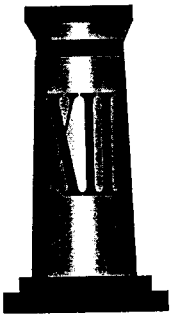

Economic Reason: The Interplay of Individual Learning and External Structure

Andy Clark

XII.1 INTRODUCTION

Much work in economics, the social sciences, and elsewhere takes as its starting point a somewhat unrealistic conception of rationality—a conception that ignores or downplays both the temporal and the situated aspects of human reason. Biological reason, I shall argue, is better conceived as an iterated process of adaptive response made under extreme time pressure and exquisitely keyed to a variety of external structures and circumstances. These external structures and circumstances act as filters and constraints on the spaces of possible real-time responses. Paramount among such structures and circumstances, in the case of human reason, are the cultural artifacts of language and of social and economic institutions. Models of rational decision making need to situate the reasoning agent as just one element in a complex and time-sensitive feedback system in which such external structures play a major role. It is therefore crucial that we understand the complex and mutually modulatory interplay between individual cognition and the extended environmental loops in which it participates. I shall explore a few potential avenues for developing such an understanding including neural network research and multiple time scale simulations.

The Frontiers of the New Institutional Economics

XII.2 RATIONALITY ON THE HOOF

Human minds were not designed as instruments of unhurried, fully informed reason. They were not designed to adhere to rigid or even consistent preference orderings or to act so as to maximize reward on all occasions. Human minds, it is now widely agreed, are better understood as loci of only *bounded* rationality: reason restricted by a variety of evolutionary and pragmatic factors. There is probably no need to labor this point. But a brief listing of some of the central considerations in favor of a conception of bounded rationality would surely include the following:

1. The observation that human brains often deploy cost-cutting, problem-solving strategies that were more or less reliable in their original ecological settings [e.g., the Wason selection test examples in which abstract and concrete versions of the same logical problem yield very different success rates, (see, e.g., Wason and Johnson-Laird, 1970; Holland *et al.,* 1986), or the Kahneman and Tversky work on persistent errors in statistical reasoning (see, e.g., Kahneman *et al.,* 1982)].
2. The (related) observation that ecologically realistic problems often demand rapid, *real- time responses* (e.g., work on animate vision; see Ballard, 1991; Churchland *et al.,* 1994; Clark, 1995) which may preclude large-scale information gathering, integration, and deduction.
3. The (again related) observation that real-time adaptive behavior may be best served by the construction of *multiple partial models* of the environment without attempting to integrate the information gathered into a consistent whole, or even to translate it into a common central code (e.g., work on autonomous agents and mobile robots; see Brooks, 1991; Clark and Toribio, 1994).

One upshot of this recent explosion of work on bounded, real-time, embodied and environmentally embedded reason is a growing sense of discontent with the image of rational choice as maximizing reward relative to a complete and consistent ordering of preferences (see, e.g., Friedman, 1953; although Friedman insisted that the model should *not* be taken as a claim about individual psychology). This image (which can be extended to embrace conditions of incomplete information and risk; see discussion in Denzau and North, 1996) has been termed the paradigm of *substantive rationality* (see Denzau and North, 1996).

Taken as a psychological theory of individual choice, the substantive rationality model is almost certainly incorrect. Individuals seldom complete a complex and consistent ordering of preferences. Even if they did, reliance

on cost-cutting evolved reasoning strategies and the demands of rapid, real-time decision making would probably conspire to undermine the on-line use of such an ordering so as to maximize reward. The individual reasoner, as Herbert Simon (1982) has persuasively argued, is better cast as a quick and dirty satisficer than a cool, well-informed, unhurried classical reasoner.

It is perhaps surprising, then, that neoclassical economics has done as well as it has. Why, given the gross psychological irrealism of its model of human choosing, has traditional economics yielded at least moderately successful and predictive models of, for example, the behavior of firms (in competitive posted price markers) and of political parties, and of the outcomes of double auction experiments (for illuminating discussion, see Satz and Ferejohn, 1994; Denzau and North, 1996)? Why also—on a less optimistic note—has it *failed* to illuminate a whole panoply of other economic and social phenomena? Such notable failures are the failure to model large-scale economic change over time, and the failure to model choice under conditions of strong uncertainty (i.e., cases where there is no preexisting set of outcomes that can be rank ordered according to desirability) (see Danzau and North, 1996). These are fundamental failures insofar as they ramify across a wide variety of more specific cases, such as the inability to model voter behavior, the inability to predict the development of social and economic institutions, and the inability to address the bulk of the choices faced by public policy makers.[1]

The pattern of successes and failures is both fascinating and informative. For the best explanation of the pattern looks to involve a dissociation between cases of what may be termed *highly scaffolded choice* and cases of more weakly constrained individual cogitation. The substantive rationality paradigm, as several authors have recently argued,[2] seems to work best in the highly scaffolded case and to falter and fail as the role of weakly constrained individual cogitation increases. A fully successful economic theory will thus need to address both types of situations and (and this is the crucial move) do so within some overarching framework that enables us to plot the delicate interrelations between the two, that is, to understand how active agents come to construct elaborate forms of social and political scaffolding, which then fundamentally inform their collective problem-solving behavior, and how such scaffolding can itself become transformed by being also the subject of episodes of weakly constrained individual cogitation. In the next section I rehearse the arguments in favor of the hypothesis that neoclassical economics works (insofar as it works at all) only for the highly scaffolded cases. The remaining two sections attempt to sketch a single

[1] The point about voter behavior is forcefully made in Satz and Ferejohn (1994).
[2] I am especially influenced by Satz and Ferejohn (1994) and by Denzau and North (1996).

framework in which interactions between weakly constrained and highly scaffolded choice can begin to be addressed.

XII.3 SCAFFOLDED CHOICE AND MULTIPLE EQUILIBRIA ECONOMICS

Two fundamental distinctions have been invoked to explain the patterns of success and failure reported here:

1. The distinction between cases involving highly scaffolded choice and ones involving only weakly constrained individual cogitation.
2. The distinction between cases involving negative feedback and classical equilibria and those involving positive feedback and multiple or unstable equilibria.

It would simplify matters if these were the same distinctions, differently expressed. Alas, they are at least partially independent, as we shall see.

The idea of highly scaffolded choice is at the heart of important recent treatments by Satz and Ferejohn (1994) and Denzau and North (1996). The common theme is that neoclassical economic theory works best in situations in which individual rational choice has been severely limited by the quasi-evolutionary selection of constraining policies and institutional practices. The irony is explicitly noted by Satz and Ferejohn: "the [traditional] theory of rational choice is most powerful in contexts where choice is limited" (1994, p. 72).

How can this be? According to Satz and Ferejohn, the reason is simple: what is doing the work, in such cases, is not (so much) the individual's cogitations as the larger social and institutional structures in which she is embedded. These structures have themselves evolved and prospered (in the cases where economic theory works!) by promoting the selection of collective actions that do indeed maximize returns relative to a fixed set of goals. For example, the competitive environment of capital markets ensures that, by and large, only those firms that maximize profits survive. It is this fact, rather than facts about the beliefs, desires, or other psychological features of the individuals involved, that ensures the frequent success of substantive rationality models in predicting the behavior of firms. Strong constraints imposed by the larger scale market structure result in firm-level strategies and policies that maximize profits. In the embrace of such powerful scaffolding, the particular theories and worldviews of individuals may at times make little impact on overall firm-level behavior. Where the external scaffolding of policies, infrastructure, and customs is strong and (importantly) is a result of competitive selection, the individual members are, in effect, interchangeable cogs in a larger machine. The larger machine extends way outside the individual and incorporates large-scale social, physical and even

geopolitical structures. It is the behavior of this larger machine that traditional economic theory often succeeds in modeling. A wide variety of individual psychological profiles are fully compatible with specific functional roles within such a larger machine. As Satz and Ferejohn have remarked:

> Many sets of individual motivations are compatible with the constraints that competitive market environments place on a firm's behavior. In explaining firm behavior, we often confront causal patterns that hold constant across the diverse realizations of maximizing activity found in Calvinist England and the Savings and Loan Community in Texas.
>
> (1994, p. 79)

By contrast, the theory of consumer behavior is weak and less successful. This is because individual worldviews and ideas loom large in consumer choice and the external scaffolding is commensurately weaker. Similarly, the theory of voting behavior is weak beside the theory of party behavior in electoral competitions. Once again, the parties only survive subject to strong selection pressures that enforce vote-maximizing activity. By comparison, individual choice is relatively unconstrained (see Satz and Ferejohn, 1994, pp. 79–80).

Satz and Ferejohn have suggested that the crucial factor distinguishing the successful and unsuccessful cases (of neoclassical, substantive-rationality assuming theory) is the availability of a structurally determined *theory of interests.* In cases where the overall structuring environment acts to select in favor of actions that are restricted so as to conform to a specific model of preferences, neoclassical theory works. It works because individual psychology no longer matters: the "preferences" are imposed by the wider situation and need not be echoed in individual psychology. Thus, the competitive environment of posted price markets *externally* determines the types of action patterns that will lead to survival and success, and a theory of the "preferences" of the players can be seen not as a theory of individual psychologies but as a reflection of these constraints. Individual psychology is of reduced importance—all that matters is that *whatever* the individual psychological profiles of the players, they must be compatible with this pattern of action selection. The case of political party behavior, likewise, is one in which the overall situation (at least in democratic, two-party electoral systems) selects for parties that act to maximize votes. This external structuring force allows us to impute "preferences" on the basis of the constraints on success in such a larger machine. The constraints on individual voters are much weaker. Hence, real psychological profiles come to the fore, and neoclassical theory breaks down (see Satz and Ferejohn, 1994, pp. 79–80).

This general diagnosis is supported by the analysis of Denzau and North (1996). They have noted that traditional economic theory nicely models choice in competitive posted price markets and in the experimental

studies of double auctions. In such cases, they have suggested, certain institutional features play a major role in promoting "maximizing-style" economic performance. By way of illustration, Denzau and North cited fascinating studies by Gode and Sunder (1992). These studies invoke so-called zero intelligence (ZI) traders, that is, modeled agents who do not actively theorize, recall events, or attempt to maximize returns. By constraining such simple agents to bid only in ways that do not yield an immediate loss, an efficiency of 75 percent (measured as "the percentage of sum of potential buyer and seller rents"; Denzau and North, 1996, p. 5) was achieved. Replacing the ZI traders with humans increased efficiency by a mere 1 percent! But altering the institutional scaffolding (e.g., from collecting all bids in a double auction before contracting to allowing simultaneous bidding and contracting) yielded a 6 percent improvement in efficiency. The strong conclusion is this:

> Most efficiency gains in some resource allocation situations may be attributed to institutional details, independent of their effects on rational traders.
>
> (Denzau and North, 1996, p. 5)

The results of the zero intelligence trader experiments are revealing. They clearly demonstrate the power of institutional settings and external constraints to promote collective behaviors that conform to the model of substantive rationality. Such results fit nicely with the otherwise disquieting news that the bulk of traditional economics would be unaffected if we assumed that individuals chose randomly (see Alchian, 1950, cited in Satz and Ferejohn, 1994, p. 76) rather than by maximizing preferences, and that pigeons and rats can often perform in ways consistent with the theory of substantive rationality (see Kagel, 1987, also cited in Satz and Ferejohn, 1994, p. 77)! Such results make sense if it is the scaffolding of choice by larger-scale constraining structures which acts as the strongest carrier of maximizing force. In the extreme limiting cases of such constraint, the individual chooser is indeed a mere cog—a constrained functional role played as well by a ZI trader, a pigeon, a rat, a human trader or, in the worst cases, by a coin-flipping device!

It is important to note that this is *not* to claim (falsely) that highly scaffolded choices will always conform to the norms of substantive rationality. This will only be the case if the institutional scaffolding has itself evolved as a result of selective pressure to maximize rewards and if the economic environment has remained stable or if the original institutional scaffolding itself built in sufficient adaptability to cope with subsequent change. A major question—one at the heart of Douglass North's research agenda—is thus: Under what conditions can such efficient and adaptable institutions evolve?

In sum, traditional economic theory (invoking the substantive rationality paradigm) succeeds wherever individual choice is strongly constrained

by social and institutional scaffolding that has *itself* evolved subject to selective pressures to maximize rewards. Outside such highly constrained settings, genuine individual thought plays a greater role, and the psychological irrealism of the substantive rationality model takes its toll.

A seemingly different diagnosis of the pattern of failures of neoclassical economics has been proposed by Brian Arthur. Arthur (1990) has suggested that the successes of neoclassical theory are marked by the operation of processes of negative feedback leading to equilibrium solutions in which buyers and sellers settle on a price that constitutes the most efficient solution for the overall economy. The dynamics of such cases are those of single fixed-point attractors in economic space. Arthur noted that other types of dynamical description may, at times, be more appropriate. The now-classic example is the development of high-technology microelectronics. Here, positive feedback (increasing rather than diminishing returns) can result in economic spaces characterized by multiple equilibria. The process by which a resting point is selected is, in these cases, not one that looks likely to settle at the best solution. Instead, small early biases, amplified by positive feedback, can set the overall system onto a path that becomes increasingly fixed as positive feedback continues to operate. As Arthur pointed out, Betamax video technology may indeed have been superior to VHS format, but once—for whatever reason—the market share of the VHS format edged ahead, processes of positive feedback took over. As video shops stocked more VHS format tapes in response to early market lead, so the choice for new buyers of video equipment became clearer. And as more new buyers chose VHS equipment, so the shops stocking policies became clearer still. The upshot being that VHS cornered the market. In such scenarios, small early advantages (due to luck, advertising, whatever) initiate a tidal wave of gains that quickly locks the market into a particular course—a course largely independent of the objective quality of the product relative to its rivals. In such cases, the initial market system is not characterized by a single point attractor representing an optimal solution enforced by strong negative feedback pushing us away from inferior choices. Instead, the initial space may contain multiple attractors, any one of which could win out if some early chance perturbation should place us within its orbit. Such small chance events are bound to confound economic prediction. At best, we may be better able to formulate science and economic policy, once we recognize the potential for the operation of positive feedback and increasing returns in specific markets.

How, though, do these considerations fit in with our earlier observations concerning highly scaffolded versus weakly constrained choice? In one sense, the positive feedback cases are clearly only weakly constrained. The institutional infrastructure, in such cases, leaves a system's future trajectory quite underdetermined. But the lack of institutional constraints on the outcomes is not, at least in the example cases, generally resolved by a

commensurately increased role for individual psychological profiles. Instead, factors external to individual psychological profiles still tend to bear the explanatory burden. For example, Arthur (1990, p. 95) cited the concentration of the electronics industry in Silicon Valley as a possible consequence of a "snowball" effect initiated by the chance geographical positioning of the first few firms to enter that arena. No facts about formal or informal norms of practice determined that location, but neither is it best explained by the individual psychological profiles of the players. Here, it is pure historical and geographical chance, and not institutional scaffolding *or* individual psychology, which sowed the seed that positive feedback dynamics amplified into the dense geographical concentration of expertise we see today.

In one respect, at least, these positive feedback cases thus resemble the highly scaffolded ones: individuals, in both cases, are functioning as cogs in a larger machine. It is the dynamics of the larger machine that is selecting economic outcomes. The difference is just that in the cases where neoclassical theory works best, the dynamics of the larger machine is one of single fixed-point attractors, whereas in the positive feedback cases, it is one of multiple attractors operating in spaces where small early perturbations can lead to strong path fixation. In these latter cases, the events that bear the explanatory burden are the small early perturbations. Such perturbations may be random, due to wider social or geopolitical events, or due to, for instance, a neat marketing idea had by one individual. Even in the case of the neat idea, however, the main explanatory construct will still be the positive feedback dynamics of the overall system. A large proportion of the (important) new range of cases that Arthur and others are studying is thus—like the classical cases—essentially *individual psychology transcending.* The clear indicator of this is once again the fact that the explanatory burden is borne by overall system dynamics in which the microdynamics of individual psychology is relatively unimportant. The multiple equilibria model *could* of course, be applied to the other class of cases too—that is, to cases in which the gross nonpsychological scaffolding is *not* bearing all of the weight. Indeed, it is precisely this class of cases that Arthur turned to in the paper that appears in the present volume. There the focus is on the economic significance of the complex coevolving ecology of *beliefs.* In such an ecology, agents must often try to act in ways that take account of how they predict that others will interpret their own actions, and so on. The simulation tools of complex systems theory, as Arthur noted, may provide our best measure of plotting and understanding such complex and iterated processes of psychological interaction.

How, then, *should* we conceive the role of individual psychology in economic explanation? The key to such an understanding, I believe, lies in seeing how to *couple* psychological effects with the kinds of larger scale system

dynamics identified in the preceding discussion. To do so, we must first get a better view of the nature of individual cognition itself.

XII.4 WHAT KIND OF MIND NEEDS A SCAFFOLD?

The moral so far is that the scaffolding matters: the external structuring provided by institutions and organizations bears much of the explanatory burden for explaining current economic patterns. To see where human psychology fits in, let us begin by asking, what kind of individual mind *needs* an external scaffold?

A vital role for external structure and scaffolding is, in fact, strongly predicted by recent work on individual cognition. *Classical* artificial intelligence and information processing psychology tended to downplay any such role. But more recent work in the field of *time-critical adaptive response* posits a much greater reliance on real-world structures, external data stores, and active interventions. (See Clark, 1997, for a review.)

Simon's (1982) notion of bounded rationality was probably the first step in this direction. But although Simon (rightly) rejected the view of human agents as perfect logical reasoners, he remained committed to a basically classicist model of computation as involving explicit rules and quasi-linguistic data structures. The major difference was just the use of *heuristics*, with the goal of *satisfying* rather than optimizing—that is, the use of rules of thumb to find a workable solution, with minimal expenditures of time and processing power.

The reemergence of connectionist (also known as artificial neural networks, or parallel distributed processing) ideas in the last decade or so took us further by challenging classical models of internal representation and of computational process. In place of explicit rules operating on language-like data structures, connectionists posit an intermingling of data and processing supported by a dense parallel architecture of idealized "neurons." Such systems are often trained on examples of a desired input-output mapping and learn (by a process of gradient descent learning) to assign weights (positive and negative numeric values) to the interconnections between the idealized neurons. These weights cause the network to behave as an interpolating associative memory: they enable it to re-create patterns of designed output given fragments of the original inputs and to generalize its knowledge so as to deal with novel cases on the basis of their partial similarity to the training cases.

One feature of neural network learning, to which we will later return, is its rampant path dependence. A compelling series of experiments by Jeff Elman and others (e.g., Elman, 1994) shows that connectionist learning is heavily dependent on the sequence of training cases. If the early training

goes wrong, the network is often unable to recover. Thus a specific network proved able to learn complex grammatical rules only if it was previously trained solely on more basic examples highlighting, for example, verb-subject number agreement. Early exposure to the more complex cases of long-distance dependencies, and so forth led it into bad early "solutions" (local minima) from which it was unable subsequently to escape (for a detailed treatment of this case, see Clark, 1993).

For present purposes, however, what matters most is that such networks constitute *fast but limited* systems that in effect substitute *pattern recognition* for reasoning. This, as might be expected, is both a boon and a burden. It is a boon insofar as it provides just the right resources for the tasks humans perform best and most fluently: tasks such as motor control, face recognition, deciphering handwritten zip codes, and so on (for more on these examples, see Jordan *et al.*, 1994; Churchland (1996), Le Cun *et al.*, 1989). But it is a burden insofar as it does not provide an ideal substrate on which to build processes of careful, sequential reasoning, or of long-term planning (see Clark, 1989; Norman, 1986). This is not in itself a bad thing. If our goal is to model human cognition, computational underpinnings that yield a profile of strengths and weaknesses similar to our own are to be favored. And we *are* generally better at Frisbee than at logic. Nonetheless, we *are* able to achieve impressive results using logic, sequential reason, long-term planning, and so on. If we are at root associative pattern recognition devices, how is this competence ever achieved?

Two factors, I suggest, conspire to enable us to rise above our computational roots. The first is our old friend, external scaffolding. The second involves a more tendentious idea, namely, that we are internally driven to reorganize our own knowledge continually in purely exploratory ways. It is the combination of these two factors that (I believe) explains our peculiar degree of communal cognitive success.

Connectionist minds are ideal candidates for extensive external scaffolding. A simple example, detailed in Rumelhart, Smolesky, McClelland, and Hinton (1986), concerns long multiplication. Most of us, they argue, can learn to know at a glance the answer to basic multiplications. Thus we know at a glance that 7×7 is 49. Such knowledge could easily be supported by a basic on-board pattern-recognition device. But longer multiplications (for most of us) present a different kind of problem. Asked to multiply 7,222 by 9,422, most of us resort to pen and paper (or calculator). What we achieve, using pen and paper, is a reduction of the complex problem to a sequence of simpler problems beginning with 2×2. We use the external medium (paper) to store the results of these simple problems, and by an interrelated series of simple pattern completions coupled with external storage, we finally arrive at a solution. At this point the authors

comment that

> This is real symbol processing and, we are beginning to think, the primary symbol processing that we are able to do. Indeed, on this view, the external environment becomes a key extension to our mind.
>
> (Rumelhart, Smolesky, McClelland, and Hinton, 1986, p. 46)

Some of us, of course, go on to learn to do such sums in our heads. The trick in these cases, it seems, is to learn to manipulate a mental model in the same way as we originally manipulated the real world. In such cases we are able to mentally simulate the external arena, and hence, at times, internalize cognitive competencies that are nonetheless rooted in manipulations of the external world (here, cognitive science meets Soviet psychology; see Vygotsky, 1962).

Institutions, firms, and organizations seem to me to share many of the key properties of pen, paper, and arithmetical practice in this example. Pen and paper provide an external medium in which we behave (using basic online resources) in ways dictated by the general policy or practice of long multiplication. Most of us do not know the mathematical justification for the procedure. But we use it, and it works. Similarly, firms and organizations provide an external resource in which individuals behave in ways dictated by norms, policies, and practices; norms, policies, and practices that may even become internalized as mental models. Daily problem solving, in these arenas, often involves locally effective pattern recognition strategies that are invoked due to some external originating prompt (a green slip in the in-tray), discharged in a preset manner, and that leave their mark as further traces (different slips of paper, E-mails, whatever) in the overarching machinery of the firm. It is in these contexts that, in the short term at least, the role of individual rationality can become somewhat marginalized. If the overall machinery has been selected so as to maximize profits, the fact that the individuals are cogs with very bounded forms of rationality will not matter (individual neurons are, if you like, even more restricted cogs, but once organized into brains by natural selection, they too support a grander kind of reason).

Much of what goes on in the complex world of humans may thus, somewhat surprisingly, be understood in terms of so-called *stigmergic* algorithms. A stigmergic algorithm (see Beckers *et al.*, 1994) is one in which actions are strongly determined by external structures that are themselves the *operands* or objects of the actions. The notion originates from studies of termite nest building (see Grassé, 1959; for a more recent discussion, see Bonabeau *et al.*, 1994) in which the termites action is controlled by local nest structure yet results in modifications of that structure, which in turn controls other actions (by itself or others). A nice computational

treatment of stigmergy is to be found in the work of Bonabeau and colleagues (1994).

Stigmergic, scaffolded reason is, however, at best, part of the true story. For there is a delicate balancing act in human cognition between stigmergy and scaffolded response, on the one hand, and innovation and integration on the other. It is in understanding the details of this balancing act that we may one day obtain real insight into the longer-term processes by which the external scaffolding of institutions, norms, and practices itself changes and develops. This is the topic of the next and final section.

XII.5 LEAPS AND LEVELS

Processes of innovation and integration appear to operate both at the level of individual learning and, in a curious way, at the level of the cultural scaffolding itself. There is also a very interesting interplay between the shape of individual learning and the various structures that constitute the scaffolding. In this final section, we scout each of these effects in turn.

XII.5.1 Innovation and Integration in Individual Learning

At the very simplest level, standard connectionist gradient descent learning gives rise to what, *from the outside,* can look like a series of radical leaps or discontinuities in acquired knowledge. Nonetheless, such shifts are often a direct consequence of the gradual, gradient descent learning process itself.

The basic way in which simple gradient descent learning can lead to "cognitive leaps" is where the error surface itself contains a sharp gradient. In such cases, a very small change to the internal weights can suddenly lead the network from the top to the bottom of a cliff, with dramatically improved performance making a sudden rapid appearance. In addition, the use of a nonlinear activation function means that small changes to the early weights (between input and hidden units) can make a very large difference to the outputs and hence lead to sudden, dramatic falls in the total sum squared error. Sudden improvements in gross performance can thus result from small changes to the early weights. Nice examples of these and related phenomena are to be found in the work of Plunkett and colleagues (1990) and are further discussed in Clark (1993, Chapter 8).

Such phenomena go some way toward accounting for the rather sudden leaps that are evident in human learning (see especially the developmental psychological literature, e.g., Karmiloff-Smith, 1992). One example of such a leap is the sudden transition from a rote-style representation of a

few past tenses to a more systematic and productive representation—a transition that Plunkett and colleagues (1990) successfully modeled using a gradient descent network whose output suddenly alters once the training outputs reaches a critical mass of 40 to 50 verbs. Another example concerns the sudden shifts observed in a beam-balancing task. These include a shift between balance judgments based solely on the weights of items to ones that factor distance from a fulcrum into the judgment. These shifts have been recapitulated in a connectionist network studied by McClelland (1989) and more recently by Schulz (1994).

In such cases, discontinuities in output ("radical new ideas") are undergirded by a smooth continuum of inner computational transitions. But this may not exhaust the tricks of the canny human cognizer. One conjecture [due to Karmiloff-Smith (1992) and pursued further by Clark and Karmiloff-Smith (1994), and Clark (1993)], concerns the possible operation of a process termed "representational re-description." The conjecture is that human learning incorporates an endogenous drive to go "beyond success." Thus imagine a being who has achieved a good working solution to a specific problem. If that being were a standard connectionist network, no further learning would occur. In the absence of error messages, the learning algorithm remains effectively inert. But if that being were a re-describer (it is conjectured), she would take the absence of error as a signal to seek *reorganizations* of her knowledge—reorganizations that might permit greater control, or involve integrating this ability/knowledge with other aspects of her skills and understanding. Karmiloff-Smith has gathered an impressive body of evidence in support of the hypothesis that such processes of redescription occur (evidence from the domains of drawing, number, and physical and social understanding; see Karmiloff-Smith, 1992). As an illustrative example, consider the work in drawing. In a classic series of experiments, Karmiloff-Smith showed how children first internalize a procedure for drawing some familiar and much practiced item (e.g., a man) and then proceed through a series of reorganizations of that knowledge. Each such reorganization increases the flexibility of use of the drawing knowledge allowing more and more innovative changes to be made. Early drawing allows for the deletion and insertion of components (e.g., limbs on a person), but not for the reorientation of parts or the integration of items from another category (e.g., a combination man and airplane). Later on, these more complex changes become possible. The underlying explanation, Karmiloff-Smith has argued, involves the child (or adult) engaging in a process of redescription in which one form of information coding and/or control is reconfigured so as to yield another, more flexible encoding (think, for example, of going from a push-down stack to a random access memory).

It remains unclear how best to model such a process using connectionist computational resources. But one possibility currently under investiga-

tion involves the use of multiple networks, with successful "expert" networks being subsequently coordinated by "gating" networks so as to support increasingly complex behaviors and to facilitate the reuse of acquired skills and knowledge in new problem domains [see Jacobs *et al.* (1991) for a partial connectionist model, Anderson and Van Essen (1994) for a neurally plausible vision of gating, Damasio (1994) for a link between the gating hypothesis[3] and patterns of cognitive deficit, and Clark (1993, Chapter 8, Clark and Thornton (1996), and Clark and Karmiloff-Smith (1994) for further general discussion of the space of computational options)].

XII.5.2 Incremental Learning and the Second Role of External Scaffolding

One way or another, individual learning is clearly a source of some discontinuities in human ideas and mental models. We have already seen how the presence of various kinds of environmental scaffolding (organizations, institutions, language, external memory devices, etc.) may help to reduce individual cognitive loads and to restrict the scope of individual choice. It seems likely, however, that such scaffolding plays a dual role: it both restricts and *expands* our intellectual horizons. One major way in which it expands them is by enabling the collective exploration of multiple learning trajectories. In this respect, the scaffolding itself can act as the vehicle of a novel kind of cognitive evolution.

Recall the strong path dependence of individual connectionist learning. The use of external memory systems helps ameliorate some of the effects of this path dependence by allowing achieved innovations ("redescriptions") to be transmitted between individuals. This allows the collective construction of representational trajectories that crisscross individual cognizers and hence increase the chance of a good idea finding a viable niche for further development. This is, of course, an old idea. But it is one whose value cannot be fully appreciated except in the context of our increasing understanding of the boundedness and extreme path dependence of individual reason.

The emerging picture fits neatly with Merlin Donald's (1991) exploratory work on the evolution of culture and cognition. Donald recognized very clearly the crucial role of forms of external scaffolding (particularly, of external memory systems) in human thought. But he distinguished two major types of scaffolding, which he termed the mythic and the theoretic. Before the Greeks, Donald claimed, various external formalisms were in use but were deployed only in the service of myths and narratives. The key innovation of the Greeks was to begin to use the written medium to record the *processes* of thought and argument. Whereas previous written

[3] Damasio's term is "convergence zones"—such zones control patterns of activity in multiple spatially distant neural networks.

records contained only myths or finished theories (which were to be learnt wholesale and passed down relatively unaltered), the Greeks began to record partial ideas, speculations with evidence for and against, and the like. This new practice allowed (in our terms) *intermediate level recodings* to be passed around, amended, completed by others, and so forth. According to Donald, what was thus created was

> Much more than a symbolic invention, like the alphabet, or a specific external memory medium, such as improved paper or printing. They founded the *process* of externally encoded cognitive change and discovery.
>
> (1991, p. 343)

One effect of such a process is precisely to ameliorate the detrimental effects of path dependency in individual learning. The path to a good idea can now crisscross individual learning histories, so that one agent's local minima becomes another's potent building block. Even a blind and unintelligent search for productive recodings of stored data will now and again yield a powerful result. By allowing such results to migrate between individuals, culturally scaffolded reason is able to incrementally explore spaces that individual reason could never hope to penetrate (for a detailed, statistically based investigation of this claim, see Clark and Thornton (to appear)).

XII.5.3 The Interplay of Individual Learning and External Structures

The external scaffolding of which we have made so much is, in truth, a strange and various beast. But some of its most interesting manifestations share a core property. Such manifestations relate to their human users in much the way a beneficial parasite relates to its host. For the human user provides a niche to which the scaffolding must adapt. And as the human user changes, so too must the scaffolding, if it is to prosper.

Such vague but suggestive ideas can now be given some admittedly simplistic quantitative and computational flesh. In a recent study, Hare and Elman used a "cultural phylogeny" of connectionist networks to model, in some detail, the series of changes that characterized the progression from the past-tense system of Old English (circa 870) to the modern form. In so doing, they explicitly addressed a number of questions that have clear analogies in the social and economic realms. Thus they wrote:

> The complex inflectional system of OE [Old English] existed for hundreds of years—what permitted that stability in the face of an apparent drive toward simplification? What eventually disrupted that stability and caused the system to change? Can the direction of change be predicted?
>
> (Hare and Elman, 1994, p. 4)

Hare and Elman showed that the historical progression can be modeled, in some detail, by a series of neural networks in which the output from one generation is used as the training data for the next. This process yields changes in the language itself as the language alters to reflect the learning profiles of its users.

Briefly, this is what happens. An original network is trained on the Old English forms. A second network is then trained (though not to asymptote) on the forms produced by the first. This output is then used to train a further network, and so on. Crucially, any errors that one network makes in learning to perform the mappings becomes part of the next network's data set. Patterns that are hard to learn, or items whose form is close to that of other, differently inflected items, tend to disappear. As Hare and Elman put it:

> At the onset, the classes [of verbs] differ in terms of their phonological coherence and their class size. Those patterns that are initially less common or less well-defined are the hardest to learn. And these tend to be lost over several generations of learning. This process snowballs as the dominant class gathers in new members and this combined class becomes an ever more powerful attractor.
>
> (1994, pp. 19–20)

In short, we find a process of positive feedback of the kind described by Arthur (see Section XII.3). But the source of the feedback is located squarely in the specific learning profile of the individual (or, in this case, of the network). By thus studying the interplay between the external data set and the processes of individual learning, Hare and Elman were able to make some very fine-grained,[4] quantitative predictions about the historical progression from Old English to the modern forms. Their account thus goes considerably beyond the common observation that morphological change is somehow driven by analogy, class size, and so on. By factoring a specific model of individual learning into the equation, it becomes possible to understand "what seem to be idiosyncratic analogical effects as inevitable and predictable results of gradient descent learning" (Hare and Elman, 1994, p. 32).

The distance between this kind of work and the social-economic case should not, of course, be underestimated. Where the external structure consists of ideas, or of institutions, the dynamics of change will be importantly different. But the same phenomena of interaction will surely emerge. The content-carrying structures will respond to the particular learning profile of the human agent in ways that favor some ideas over others. Also, processes of positive feedback will kick in to amplify such effects into radical changes

[4] For example, predictions of rare but significant cases of irregularization, in which a once-regular verb falls under the inflectional influence of a phonologically similar, high-token frequency irregular.

in the external intellectual and institutional climate. By better understanding the dynamics of individual learning, we will at a minimum increase our grasp of the possible trajectories of such overall systems.

Crucial to any such understanding will be the use of computational models that incorporate multiple time scales and multiple levels of organization. It is only quite recently that the computational tools, and computational power, necessary for these types of study have become widely available. Work in the area known as "artificial life" is especially pertinent here, combining as it sometimes does the use of individual neural network learning with processes of simulated genetic evolution in small populations of "creatures" (see, e.g., Ackley and Littman, 1994). Such work addresses, albeit in necessarily skeletal form, the interplay between individual learning, collective fitness, and longer-term processes of evolutionary change. The skeptical reader will surely object that it is premature to confront these complex issues while our understanding of the component processes, such as individual learning, remains so poor. But this objection is, I suggest, misplaced. For it may be that even very simple experiments combining multiple time scales and levels of organization will yield a preliminary understanding of the types of dynamics that characterize collections of environmentally embedded learning agents. Such global and transtemporal dynamics may be as fundamental, in their way, as the local dynamics of individual human brains.

XII.5.4 Communities and Communication

As a final example of how computational and cognitive scientific theorizing may shed light on the interplay between pattern-completing individual cognition and the larger scaffolding provided by social organizations and institutions, consider the case described by Hutchins (1995).

Hutchins investigated the collective problem-solving capacity of a small "community" of simple neural networks. Each individual in this community comprised a small number of linked processing units. Each unit coded for some specific environmental feature. Excitatory links connected mutually supportive features, whereas inhibitory links connected mutually inconsistent features. A feature like "is a dog" would thus be coded by a single unit with excitatory links to, for example, "barks" and "has fur" units, and inhibitory links to, for example, "meows" and "is a cat" units (the latter being themselves linked by an excitatory connection). Such networks are known as constraint satisfaction networks.

Once a constraint satisfaction is set up (either by learning or by hand coding), it exhibits nice properties of pattern-completion style reasoning. Thus imagine that the various units receive input signals from the environment. Activation of a few units that figure in a linked web of excitatory connections will yield activity across all the linked units. The input "barks" will

thus yield a global activation profile appropriate to the category "dog," and so on. Individual units often "choose" whether or not to respond (become active) by summing the inputs received along various channels and comparing the result to some threshold level. As a result, once a constraint satisfaction network settles into an interpretation of the input (e.g., by having all the dog-feature units become active), it becomes very hard to dislodge it, as the units lend each other considerable mutual support.

This feature of such networks, as Hutchins pointed out, corresponds rather nicely to the familiar psychological effect of confirmation bias, namely, the tendency to ignore, discount, or creatively reinterpret evidence (like a solitary "meows" input), which goes against some hypothesis or model that we already have in place (see, e.g., Wason, 1968). Now imagine a community of constraint- satisfaction networks in which each network has different initial activity levels ("predispositions") and different access to environmental data. Hutchins showed that in such cases, the way in which the internetwork communication is structured makes a profound difference to the kind of collective problem solving displayed.

Surprisingly, what Hutchins (1995, p. 252) found was that in such cases more communication is not always better than less. In particular, if from the outset all the networks are allowed to influence each other's activity (to communicate), the overall system shows an extreme degree of confirmation bias—much more than any one of the individual nets studied in isolation. The reason is that the dense communication patterns impose a powerful drive to rapidly discover a shared interpretation of the data, that is, to find a stable pattern of activity across all the units. The individual nets, instead of giving due weight to the external input data, focus instead on these internal constraints (the need to find a set of activation patterns that does not disrupt the others). As a result, the social group rushes "to the interpretation that is closest to the center of gravity of their predispositions, regardless of the evidence" (Hutchins, 1995, p. 259).

By contrast, if you restrict the level of early communication, this gives each individual network time to balance its own predispositions against the environmental evidence. If internetwork communication is *subsequently* enabled, then overall confirmation bias is actively reduced, that is, the group is more likely to fix on a correct solution than is the average member.

Such results suggest that the collective advantage of a jury over an individual decision may dissipate proportionally to the level of early communication between members.[5] More importantly, however, the example illustrates one more way in which we may begin to understand, in a rigorous

[5] Compare to the Condorcet Jury theorem, which states that if (among other things) juror choices are *independent,* then a majority vote by a jury will be correct more often than an average juror.

manner, some aspects of the delicate interplay between individual cognition and group level dynamics. Such understanding will surely be crucial to a better appreciation of the role of institutional and organizational structures in determining social and economic outcomes, and of the balance between individual cognition and the external scaffolding that it both shapes and inhabits.

XII.6 CONCLUSION: EVERY BRAIN IN ITS PLACE

Classical rule-and-symbol-based artificial intelligence made a fundamental error. It mistook the cognitive profile of the agent *plus* her environment for the cognitive profile of the brain (see Clark, 1989; Hutchins, 1995). The neat classical separation of data and process, of symbol structures and CPU, reflected nothing so much as the separation between the agent and an external scaffolding of ideas persisting on paper, in filing cabinets, or electronic media. Recent models of individual cognition depict the brain as a very different kind of device. They depict it as a loose coalition of pattern completing devices geared to solving real-world problems in real time. To that end, the brain readily exploits external structures to support and amplify its own basic problem-solving capacities.

Biological reason, at the individual level, is thus revealed as a process of iterated, local responses, whose computational costs are reduced by extensive reliance on external structures and circumstances. The successes of classical economics emerge, within this paradigm, as cases that depend largely on the short-term dynamics of responses strongly determined by particular kinds of institutional or organizational structure, namely, structures that have *themselves* evolved as a result of selective pressure to maximize rewards of a certain kind. Individual human reason, in such cases, bears but little of the explanatory burden.

Nonetheless, these external scaffoldings are, in most cases, themselves the products of collective human thought and activity. It is thus in the generation and evolution of the institutional and organizational scaffolding that human psychology may be playing the greatest role. We sketched a few ways in which recent research in computational cognitive science may help us better to understand both the role and gradual evolution of such external structures. And we saw hints of how individual cognitive profiles may interact with different types of social structuring (e.g., various patterns of interpersonal communication) so as to yield different types of collective outcome.

Such hints and sketches barely scratched the surface of a large and difficult project: understanding the way the mind both structures and inhabits a world populated by cultures, countries, organizations, institutions, political parties, E-mail networks, and all the vast paraphernalia of external structures and scaffoldings that guide and inform our daily actions. This large

and difficult project, I suggest, demands the application of a wide range of new computational and analytic tools: tools, which, at least in simplified cases, can begin to transform our hints, sketches, and speculations into quantitative demonstrations and to suggest more general overarching dynamical principles. One key to such an improved understanding is the use of computational models that span multiple time scales (individual learning and genetic or cultural evolution) and multiple levels of organization (individuals, populations, and populations-plus-external-scaffolding structures). Many of the tools and skills needed to conduct such investigations are now in place, and a few restricted simulations have already yielded interesting results. At the end of the day, the frontiers of institutional economics may turn out to border rather closely on those of cognitive psychology, cognitive science, and the theory of complex nonlinear systems and neural networks. There is clearly much to learn. Perhaps we should learn it together.

ACKNOWLEDGMENTS

Thanks to Douglass North, Arthur Denzau, Norman Schofield, and John Drobak for support, suggestions, and criticisms.

REFERENCES

Ackley, D., and Littman, M. (1994). Altruism in the evolution of communication. In *Artificial Life* (R. Brooks and P. Maes, eds.), Vol. IV. pp. 40–48 MIT Press; Cambridge, MA.

Alchian, A. (1950). Uncertainty, evolution and economic theory. *Journal of Political Economy* **57,** 211–221.

Anderson, C., and Van Essen, D. (1994). Dynamic routing strategies in sensory, motor and cognitive processing. In *Large-Scale Neuronal Theories of the Brain* (C. Koch and T. Davis, eds.), pp. 271–299. MIT Press, Cambridge, MA.

Arthur, B. (1990). Positive feedbacks in the economy. *Scientific American,* **92,** 99.

Ballard, D. (1991). Animate vision. *Artificial Intelligence* **48,** 57–86.

Beckers, R., Holland, O., and Deneubourg, J. (1994). From local actions to global tasks: Stigmergy and collective robotics. In *Artificial Life* (R. Brooks and P. Maes, eds.), Vol. IV. pp. 181–189. MIT Press, Cambridge, MA. pp. 181–189.

Bonabeau, E., Theraulaz, G., Arpin, E., and Sardet, E. (1994). The building behavior of lattice swarms. In *Artificial Life* (R. Brooks and P. Maes eds.), Vol. IV. pp. 307–312. MIT Press, Cambridge, MA.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence* **47,** 139–159.

Churchland, P. M. (1996). The Engine of Reason, the Seat of the Soul (MIT Press: Cambridge MA).

Churchland, P., Ramachandran, V., and Sejnowski, T. (1994). A critique of pure vision. In *Large-Scale Neuronal Theories of the Brain.* (C. Koch and J. Davis, eds.). 23–60. MIT Press, Cambridge, MA.

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

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

Clark, A. (1995). Moving minds: Situating content in the service of real-time success. In *Philosophical Perspectives* (J. Tomberlin, ed.), Vol. 9. 89–104. (Ridgeview, CA)

Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again.* MIT Press, Cambridge, MA.

Clark, A., and Karmiloff-Smith, A. (1994). The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind & Language,* 8:487–519.

Clark, A., and Thornton, C. (to appear). Trading spaces: Connectionism and the limits of learning. *Behavioral and Brain Sciences.*

Clark, A., and Toribio, J. (1994). Doing without representing? *Synthese* **101,** 401–431.

Damasio, A. (1994). *Descartes' Error.* Grosset, Putnam, NY.

Denzau, A., and North, D. C. (1996). *Shared Mental Models: Ideologies and Institutions.*

Donald, M. (1991). *Origins of the Modern Mind.* Harvard University Press, Cambridge, MA.

Elman, J. (1994). Learning and development in neural networks: The importance of starting small. *Cognition* **48,** 71–99.

Friedman, M. (1953). *Essays in Positive Economics.* University of Chicago, Chicago.

Gode, D., and Sunder, S. (1992). *Allocative Efficiency of Markets with Zero Intelligence (ZI) Traders: Markets as a Partial Substitute for Individual Rationality,* Working Paper No. 1992–16. Carnegie-Mellon Graduate School of Industrial Administration, Pittsburgh, PA.

Hare, M., and Elman, J. (1994). *Learning and Morphological Change.* Center for Research in Language, University of California at San Diego.

Holland, J., Holyoak, K., Nisbet, R., and Thagard, P. (1986). *Induction.* MIT Press, Cambridge, MA.

Hutchins, E. (1995). *Cognition in the Wild.* MIT Press, Cambridge, MA.

Jacobs, R., Jordan, M., and Barto, A. (1991). Task decomposition through competition in a modular connectionist architecture: the what and where visual tasks. *Cognitive Science* **15,** 219–250.

Jordan, M., Flash, T., and Arnon, Y. (1994). A model of the learning of arm trajectories from spatial deviations. *Journal of Cognitive Science* **6,** 359–376.

Kagel, J. (1987). Economics according to the rat (and pigeons too). In *Laboratory Experimentation in Economics: Six Points of View* (A. Roth, ed.). pp. 94–115. Cambridge University Press, New York.

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgement Under Uncertainty.* Cambridge University Press, Cambridge, UK.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science.* MIT Press/Bradford Books, Cambridge, MA.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R., Hubbard, W., and Jackal, L. (1989). Back propagation applied to handwritten zip code recognition. *Neural Computation* **1** (4), 541–551.

McClelland, J. L. (1989). Parallel distributed processing—Implications for cognition and development. In *Parallel Distributed Processing—Implications for Psychology and Neurobiology.* (R. Morris, ed.). pp. 104–139. Clarendon Press, Oxford, UK.

Norman, D. (1986). Reflections on cognition and parallel distributed processing. In *Parallel Distributed Processing* (J. McClellan, D. Rumelhart, and T. P. R. Group, eds.), (Vol. 2, pp. 110–146). MIT Press, Cambridge, MA.

Plunkett, K., Marchman, V., and Knudsen, S. L. (1990). *From rote learning to system building: Acquiring verb morphology in children and connectionist nets.* Paper presented at the Proceedings of the 1990 Connectionist Models Summer School.

Rumelhart, D., and McClelland, J. (1986). On learning the past tenses of English verbs. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (J. McClelland, D. Rumelhart, and T. P. R. Group, eds.), Vol. 2, pp. 216–271. MIT Press, Cambridge, MA.

Rumelhart, D., Smole-sky, P., McClelland, T., and Hinton, G. (1986). Schemata and Sequential Thought Processes. In PDP Models in J. McClelland and D. Rumelhart (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT Press, Cambridge, MA) pp. 7–58.

Satz, D., and Ferejohn, J. (1994). Rational choice and social theory. *Journal of Philosophy* Vol. 9:102:71–87.

Schulz, T. (1994). The challenge of representational redescription. *Behavioral and Brain Sciences* **17,** 728–729.

Simon, H. (1982). *Models of Bounded Rationality,* Vols. 1 and 2. MIT Press, Cambridge, MA.

Vygotsky, L. (1962). *Thought and Language.* MIT Press, Cambridge, MA.

Wason, P. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology* **20,** 273–281.

Wason, P., and Johnson-Laird, P. (1970). A conflict between selecting and evaluating information in an inferential task. *British Journal of Psychology* **61,** 509–515.