Artificial Intelligence and the Many Faces of Reason² (Refereed book chapter) in S. Stich and T. Warfield (eds) THE BLACKWELL GUIDE TO PHILOSOPHY OF MIND (BLACKWELL, 2003)

Artificial Intelligence and The Many Faces of Reason

Andy Clark
Philosophy/Neuroscience/Psychology Program
Department of Philosophy
Campus Box 1073
Washington University in St. Louis
St. Louis, MO 63130
USA
andy@twinearth.wustl.edu

Note: Correspondence address after June 1st, 2000 will be
School of Cognitive and Computing Sciences
University of Sussex
Brighton
BN1 9QH
England
UNITED KINGDOM

The old email address will remain active.

0. Pulling A Thread

I shall focus this discussion on one small thread in the increasingly complex weave of Artificial Intelligence and Philosophy of Mind: the attempt to explain how rational thought is mechanically possible. This is, historically, the crucial place where Artificial Intelligence meets Philosophy of Mind. But it is, I shall argue, a place in flux. For our conceptions of what rational thought and reason *are*, and of what kinds of mechanism might explain them, are in a state of transition. To get a sense of this sea change, I shall compare several visions and approaches, starting with what might be termed the Turing-Fodor conception of mechanical reason, proceeding through connectionism with its skill-based model of reason, then moving to issues arising from robotics, neuroscientific studies of emotion and reason, and work on "ecological rationality". As we shall see there is probably both more, and less, to human rationality than originally met the eye.

First, though, the basic (and I do mean basic) story…

1. The Core Idea, Classically Morphed

One core idea, common to *all* the approaches I'll consider today, is that sometimes form can do duty for meaning. This is surely the central insight upon which all attempts to give a mechanical account of reason are based. Broadly understood, it is this same trick that is at work in logic, in the Turing Machine, in symbolic Artificial Intelligence, in connectionist artificial intelligence, and even in "anti-representationalist" robotics. The trick is to organize and orchestrate some set of non-semantically specifiable properties or features so that a device thus built, in a suitable environment, can end up displaying "semantic good behavior". The term 'semantic good behaviors covers, intentionally, a wide variety of things. It covers the capacity to carry out deductive inferences, to make good guesses, to behave appropriately upon receipt of an input or stimulus, and so on. Anything that (crudely put) *looks like it knows what it is doing*, is exhibiting semantic good behavior: cases include the logician who infers –A from (C? - A,C), the person who chooses to take out an umbrella because they believe it will rain and desire to stay dry, the dog who chooses the food rather that the toxin, the robot that recovers its balance and keeps on walking after one leg is damaged. There's a *lot* of semantic good behavior around, and we understand some of it a whole lot better than the rest. Where, though, does *reason* come into the picture?

Reason-governed behavior is, arguably at least, a special subset of what I am calling semantic good behavior. It is Jerry Fodor's view, for example, that is was not until the work of Turing that we began to have a sense of how *rationality* (which I'll assume to mean reason-governed behavior) could be mechanically possible (for a nice capsule statement, see Fodor (1998 p. 204 – 205)). Formal logic showed us that truth preservation could be ensured simply by attending to form, not meaning. B follows from A & B *regardless* of what A means and what B means, and if your keep to rules defined over the shapes of symbols and connectives you will never infer a falsehood from true premises, even if you have no idea what either the premises or the conclusions are about. Turing, as Fodor notes, showed that for all such formally ("by shape") specifiable routines, a well-programmed machine could replace the human.

It is at about this point that what was initially just an assertion of physicalist faith (that somehow or other, semantic good behavior has always and everywhere an explanatorily sufficient material base) morphs into a genuine research program targeting reason-governed behavior. The idea, rapidly enshrined in the research program of classical, symbolic Artificial Intelligence, was that reason could be mechanically explained as the operation of appropriate computational processes on symbols, where symbols are non-semantically individable items (items typed by form, shape, voltage, whatever) and computational processes are mechanical, automatic processes that recognize, write and amend symbols in accordance with rules (which themselves, up to a certain point, can be expressed as symbols). In such systems, as Haugeland (1981, p. 23) famously remarks, “if you take care of the syntax [the non-semantic features and properties] the semantics will take care of itself”. The core idea, as viewed through the lens of both Turing’s remarkable achievements and then further developments in classical Artificial Intelligence, thus began to look both more concrete, and less general. It became the idea, in Fodor’s words, that:

> “…some, at least, of what makes minds rational is their ability to perform computations on thoughts; when thoughts…are assumed to be syntactically structured, and where ‘computation” means formal operations in the manner of Turing”
>
> Fodor (1998) p.205.

The general idea of using form (broadly construed) to do duty for meaning, thus gently morphed into the Turing Machine dominated vision of reading, writing and transposing symbols: a vision which found full expression in early work in Artificial Intelligence. Here we encounter Newell and Simon’s (1976) depiction of intelligence as grounded in the operations of so-called *physical symbol systems*: systems in which non-semantically identifiable entities act as the vehicles of specific contents (thus becoming “symbols”) and are subject to a variety of familiar operations (typically copying, combining, creating and destroying the symbols, according to instructions). For example, the story understanding program of Schank (1975) used a special event description language to encode the kind of background knowledge needed to respond sensibly to questions about simple stories, thus developing a symbolic data-base to help it “fill in” the missing details.

Considered as stories about how rational, reason-guided thought is mechanically possible, the classical approach thus displays a satisfying directness. It explains semantically sensible thought-transitions (“they enjoyed the meal, so they probably left a tip,” “it’s raining, I hate the rain, so I’ll take an umbrella”) by imagining that each participating thought has an inner symbolic echo, and that these inner echoes share relevant aspects of the structure of the thought. As a result, syntax-sensitive processes can regulate processes of inference (thought-to-thought transitions) in ways that respect semantic relations between the thoughts.

2. The Core Idea, Non-classically Morphed

The idea that reason-guided thought transitions are grounded in syntactically driven operations on inner symbol strings has a famous competitor. The competing idea, favored by (many) researchers working with artificial neural networks, is that reason-guided thought-transitions are grounded in the vector-to-vector transformations supported by a parallel web of simple processing elements. A proper expression of the full details of this contrast is beyond the scope of this paper (see Clark (1989)(1993) for my best attempts). But we can at least note one especially relevant point of (I think) genuine contrast. It concerns what I’ll call the “best targets” of the two approaches. For classical (Turing Machine-like) Artificial Intelligence, the best targets are rational inferences that can be displayed and modeled in *sentential space*. By ‘sentential space’ I mean an abstract space populated by meaning-carrying structures (interpreted syntactic items) that share the logical form of sentences: sequential strings of meaningful elements, in which different kinds of syntactic item reliably stand for different things, and in which the overall meaning is a function of the items (tokens) and their sequential order, including the modifying effects of other tokens (e.g. the “not” in “it is not raining”). Rational inferences that can be satisfyingly reconstructed in sentential space include all of Fodor’s favorite examples (about choosing to take the umbrella, etc.), all cases of deductive inference defined over sentential expressions, and all cases of abductive inference (basically, good guessing) in which the link between premises and conclusions can be made by the creative retrieval of deployment of additional sentences (as in Schank’s story understanding program mentioned earlier).

The best targets for the artificial neural network approach, by contrast, are various species of reasonable ‘inference’ in which the inputs are broadly speaking perceptual and the outputs are (often) broadly speaking motoric. Reasonable inferences of this kind are implicit in, for example, the cat’s rapid assessment of the load-bearing capacity of a branch, leading to a swift and elegant leap to a more secure resting point, or the handwriting expert’s rapid intuitive conviction that the signature is a forgery, a conviction typically achieved in advance of the conscious isolation of specific tell-tale signs.

This is not to say, however, that the connectionist approach is limited to the perceptuo-motor domain. Rather, the point is that its take on rational inference (and, more broadly, on rational choice) is structurally continuous with its take on perceptuo-motor skill. Reasoning and inference are reconstructed, *on all levels*, as (roughly speaking) processes of pattern-completion and pattern-evolution carried out by cascades of vector-to-vector transformations between populations of simple processing units. For example, a network exposed to an input depicting the visual features of a red-spotted young human face may learn to produce as output a pattern of activity corresponding to a diagnosis of measles. This diagnosis may lead, via a similar mechanism, to a prescription of penicillin. The vector-to-vector transformations involved are perfectly continuous(on this model) with those by which we perform more basic acts of recognition and control, as when we recognize a familiar face or co-ordinate visual proprioceptive inputs in walking. Such pattern completing processes, carried out in networks of simple processing units connected by numerically weighted links, are *prima facie* quite unlike the sentential Artificial Intelligence models in which a medical judgment (for example) might depend on the consultation of a stored set of rules and principles. One important source of the difference lies in the way the connectionist system typically *acquires* the connection weights that act both as knowledge-store and processing-engine. Such weightings are acquired by exposing the system to a wide range of exemplars (training instances): a regime which leads, courtesy of the special learning rules deployed, to the development of a *prototype-dominated knowledge base* (see Churchland (1989)). What this means in practice is that the system learns to 'think about' a domain in terms of the most salient features of a body of exemplar cases, and that its responses, judgments and actions are guided by the perceived similarity of the current case to the patterns of features and responses most characteristic of the exemplars. And what this means, in turn, is that what such a system knows is seldom, if ever, neatly expressible as a set of sentences, rules, or propositions about the domain. Making the expert medical judgment, on this model, has more in common with knowing how to ride a bicycle than with consulting a set of rules in a symbolic date-base. A well-tuned connectionist network may thus issue judgments that are rationally appropriate but that nonetheless resist quasi-deductive sentential reconstruction as the conclusion of an argument that takes symbolic expressions as its premises. Such appropriate responses and judgments are, on this view, the fundament of reason, and of rationality. Linguaform argument and inference is depicted as just a special case of this general prototype-based reasoning capacity, different only in that the target and training domain here involves the symbol strings of public speech and text.

Connectionism and classicism thus differ (at least in the characteristic incarnations I am considering) in their visions of reason itself. The latter depicts reason as, at root, symbol-guided state transitions in quasi-linguistic space. The former depicts reason as, at root, the development of prototype-style knowledge guiding vector-to-vector transformations in the same kinds of (typically) non-sentential space that also underlie perceptuo-motor response. Beneath this contrast, however, lies a significant agreement. Both camps agree that rational thoughts and actions involve the use of inner resources to represent salient states of affairs, and the use of transformative operations (keyed to non-semantic features of those internal representations) designed to yield further representations (in a cascade of vector-to-vector transformations in the connectionist case) and, ultimately, action.

3. Robotics: Beyond The Core?

Is it perhaps possible to explain reasoned action without appeal to inner, form-based vehicles of meaning at all? Might internal representations be tools we can live without?

Consider the humble house-fly. Marr (1982, p. 32-33, reported by McLamrock (1995) p. 85) notes that the fly gets by without in any sense encoding the knowledge that the action of flying requires the command to flap your wings. Instead, the fly’s feet, when not in contact with ground, automatically activate the wings. The decision to jump thus automatically results (via abolition of foot contact) in the flapping of wings.

Now imagine such circuitry multiplied. Suppose the “decision to jump” is *itself* by-passed by e.g. directly wiring a “looming shadow” detector to the neural command for jumping. And imagine that the looming shadow detector is itself nothing but a dumb routine that uses the raw outputs of visual cells to compute some simple, perceptual invariant. Finally, imagine if you will a whole simple creature, made up of a fairly large number of such basic, automatic routines, but with the routines themselves orchestrated –

by exactly the same kind of tricks – so that they turn each other on and off at (generally speaking) ecologically appropriate moments. For example, a 'consume food" routine may be overridden by the "something looming-so-jump" routine, which in turn causes the "flap wings" routine, and so on. What you have imagined is, coarsely but not inaccurately, the kind of "subsumption architecture" favored by robotists such as Rodney Brooks (1991), and responsible for provocative paper titles such as "Intelligence Without Representation" and slogans (now co-opted as movie titles!) such as "Fast, Cheap and Out of Control."

It is not *at all* obvious, however, that such a story could (even in principle) be simply scaled-up so as to give us "rationality without representation". For one thing, it is not obvious when we should say of some complex inner states that it constitutes at least *some kind* of representation of events, or states of affairs. The house-fly wing-flapping routine looks like a simple reflex, yet even here there is room for someone to suggest that, given the evolutionary history of the reflex circuit, certain states of that circuit (the ones activated by the breaking of foot-surface contact) represent the fact that the feet have left the surface. What Brooks and others are really suggesting, it often seems, is rather the absences of a certain *type* of internal representation viz the broadly linguaform representations favored by classical Artificial Intelligence.

A more fundamental difficulty, however, (which goes well beyond the vagueness of the term "internal representation") concerns the *kinds* of behavior that can plausibly be explained by any complex of reflex-like mechanisms. The problematic cases here are obviously deliberative reason and abstract thought. The kinds of behavior that might be involved include planning next years family vacation, thinking about U. S. gun control issues (e.g. "should gun manufacturers be held responsible for producing more guns than the known *legal* market requires?"), using mental images to count the number of windows in your spanish apartment while relaxing on the river Thames, and so on. These cases are by no means all of a piece. But they share at least one common characteristic: they are all “represention hungry” (to use a term from Clark and Toribio (1994)) in quite a strong sense. All these cases, on the face of it, require the brain to use internal *stand-ins* for external states of affairs, where a “stand-in”, in this strong sense (see Clark & Grush (1999)) is an item designed not just to carry information about some state of affairs (in the way that, e.g., the inner circuit might carry information about the breaking of foot-surface contact in the fly) but to allow the system to key its behavior to features of specific states of affairs *even in the absence of direct physical connection*. A system which must coordinate it’s activity with the distal (the windows in my spanish apartment) and the non-existent (the monster in the tool-shed) is thus a good candidate for the use of (strong) internal representations: inner states which are meant to act as full-blooded stand-ins, not just as ambient information-carriers. (For some excellent discussion of the topics of connection and disconnection, see B. C. Smith (1996). By contrast, nearly all (but see Stein (1996) and Beer (2000)) the cases typically invoked to show representation-free adaptive response are cases in which the relevant behavior is continuously driven by, and modified by, ambient input from the states of affairs to which the behavior is keyed.

Rational behavior is, in some sense, behavior that is guided by, or sensitive to, reasons. Intuitively, this seems to involve some capacity to step back, and assess the options; to foresee the consequences, and to act accordingly. But *this* vision of rationality (‘deliberative rationality’) places rational action squarely in the “representation-hungry” box. For future consequences, clearly, cannot directly guide current action (in the way that, say an ambient light source may directly guide a photosensitive robot). Such consequences will be effective only to the extent that the system uses something else to stand-in for those consequences during the process of reasoning. And that, at least on the face of it, requires the use of internal representations in some fairly robust sense.

4. Emotions and Reason

A mechanical explanation of our capacities to display reason-guided behavior cannot, it seems, afford to dispense with the most basic notion of inner stand-ins capable of directing behavior and inference in the absence of the events and states of affairs concerned. Work in connectionism and real-world robotics is best viewed (I believe) as expanding our conceptions of the possible *nature* of such stand-ins, and as highlighting the many ways in which bodily and environmental structures, motion, and active intervention may all serve to transform the *problems* that the brain needs to solve. The use of pen and paper, for example, may greatly alter the problems that the brain needs to solve when confronting complex arithmetical tasks, when planning a long-term strategy, and even when reasoning about gun control. But such transformations do not by-pass the need for internal structure-sensitive operations defined over inner content-bearing vehicles: rather, they re-shape the problems that such an inner economy needs to solve.

The stress on reason-sensitive thought and inference can, however, blind us to the crucial importance of a further dimension of human cognition. For human reason is tightly, perhaps inextricably, interwoven with human emotion. Doing justice to this significant interaction is one of the two major challenges for the next generation of Artificial Intelligence models.

Emotions were long regarded (at least in a broadly Kantian tradition) as the enemy of reason. And we certainly do speak of (for example) judgments being clouded by envy, acts as being driven by short-lived bursts of fury and passion rather than by reasoned reflection, and so on. It is becoming increasingly clear, however, that the normal contributions of emotion to rational response are far from detrimental. They are, in fact, best seen as part of the mechanism of reason itself. Consider, to take a famous example, the case of Phineas Gage. Gage was a 19th century railway worker whose brain was damaged when an iron rod was driven through his skull in an explosion. Despite extensive damage to prefrontal cortex, the injury left Gage's language, motor skills, and basic reasoning abilities intact. It seemed as if he had escaped all cognitive compromise. Over, subsequent years, however, this proved sadly incorrect. Gage's personal and professional life took noticeable turns for the worse. He lost jobs, got into fights, failed to plan for the future and to abide by normal conventions of social conduct, became a different and markedly less successful person. The explanation, according to H. Damasio, et. al. (1996) was that the damage to prefrontal cortex had interfered with a system of (what they termed) "somatic markers" – brain states that tie the image/trace of an event to a kind of gut reaction (aversion or attraction, according to the outcome). This marker system operates automatically (in normal subjects) influencing both on-the-spot response and the array of options that we initially generate for further consideration and reflection. It is active also – and crucially- when we imagine an event or possible action, yielding a positive or negative affective signal that manifests itself in (among other things) galvanic skin response. Gage, it is hypothesized, would have lacked such responses, and would not have had his reasoning and deliberations constrained by the automatic option-pruning and choice-influencing operations of the somatic marker system gradually acquired during his lifetime's experience of social and professional action. Contemporary studies seem to confirm and clarify this broad picture. E. V. R. (a patient displaying similar ventromedial frontal damage) shares Gage's profile. Though scoring well on standard I.Q. and reasoning tests, E. V. R. likewise lost control of his professional and social life. In an interesting series of experiments (Bechera, Damasio, *et al* (1997)) normal controls and prefrontally lesioned patients played a card game involving (unbeknownst to the subjects) two winning decks and two losing decks.

Subjects could choose which deck (A, B, C, or D) to select cards from. After a little play, the normal controls fix on the better decks (smaller immediate rewards, but less secure penalties and more reliable long-term) and rapidly show a heightened galvanic skin response when reaching for the “bad” decks. This skin response, interestingly, appears before the subjects could articulate any reasons for preferring the better decks. E. V. R., by contrast, shows no such skin response. And this absence of somatic cues seems to interfere with his capacity to choose the better decks *even once his conscious mind has figured it all out* – he will know that A and B are losing decks, yet continue to favor them during play.

There is obviously much to discuss here. Are these cases best understood, as P. S. Churchland ((1998) p. 241) suggests, as arising from “the inability of emotions to affect [the patient’s] reason and decision-making”. Or is it a case of *inappropriate* emotional involvement – the triumph of short-term reward over deferred (but greater) gratification. Perhaps these are not really incompatible: either way it is the lack of the on-the-spot unconscious negative responses (evidenced by the flat galvanic skin responses) that opens the door to cognitive error.

Human reason, it seems fair to conclude, is not best conceived as the operation of an emotionless logic engine occasionally locked into combat with emotional outbursts. Instead, truly rational behavior (in humans) is the result of a complex and iterated series of interactions in which deliberative reason and subtle (often quite unconscious) affect-laden responses conspire to guide action and choice. Emotional elements (at least as suggested by the somatic marker hypothesis) function, in fact, to help rational choice operate across temporal disconnections. Somatic markers thus play a role deeply analogous to internal representations (broadly construed); they allow us to reason projectively, on the basis of past experience. What could be more appropriately deemed part of the mechanism of reason itself than something that allows us to imaginatively probe the future, using the hard-won knowledge of a lifetimes choices and experiences all neatly distilled into a network of automatic affective reverberations?

5. Global Reasoning

A further source of complication concerns what Fodor ((1983) p. 111) calls “global properties of belief systems”. Artificial Intelligence according to Fodor, confronts a special problem hereabouts. For the Turing Machine model of rational inference (recall section 1 above) is said to be irredeemably *local*. It is great at explaining how the thought (syntactically tokened) that it is raining gives way to the thought that an umbrella is indicated. It is great, too, at explaining (given a few classical assumptions – see Fodor and Pylyshyn (1998)) why the space of possible thoughts (for an individual) exhibits a certain kind of closure under recombination – the property of ‘systematicity’, wherein those who can think aRb typically also think bRa, and so on. But where current Artificial Intelligence based models crash and burn, Fodor insists, is when confronting various forms of more globally sensitive inference. For example, cases of abductive inference in which the best explanation for some event might be hidden anywhere in the entire knowledge base of the system: a knowledge-based deemed too large by far to succumb to any process of exhaustive search. Fodor rejects classical attempts to get around this problem by the use of heuristics and simplifying assumptions (such as the use of “frames” – see Minsky (1975), Fodor (1983) p. 116) arguing that this simply relocates the problem as a problem of “executive control” viz how to find the *right* frames (or whatever) at the right time. Since even the decision to take the umbrella against the rain is potentially sensitivity to countervailing information coming from anywhere in the knowledge base, Fodor is actually left with a model of mechanical rationality which (as far as I can see) can have nothing to say about any genuine but non-deductive case of reasoning whatsoever. The Fodor-Turing model of rational mechanism works best, as Fodor frequently seems to admits, only in the domain of “informationaly encapsulated systems" – typically, perceptual systems that process a restricted range of input signals in a way allegedly insensitive to all forms of top-down knowledge-driven inference. Hardly the seat of reason, one cannot help but feel.

Give this pessimistic scenario (enshrined in Fodor's "first law of the non-existence of cognitive science": "the more global…a cognitive process is, the less anybody understands it. Very global processes…aren't understood at all" Fodor (1983) p. 107), it is not surprising to find some theorists (Clark (1993) p.111, Churchland (1989) p. 178) arguing for connectionist approaches as one solution to this problem of "globally sensitive reason". Such approaches are independently rejected by Fodor for failing to account for systematicity and local syntax-sensitive inference. But it now seems to me (though this is a long story – see Clark (in progress)) that the problem of global abductive inference really does affect connectionist approaches too. Very roughly, it emerges therein as a problem of routing and searching: a question of how to use information, which could be drawn from anywhere in the knowledge-base, to sculpt and redirect the flow of processing itself, ensuring that the right input probes are processed by the right neural sub-populations at the right times.

Churchland (1989) and Clark (1993) depict this problem as solved (in the connectionist setting) because "relevant aspects of the creature's total information are automatically accessed by the coded stimuli themselves" (Churchland, op cit p. 187). And certainly, input probes will (recall section 2 above) automatically activate the prototypes that best fit the probe, along whatever stimulus dimensions are represented. But this, is at best a first step in the process of rational responsiveness. For having found these best syntactic fits (for this is still, ultimately, a form-driven process) it is necessary to see if crucially important information is stored elsewhere, unaccessed due to lack of surface matching to the probe. And it is this step which, I think, does most of the work in the types of cases with which Fodor is (properly) concerned.

The good news, which I make much of in Clark (in progress) but cannot pursue here, is that this second step now looks potentially computationally tractable, thanks to an odd combination of neuro-connectionist research and an innovative "second-order" search procedure developed for use on the world wide web (Kleinberg (1999)). The idea is to combine a first pass (dumb, pattern-matching, syntax-based) search with a follow-up search based on the patterns of connections into and away from the elements identified on the first pass. But the point, for present purposes, is simply to acknowledge the special problems that truly globally sensitive processing currently presents to all existing models of the neural computations underlying human reason.

6. Fast and Frugal Heuristics

It might reasonably be objected, however, that this whole vision of human rationality is wildly inflated. Very often, we *don't* manage to access the relevant items of knowledge; very often, we *don't* choose that which makes us happiest, or most successful; we even (go on, admit it) make errors in simple logic. What is nonetheless surprising is that we very often do as well as we do. The explanation, according to recent theories of "ecological rationality" is our (brains) use of simple, short-cut strategies designed to yield good results given the specific constraints and opportunities that characterize the typical contexts of human learning and human evolution. A quick example is the so-called "recognition heuristic". If you ask me which city has the largest population, San Diego or San Antonio, I may well assume San Diego, simply because I have *heard of* San Diego. Should I recognize both names, I might deploy a different fast and frugal heuristic, checking for other cues. Maybe I think a good cue is "have I heard of their symphony?", and so on. The point is that I don't try any *harder* than that. There may be multiple small cues and indicators, which I could try to "factor in". But doing so, according to an impressive body of recent research (see e.g. Chase, Hertwig and Gigerenzer (1998)) is likely to be both time-consuming and (here's the cruncher) unproductive. I'll probably choose *worse* by trying to replace the fast and frugal heuristic with something slower and (apparently) wiser.

It is not yet clear how (exactly) this important body of research should impact our vision of just *what* you need to explain in order to explain how rationality is mechanically possible. A likely alliance might see fans of robotics and Artificial Life based approaches (section 3) using relatively simple neural network controllers (section 2) to learn fast and frugal heuristics that maximally exploit local opportunities and structures. The somatic marker mechanism (section 4), might be conceived as, in a sense, implementing just another kind of fast and frugal heuristic enabling current decision-making to cheaply profit from past experience. Under such an onslaught, it is possible that much of the worry about global abductive inference (section 5) simply dissolves. My own view, as stated above, is that something of the puzzle remains. But the solution I favor (see Clark, in progress) can *itself* be seen as a special instance of a fast and frugal heuristic: a cheap procedure that replaces global content-based search with something else (the second pass, connectivity-pattern based search mentioned earlier).

7. Conclusions: Moving Targets and Multiple Technologies

Rationality, we have now seen, involves a whole lot more, and a whole lot less, than originally met the eye. It involves a whole lot more than local, syntax-based inference defined over tractable sets of quasi-sentential encodings. Even Fodor admits this - or at least, he admits that it is not yet obvious how to explain global abductive inference using such resources. It also involves a whole lot more than (as it were) the dispassionate deployment of information in the service of goals. For human reason seems to depend on a delicate interplay in which emotional responses (often unconscious ones) help sift our options and bias our choices in ways which *enhance* our capacities of fluent, reasoned, rational response. These emotional systems, I have argued, are usefully seen as a kind of wonderfully distilled store of hard-won knowledge concerning a lifetime’s experiences of choosing and acting.

But rationality may also involve significantly *less* than we tend to think. Perhaps human rationality (and I an taking that as our constant target) is essentially a quick-and-dirty compromise forged in the heat of our ecological surround. Fast and frugal heuristics, geared to making the most of the cheapest cues that allow us to get by, may be as close as nature usually gets to the space of reasons. Work in robotics and connectionism further contributes to this vision of less as more, as features of body and world are exploited to press maximal benefit from basic capacities of on-board, prototype-based reasoning. Even the bugbear of global abductive reason, it was hinted, just might succumb to some wily combination of fast and frugal heuristics and simple syntactic search.

Where then does this leave the reputedly fundamental question “how is rationality mechanically possible?”. It leaves it, I think, at an important crossroads, uncertainly poised between the old and the new. If (as I believe) the research programs described in sections 3-7 are each tackling important aspects of the problem, then the problem of rationality becomes, precisely, the problem of explaining the production, in social, environmental and emotional context, of broadly appropriate adaptive response. Rationality (or as much of it as we humans typically enjoy) is what you get when this whole medley of factors are tuned and interanimated in a certain way. Figuring out this complex ecological balancing act just *is* figuring out how rationality is mechanically possible.

References

Bechera, A., Damasio, H., *et al* (1997). Deciding Advantageously Before Knowing The Advantageous Strategy. *Science,* 275, 1293-1294.

Chase, V., Hertwig, R., *et al* (1998). Visions of Rationality. *Trends in Cognitive Sciences* 2(6), 206-214.

Churchland, P. M. (1989). *The Neurocomputational Perspective*. Cambridge: MIT/Bradford Books.

Churchland, P. S. (1998). Feeling Reasons. In P. M. Churchland and P. S. Churchland (eds), *On The Contrary*. MIT Press, 231-254.

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing*. Cambridge: MIT Press.

Clark, A. (1993). *Associative Engines: Connectionism, Concepts and Representational Change*. Cambridge: MIT Press.

Clark, A. (1996). Connectionism, Moral Cognition and Collaborative Problem Solvin". In L. May, M. Friedman, and A. Clark (eds), *Minds and Morals*. Cambridge, MA: MIT Press, 109-128.

Clark, A. and Thornton, C. (1997). Trading Spaces: Connectionism and the Limits of Uninformed Learning. *Behavioral and Brain Sciences* 20(1), 57-67.

Clark, A. and Grush, R. (in press). Towards a Cognitive Robotics. *Adaptive Behavior*.

Clark, A. (in progress). Global Abductive Inference and Authoritative Sources.

Damasio, H., Grabowski, T., *et al* (1994). The Return of Phineas Gage: Clues about the Brain from the Skull of a Famous Patient. *Science* 264, 1102-1105.

Fodor, J. and Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition* 28, 3-71.

Fodor, J. (1998). *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J. and Lepore, E. (1993). Reply to Churchland. *Philosophy and Phenomenological Research* 53, 679-682.

Haugeland, J. (1991). Semantic Engines: An Introduction to Mind Design. In J. Haugeland (ed), *Mind Design: Philosophy, Psychology, Artificial Intelligence*. Cambridge, MA: MIT Press, 1-34.

Kleinberg, J. (1997). Authoritative Sources in a Hyperlinked Environment" *IBM Research Report,* RJ 10076

Marr, D. (1982). *Vision*. San Francisco, CA: W.H. Freeman.

McClamrock, R. (1995). *Existential Cognition*. Chicago, IL: Chicago University Press.

Minsky, M. (1975). A Framework For Representing Knowledge. In P. Winston (ed), *The Psychology of Computer Vision*. New York: McGraw-Hill.

Newell, A. and Simon, H. (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the Association for Computing Machinery*, 19, 113-126.

Schank, R. (1975). Using Knowledge to Understand. *TINLAP*, 75.

Smith, B. C. (1996). *On the Origin of Objects*. Cambridge, MA: MIT Press.