, to build a yet smarter AGI running on femtocomputers, which, finally, will build

things wholly inexpressible to the human mind.

Would Femtotech Superminds Bother to Exterminate Humans?

Supposing such femtotech-based superminds were created—why would they bother with us? Why would they bother to destroy us, or help us, or mess with us at all... Any more than we ponder the fate of various bacterial colonies living in various mud puddles around the Earth?

You could argue that any system will naturally have the desire to expand and dominate others to fight for its survival. That has been the case on Earth due to evolution by natural selection, although cooperation has been equally important as competition in the evolution of species (the two have been richly interwoven together, along with other more complex sort of self-organizing dynamics spanning the organism and ecosystem level).

Maybe a population of superhuman AGIs of roughly equal intelligence would sink into conflict over basics like survival and resources. Perhaps one would decide that the molecules now utilized by human beings are required to increase its processing powers and obtain a competitive advantage. This is not impossible.

On the other hand, what if one very powerful AGI mind develops faster and succeeds in outclassing the others? Is it necessarily going to want to grab all the available processing power in order to stymie rivals? That's far from obvious. AGI minds would not be constrained by biology, only by physics, and possibly not even the physics with which we are familiar. To these advanced minds, individuality itself could be passé, leading them beyond the categories of individual social beings, as we now understand them. How can we be confident that animal-level motivations like competition will still be remotely relevant to such beings?

Another possibility is that once AGIs become sufficiently intelligent, they will come to the attention of intelligences spawned by other races amongst the stars.

Very intelligent creatures could be out there in the universe, monitoring us through means we can't understand. Once they become aware of equivalent intelligences on Earth, they may be tempted to make contact. That may sound outlandish—but the bottom line is that we just don't know.

In trying to understand AGIs, we're a bit like early humans communicating at a simple level in the first language ever invented. These early humans, the first ones to invent language, would have realized they were onto something important—but they wouldn't have been able to foresee the real consequences of their invention. The imaginative flights of William Shakespeare and Marcel Proust, and the rigorous logic of differential calculus and computer programming, would exceed the power of these early humans to communicate in their caveman grunt language. And similarly, AGIs could transcend life, the universe and everything as we know it. We can only appreciate the immensity of their potential powers, and speculate about the road to superintelligence.

Hugo says there will be no cyborgs. I think there probably *will* be cyborgs after all—but I also think that, fascinating as cyborgs will be, they probably won't be the most interesting new beings generated by humanity. My *hope* is that, once suitable technologies are available, each human will have the choice to remain an ordinary human, to become a cyborg – or to expand their intelligence more dramatically. Some people may choose to remain at the human or cyborg level, but plenty of others will anxiously push the boundaries of the possible and vanish over the horizon, en route to a higher state of being.

The Global Brain

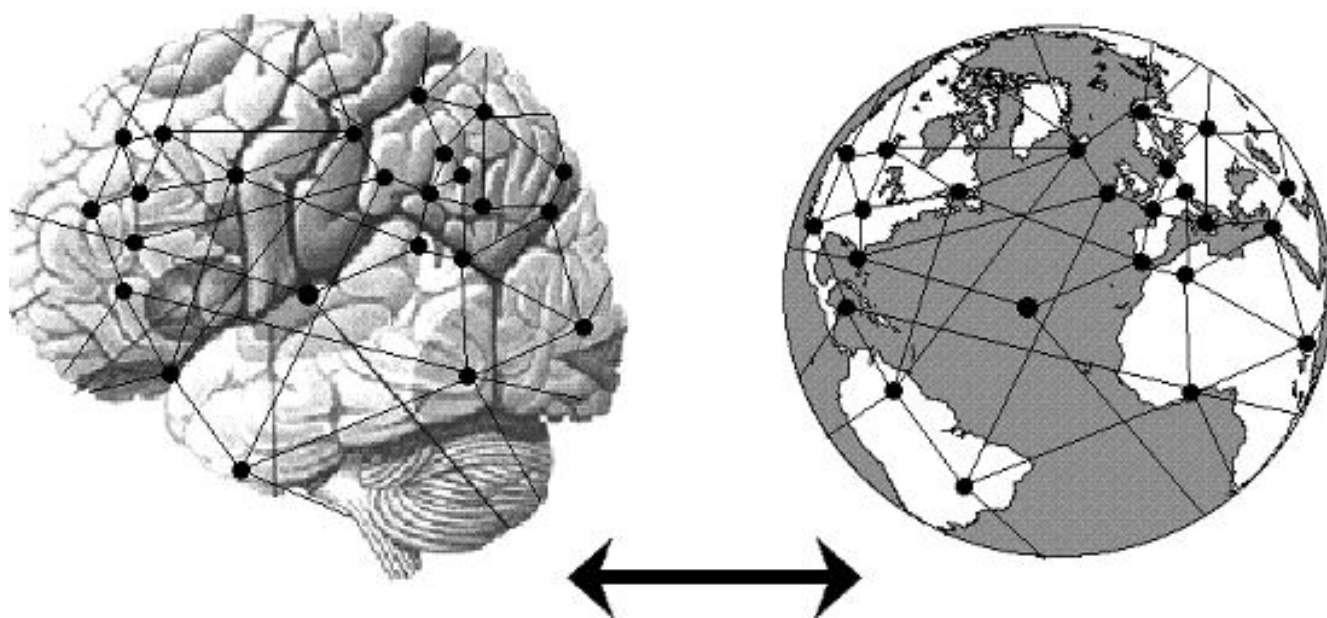


Figure 129: Artistic depiction of the Global Brain – of the analogy between people in society, and neurons in the brain.
<http://pespmc1.vub.ac.be/Images/Brain-Earth.GIF>

AGI robots, cyborgs, and incomprehensible femtotech superminds are not the only possible futures for advanced posthuman intelligence. There's also a potential that feels quite close to home – the Global Brain; the notion that computers and communications technology, already one and the same, will gradually meld with and link humanity through all manner of devices, forming a networked community of people and machines that will give rise to a higher order of intelligence. I thought about this sort of thing a lot in the late 1990s, when the Web was just beginning. It still seems a relevant paradigm for conceptualizing many aspects of the future.

People and networked machines across the planet would be like neurons, the brain cells, in this collective brain that would think via trends, essentially patterns of ideas spreading across a network. Individuals, even if they were integrated into the Global Brain, might not be able to sense these thoughts in detail any better than a single neuron can understand the thoughts flashing across our mindscapes.

The Global Brain might seem a threat to human freedom, in a different way from robot overlords or indifferent femtotech superminds – but it ain't necessarily so. A neuron may be free to be a neuron, firing when it wishes and living on the level of a neuron in a way that's completely natural to it. Nevertheless, from a different point of view, a neuron is part of the coherent coordinated activity of the human brain. In a similar way, we may go about our lives as humans in a manner that seems natural to us, exercising what seems to us to be free will. Yet on another level, we may be part of a higher-order coordinated intelligence.

While not as prominent in the popular eye as AGI robots, the Global Brain concept is gaining a bit more attention lately. My friend Francis Heylighen, a professor at the Free University of Brussels in Europe, has done more than anyone else to promote the notion. In 2001, I suggested he organize a Global Brain conference at his university, and he thankfully followed through. We called it Global Brain 0, and it went pretty interestingly. More recently, in early 2012, Francis received funding from a retired Internet entrepreneur to start a Global Brain Institute at his university.

One thing I realized during Global Brain 0 was that the conference participants had a huge variety of views on the Global Brain. Essentially, their views could be divided into three categories.

- **The Global Brain Is Here!** One group believed that the Global Brain is already here. We may not recognize or understand it fully, but it's here. The Internet is an intelligent mind, different from what we are, maybe smarter in some ways and less so in others.

- **The Global Brain Will Spontaneously Emerge.** Another group believed that the Global Brain is not yet present in full force and can't be until it undergoes a phase transition into an impressively intelligent and organized form, one that will emerge naturally without our help. It will develop on its own as computing, communication technologies and human culture advance.
- **The Global Brain Will Come, Because Someone Will Build It.** Finally a third group, to which I belonged, argued that the Global Brain will emerge, but as a purposeful consequence of Global Brain engineering, alongside the emergent dynamics of communication systems and cultures arising for other reasons.

More specifically, I believe that – if something else, like a destructive world war or an AGI robot supermind, doesn't come first – a powerful Global Brain will come about after we release fairly advanced autonomous or semi-autonomous AGIs on the Internet. Imagine AGIs let loose on the Net, with the goal of interacting with people, mediating communications, ingesting information, then summarizing and presenting it in a more comprehensible form, and finally, placing new information on the interwebs for people and other AIs to ingest. A network of such AGIs could serve as the central cortex of a Global Brain, with humans and other software systems (including a diverse spectrum of narrow AIs) filling supporting roles. This is the vision I laid out in my 2001 book *Creating Internet Intelligence*, and I still think it's a plausible future scenario.

Eventually, some AGIs may outgrow a Global Brain in which humans can play a role, moving on to some higher form of intelligence beyond human comprehension. But even after such a transcension, this planet might still be dominated by some sort of collective Global Brain consciousness shared between AGIs and humans.

It's easy to see how some of today's technologies are explicitly pushing us in this direction. For instance, social media; systems like *Digg*, *Reddit* and *Slashdot* (or whatever their new, fashionable analogues are at the time that you're reading this) consist of people collectively summarizing news and pushing stories to the top of listings, collaborating to do the job of an editor and reporter simultaneously.

Social networking systems like *Twitter* (again, insert your favorite social network of the day!), although often frivolous, also display facets of collective intelligence. Emergence is visible in Twitter when relevant information bubbles up through a mass of retweets, eventually becoming a trend for reasons independent of any one person's actions.

Consumer reviews on sites like *Amazon* also represent collective intelligence, since individuals, sellers and products can build up good reputations based on others' ratings, forming active collaborative filtering systems with some simple narrow AI components. The system then presents visitors with information they wouldn't have read otherwise, inspiring them to ponder and purchase new things, reinforcing trends further.

Open-source software projects are another interesting example of group intelligence. Each starts with a body of software, which grows and evolves over time like a life form, thanks to the input of hundreds of people. A handful of people may have written the original code, but it can quickly grow beyond their intentions, splitting into multiple projects and incorporating previously developed or fresh ideas in a process akin to symbiogenesis and the emergence of organisms. The result is a whole new layer of life forms. I've even seen limited examples of this in *OpenCog*, the open-source AGI project that I'm involved with.

So... Given that Global Brain-related technology is already in active use on the Internet, what happens when an AGI takes over and uses this infrastructure to help manifest its own ideas, which derive from the Net? What happens when an AGI starts reading books and recommending them to people, weaving new linkages between strangers? What happens when an AGI contributes to open-source software projects, or starts them and recruits new developers?

What happens is: We end up with a whole new kind of collective intelligence among humans and computer software, one that is not purely self-organized from the ground up, but rather mediated by AGIs who are working toward their own goals. With any luck, their goals will mesh with ours because they'll be trained to help our collective mind emerge, grow and flourish. Eventually, some of these AGIs may want to journey on, to a place where humans can't accompany them – but others may be motivated to stay around at a relatively stable level of intelligence and persist their communion with humans.

Does the Global Brain provide a way around the potential dangers of advanced AGI? Not exactly. But it does provide a unique perspective, and highlights a set of pathways that are commonly ignored. If AGI matures in the context of the Global Brain, then it will grow up feeling tightly interlinked with humanity. This will certainly affect the AGI's attitudes as it grows beyond its human origins— though precisely how is hard to say at this point.

The Risks and Rewards of Advanced AGI

So what's the take-away message on AGI safety? Are the super-AIs going to kill us all or not?

The real truth is that nobody knows; we don't know enough about AGI yet. I believe I've found a way to build a human-level AGI, superior to humans in some ways. But that certainly doesn't mean I can foresee all the consequences of such an invention!

To use the language analogy again, this situation is equivalent to an especially prescient early human at the dawn of language. Suppose this hypothetical far-sighted caveman is seeing into the future and struggling to comprehend all the terrible things the development of language might bring —systematic hatred, violence and war. And suppose the same early human also dimly foresees the possibility of literature, science and mathematics. With his restricted view, how could he possibly balance the costs and benefits? Suppose he also has the wisdom to realize the deep limitations of his present perspective. What will be his ultimate conclusion? Maybe he'll think: *Something really big is happening, I'm playing a small part in it, and nobody right now is really equipped to understand what's going to come of it. There's a lot of risk here, but all I can do is hope for the best, and do my utmost to push things in my own local sphere in a positive direction.*

Yes, there is unquestionably the possibility that an advanced AGI could destroy humanity. And, yes, as a human I would like to minimize the risk of this happening. I'm doing my best to create a beneficial

AGI before somebody else creates a malevolent or indifferent one. But I'm also acutely aware of how little we know about the massive transformational processes going on in which each of us just plays a tiny part. The bottom line is that AGI is an inevitable consequence of human technological development. Someone is going to build it and no one really has the capability to stop that from happening. Computers will keep developing because we need them for increasingly demanding applications. We must understand the human brain in order to understand ourselves, and we're going to keep developing ever more powerful tools to do that. And our computational algorithms are going to be developed further because we want software to do more. Robotics will progress because we want relief from tedious tasks. Games will advance as we come to expect more from our entertainment. And although all this development is only indirectly related with AGI, it's still leading us there, bit by bit – and at an accelerating pace. Eventually, someone will put the right pieces together—maybe my OpenCog colleagues and me, maybe Demis Hassabis's team at Deep Mind, maybe some other folks—and we'll have the first advanced thinking machines.

The development of AGI is one strand in an intricate web of technological, scientific and human developments, one you can't disentangle. AGI will emerge as part and parcel of a much broader story of scientific, engineering and cultural advancement. No political or military organization has much chance of stopping it. The only way to pause progress toward AGI really convincingly, would be to take over the entire world with a single government, and pause general technological development globally. But obviously such a move would cause a lot of problems, and doesn't seem at all likely to happen.

Since stopping AGI is infeasible, it makes little sense to harp on the issue of WHETHER to build AGI, and more sense to focus attention WHEN and HOW to create AGI.

Does Humanity Need an AGI Nanny?

In thinking about the future of AGI, it's worth remembering there may be other potentially super-dangerous technologies on the horizon. For instance, nanotechnology and synthetic biology are advancing fast. How long until someone makes incredibly potent bio-weapons? These aren't as striking to visualize as the Terminator, but they may be more realistic threats. Arguably, even discounting the potential threats from AGI, humanity is heading for disaster in the next century unless some radical solution is undertaken. By the time anyone with access to basic bio lab equipment can make synthetic viruses capable of poisoning billions of people— we're in trouble.

Perhaps the development of countermeasures will keep up with the rate of advancement, thus neutralizing the threat. Artificial nanotech-based immune systems could emerge to counteract the synthetic viruses. But this seems difficult to count on.

What kind of radical solution would work?

One possibility is improving human nature by reining in our nastier impulses.

Jeffrey Martin and Mikey Siegel, two colleagues of mine in Hong Kong, have been working on exactly this. They've proposed *neurofeedback*—a method of discovering the aspects of the human brain correlated with compassion and enlightened states of consciousness. The idea is to let people look at a real-time computer image of what's happening in their brain and try to exert control over these positive states of awareness. Their hope is that this sort of technology can enlighten the population—educating people to be more compassionate, loving and cooperative. I'm personally skeptical that salvation could come in time to save us from the threats posed by advanced technologies, but I admire the effort.

I've also tossed around the idea, tongue partly in cheek, of an AI Nanny—a system tasked to watch over the human race and ensure nothing bad happens. Ideally, you would make an AI Nanny that was a couple times smarter than human beings, but which lacked the motivation to augment itself and attain superhuman AI levels. It would have to remain a relatively dumb AI, one passionately excited about, and dedicated to its mission in life — protecting the human race from external threats, natural disasters and, well, humanity itself. This may sound like an Orwellian parody or a Big Brother scenario, but if done right, the AI Nanny would stay in the background, while subtly aiding us to make life better, intervening directly only when something really bad is about to happen.

But what about a more moderate solution than the AI Nanny, something less of a caricature or archetype? One option is global democratic governance, an open and accountable force for good that would help monitor world events and prevent bad things from happening. This relates to an interesting concept called *sousveillance*, which science fiction writer and social theorist David Brin brought to my attention. *Sousveillance* is the idea that, rather than giving a leader the power to watch the peons, everyone should have the power to watch. All the information about everything happening in the world would be available to everybody, including the AI nanny, if such a thing were to exist.

This may seem like a disturbing idea. However, the analogy David Brin uses is people having private conversations in a restaurant. We could eavesdrop on everything that's being said, but we typically

don't because other people's conversations aren't very interesting.

The same is true when it comes to spying on other people's most intimate activities. Voyeurism would get old pretty fast and people would likely just get on with their lives, even if they knew they were being watched. *Sousveillance* would arguably even have a positive impact on us, making us less self-conscious, less dishonest, more open and tolerant. It would also have the benefit of making something like an AI Nanny seem less oppressive since it wouldn't just be this AI overlord aware of everything we do; everyone would be looking at everything, including the AI Nanny's activities.

But still, no matter how prettily one paints it, the idea of an AI Nanny or universal *sousveillance* mechanism watching over us to prevent untoward progress still has a certain disappointing aspect to it. I mean, don't we want the human race to be something amazing? Don't we want to go way beyond anything we can presently imagine -- and transcend to something resembling godhood? I certainly do, but I'm not sure about the right way to get there. Advancing from here to godlike AGIs in a more measured fashion may be the most intelligent thing to do, but removing the element of risk could really slow us down. What if the safest route —something like an AGI Nanny system— dragged out the path from here to godhood so it took two thousand years instead of two? A better, kinder future could lie within our reach, veiled by a fog of perceived but avoidable or even non-existent danger.

I'm skeptical there will ever be an unquestionably friendly AI or AGI system. However, we could increase the odds of a beneficial Singularity by having a better theory of intelligence before it gets here. Devising a solid theory of intelligence, however, would require experimentation with a bunch of AGI systems. So how do you stop these experimental AGI systems from advancing too fast and causing destruction, even as they lead to an enhanced theoretical understanding?

An intelligent control mechanism like the archetype of the AGI Nanny is potentially valuable, but ultimately we may have to work out a solution as we go along. I like to think of this process as a net positive enabling the human race to mature in its self-judgments. But who knows? Maturity is a complex thing, a delicate balance, and that's just as true on the societal level as it is on the individual level.

In the long run, one single solution to AGI safety seems unlikely. To avoid the existential risks of AGI development we must spread the needed maturity through society and the community of AGI researchers first, so it's inculcated in the way AGIs are engineered. We need to get the basics of ethics right ourselves, so we can pass them on to AGI systems that will learn from us in a loving but clear

way.

We'll also have to choose appropriate vessels for the first AGIs. Rather than killer military robots, for example, a wiser option would be helping, loving applications – teachers, doctors, medical researchers, AGI philosophers, authors, artists, scientists and engineers.

We'll have to get some basic stuff right to have a positive Singularity— but there's a lot that we're not going to be able to understand for a long time. Experimentation must unfold gradually along with ethics and a theory of general intelligence. We'll have to handle this with some level of maturity and wisdom. It'll be a challenge, but an exciting one.

Why AGI?

I believe we could create human-level AGI fairly rapidly with the OpenCog project, given sufficient funding. But if that doesn't end up happening, for one reason or another, then somebody else is going to do it. Even if AGI funding remains relatively hard to come by, and even if brain scanning accuracy lags behind Kurzweil's expectations — still, sometime in this century, somebody's going to do it.

As I've emphasized repeatedly, I don't think AGI can be stopped, except by a calamity that wipes out humanity or destroys civilization. AGI is just the next step in the evolution of complexity and intelligence. It's what the universe has in store next.

From that point of view, asking “Why AGI?” is irrelevant. It's sort of like a one-celled organism, floating around in the primordial soup, asking its neighbor “Why bother with multicellular life? Why not leave well enough alone?”

That's the grand, cosmic view. It's the view I inevitably fall into when I'm hiking in the mountains near my house in rural Hong Kong, looking out at the islands and the ocean and the strange-looking trees atop the hills, and thinking about how much more vast is the universe than anything I or any person can conceive. But no human always thinks that way, not even a wild-eyed transhumanist like me. Sometimes when I'm looking at my wife smile, or listening to my son play piano, I can't help feel glued to the human world. I sink into the role of the confused one-celled organism and ask myself the question: *Why create AGI? Especially given the potential dangers— why not leave well enough alone? Humanity's not so bad!*

But, of course, I know the answers. First: because that's what the universe does. It creates amazing new things – like molecules, bacteria, animals, plants, humans, computers, Internets – and then transcends them. Second: Because, even within a narrowly human-focused perspective, forbidding AGI wouldn't keep things safe and stable anyway. Without AGI to help protect us, the various other technologies we're developing -- synthetic biology, nanotech, etc. – would stand a good chance of harming us badly badly, maybe even killing us all. AGI may be our only chance of NOT getting destroyed by something else. Third: Because there is so damn much apparently unnecessary human suffering in the current and historical human condition, alongside all the beauty.

My wife's lovely smile will fade one day into the macabre grinning rictus of a skeleton – unless we do something about aging and death with advanced technology. The fingers of my son's decomposing skeleton won't be able to play Scriabin's music any better than the skeleton of Scriabin – who died at age 43, in the midst of composing a masterwork, from an infected sore on his upper lip. There were no antibiotics in the time of Scriabin, so he died young from something that could easily be cured today. It isn't human nature to just accept terrible waste and suffering as inevitable and natural – it's human nature to strive to make things better, even when this means transforming human nature, or even transcending it.

As a human being working hard on AGI, my motivation is largely driven by the following belief— AGI has the potential to make the life-experience of sentient beings a heck of a lot *better*!

Being human can be wonderful and beautiful, but – as Scriabin learned the hard way at 43 years old – it can also be a hell of a pain. And it can be extremely limiting— so many of our dreams and ideas just never get realized, for boring old practical reasons.

According to Buddhism, all existence is suffering, which doesn't mean that life totally consists of pain, but that suffering is interwoven into everything. Everything has a little bit of suffering. As Nietzsche said, "If you have experienced one joy, then you have experienced all woe as well." All this may be a bit overstated but there's certainly something to it— almost every aspect of human life is characterized by suffering of some sort.

Now, it may be that pain is just a universal feature of existence and there's no way to eliminate the experience of suffering altogether. I wouldn't be shocked to learn this was true, though I'm not willing to commit to it at this point. I do think, though, there is a *lot more* suffering in human nature than is necessary, and far too much misery in the world.

There are many obvious horrors in the world today -- child abuse, rape, starvation amidst plenty and agonizing death from disease. There's also a high density of less spectacular frustrations like headaches, telemarketers, and the trauma of failed relationships. We accept these things because we're accustomed to them, but that's just like cavemen accepting rotting teeth because of ignorance about dentistry, and Scriabin's friends and family accepting him dying of a simple infection because antibiotics hadn't been invented yet.

You may say all the pain of human is necessary, and that without pain there wouldn't be pleasure. There's some truth underlying this perspective; all things are connected. But I bet there are ways of living that involve a lot less suffering and a far higher percentage of joy, as compared to current human life. I think you could get rid of death, disease, rape and all the negative feelings that accompany them without eliminating the fundamental joy, pleasure and wonder of being alive, of living, loving and existing in the world.

I see the creation of a positive Singularity, including AGI and radical human enhancement, as the short path to a better life, a way of breaking out of this existence into something profoundly more joyful and fulfilling. That doesn't mean I consider present life to be a horrible way of living. Current life is not a torture – I basically enjoy my life and I see many others who do, too. But the fact remains that many others are less fortunate, mired in poverty, trapped by violence or bereft of opportunities for improvement, their happiness and potential accomplishments stifled through no fault of their own. And even in a relatively good life like mine, there's a disturbing amount of everyday discomfort and worry. Of course, human life isn't uniformly bad— in spite of all the problems, there's a glorious joy at the center of it, and most of us have plenty of wonderful moments. But one of the most amazing things about human life is that it contains the potential to create new kinds of minds with new ways of living, far better than anything humanity has ever experienced.

Everything that has an upside also has a downside, and with AGI just like everything else, it's up to us to do our best manage the balance between the two. The odds that we'll ever develop a technology that is wholly good are slim. We just have to do our best to be smart about innovation and inject as much wisdom and foresight as we can into the proceedings. On the whole, technology has vastly improved human life; it has certainly made the human experience richer, more complex and interesting – and I believe this will continue.

Transcending the Discontents of Civilization

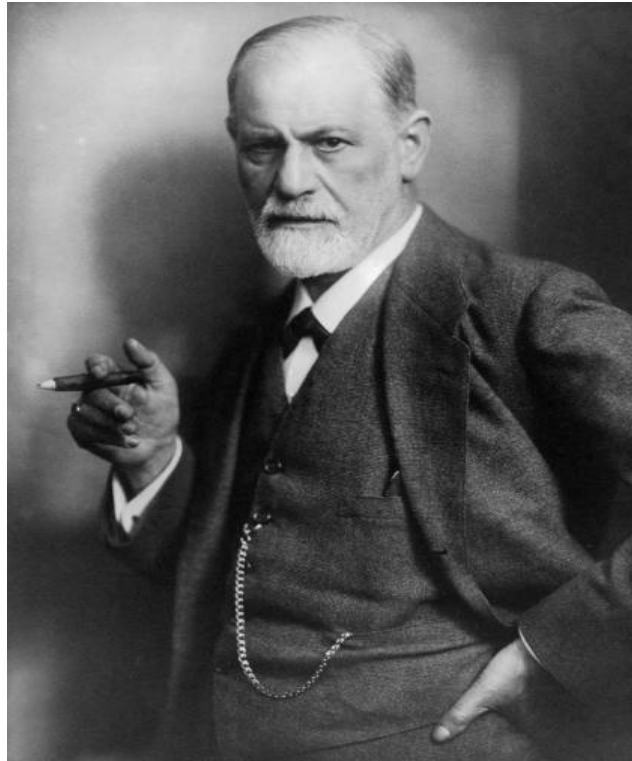


Figure 130: *While some of Sigmund Freud's ideas have been discredited by modern science, others remain as acute and insightful today as they were when he formulated them. His analysis of how the restrictions posed by civilization are the root of much of the unhappiness of modern humans, falls clearly into the latter category. Following up earlier related ideas by Nietzsche and others, he was a careful observer of the processes by which we take the constraints and criticisms of our parents and our society at large, and transform them into our own internal controllers and critics, thus setting up complex, often emotionally troublesome dynamics within our own minds. He noted that at bottom, most of our troublesome inner dynamics are related to the struggle between our evolved animal urges, and the demands civilization places on us to act less like typical animals. Unfortunately his proposed methods for eliminating the sources of unhappiness he identified (mainly, certain forms of talk therapy), did not prove extraordinarily effective. One wonders what Freud would have thought about the possibility of eliminating or drastically reducing the psychological problems induced by civilization, via engineering out some of the “animal” aspects of the human mind, and via lessening the restrictions posed by society via massively reducing the scarcity of human-relevant resources.*http://en.wikipedia.org/wiki/File:Sigmund_Freud_LIFE.jpg

In his book *Civilization and Its Discontents*, the great psychologist Sigmund Freud has an interesting

view of the evolution of humanity and culture—all our varied human neuroses and psychological problems, all the forces that disturb our mental health, are ultimately a consequence of civilization. Prehistoric people didn't have all these problems. They had others: Child mortality was sky high, most serious diseases were fatal, and life spans were short. But they didn't suffer from the deep psychological unhappiness that afflicts so many of their descendants because they lived according to their impulses. All too often, modern society compels us to repress ours. We end up being tormented by strong desires we can't possibly fulfill because the rules we live by won't allow it. We're too intimately bound up with other people and doing these things risks harming others.

Freud wasn't saying this is necessarily a bad thing; he wasn't proposing that we return to the Stone Age. Doing so might, admittedly, rid us of most of our neuroses and psychoses, but in the process we would lose many of the things that make our modern lives precious —language, culture, literature, science, mathematics, cinema, and do forth. Freud's solution to the dilemma he identified was psychoanalysis, which hasn't worked all that well so far. Most modern psychiatrists rely on pharmacological rather than talk-therapy solutions these days, and the business of prescribing psych meds is surprisingly reliant on trial and error.

Relative to Freud's perspective, I see the Singularity as a way of resolving the problems that have festered since the advent of civilization. Rather than turning our backs on development, we can embrace it as something that will lead us to an advanced and more enlightened state of being, one that eases the restrictions and shackles that civilization has placed upon us, while retaining all the positives derived from the progress we've made so far.

While Freud didn't explicitly view it this way, I believe the problems of civilization are at bottom mainly problems of resource scarcity — scarcities of space, time, energy and intelligence that make us step on each other's toes and require us to restrict ourselves for the good of others. Advanced technology can liberate us from these bottlenecks, opening up new and previously unknown vistas of freedom and fulfillment. The modern move toward Singularity could be viewed as a process of transitioning away from the constraints on knowledge and consciousness long imposed by modern society.

And after the civilization we know is gone, the post-Singularity phase will usher in new freedoms — new ways to grow, to feel, to think, to interact, to enjoy, to create. This promise of a dramatically better, wider, more richly and diversely fascinating and satisfying tomorrow is what makes the

Singularity so exciting. That's why it's worth accepting dangers and risks, even the existential ones; they're worth it because of the payoff. Just as the move to multicellular life was worth it, in the big picture, even though it shook up the everyday lives of a large number of amoebas, paramecia and Euglena. New ways of thinking, being, feeling and living, with others and ourselves – and going beyond the distinction between “others” and “ourselves” -- are beckoning us from the peaks ahead. We must find the courage to ascend and risk falling, for the rewards of success may be LITERALLY beyond anything we can imagine.

The Cosmist Perspective

It's clear that our contemporary human belief systems leave us with an incomplete understanding of the implications surrounding AGI and the Singularity. New sciences and technologies require new ways of thinking.

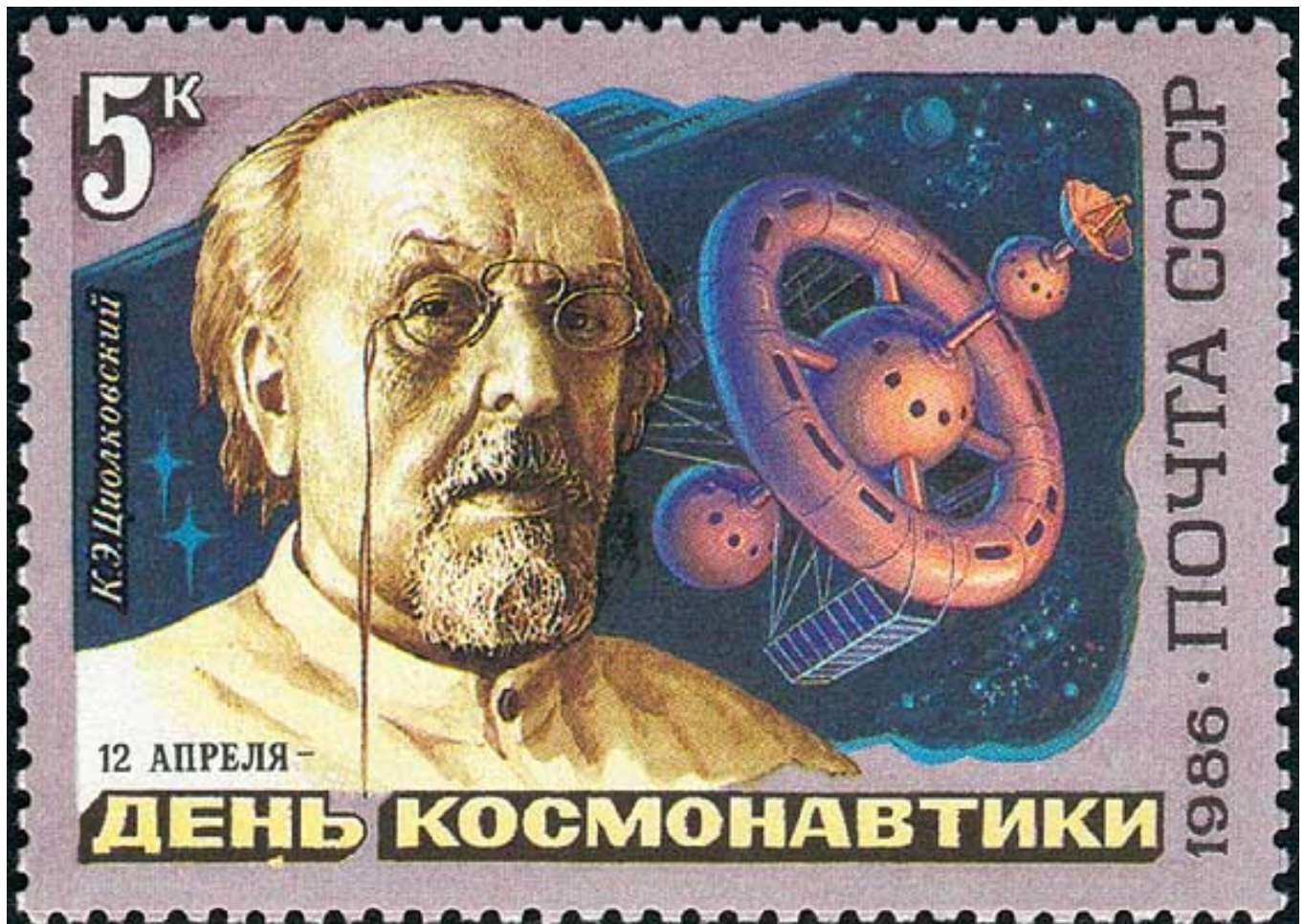


Figure 131: Russian postage stamp commemorating rocketry pioneer and Cosmist philosopher Konstantin Tsiolkovsky (who lived 1857-1935, foresaw humanity colonizing the galaxy, and pursued a sophisticated panpsychist philosophy in which the expansion of humanity and technology was a manifestation of the conscious will of the physical universe).
http://ferrebeekeeper.files.wordpress.com/2011/09/sm_tsiolkovsky.jpeg



Figure 132: Giulio Prisco, modern-day Cosmist, who has been one of the strongest advocates of the relationship between transhumanism and spirituality. His Turing Church blog has been influential in the transhumanist community, as have his online technology/spirituality discussion groups, held in virtual worlds (such as depicted above). http://1.bp.blogspot.com/_nUDJsVoQIP8/S4Vbzg_38MI/AAAAAAAAAbY/NZysLMFaxxQ/s320/streamavatars.jpg

With this in mind, I wrote a short book in 2010 called *A Cosmist Manifesto*, outlining a philosophy of life, the universe and everything that would still make sense in the context of a Singularity, and all of the associated transformative technologies. I called that philosophy *Cosmism*. I chose the name because I was inspired by the Russian Cosmist philosophers of the early 19th century, who already had many ideas common in the transhumanist community today, at a time when the state of technology made them a lot less obvious. These folks, Tsiolokovsky and Federov and the others, were true futurist visionaries! They saw far beyond their time, just as I am attempting to do here in these pages.

“The Earth is the Cradle of the Mind -- but one cannot eternally live in a cradle.”

- Konstantin Tsiolokovsky

... Advocating mankind exploring the universe

“How unnatural it is to ask, ‘Why does that which exists, exist?’ and yet how completely natural it is to ask, ‘Why do the living die?’ “

- Nikolai Federov

... Speaking against philosophy and in favor of immortality for all

At the start of the *Cosmist Manifesto* I give a list of ten “Cosmist Convictions”—ten basic principles of Cosmist thinking. My friend and colleague Giulio Prisco mostly came up with them, but I edited and

augmented his list. They don't actually tell you all that much about Cosmist philosophy— if you want to know about that, read the *Cosmist Manifesto*!— but they do articulate the general Cosmist perspective, which serves as a conceptual foundation for Cosmist philosophy.

Without further ado, here are the Ten Cosmist Convictions:

- 3) Humans will merge with technology to a rapidly increasing extent. This is a new phase of the evolution of our species, just picking up speed about now. The divide between natural and artificial will blur and then disappear. Some of us will continue to be humans, but with a dramatically expanding range of options, creating radically increased diversity and complexity. Others will grow into new forms of intelligence far beyond the human domain.
- 4) We will develop sentient AI and mind uploading technology. Mind uploading technology will permit an indefinite lifespan to those who choose to leave biology behind and upload. Some uploaded humans will choose to merge with each other and AIs, requiring reformulations of current notions of self.
- 5) We will spread to the stars and roam the universe. We will meet and merge with other species out there. We may roam to other dimensions of existence as well, beyond the ones of which we're currently aware.
- 6) We will develop interoperable synthetic realities (virtual worlds) able to support sentience. Some uploads will choose to live in virtual worlds. The divide between physical and synthetic realities will blur and then disappear.
- 7) We will develop spacetime engineering and scientific "future magic" much beyond our current understanding and imagination.
- 8) Spacetime engineering and future magic will use scientific means to fulfill most of the promises of religion— and many amazing things that no human religion ever dreamed. Eventually we will be able to resurrect the dead by "copying them to the future.”³
- 9) Intelligent life will become the main factor in the evolution of the cosmos and guide its path.
- 10) Radical technological advances will reduce material scarcity drastically, so that abundances of

³ Recreating a deceased person in the future, based on all known data about the person – their writings, videos of them, memories of them in the minds of still living people – and advanced knowledge of the brain and body.

wealth, growth and experience will be available to all desiring minds. New systems of self-regulation will emerge to mitigate the possibility of mind-creation running amok and exhausting the ample resources of the cosmos.

- 11) New ethical systems will emerge, based on principles including the spread of joy, growth and freedom through the universe, as well as new principles we cannot yet imagine.
- 12) All these changes will fundamentally improve the subjective and social experience of humans, our creations and successors, leading to states of individual and shared awareness possessing depth, breadth and wonder far beyond that accessible to "legacy humans."

Giulio Prisco, who formulated the first draft, made the following comment on the use of the word "will" in these principles: "*... 'will' is not used in the sense of inevitability, but in the sense of intention: we want to do this, we are confident that we can do it, and we will do our f**king best to do it.*"

It's worth keeping in mind that the general Cosmist perspective and the particular points of Cosmist philosophy— like the technical and conceptual specifics of AGI—are basically independent of the notion of a *Singularity* per se. If we have a slower, more gradual advent of advanced technology, the philosophical and technical issues involved will basically be the same.

Cosmist philosophy is a world-view within which these sorts of principles are natural and sensible, rather than weird or counterintuitive. Cosmism views the world – like the mind – as an entity with multiple aspects, going beyond any one perspective, and defying simplistic categories like “objective” and “subjective”. Among the perspectives it adopts are three delineated by philosopher Charles S. Peirce:

- **Firstness**: pure experience and Being; raw awareness
- **Secondness**: reaction, interaction and movement
- **Thirdness**: pattern and relationship
- **Fourthness, synergy and emergence** (not emphasized by Peirce but added on by later thinkers like Carl Jung and Buckminster Fuller)

The Thirdness perspective views the world as a web of interlocking patterned relationships – this is the view that science leads us to. Each scientific observation posits a relationship between certain observations. Observations themselves, viewed subjectively, are Firsts or Seconds; but when multiple

observations are woven together into a repeatable pattern, one has a Third, a relationship.

In the Cosmist view, individual minds, societies, physical objects, and even space and time themselves may be viewed as patterns, as regularities occurring among elementary observations. The “observations” themselves are primary, more so than the notion of a “self” doing the observing – since every psychologist knows that the Self is an abstraction a mind builds for itself.

There are echoes of quantum theory here, as in quantum mechanics the physical world is viewed as having reality only relative to acts of observation. But Cosmism is a philosophy whereas quantum mechanics is a scientific theory. Cosmism can help us interpret quantum theory, but will retain its philosophical validity even if quantum theory is obsoleted by very different physical ideas.

In the Cosmist view, humans – like apes, rodents, bacteria and molecules – are best viewed as particular patterns of organization, emergent from elementary observations. Each of these patterns of organization has a certain coherence, a certain synergy. The universe, if you view it from a perspective that embraces the linear flow of time, reveals a pattern of progressive growth from simpler to more complex emergent wholes. There is also a pattern of creative destruction that occurs when new wholes arise to incorporate and disrupt aspects of previous ones. The issues humanity currently faces regarding the Singularity instantiate this larger process.

Human ethics, viewed with Cosmist eyes, is in part just a particular set of patterns that certain societies of intelligent entities have adopted in order to maintain their stability and/or promote their growth. However, there are also some universal principles underlying the diversity of human ethical precepts. For instance, the value of joy above suffering is important, and goes beyond particular cultures or organisms. The value of growth above stagnation is intrinsic to the universe, and seems to come along with any perspective that views time as going forward. The value of choice is critical from any perspective that involves minds distinguishing themselves from the universe – including entities like group minds and global brains. While classical “free will” is largely illusory, a broader notion of choice is essential to the existence of separate minds as distinct entities.

It seems unlikely that particular human ethical precepts like “don’t covet thy neighbor’s wife” are going to have broad meaning after the Singularity. They may still be useful within particular human communities that persist at that point, but they won’t play a major role across the scope of intelligences in the Cosmos. But extensions of deeper human values like Joy, Growth and Choice may continue to be critical. New kinds of minds will create their own values – as Nietzsche foresaw the hypothetical

Superman coming after humanity would do – but these values may well still be compatible with broad principles such as these.

The Cosmist perspective— or whatever name you want to attach to it— is something I gleaned at an early age from reading science fiction and Buddhist philosophy, and it’s central to my personal view of the world, more so than the Singularity or even AGI. It’s a general view of the mind and the world that underlies all the particular matters I discuss here and in my other writings. It helps me to grapple with the possibilities the future offers, without getting my mind blown.

Cosmist philosophy implies the conceptual POSSIBILITY of amazing advancements like AGIs, cyborgs, teleportation, radical life extension, group minds and so forth – and it posits a drive for intelligence to grow and expand -- but it doesn’t intrinsically imply any specific time-scale for these things to develop. The “Singularitarian” perspective, as I articulated it at the start of this book, adds an ingredient that’s very important from our contemporary human perspective: That not only are all these amazing things feasible and reasonably likely to happen— the serious fun and creative destruction may start this century, and at a certain point may start to unfold extremely fast relative to human experience.

Shaping the Singularity

Amazing things are going to happen this century— and as the title of this book says, I think they’re likely to happen *faster than you think*.

From the big picture perspective, the accelerating progress of science and technology is essentially unstoppable, as is the passage from humans on to engineered superintelligences. And from the Cosmist view, the particulars of how all this happens may not matter so much. It’s not clear how sensitively the medium-term future of superhuman superminds depends on the specifics of how the human Singularity goes down.

On the other hand, from the perspective of you and me as individual humans, the specific way that AGI emerges and the Singularity unfolds may matter a lot. For instance, it matters to me whether my 3 kids get their molecules absorbed into some super-AGI’s processing unit, or whether they get to choose their own future. I would like to see them have the choice whether to live on in legacy human form, while others explore transhuman domains; or to expand their minds through uploading or brain implants, and gradually become something more than human. What we do now, how we handle the development of AGI and other advanced technologies, might not impact the ultimate development of

intelligence throughout the Cosmos, but it may well impact what happens to ourselves and our loved ones during the coming transition.

My own feeling is that to make the transition as smooth as possible, we must develop a reasonably empathic, beneficial human-like AGI system as soon as possible – before other advanced technologies exacerbate a situation where AGI somehow goes awry in the early stages. I’ve tried to get this started with OpenCog, in primitive and simple ways. For instance, one can give a virtual agent positive reinforcement signals when it does something to help another agent in its virtual world (say, holding the door open or moving an object obstructing their path). In this sort of way one gradually lays a foundation for AGI systems with the desired sort of goals and values. As technology and science advance and powerful AGI systems get created, we will need to work with these systems to figure out the next steps forward— to figure out how to move on toward progressively smarter AGIs in a way that benefits everyone. It’s not likely to be easy, and it certainly involves a lot of uncharted territory.

But hey— as the ancient Chinese curse says, we live in interesting times!

And I for one am glad of it.

FURTHER READING

A Few Relevant Websites

Here are a few websites you may want to look at, related to the themes of the book, based on their content or links circa 2013. Some of these represent perspective I agree with, others perspectives I don't embrace but consider relevant and interesting – as always you'll need to use your own judgment! Omission of a site does not imply any negative evaluation of its contents; this is a partial and somewhat ad hoc list.

[links to be inserted]

- AGI Society
- Artificial General Intelligence Conference Series
- Future of Humanity Institute
- H+ Magazine
- Humanity+
- Kurzweilai.net
- Machine Intelligence Research Institute
- Next Big Future
- OpenCog.org
- Radical Futurism for Newbies
- Singularity Hub
- MORE

A Few Relevant Books

The number of books relevant to the ideas presented here is obviously tremendous. This is a short list of a handful of highly relevant works. Many works left off this short list are also very relevant and valuable.

[details to be inserted; list to be extended]

- The Age of Spiritual Machines
- A Cosmist Manifesto
- Mind Children
- The Phenomenon of Science
- The Self-Organizing Universe
- The Singularity is Near
- The Spike
- MORE

Some Online Background Reading

In 2011 I put together a webpage called *Radical Futurism for Newbies*, linking to a variety of articles by futurist visionaries and scientists, covering the gamut of critical aspects of transhumanist thinking. Some of the links may get stale, but if so the contents should still be locatable online via author and title. I don't agree with all the ideas at all these links – but I think it's all worth understanding.

[insert material from: http://wp.goertzel.org/?page_id=310]

Acknowledgements

This book summarizes some very important aspects of an intellectual and practical quest I've been pursuing for many decades – a quest to understand how mind works well enough to build an advanced intelligent system. While this quest of mine began in the early 1980s as an almost entirely solitary pursuit (inspired of course by many books from the library, recounting prior works and ideas of others), as time has passed it has become more and more of a social occupation. As a result, the number of people to whom I owe thanks for one or another aspect of this book is truly tremendous, and there's no possible way I can acknowledge them all here. All I can do is give a sampling of some of the folks who have been most important, and hope for the forgiveness of the many whom I will inevitably, inadvertently leave out.

Let me start with those who directly helped with the creation of this book itself. In the process I'll say a little bit about how the book originated.

In 2010 I had the idea to write a non-technical “trade” book on the future of technology, humanity and intelligence -- but I didn't feel I had the time to create one from scratch, so I asked my friends Stephan Bugaj and Lisa Rein to help out. Specifically, I asked them to help me update and re-focus Stephan's and my earlier future-of-tech book *The Path to Posthumanity*. The idea of updating P2P was gradually scrapped when we realized just how much rewriting would have to be done. So – in a series of often pretty interesting F2F and email discussions – the three of us began plotting a new book.

As it happened, that intended collaboration eventually fell apart – mostly because I realized that producing another broad-based book about the future would be a massive amount of work that I simply didn't have time for. Instead, I realized, what I just might BARELY have the time for, was to write a book more focused on *AGI* and its future, and to do it from a first-person, idiosyncratic perspective. And when I launched into working on this sort of book, it became clear that it was going to become more of a personal project than a three-way collaboration.

But for sure, even though Stephan, Lisa and I never wrote the broader book on the future that we were thinking of, the discussions that we had in the process of planning that hypothetical book, were instrumental in shaping this one.

The next step in the path to this book was triggered by a suggestion from my friend Meg Heath, who was staying at my house in Maryland for a while in 2011-2012 (during which time I was in the process of relocating to Hong Kong, but still living in Maryland part-time). Meg noted that I tended to explain complex ideas much more clearly orally than in writing. I thought, “Hmmm, she may have a point there. Maybe I should record the first draft of the book in audio form, and then get it transcribed and edit the transcription.”

That worked out reasonably well, especially since I was doing some solo commuting in Hong Kong, from the village where Ruiting and I lived to and from the office of my AI-based hedge fund Aidya, which was in a more urban area. It worked out well to record some of my thoughts about AGI and the future into my Iphone while driving. And after I had recorded a sufficiently large batch, I found a very competent transcriber named Mara Albea on oDesk, who transformed my ramblings into text.

I then got a writer I'd worked with on H+ Magazine before, Tim Giannuzzi, to clean up the text a bit – since Mara's transcription contained lots of umms and errs and informal speech constructions that wouldn't fit well in a book. And then came the hard work for me -- editing and dramatically rewriting

the result of Tim's cleanup – and adding a lot more text. Probably 40% of the text was written at this phase, whereas 60% came from the original audio recordings.

At this point I began working with Lisa Rein again, who helped edit some of the earlier chapters. But after doing some very useful work on the book, Lisa got busy with other pursuits, and so I brought in Darren Lurie, who did an amazingly careful and thorough job of editing the text, including some substantial structural changes. He also found a number of the pictures in the book, and wrote some of the captions. And then it was back to me to edit one more time, and add more things that Darren or I felt were missing.

That takes care of those who helped with the production of the book itself. The list of people who helped me develop the ideas described here is much, much longer, and would be impossible to give in any thorough way. I will just give a few very incomplete lists here...

My colleagues at futurist nonprofit Humanity+: XXXX

My colleagues at the AGI Society, who have helped e.g. with the AGI conference series: XXX

My colleagues on the OpenCog project: XXX

My colleagues at Novamente LLC: XXX

Some of my many colleagues at Intelligenesis Corp. / Webmind Inc.: XXX

My colleagues in the Addis AI Lab: XXX