Moving AI Out of its Infancy: Changing Our Preconceptions

Steve Grand

The other day, it was my turn to answer stupid questions about the movie *I, Robot*. "Do you think it's about time we started incorporating Asimov's three laws into real robots?" a journalist asked. I replied that Asimov's laws are about as relevant to real robotics as leechcraft is to modern medicine. Yes, before anyone writes me smug emails, I know that leeches are very useful in modern medicine, but I said "leechcraft." Leeches might be useful, but the paradigm of thought that originally led to their use is a ridiculous anachronism.

The same is true for Asimov's laws. Of course, he invented them as a plot device, useful precisely because of their many paradoxes and faults. Nevertheless, they belong to a bygone age: a world created by Boole and Babbage, in which people seriously thought that intelligence was some form of logical calculus. At the time, it seemed reasonable that we could give a robot explicit rules for every aspect of its behavior. We could "program" it to recognize when it was in danger of breaking any laws and could guarantee it wouldn't "break its programming" by doing anything wrong. Within a few years of Asimov's stories, the digital computer had been invented and people seriously set to work trying to program in such rules for intelligence.

How quaint! Luckily nobody believes this nowadays. Or do they?

Traditional AI: Intelligence not required

Explicit symbolic logic has faded from prominence, but the close coupling of AI and the digital computer, and of thought and the stepwise algorithm, seem about as strong and unquestioned as ever.

Of course there's connectionism, but this too is mired in false assumptions that date back a long way. And it seems to have dragged neuroscience down with it to the extent that we now seem unable to think about real brains without resorting to models that owe too much of their inspiration to the three-layer perceptron. Brains really aren't like that.

Traditional AI has excelled at solving certain kinds of problems. It can build machines that play chess but not ones that can pick up the pieces when they fall over. It can create machines that read text but not ones that can recognize objects from arbitrary angles. It can make systems that learn but not in any generally applicable way.

This isn't a criticism—traditional AI mostly follows John McCarthy's dictum that AI is about making machines do what humans use intelligence to do, and often this doesn't actually require the machines to show any intelligence at all. But for many tasks, especially in robotics, the ability to see, learn, and perform complex motor actions is a prerequisite that the traditional approach has utterly failed to fulfill. If robots are any kind of a threat to humanity, it's only because they tend to be heavy and fall over a lot. Even if a machine could contemplate murder, it wouldn't be able to pick up the knife or locate the victim.

In search of AI's periodic table

The New AI fares a little better at solving some of these supposedly lower-level tasks. But where good old-fashioned AI was inspired by the logical thought processes of advanced mathematicians, New AI is inspired by the nervous systems of the simplest invertebrates (and no, that's not the same thing). The snag is that these extreme bottom-up and top-down approaches don't meet in the middle. There's a huge gulf precisely where the interesting behavior lies. Neither approach tells us much about how to make machines that can perceive or make complex movements in the way that even the most primitive mammals can, yet these are the very competencies that robots so desperately need. Despite what some people seem to assume, you can't simply combine techniques from both approaches. A human being is not an ant with a natural language interface.

AI currently stands in relation to real intelligence much as alchemy once did to chemistry. Without alchemists, we'd never have developed chemistry, so I mean no insult by this. But until the discovery of the periodic table, everyone was essentially stabbing in the dark. Fundamentally, what we've learned over the past 50 years is a lot about how not to build intelligent machines, but we still haven't made the critical breakthrough. There's one class of machine that we know

1541-1672/04/$20.00 © 2004 IEEE
Published by the IEEE Computer Society

for sure can solve all these problems of perception and complex action: the mammalian brain. But we simply don't know its fundamental operating principles (although I'm quite certain it has some). We have no periodic table of neural function to help us see the underlying logic. Turing machines and neural networks were hopeful candidates for a theory of intelligence, but they simply don't cut the mustard.

Because of how science works, we have a tendency to hop on bandwagons, for the most part making only incremental improvements to ideas that already exist. But a beautifully polished and optimized bad idea is still a bad idea. The digital computer was inspired by one abstract view of the thought process, but it turns out to have been the wrong one as far as general intelligence is concerned. Connectionists tried to pay more attention to the neural hardware, but they did so within a paradigm drawn from electronic circuitry, causing them to make rash assumptions about the roles of nerve cells and synapses. Neither approach has worked, so we should abandon these paradigms and look for other models. What we need are new and radical ideas at the most fundamental level.

Po statements

Edward de Bono, the champion of lateral thinking, has a technique that he calls "Po." It involves making deliberately provocative statements ("the best place to sell ice cream is the North Pole") to shake us out of our preconceptions and find new paths. Suppose de Bono were to take up AI. What kinds of Po statements might he make?

I really couldn't say, but the following are some of the deliberate provocations that stimulate my own research. To my mind, they're nowhere near as radical as marketing ice cream to Eskimos, although some might think so. Nevertheless, I think they're sufficiently misaligned with established wisdom to suggest interesting new directions. I ask you to treat them in the spirit of Po: not as something to criticize but as ideas to run with just to see where they might lead.

Po 1: Brains exist to compensate for the slowness of nerves

AI sometimes has a terrible tendency to treat intelligence as a reactive, even passive, process. New AI has a particular mistrust of internal models and top-down mechanisms, arising from a justifiable disenchantment with symbolic representation. But for animals larger than a pinhead, prediction is an essential part of intelligent behavior, and reactive solutions simply won't do. Turning your eyes toward the point where a fast-moving object was when the light from it hit your retina will cause you to miss it by miles. Worse still, waiting until after the lion has actually eaten you isn't the best time to think about a response.

This principle applies universally. Indeed, perhaps an animal's intelligence is by definition proportional to its degree of predictive power. Relatively primitive animals live for the moment, but even they must be able to extrapolate their prey's movements or predict a social rival's likely response. Humans can form predictions many years into the future or many steps into a tree of possibilities. Intelligence is all about prediction. Brains exist fundamentally to ask "What next?" and, in some animals, "What if?"

If robots are any kind of a threat to humanity, it's only because they tend to be heavy and fall over a lot. Even if a machine could contemplate murder, it wouldn't be able to pick up the knife or locate the victim.

Both questions imply a mental model of the world—not a symbolic model or even an explicit one, but a model nonetheless. Without some means of fast-forwarding the present, it's impossible to anticipate the future, especially when that future is highly conditional. Extreme reactivists might disagree with this, but finding reactive alternatives often requires absurd contortions and flies in the face of a large body of evidence. Thinking about where and how the brain could develop, store, and use such a model or models can be a remarkably productive exercise when freed from any historical baggage. Furthermore, it can suggest unifying principles that extend from simple reflexes right through to conscious imagery.

Po 2: Brains don't make decisions

Brains simply try to reduce the tension between how things are and how we expect or would like them to be.

AI has shown a lamentable tendency to slide from reasonable observations to overstylized, formal solutions. Of course brains make decisions, but it doesn't follow that there's an explicit decision-making mechanism in the brain in the sense used in action-selection networks, for example. Much of what the brain does requires analog—"left hand down a bit"—kinds of responses and yet so many AI techniques (and quite a lot of behaviorist psychology) presume that decisions are all-or-nothing, discrete choices.

The brain must contain an anticipation of the world's future state to act in good time, and to construct this, it must also have a representation of the present (or more strictly the recent past) produced by the senses. So, at any one moment, the brain contains two complex vectors: one pattern of nerve activity representing how things are, and the other representing how things might be soon. It makes good neurological and psychological sense to assume that these two state vectors map onto the same brain territory. If so, the two can be directly compared point for point, and the comparison can yield useful consequences. My present research started with the assumption that brains are in essence arrays of servomotors—each comparing one pair of "values" from the two state vectors and producing an output designed to reduce the difference between them, either by driving muscles or by becoming the "intention" value for another servo in the network.

When you think about it, an intention is a kind of prediction about how the world will look if things go as planned. The difference between how things are now and how we would like them to be tells us something useful about what we need to do. Our goal is to bring the state of the world in line with our prediction, which requires us to perform a servo action. But equally, sometimes we need to bring our predictions back in line with reality. This is what happens when we update our beliefs in light of new information. Beliefs, hypotheses, expectations, attentions, plans, and intentions all become the same thing when seen in this light.

Po 3: Brains perform coordinate transforms

If brains are networks of servos, each servo must operate in a particular coordinate space (for example, retinotopic, somatotopic, or tonotopic), and the links between them must therefore carry out a conversion from one space to another. If something that we see is to have an effect on what we do, then it follows that information mapped in retinal coordinates eventually must produce changes in the brain that are mapped in muscle coordinates.

The more I think about this principle, the more I find that we can describe other kinds of mental process as coordinate transforms too—even object recognition and abstract symbol manipulation. One especially interesting example might explain how we can recognize shapes by sight or touch, regardless of their scale, rotation, and position on the visual field or the skin. Such invariance is easily the most striking and challenging aspect of perception (and, indeed, of motor action). Without the ability to replicate this property, we have no hope of making robots that can see like animals do.

So, is there a coordinate frame in which a banana, for the sake of argument, looks exactly the same shape, regardless of its position, orientation, and scale? Yes. It's a rather abstract frame, but try this: Imagine the image of a banana falling on your retina. Now mentally project yourself until you are inside the banana, looking outward. What you have just performed is a conversion from eye-centered coordinates into banana-centered coordinates. From inside, the banana remains exactly the same shape, regardless of its location and orientation with respect to your original viewpoint. So if brains can perform on-the-fly transforms from egocentric to object-centered coordinate space, they have the means to develop visual invariance.

Exactly how this might happen is an unsolved problem, but it's something I've been working on with a modicum of success—enough to suggest that it's a meaningful idea. Significantly, if a general mechanism can be found, it'll bring the two key visual data streams (the "where" pathway of the parietal lobes, and the "what" pathway of the temporal lobes) into a common level of explanation.

Po 4: Nervous tissue is a new state of matter

The more I think about concepts such as servos and coordinate spaces, the more irrelevant the traditional view of the neuron seems to become. The stereotypical neural network is a sparsely connected, discrete signaling system, but real neurons are nothing of the kind. They're so densely interconnected and leaky that an unbiased appraisal of the facts would suggest that nervous tissue is more like a wobbly jelly than a printed circuit board. Signals spread out rapidly, and large-scale phenomena such as waves build up on the neural surface. Yet at the same time, neurons have the ability to make changes to signal propagation on a very fine scale. It seems to me that nervous tissue is a substance with some of the properties of a discrete network of wires and some of a continuous solid.

On such a medium, patterns of nerve activity need to be interpreted differently than in conventional models based on very small networks. Perhaps it isn't the neurons that perform the computations in the brain at all. Perhaps they provide the surface upon which the *patterns of nerve activity* perform computations (see Steven Lehar's "Harmonic Resonance Theory" at http://cns-alumni.bu.edu/~slehar/webstuff/hr1/hr1.html for one interpretation of such second-order computation). You wouldn't learn anything about the behavior of the Niagara Falls by studying a handful of water molecules, so it seems ridiculous to try and mimic nervous systems using simulations built from 16 neurons. If the scale of activity is as large as I suspect, we simply won't be able to see the wood for all the trees.

Po 5: The more complex the robot, the easier it is to make progress

A similar argument about scale applies to robots and to the segregation of disciplines in AI. Toy environments are often far too stylized and reduced to capture the essential features of a problem, and with robots, that's especially true. How intelligent would *you* have become if you'd been born equipped with only two wheels and a handful of bump sensors?

But there are other reasons to think big. Hearing, seeing, planning, and moving seem on the surface to be radically different problems, and yet the brain tissues involved in each of these processes are fundamentally similar. Motor cortex looks slightly different from primary visual cortex, but the essential architecture is the same and most of the differences are likely to be a consequence of adaptation. So if the same brain architecture can perform all these different tasks, there must be a level of description at which they're the same task.

Moreover, so much of development and learning is multimodal. How can we learn to see depth unless we have the ability to reach out and touch things to confirm how far away they are? How can we learn to reach out and touch things unless we can see their depth? Learning is a process of integration, correlation, and confirmation between all the senses and motor systems at once, so we need to study them together.

Lucy: Building a somebody

So, to help me develop my Po ideas about possible new neural computing architectures, I decided to build as complex a robot as my limited resources would allow (see Figure 1). Her name is Lucy, and she's ostensibly a robot orangutan, although the similarity is minimal because I'm not much of an artist. Nevertheless, I gave her a face and refer to her as "her" rather than "it" to remind me that I'm trying to make a somebody, not a something—a complete integrated organism.

Physically, she has vision, hearing, proprioception, a virtual model of the vocal tract, and enough degrees of freedom to make movement a challenge. My goal is to build Lucy a brain from scratch, facing the same problems that nature must have faced, armed with the same tools and equipment (as far as I can manage). What I'm trying to find is a common level of description that marries all the apparently disparate tasks that brains carry out. I'm essentially on the lookout for a protomachine: a generalized neural architecture that can spontaneously self-organize into a variety of specialized machines, driven only by the nature of the signals supplied to it.

So far, Lucy's only party trick is that she's learned how to point at bananas. I hold up an apple and a banana and she can point at the banana. Impressive, huh? And all that this feat requires is a neural network composed of around 50,000 complex neurons. Not the most efficient way to recognize a banana, when "point at the yellow bit" would suffice. But I don't care, because I'm only using digital computers as an interim solution. My aim is to find radically new kinds of computing devices that work more like I think brains do.

Lucy's virtual brain is composed of a series of neural surfaces, each performing a different aspect of looking, recognizing, or pointing toward things. The important point is that all these surfaces have a lot in common, despite the differences in their function. At some level, we can describe each as a servomotor, which performs some sort of coordinate transform and computes its results using the properties of large-scale patterns of nerve activity.

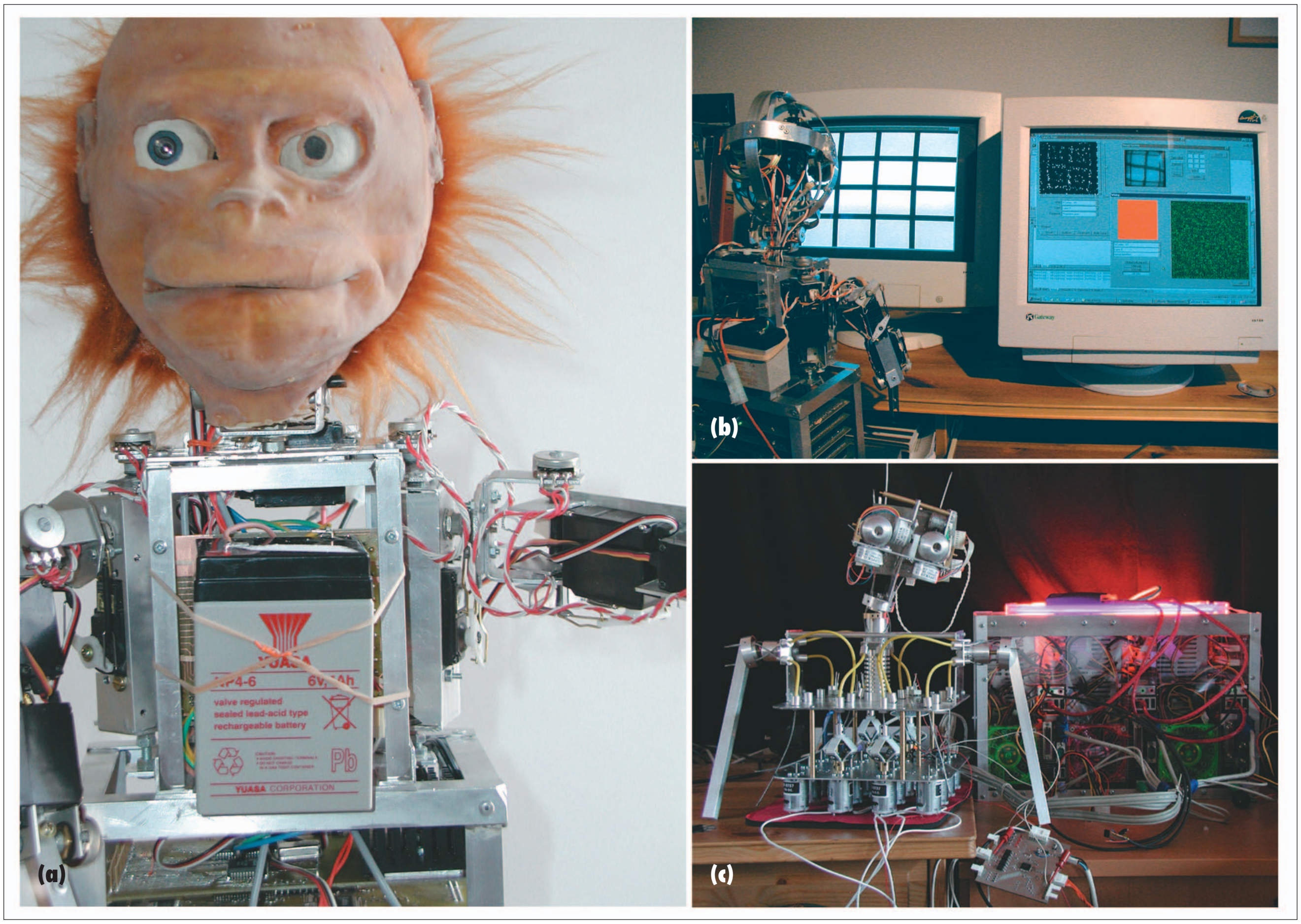
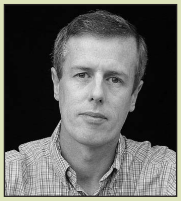


Figure 1. (a) Lucy Mk. I; (b) Lucy Mk. I staring at patterns on a computer screen while her visual cortex spontaneously organizes itself into an orientation detecting system (as seen in the "brain scans" on the right-hand monitor); (c) Lucy Mk. II.

There's an awfully long way to go until these ideas gel into a unified and powerful mechanism capable of being used in real applications. But all adults start out as babies, and I don't see any reason why AI should be different. I'm convinced that there *is* such a universal architecture for creating mammal-like general intelligence and that this bears little resemblance to existing neural networks and none whatsoever to the concepts underlying the digital computer. Until we find such a radical new way forward, I don't think we'll ever build robots for which Asimov's three laws of robotics have the slightest relevance.

I'd like to point out that I'm sure many of the ideas I've outlined have already been proposed in some form by other researchers. If so, I apologize for not citing them. My aim isn't to plagiarize; it's simply that I pay no attention to the literature. One of the best ways to kill off a promising line of thought is to say, "Well, I know that so-and-so tried that in 1978 and it didn't work." In reality, the likelihood is that so-and-so didn't have exactly the same ideas in mind, wasn't thinking about them in precisely the same way, and wasn't driven by quite the same motives, so it's better not to know.

I've no doubt Edward de Bono would agree that most breakthroughs arise when people doggedly plow their own furrow, unknowingly attempting things that wiser people "know" to be impossible. Ignorance can therefore be a huge asset, and as an unfunded amateur, I have no obligation to stick to the rules and etiquette of professional science, so the less I know about other people's ideas, the better. In fact, if I were you, I wouldn't be reading this article at all—it would only color my thoughts and reduce the potential for novelty. But perhaps the final paragraph wasn't the ideal place to mention this.



Steve Grand is an independent scientist. He's also an honorary research fellow at Cardiff University and Bath University. He's author of *Growing Up with Lucy: How to Build an Android in Twenty Easy Steps* (Weidenfeld & Nicolson, 2004). Contact him at steve@cyberlife-research.com.