Rules: The Basis of Morality… ?

Paul M. Churchland

Most theories of moral knowledge, throughout history, have focused on behavior-guiding *rules*. Those theories attempt to identify which rules are the morally *valid* ones, and to identify the *source or ground* of that privileged set. The variations on this theme are many and familiar. But there is a problem here. In fact, there are several. First, many of the higher animals display a complex social order, one crucial to their biological success, and the members of such species typically display a sophisticated knowledge of what is and what is not acceptable social behavior —but those creatures have no *language* at all. They are unable even to *express* a single rule, let alone evaluate it for moral validity. Second, when we examine most other kinds of behavioral skills—playing basketball, playing the piano, playing chess—we discover that it is surpassingly *difficult* to articulate a set of discursive rules, which, if followed, would produce a skilled athlete, pianist, or chess master. And third, it would be physically impossible for a biological creature to identify *which* of its myriad rule are relevant to a given situation, and then apply them, in real time, in any case. All told, we would seem to need a new account of how our moral knowledge is stored, accessed, and applied. The present paper explores the potential, in these three regards, of recent alternative models from the computational neurosciences. The possibilities, it emerges, are considerable.

Keywords
Moral character | Moral knowledge | Moral perception | Moral rules | Neural networks | Non-discursive knowledge | Skills

Author

Paul M. Churchland
pchurchland@ucsd.edu
University of California
San Diego, CA, U.S.A.

Commentator

Hannes Boelsen
hboelsen@students.uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

An old college teacher of mine once remarked to me that "[a] philosopher's fundamental mistakes often appear on the very first page of his major treatise". A possible instance of this eyebrow-raising historical insight is the opening page of the long section on moral philosophy found in the prominent undergraduate philosophy textbook entitled *Introducing Philosophy*—from *Oxford University Press*, no less—skillfully edited by Robert C. Solomon (2001). Solomon there begins his broad survey of this profound and important topic with the following explanatory definition:

> The core of ethics is morality. Morality is a set of fundamental rules that guide our actions.

You may well wonder how there could be anything controversial about this lucid statement, for it does indeed capture the focus of at least ninety percent of the moral philosophers' writing in the Western traditions of religious and academic philosophy. It also captures the focus of most contemporary moral discussions, even in the marketplace and at the dinner table. We are all familiar with, and frequently argue about, presumptive "moral rules," both major and minor. We are all familiar with the competing rationales often offered in explanation of the presumed authority of such rules—that they come from God, or that they are part of the social contract, or that (when followed) they serve to maximize collective welfare, and so forth. How *else* should we focus and pursue our concern with moral reality? How else might one even *begin* to address the topic?

Hereby hangs a tale. For there are indeed other ways of approaching the topic, both as engaged citizens and as theorizing philosophers. A monomaniacal fixation on *rules* and on the source of their *authority* may reflect a fundamental misconception of what is actually going on inside successful moral agents when they engage in typical moral cognition. It may misrepresent the underlying nature of anyone's precious moral virtue. It may misrepresent the learning process by which the moral virtues are acquired. And it may misrepresent the ways in which those virtues are actually exercised in our day-to-day moral reasoning.

Before citing historical/moral authorities in hopes of winning some credence for this admittedly audacious suggestion, let us survey some of the many *non*-moral, *empirical*, or *factual* reasons for entertaining an approach to understanding morality that is not focused on rules. Such extra-moral reasons are not hard to find.

First, and perhaps foremost, rules in the literal sense require a language in which they can be expressed (and taught, and imposed, and discussed, and modified). But none of the many social creatures on this planet—excepting only humans—possess any language at all, and certainly none equal to the task of expressing even the simplest of social rules. Chimpanzees, wolves, baboons, and lions, for example, are quite innocent of language, and yet their collective behavior displays a complex social order that the adult animals must respect—on pain of punishment or retribution from their peers—and which the juveniles must learn to recognize, understand, and eventually protect with their own watchful behavior. They, too, live within a more-or-less stable moral order that serves many if not most of the same functions served by our own moral order. An adult chimp will chide, sometimes severely, a juvenile chimp that steals food from the hands of an infant chimp, and will even return the stolen food to the aggrieved victim. Wolves, and even domestic dogs, will offer comfort and solace to a wounded compatriot and will spring to defend it against future threats. The trust, social foresight, and mutual dependence displayed by a pack of lions organizing and executing a hunt to bring down a gazelle is a marvelous example of collective purposeful activity. And the subsequent sharing of the spoils among all who participated in the hunt is a striking example of distributive justice, even if momentary squabbles occasionally break out over access to the choicest bits of the kill. (Nobody is morally perfect, especially a tired and hungry lion.)

In sum, moral perception, moral reasoning, moral activity, moral norms, moral defense, and moral retribution all exist elsewhere in the animal kingdom (presumably for many of the same reasons that they exist in us), but in none of those other cases do language or discursive rules play any role at all in the moral phenomena at issue. The whole thing happens—most of it, anyway—but without language.

So what is going on? What is it that regulates or steers their behavior, if not rules? Before canvassing possible answers to this question, let us ponder some additional data, this time concerning humans. Adult humans occasionally fall victim to something called *global aphasia*, a stroke-induced brain malady in which the cortical areas responsible for the manipulation, production, and comprehension of *language*—in any form: spoken, written, or printed—are totally destroyed. The loss of this critically important neuronal machinery (roughly, Broca's area and Wernicke's area, typically on the left side of the brain) leaves the victim without any capacity to formulate, process, or comprehend any linguistic structures whatsoever. That dimension of cognitive representation is now completely out of business. There is nothing wrong with the victim's sensory inputs or motor outputs; these peripheral systems remain entirely functional. The cognitive deficit lies deeper. The capacity for even *forming* linguistically structured thoughts has disappeared entirely. The victim cannot formulate or comprehend any declarative sentence, nor any interrogative sentence, nor any imperative sentence, nor any rule. These elements, so familiar to the rest of us, no longer play any role in their cognitive lives.

And yet their cognitive lives in other respects remain surprisingly unaffected, despite this disaster where specifically *linguistic* structures are concerned. Some three decades ago, we had such a left-brain stroke victim in our own extended family. Aunt Betty, as she was fondly called, could still drive a car around town, shop for the groceries, cook a dinner, and watch a football game on TV with understanding and enjoyment. More to the point, her basic trust in other humans, and her own basic trustworthiness, were quite intact. During visits, her comprehension of the moral flux around her, especially where the adventures and interactions of our youngish children were concerned, seemed quite undiminished, as were her skills in providing comfort for the teary-eyed and fairness in the distribution of small pastries at lunch. Her moral cognition was up and running smoothly, evidently, much as before—but without the benefit of any rules to tell her what to do. She could no longer comprehend or even contemplate them, and yet somehow, she didn't need them.

Another illustration of the superfluity of rules to moral character emerged, without warning and to much amusement, in an interview of a moderately charming Georgia Congressman on the TV comedy show *The Colbert Report.* The topic of their extended discussion was a recent higher-court ban on the public display of the Judeo-Christian Ten Commandments in the foyer of a Louisiana courthouse, and the justice/injustice of their subsequent court-ordered removal from that public venue. The congressman, a Mr. Lynn Westmoreland, was defending their public, cast-bronze-on-granite display on a variety of grounds, but most trenchantly on the grounds that, collectively, those ten rules constitute the very foundation of our morality, insofar as we have any morality. Their public display, therefore, could only serve to enhance the level of individual morality.

Sensing an opportunity, Steven Colbert nodded his presumptive assent to this claim, and asked his guest, "Could you please cite them for us, congressman?" Westmoreland, plainly taken aback by the request, gamely began, "Don't lie, . . . don't steal, . . . don't kill, . . ." as Colbert, with his eyebrows raised in expectation, held up first one finger in response, then two, then three. After an awkward pause at that point, the congressman, who had plainly drawn a blank beyond those three, bravely and with evident honesty said, "No, I'm sorry. I can't name them all". My immediate reaction (oh, alright, my second) was sympathy for the congressman, because I don't think I could have named them all, either. At which point Colbert ostentatiously thanked his guest for his wisdom and brought the interview, before a large audience, to an uproariously received and laughter-filled conclusion.

The comedic point was plain enough and doesn't need any further elaboration from me. But there is a deeper lesson to be drawn from this exchange. The fact is, the congressman is probably as good an example of worthy moral character as one is likely to encounter at one's local post office or grocery store. After all, he inspired sufficient public trust to get himself elected, and he thinks morality important enough to defend it, with some passion and resourcefulness, on television. He is a presumptive example of a conscientious man with a morally worthy character. But if he is, these welcome virtues are clearly *not* owed to his carrying around, in memory, a specific list of discursive rules, rules at his immediate command, rules that he literally consults in order to guide his ongoing social behavior. He could remember only three of the ten "commandments" at issue, and, if you check the bible, he didn't get two of those three quite right in any case. If we are looking (and we *are*) for an explanation of the actual ground or source of people's moral behavior, the proposal that we are all following a specific and finite set of discursive *rules* in order to produce that behavior is starting to look strained and threadbare, to put it mildly.

Before addressing an alternative explanation, let us note one further domain of empirical evidence, relevant to our issue concerning the role of rules. Moral expertise is among the most precious of our human virtues, but it is not the only one. There are many other domains of expertise. Consider the consummate skills displayed by a concert pianist, or an all-star basketball player, or a grandmaster chess champion. In these cases, too, the specific expertise at issue is acquired only slowly, with much practice sustained over a period of years. And here also, the expertise displayed far exceeds what might possibly be captured in a set of discursive rules consciously followed, on a second-by-second basis, by the skilled individuals at issue. Such skills are deeply inarticulate in the straightforward sense that the expert who possesses them is unable to simply *tell* an aspiring novice *what to do* so as to be an expert pianist, an effective point guard, or a skilled chess player. The knowledge necessary clearly cannot be conveyed in that fashion. The skills cited are all cases of knowing *how* rather than cases of knowing *that*. Acquiring them takes a lot of time and a lot of practice.

To be sure, the point-guard can instruct the novice, "When you get possession of the ball at your end, dribble it down the floor toward the opposition's basket, and when the defense starts to resist, pass the ball to whichever of your teammates has the best chance of sinking a shot." But this rule, even if it is tatooed on the novice's forearm, will hardly make him an effective player. It doesn't tell him how to *dribble* effectively, nor could any other list of rules. It doesn't tell him how to *recognize* a teammate's fleeting opportunity to take a high-percentage shot, or perhaps set one up for yet a third player. It doesn't tell him how to *pass* the ball so as to avoid interception, or how to *deceive* the defense with various kinds of fakes and feints. It doesn't even address the issue of how to execute any one of the dozen or so different kinds of shots he himself might have to take, or when to take them. It doesn't tell him .01 percent of what he needs to know to be a skilled player. And even if he did somehow memorize 10,000 rules on all of these diverse topics, he couldn't possibly recall, from that vast store, exactly the rule relevant at any instant and then apply it swiftly enough to steer his ongoing play. The game unfolds much too quickly for that plodding strategy to be effective. Something else is going on inside the basketball player's head. Something else entirely.

The game of chess is much slower, of course, and simpler too. But the same lesson emerges here as well, although from an unexpected direction. Unlike the basketball case, and because of the discreteness and comparative simplicity of chess, computer programmers have indeed written computer programs—that is, large sets of literal rules for the computer to consult and follow— that will enable a computer to play a creditable game of competitive chess. These programs were common by the early 1980s, and they were competent enough to defeat non-expert human chess players (such as me) quite regularly.

The computer-guiding rules were written so as to address any arbitrary configuration of chess pieces on the board, as might emerge in the course of a game, and to evaluate, in sequence, the cost or benefit of each of the perhaps thirty legal moves (or something in that neighborhood—it will vary) then available to the computer. To be at all effective, this strategy requires that the computer also considers the potential cost/benefit (to the computer) of its opponent's possible *responses* to each of those contemplated moves. Each such response would of course present the computer with a new set of possible moves of its own, each requiring evaluation, and so on, for another cycle of possible moves-and-responses. If the computer is to look ahead in this fashion for only two cycles of play, it will already be evaluating something like $(30 \times 30) \times (30 \times 30) =$ 810,000 or almost a million possible move-sequences! And if it presumes to look forward, in this brute-force evaluative fashion, a mere *four* cycles of play, its task explodes to examining the cost/benefit ratio for almost a *trillion* possible move-sequences.

Now you and I could never hope to execute a game-strategy of this kind, but a computer can, although just barely. Let us assume that the computer's central processing unit (CPU) has a clock-frequency of, say, 100 Megahertz (= 100 million elementary computations per second), a fairly modest machine, these days. Such a computer will take only (1 trillion moves to be evaluated) / (100 million evals/sec) = 10,000 seconds, or about three hours to complete its evaluation of four cycles of play, assuming that the cost/benefit estimate for each move-sequence (a comparatively simple matter) can be calculated in a single elementary computation.

"But this is still ridiculous," you might say. "Three *hours* of mulling per turn!? That's not even legal. And looking ahead only four move-cycles? *That's* not going to defeat a really good human chess player." And you would be right. But in fact, some artful pruning of the decision-tree constructed by the computer's program (e.g., through ignoring some possible moves, on both sides, that are likely to be irrelevant) will substantially reduce the combinatorial explosion in the number of moves that need to be evaluated. This can reduce the time of evaluation from three hours to perhaps three minutes, though at some cost to security. A somewhat faster CPU might further reduce it to less than three seconds. And the occasional deployment of a slightly more penetrating five or six-cycle lookahead evaluation for the occasional moves of potentially great value, positive or negative, can add some deeper, if localized, insight without adding too much in the way of a computational burden. In these ways the programmed computer can be brought into the range of real-time chess competence, even excellence.

Still, it is worth remarking that it took over three decades of program and computer development before a chess-playing computer was finally able to defeat a world-champion human chess master. The Russian master Gary Kasparov (poor devil) finally went down to an IBM monster computer named "Deep Blue" in 1997, to the celebration of nerds and technophiles everywhere (Campbell et al. 2002). That is, the gross strategy of applying discursive rules, again and again at blistering speeds, finally paid off. But it did so only because the computer CPU's clock-speed was roughly a million times faster than any cyclic process in a human brain (which maxes out at a mere one hundred cycles per second) and only because the conduction velocity of the electrical signals inside the computer (almost the speed of light) was roughly a million times faster than the conduction velocity in a human nerve fiber (about the speed of a fast bicycle rider). These make the computer about (a million times a million =) a *trillion* times faster than we are. Without these singular and *superhuman* physical advantages, the computer and its list of rules—its program—would be dead in the water. And so would we humans, if the rule-based strategy were how human chess-playing competence is grounded. But plainly it is not. It couldn't possibly be. Something else is going on inside the human chess-master's head. Something else entirely.

2 An alternative account of moral skill

We have only recently begun to understand what that "something else" is. It has to do with the peculiar way the brain is wired up at the level of its many billions of neurons. It also has to do with the very different style of representation and computation that this peculiar pattern of connectivity makes possible. The basics are quite easily grasped, so without further ado, let us place them before you.

The first difference between a conventional digital computer and a biological brain is the way in which the brain *represents* the fleeting states of the world around it. The retinal surface at the back of your eye, for example, represents the scene currently before you with a pattern of simultaneous (repeat: simultaneous) activation or excitation levels across the entire population of rod- and cone-cells spread across that light-sensitive surface. Notice that this style of representation is entirely familiar to you. You confront an example of it every time you watch television. Your TV screen represents your nightly news anchor's face, for example, by a specific pattern of brightness levels ("activation" levels) across the entire population of tiny pixels that make up the screen. Those pixels are always there. (Tiptoe up to the screen and take a closer look.) What changes from image to image is the *pattern* of brightness levels that those unmoving pixels collectively assume. Change the pattern and you change the image.

It is the same story with any specialized population of neurons, such as the retina in the eye, the visual cortex at the back of the brain, the cochlea of the inner ear, the auditory cortex, the olfactory cortex, the somatosensory cortex, and so on and so on. All of these neuronal areas, and many others, are specialized for the representation of some aspect or other of the reality around us: sights, sounds, odors, tactile and motor events, even features of social reality, such as facial expressions. These neuronal activation-patterns need not be literal *pictures* of reality, as they happen to be in the special case of the eye's retinal neurons. But they are *representations* of the fine-grained structure of some aspect of reality even so, for each activation-*pattern* contains an enormous amount of *information* about the external feature of reality that, via the senses and internal brain pathways, ultimately produced it.

Just *how much* information is worth noting. The retina contains roughly 100 million light-sensitive rods and cones. (In modern electronic camera-speak, it has a rating of one hundred megapixels. In other words, your humble retina still has *ten times* the resolution of the best available commercial cameras.) Compare this to the paltry representational power of a typical computer's CPU: it might represent at most a mere 8 bits at a time, if it is an old model, but more likely 16 bits or 32 bits for a current machine, or perhaps 64 or 128 bits for a really high-end machine. Pitiful! Even an old-fashioned TV screen simultaneously activates about 200,000 pixels, and an HDTV will have over two-thirds of a million (1,080 × 640 = 691,200 pixels). Much better. But the retina, and any other specialized population of neurons tucked away somewhere in the brain, will have roughly 100 million simultaneously activated pixels. Downright excellent. Moreover, these pixels—the individual neurons themselves—are not limited to being either on or off (i.e., to displaying a one or a zero), as with the elements in a computer's CPU. Biological pixels can display a smooth variety of different excitation levels between the extremes of 0 percent and 100 percent activation. This smooth variation (as opposed to the discrete on/off coding of a computer's bit-register) increases the information-carrying capacity of the overall population dramatically. In all, the representational technique deployed in biological brains—called *population coding* because it uses the entire population of neurons simultaneously—is an extraordinarily effective technique.

The brain's *computational* technique, which dovetails sweetly with its representational style, is even more impressive. (As with any computer, a computational operation in the brain consists in its *transforming* some input representation into some output representation.) Recall that any given representation within the brain typically involves many millions of elements. This poses a prima facie problem, namely, how to deal, swiftly, with so many elements. Fortunately, what the brain *cannot* spread out over *time*—as we noted above, it is far too slow to use that strategy—it spreads out over *space*. It performs its distinct elementary computations, many trillions of them, each one at a distinct micro-place in the brain, but all of them at the *same time*. Let us explain with a picture so you can see the point at a glance.

At the bottom of figure 1 is a cartoon population of many neurons—retinal neurons, let us suppose. As you can see, they are currently representing a human face, evidently a happy one. But if the rest of the brain is to *recognize* the specific emotional state implicit in that sensory image, it must *process* the information therein contained so as to activate a specific pattern within the secondary patch of neurons just above it. That second population, let us further suppose, has the proprietary job of representing any one of a range of possible *emotions*, such as happiness, sadness, anger, fear, boredom, and so forth. The system achieves this aim by sending the entire retinal activation-pattern upward via a large number of signal-carrying axonal fibers, each one of which branches at its upper end to make fully eighty *synaptic connections* with the neurons at this second layer. (Only some of these axonal fibers are here displayed, so as to avoid an impenetrable clutter in the diagram. But every retinal neuron sends an axon upward.)

When the original retinal activation-pattern reaches the second layer of emotion-coding neurons, you can see that it is forced to go through the intervening filter of (4,096 axons ×

80 end-branches each = 327,680) almost a third of a million synapses, *all at the same time*. Each synaptic connection magnifies, or muffles, its own tiny part of the incoming retinal pattern, so as collectively to stimulate a *new* activation-pattern across the second layer of neurons. That new pattern is a representation of a specific emotion, in this case, happiness. The third and final layer of this neural network has the job of discriminating these new 80-element patterns, one from another, so as to activate a single cell that codes specifically for the emotion still opaquely represented at the second population. That is achieved by tuning a further population of synaptic connections from every cell in the middle layer to each of the five cells in the final layer. In all, what was only *im*plicit in the original retinal activation-pattern (mostly in the mouth and eyebrows) is now represented *ex*plicitly in the top-most activation-pattern across the five cells there located.

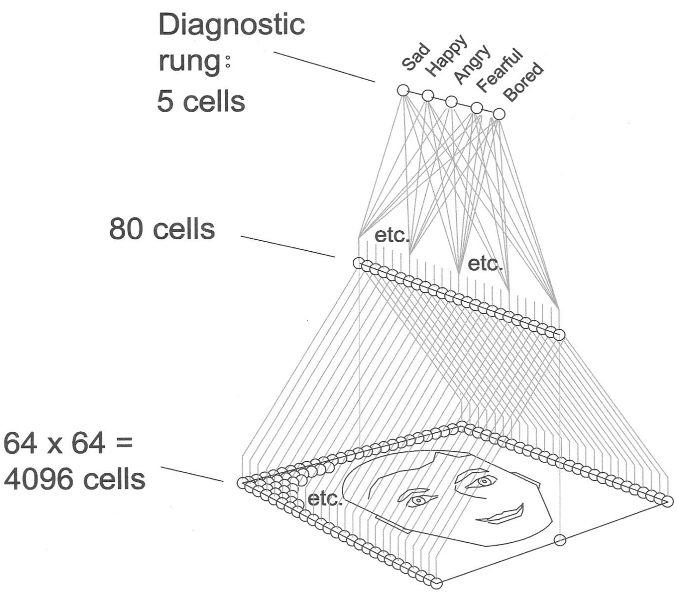

Figure 1

This trick is swiftly turned by the special configuration of the various *strengths* of each of the intervening synaptic connections. Some of them are very large and have a major impact in exciting the upper-level neuron to which it is attached, even for a fairly weak signal arriving from its retinal cell. Other connections are quite small and have very little excitatory impact on the receiving cell, even if the arriving retinal signal is fairly strong. Collectively, those 327,680 synaptic connections have been carefully adjusted or tuned, by prior *learning*, to be maximally and selectively sensitive to just those aspects of any face image that convey information about the five emotions mentioned earlier, and to be "blind" to anything else. The complex "pattern transformation" they effect plainly loses an awful lot of information contained in the original (retinal) representation. Indeed, it loses most of it. But it does succeed in making explicit the specifically emotional information that this little three-layer "neural network" was designed to detect.

This style of computation is called Parallel Distributed Processing (PDP), and it is your brain's principal mode of doing business on any topic. Even in this cartoon example, you can see some of the dramatic advantages it has over the "serial" processing used in a digital computer. A typical 8-bit CPU has a population of only eight representational cells at work at any given instant, compared to fully 4,096 just for the sensory layer of our little cartoon neural network. The CPU performs only eight elementary transformations at a time, as opposed to 327,680 for the neural network, one for each of its 327,680 synaptic connections. When we consider the human brain as a whole instead of the tiny cartoon network above, we are looking at a system that contains roughly a thousand distinct neuronal populations of the same size as the human retina, all of them interconnected in the same fashion as in the cartoon. This gives us (1,000 specialized populations × 100 million neurons per population =) a total of 100 *billion* neurons in the brain as a whole. As well, the total number of synaptic *connections* there reaches more than 100 *trillion*, each one of which can perform its proprietary magnification or minification of its arriving axonal message at the very same time as every other. Accordingly, the brain doesn't have to do these elementary computations in laborious temporal sequence in the fashion of a digital computer. As we saw, a PDP network is capable of pulling out subtle and sophisticated information from a gigantic sensory representation *all in one fell swoop*. That is the take-home lesson of our cartoon network. The digital/serial CPU is doomed to be a comparative dunce in that regard, however artful may be the *rules* that make up its computer program. They simply take too long to apply.

Enough of the numbers. What wants remembering in what follows is the holistic character of the brain's representational and computational activities, a high-volume character that allows the brain to make penetrating interpretations of highly complex sensory situations in the twinkling of an eye. You are of course intimately familiar with this style of cognition: you use it all the time. Every time you recognize frustration in someone's face, evasion in someone's voice, hostility in someone's gesture, sympathy in someone's expression, or uncertainty in someone's reply, a larger version of the neuronal mechanism in figure 1 has made that subtle information almost instantly available to you.

Now, however, you know *how*: massively parallel processing in a massively parallel neural network. Or, to put it more cautiously, almost three decades of exploring the computational properties of artificial neural networks, and almost three decades of experimenting on the activities of biological neural networks have left us with the hypothesis on display above as the best hypothesis currently available for how the brain both represents and processes information about the world. No doubt, the special network processor inside you, the one that is responsible for filtering out specifically emotional information, has more than the mere two layers depicted in our cartoon. In fact, anatomical data suggests that your version of that network has the retinal information climb through four or five distinct neuronal layers before reaching the relevant layer(s), deep in the brain, that explicitly registers the emotional information at issue. The original retinal information will thus have to go through four or five distinct layers of synaptic filters/transformers before the emotional information is successfully isolated and identified. But that still gives us the capacity for recognizing emotions in less than a few tenths of a second. (The several neuronal layers involved are only ten milliseconds apart.) On matters like this, we are *fast*, at least when our myriad synaptic connections have been appropriately tuned up.

The PDP hypothesis also gives us the best available account of how that synaptic tuning takes place, that is, of how the brain *learns*. Specifically, the size or "weight" of the brain's many transforming synapses *changes* over time in response to the external patterns that it repeatedly encounters in experience. The overall configuration of those synaptic connections and their adjustable weights is gradually shaped by the recurring themes, properties, structures, behaviors, dilemmas, and rewards that the world throws at them. The resulting configuration of synaptic weights is thus made selectively sensitive to—one might indeed say *tuned* to—the important features of the typical environment in which the creature lives. In our case, that environment includes other people, and the pre-existing structure of mutual interaction and social commerce—the *moral order*—in which they live. Learning the general structure of that pre-existing social space, learning to recognize the current position of oneself and others within it, and learning to navigate that abstract space without personal or social disasters are among the most important things a normal human will ever learn.

It takes time, of course. An infant, before his first birthday, can distinguish between sadness and happiness, but little else. A grade-school child can pick up on most of the more subtle emotional flavors listed three paragraphs ago, though probably only in the behavior of young children like themselves. But a normal adult can detect all of those flavors, and more, quickly and reliably, as displayed by almost any person she may encounter. (Only psychopaths defeat us, and that's because they have deviant or truncated emotional profiles.)

Withal, learning to read emotions is only a part of the perceptual and interpretational skills that normal humans acquire. People also learn to pick up on people's background *desires* and their current practical purposes. We learn to divine people's background *beliefs* and the current palette of factual information that is (or isn't) available to them. We learn to recognize who is bright and who is dull, who is kind and who is mean, and who has real social skills and who is a fumbling jerk. Finally, we learn to *do* things. We learn how to win the trust of others, and how to maintain it through thick and thin. We learn how to engage in cooperative endeavors and to do what others rightfully expect of us. We learn to see social trouble coming and to head it off artfully. And we learn to apologize for and to recover from our own inevitable social mistakes.

These skills of moral cognitive *output* (i.e., our moral behavior) are embodied in the same sorts of many-layered neural networks that sustain moral cognitive *input* (i.e., our moral perception). The diverse cognitive interpretations produced by our capacity for moral perceptions are swiftly and smoothly transformed—again by a sequence of well-trained synaptic filters/transformers—into patterns of excitation across our *motor* neurons (which project to and activate the body's muscles) and thereby into overt social behaviors, behaviors that are appropriate in light of the moral interpretations that produced them. Or at least, they will be appropriate if our moral education has been effective.

This weave of perceptual, cognitive, and executive skills is all rooted in, and managed by, the intricately tuned synaptic connections that intervene between hundreds of distinct neuronal populations, each of which has the job of representing some proprietary aspect of human psychological and social reality. That precious and hard-won configuration of synaptic weights literally constitutes the social and moral *wisdom* that one has managed to acquire. It embodies the unique profile of one's moral *character*: it dictates how we see the social world around us, and it dictates our every move within it. It is not an exaggeration to say that it dictates who we are. If our character needs changing or correcting, it is our myriad synapses that need to be reconfigured, at least in minor and perhaps in major ways. In all of these matters, then, don't think *rules*. Think information-transforming *configurations of synaptic weights*, for it is they that are doing the real work.

3 Reconceiving moral competence in non-classical terms

What *is* that "real work"? If the neural networks that make up our brains are not in the business of applying rules, vast libraries-full of them, just what business *are* they engaged in? How should we think of what they are doing, if not as administering rules? What is the positive alternative to this traditional construal, expressed in non-technical language?

What those networks are doing is (trying to) interpret any *new* experience or situation as being an instance of some prior category that the brain already understands. They are trying to assimilate each new social/moral situation to an already grasped prototype situation, a template or prototype that has been incrementally created by the brain's prior experience with its surrounding social/moral reality. They are trying to grasp each of the endless novelties that they encounter as being just a modified case of some kind-of-thing that they have already encountered many times, and with which they have already become familiar. They are trying to interpret the fleeting here-and-now (which is always specific) in terms of their comparatively enduring background concepts (which are always general). They are trying to identify which of their various categories, categories that past experience has constructed for them, is the one into which their current experience fits most closely and most accurately. In sum, they are trying to apply their acquired conceptual and practical wisdom to their current situation.

Why should they, or rather, you, be trying to do that? For the very good reason that your acquired concepts or prototypes are precisely what contains your accumulated information about the world, information beyond what your current and highly specific experience happens to make evident. Those abstract prototypes contain presumptive information about the wide range of *features* that any instance of an applied concept can typically be expected to display, about the wide range of *relations* it will typically bear to other things, about the ways in which it will typically unfold or *behave* over time, and about the ways in which it can typically be *controlled* or *steered*. That is the point, after all, of having a conceptual framework in the first place. It embodies your accumulated understanding of the world's enduring background structure, your grasp of the unchanging background framework within which the ephemeral and the changeable are always constrained to unfold.

Consider, for an example of moral perception in particular, the arrival of lunchtime in a typical elementary-school classroom. Every student retrieves a paper-bag lunch from the cloakroom and settles down to consume its contents. You are one of those students and, while eating, you perceive Johnny surreptitiously attempt to remove a banana from the lunch-bag next to Michael. On the face of it, you are witnessing a case of theft. And that interpretation implies many things: that the banana belongs to Michael, that Michael will be seriously aggrieved when he discovers Johnny's affront, that Johnny has inadequate self-control, that a noisy conflict will ensue if events are left to themselves, and so on and so on. This situation, as described, warrants some immediate intervention.

Most obviously, you might just openly berate Johnny in front of the other students. Or, more boldly, you might seize the banana from Johnny and quietly return it to Michael. Or you might call the teacher and rat Johnny out. These hardly exhaust your possible responses, but they are all typical sorts of responses to a typical sort of problem, and which response you choose will depend on contextual factors such as how big and mean Johnny is, how susceptible he is to collective disapproval, and how reliable the teacher is at dispensing justice. Perhaps the first path is the best response, with the second and third left as backups if the first path fails to return the situation to a just equilibrium.

And so that is what you do: berate him on the spot. All within a second of witnessing the presumed theft. Because your eight-year-old brain is already keenly tuned to that sort of possibility and to thousands of other social possibilities as well. Given your well-trained neural networks, it takes only the external perceptual situation itself to provoke the interpretation of theft. And it takes only that conceptual interpretation itself, in the context of one's ever-present character and background information, to activate your overt social response.

Your interpretation, of course, might be incorrect. Perhaps Johnny was just trying to retrieve his own banana, earlier stolen by the avaricious Michael. Perhaps your openly berating *Johnny* was inappropriate, since everyone in the class except you witnessed Michael's earlier theft but was too frightened of Michael to do anything about it. If so, Johnny has now been victimized twice over, once by Michael and once by you.

To be sure, there are many other convoluted possibilities, in addition to or beyond this one. But they are increasingly unlikely, compared to your first take on the situation. This is why your brain fell so swiftly and easily into that straightforward interpretation: *theft* is the simplest, most obvious, most probable explanation of what you have actually seen, and that's why it's the explanation that the brain tries first. Furthermore, once that explanatory assumption is in place, an immediate attempt at restitution is the most natural expression of your antecedent character and your acquired social skills.

What is impressive here is not just the swiftness with which your cognitive resources get tapped. It is the enormous *range* of alternative possibilities to which your brain is/was no less prepared to respond, and with equal swiftness, insight, and know-how. If, instead of a banana theft, you had witnessed Mary accidentally press her hand against the point of a newly sharpened pencil, your recognition of her pain and your comforting response would have been just as quick. If you saw the class's pet rabbit escape from its (poorly locked) cage, you would know to retrieve it and return it to its proper home. If you had turned to see a small fire blazing in the classroom's bookcase-corner library, with Johnny (him again!) slipping a plastic lighter into his pocket, you would grasp the significance of the event instantly and let out a loud warning to everyone in the room. If (here we deliberately choose something unlikely) Superman, with cape swirling, then bursts through the open classroom window and asks, "Which way did the fire-bug go?!," you would know to point to Johnny's fleeing backside as he hightails it out the classroom door. If . . . if . . . if . . . for a thousand thousand "ifs" and more, even your eight-year-old self would be competent to recognize the situation and to respond to it swiftly and appropriately.

This extraordinary breadth of capacity is a consequence, in part, of the *combinatorics* of the already large number of neurons the brain uses to represent any sort of social situation. It is the same trick, once again, used by your television screen, in order to display an almost endless variety of possible pictures, despite a (large but) finite set of pixels with which to portray them. The retina of the eye uses the same trick, recall, but boasts many more "pixels" than a TV screen. Your perceptual capacities, accordingly, far exceed the modest range of that familiar technology.

Of course, simply representing something at the perceptual level does not mean that you *understand* it, and that is strictly what concerns us here. To understand a perceptual input is, as we saw above, to assimilate it to one of the brain's learned *prototype* situations, to one of the standard, recurring, well-patterned kind of circumstances that one's past experience has impressed upon your memory and your habits of behavior. That memory and those habits, you will recall, are a matter of the acquired configuration of the brain's synaptic connections and their synaptic strengths or "weights." For it is those collective synaptic "filters" or "transformers"—the ones that intervene between each of the brain's many reality-portraying neuron populations— that steer the initial perceptual representations into the higher-level prototype patterns that fit those percepts most closely.

Look again at the toy network of figure 1 and note that its 327,680 synaptic connections were there adequate to steer a wide variety of possible input face images into one or other of exactly five emotional prototypes. If we suppose that this ratio (i.e., 327,680 synapses for every five categorial prototypes) is roughly typical, then a brain like yours, with 100 *trillion* synaptic connections, should be able to learn, and to deploy immediately (when appropriate), something in the neighborhood of (5 / 327,680) × 100 trillion = *1.5 billion* distinct categorical prototypes! Now, presumably we don't have quite *that* many distinct categories awaiting activation. That number is best viewed as a theoretical upper limit on what we might achieve. But in light of how our cognitive systems evidently do their jobs, it is small wonder that even your grade-school self is hair-trigger ready for such an astonishing range of situations, social and otherwise—and ready, note well, with an astonishing range of understanding and relevant skills.

4 Moral conflict and moral reasoning

Alas, our cognitive systems don't always work perfectly. Sometimes we misinterpret what we are seeing and hearing. That is, sometimes we assimilate the case at hand to a category or prototype of which it is not an instance, to which it positively does not belong. When that happens, you become the victim of the entire family of expected features, relations, developmental profile, and presumptively appropriate behavioral responses that automatically come with that prototype, but that fail to accurately characterize or suit the case at hand. Some dimensions of the activated prototype may fit (that's why you deployed it in the first place), but others do not, as you slowly come to appreciate. As the case before you unfolds, and perhaps as you learn more about its initial stages, your prototype-driven expectations are violated and your cognitive dissonance grows. You have somehow failed to understand the situation correctly.

At some point, the accumulated new input or evidence may be sufficient to kick your brain's activational activity *out* of the prototype-category that initially captured it and *into* a different and more appropriate prototype, one whose overall profile finally does fit the case at hand. At that point you may have the familiar "click" experience, where the problematic case suddenly re-presents itself in a new and coherent light, and you think to yourself, "Oh my god, I misunderstood what was happening." You may then struggle to repair the social/moral damage that your automatic but ultimately inapt behavior may have produced.

This happens to all of us, and quite often. It reflects the fact that our moral cognition is not infallible. Happily, such mistakes can be corrected, and regularly they are, sometimes by oneself and sometimes with the help of others. Unhappily, sometimes they are not corrected. We are all familiar with people who have too quickly taken a superficial interpretation of some social/moral issue and then stubbornly refuse to respond to, or even to see, its failures to capture adequately the social/moral complexities that the issue presents.

When this happens, we have a typical case of moral conflict. If the issue is pressing, we may begin a round of moral reasoning and moral argument with the person or persons who take the competing interpretation of the issue, and who propose a problematic response or policy in light of it. Such arguments, it must be admitted, often begin with both sides citing some favored "moral" or other, a rule that supposedly compels us to take their response to the situation or to embrace their policy recommendation. But this rarely settles the conflict, since the real disagreement is usually about how we should *interpret* the situation in the first place.

Classic examples are right in front of us. The public debate over abortion involves a presumptive conflict between the rule "Any innocent human person has the moral right to continue living" and the rule "Any woman has the moral right to control her own internal reproductive activities." But the debates typically focus on how these rules should be *interpreted*, what *qualifications*, if any, should limit their application, and which of these conflicting rules carries the greater *authority*. Ultimately, as both sides of the debate usually *agree*, the issue boils down to whether or not the fertilized egg and/or the early fetus that develops therefrom really is, or should be counted as, a *human person* in the first place. The right-to-life folks say "yes." The defenders of choice say "no."

Our point in rehearsing this issue is that, even in the case of this most celebrated of moral conflicts, the primary issue, once again, is not really about rules. It is about how we should interpret or categorize, rationally and accurately, the early fetus. One side will argue, "It's just a clutch of unfolding stem cells, without a brain or nervous system, without any character or personal identity, without any will or consciousness, without any of the dimensions of genuine personhood. It is no more a person than a recently-planted acorn is already an oak tree." The other side will argue, "Personhood begins at conception, at fertilization. That is when God places a human soul into the now-developing egg. Accordingly, that is when the right to life begins, a right not to be subsequently denied. (And by the way, acorns don't have immaterial souls.)"

The first side will respond, "We don't accept your utterly unverified theory of immaterial souls implanted by a divine being at conception, and we resist your attempt to thus impose your arbitrary and fantastical religious beliefs on the rest of us. (And by the way, modern science implies that humans don't have immaterial souls either.)" To which the second side will counter, "Your position acknowledges *no* clear or well-defined point at which the developing fetus begins to acquire rights. If it is acceptable to terminate the life of the developing fetus, why isn't it acceptable to terminate the life of a developing *newborn baby*? That would plainly be over the top, but the case of a fetus is different in no fundamental respect."

And so it goes. Each side of the debate typically attempts to get the other side to see the problematic case "in a different light," to interpret it as relevantly similar to a distinct but salient prototype whose moral status is not under dispute, to assimilate it to a category that is factually more adequate to the problematic case at hand. Thus the category "mindless clutch of cells" vies with the category "innocent and defenseless person" for our cognitive apprehension of the conceptus and early fetus. Arguments here are not conducted by repeatedly citing moral rules and deducing consequences therefrom. They are conducted by repeated attempts to highlight diverse factual similarities, and dissimilarities, between each of the contesting moral prototypes, on the one hand, and the conceptus/early fetus on the other.

I deploy this example of a moral disagreement and its typical discussion not to try to settle the issue in favor of either side here, but to illustrate the forms that moral disagreements and moral arguments typically display. It is, most assuredly, *not* the aim of this naturalistic and brain-focused essay to try to deduce any substantive moral rules from our growing understanding of how the brain conducts its moral cognition. Brains arrive at their moral wisdom by a long process of learning, often painful learning, whether in the lifetime of an individual or in the centuries-long development of a society, and there is no substitute for this learning process. It is rather like the development of *scientific* wisdom, if I may draw an optimistic analogy. At present, we are also learning how human brains engage in scientific cognition, but that does not obviate the need for our scientific communities to continue to generate theories and test them against our unfolding experience. Knowing *how* the brain works so as to generate and constantly improve our scientific understanding will not obviate the need to *keep it* working toward that worthy end, though it may help us to improve our pursuit thereof. Similarly, knowing *how* the brain works to generate and constantly improve our *moral understanding* will not obviate the need to *keep it* working toward that worthy end, though it may help us to improve our pursuit thereof. I will close on this hopeful note.

References

Campbell, M., Hoane, A. J. & Hsu, F.-H. (2002). Deep blue. *Artificial Intelligence*, *134* (1-2), 57-83. 10.1016/S0004-3702(01)00129-1

Solomon, R. C. (Ed.) (2001). *Introducing Philosophy.* New York, NY: Oxford University Press.