9 The Instinctive Mind Resurrected: Modularity, Reciprocity, and Blended Response

> It is Nature that causes all movement. Deluded by the ego, the fool harbors the perception that says "I did it."
> —Bhagavad Gita (3:27)

> He who understands baboons would do more towards metaphysics than Locke.
> —Charles Darwin

Before we dive into the tethered mind, we need to address the small elephant in the room. As noted in chapter 1, there is a model of human behavior in evolutionary psychology called "massive modularity," which accepts the reality of instincts (referring to them as "modules") and argues that they are sufficient to explain all or most human behaviors. It is not widely accepted beyond evolutionary psychology, but if I'm to advocate for the inclusion of lower-level systems in models of human behavior, it is necessary to address this literature. It is a little bit of a detour because this model does not recognize a blended response; it argues that all human behavior—including reasoning—is based only or largely on instincts. This is a bold, counterintuitive hypothesis worth considering and evaluating, if only to set it aside.

I will begin by introducing and reviewing the now infamous Wason card selection reasoning task and Leda Cosmides's novel "cheater detection" instinct explanation for the pattern of results. The task and her interpretation of the results do offer a clear, simple illustration of how evolutionary psychologists think we "reason" with instincts. While I'm sympathetic to her interpretation, this task has become so embroiled in controversy that it can yield few uncontentious conclusions. It is also the case that no evidence was offered by Cosmides that "cheater detection" is actually an instinct. Therefore, I set aside this task (and massive modularity) to seek

evidence for the existence of "cheater detection" instincts and their use in a broader range of reasoning and decision-making tasks.

Data from comparative animal research and experiments on young children provide evidence that cheater detection and the related traits of self-maximization, fairness, cheating, and punishment are indeed reasonable candidates for instincts, though all but self-maximization are largely confined to humans. I then turn to the work of a small group of behavioral economists and mathematical biologists who study cooperative monetary decision-making and explain their results as an interplay between these very traits of self-maximization, fairness, cheating, cheater detection, and punishment. On closer examination, it is apparent that learning, beliefs, and reason are also a critical part of their explanation. These data take us beyond anecdotal stories and provide experimental evidence for a model of tethered rationality, whereby reason interacts with instincts to guide human behavior.

Massive Modularity: Instincts All the Way Up

The term *massive modularity* is most closely associated with Leda Cosmides and John Tooby, of the University of California, Santa Barbara. Leda Cosmides (1989) began her academic career by giving an instinct-based account of a famous reasoning task, the Wason card selection task, named after its inventor, Peter Wason. The Wason card selection task involves four cards that are placed on a table in front of the participant (figure 9.1). In the standard version, the so-called no-content version, a letter is written on one side of the card and a number is written on the other side. The cards are placed such that two letters and two numbers are visible (figure 9.1a). The participants are also given a rule; for example, "if the letter on one side of the card is a vowel, then the number on the other side must be even." The four cards are placed on the table such that one vowel, one consonant, one even number, and one odd number are visible. The participant is asked to indicate which cards need to be turned over for verification of the rule. The task is a disguised form of conditional reasoning (if P then Q), with the choices of the cards corresponding to "P," "not P," "Q," and "not Q." It can be trivially completed by turning over all the cards. However, the instructions are to complete it by turning over as few cards as possible. Surprisingly, the accuracy rate on this task ranges from 10% to 25% (Goel, Shuren, Sheesley, & Grafman, 2004; Griggs & Cox, 1982). Everyone will select E to see if there is an even number on the other side (confirmation bias effect). Many people will stop at this point. Others will additionally select K or 4,

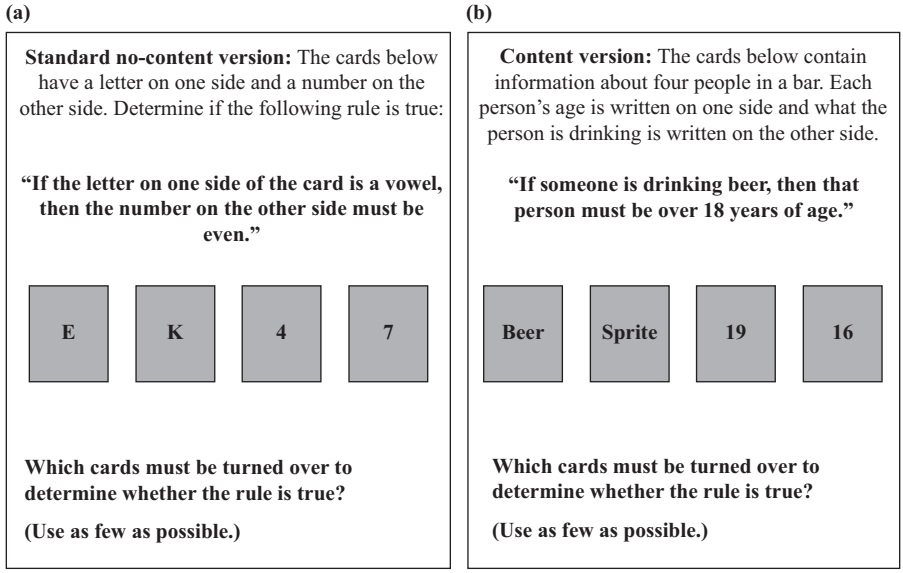


Figure 9.1
Wason card selection task.

which are both irrelevant, but very rarely is 7 selected. The selection of E (corresponding to P) and 7 (corresponding to not Q) is the correct response.

Now consider the second version of this task (figure 9.1b). Four cards are again laid out on the table. On one side of each card is written the name of a beverage and on the other side a number indicating the age of the person drinking the beverage. The layout is such that the beverages "beer" and "Sprite" are visible, along with the ages "19" and "16." The rule that must be confirmed is, "if someone is drinking beer, then that person must be over 18 years of age." Again, the task is to determine whether the rule is being violated by turning over as few cards as possible. In this logically identical version of the task, accuracy rates jump to between 65% and 90% (Goel et al., 2004; Ragni, Kola, & Johnson-Laird, 2017). Under this condition, many participants correctly select "beer" and "16." Turning over the "beer" card allows them to confirm that the individual drinking beer is over 18 years of age. Turning over the "16" card allows them to confirm that no one under the age of 18 is drinking beer. Participants rarely turn over the "Sprite" and "19" cards. The rule places no restrictions on the age for drinking Sprite, so the selection of "Sprite" provides no relevant information. Similarly, turning over "19" provides no relevant information because the rule allows for people older than 18 to drink whatever they want. This task

has spawned many hundreds of studies and at least a dozen different theories to become one of the most studied and contested tasks in the cognitive reasoning literature (Ragni et al., 2017).

The manipulation in this task falls under the content effect (chapter 7), but not all content has the same effect on all participants. For example, a rule involving content about different classes of mail and the price of postage stamps increased accuracy in participants in the United Kingdom but not in the United States (Griggs & Cox, 1982; Johnson-Laird, Legrenzi, & Legrenzi, 1972). Much of the literature on this task has been concerned with what aspects of content allow for dramatic increase in performance accuracy. Cosmides (1989) made the novel suggestion that what was in play here was not just a content effect but rather the activation of an instinct for detecting cheaters. People are aware that we have laws that establish a minimum drinking age for alcohol. If someone drinking alcohol is underage, they are breaking the law. Like many social, cooperative species, we have evolved adaptive "cheater detection" mechanisms that allow us to quickly identify anyone who might be breaking the law or violating a "fairness" norm (i.e., cheating). Notice that it is indeed the identification of the underage individual (the cheater) that accounts for the large accuracy swing. The situation simply triggers the cheater detection module—just as the swollen abdomen and the posture of the female stickleback unlocks the innate release mechanism of the male's mating behavior—and that generates the correct answer. It has nothing to do with coherence relations. Many cognitive psychologists have responded to Cosmides by questioning her data and interpretation (Atran, 2001; Carlisle & Shafir, 2005) or ignoring it (Ragni et al., 2017). I came to find the results and interpretation plausible only after observing my children interact when they were younger.

Based on such data, and steeped in evolutionary and computational ideas, Cosmides and Tooby (1994a, 1994b) went on to develop an account of cognition and reasoning different from that reported in chapter 6. Rather than a general-purpose reasoning system, based on coherence relations between proposition-like representational structures, they postulated a system of numerous domain-specific instincts or modules triggered by specific environmental cues. Each module is a specific adaptation. Adaptations are specific traits of organisms that arise based on how they improved the reproductive success of the organisms' ancestors. An organism is broken down into a collection of individual traits or adaptive solutions capable of solving specific problems, such as solicitation of assistance from parents, detection of safe and unsafe foods, coalition formation, cooperation,

cheater detection, in-group/out-group formation, inference of intentions from facial expressions, incest avoidance, mate selection, object recognition, and spatial distribution of objects in the local environment, among others. Instincts embody these adaptive solutions. There is little else to the mind. It is instincts all the way up from the bottom to the top. Importantly, in the context of these modules, information is not thought of as beliefs with propositional contents, introduced in chapter 6.

The model of the mind that massive modularity presents is one where all behavior (including reasoning) emerges from the interaction of these various individual modules or instincts. Somehow, these modules, which were developed in response to the challenges faced by our hunter-gatherer ancestors in the Pleistocene environment, interact not only to allow us to avoid incest, select suitable mates, and detect cheaters but also to reason and assess how gravity affects the fabric of space and time. Such an account may be possible but remains unrealized.[1]

I believe evolutionary psychologists are correct in reminding us that instincts play as important a role in our behavior as they do in the behaviors of bats, beavers, and baboons. This obvious insight has great value and should be embraced. It needs to be part of any model that purports to explain real-world human behavior. However, the massive modularity model itself is a nonstarter for me because it cannot accommodate propositional attitudes and coherence relations. Without accounting for these, there is no accounting for reason. The initiated reader will know that there are deep conceptual issues in play here that have been widely discussed in the literature (Buller, 2006; Fodor, 2000). I register my own conceptual critique in the appendix of this chapter. The reader more interested in tethered rationality than in the conceptual issues surrounding massive modularity can usefully skip this appendix.

The balance of this chapter delves into the following questions: (1) What is the evidence that cheater detection and related traits are instincts? (2) Are there data from tasks less controversial than the Wason card selection task to support the role of cheater detection in human decision-making? (3) What sort of model of reasoning and decision-making do the data portend? I conclude that there are indeed good reasons to regard cheater detection (along with related traits of self-maximization, fairness, cheating, and punishment) as instincts and that data from financial decision-making tasks suggest that they are clearly involved in decision-making, but importantly they are modulated by reasoning systems. This allows us to start building a model of tethered rationality.

Reciprocity and Cheater Detection in Nonhuman Animals

Darwinian selection is based on the inherently selfish mechanism of "survival of the fittest" or, as Tennyson versified, nature being "red in tooth and claw." This implies that organisms will maximize resources (food, mates, shelter) for themselves rather than for others. It is easy to find examples of this on any branch of the phylogenetic tree. However, it is also the case that humans and many nonhuman animals live in socially organized groups and will help other members of the group even at a net cost to themselves. Classic examples are food sharing among vampire bats and sentinel duty and alarm calling (presumably at greater risk to oneself) among some birds and mammals. Prima facie such altruism is problematic for the theory of evolution.

In a seminal paper, "The Evolution of Reciprocal Altruism," Robert Trivers (1971) proposed a solution to this problem. He argued that though *seemingly* altruistic, cooperative behavior actually confers fitness benefits to the donor (altruist) because it is rendered with the expectation that the recipient will reciprocate in the future. For the practice to flourish, a number of conditions need to be met, including (1) that individuals be nontransient and live in stable social groups (to maximize the number of opportunities for donors to be reciprocated); (2) that the social groups have flat dominance relations (dominant individuals can take what they want without reciprocating); (3) that individuals be able to recognize other individuals and retain memory of past interactions; (4) that the recipient must reciprocate at least once; (5) that the benefits to the recipient must be greater than the cost to the donor; and (6) that donors must be able to recognize cheaters and expel them from the system. Without this last condition, the recipient of the aid could enjoy the benefits without having to reciprocate in the future and would be at a net advantage and the donor at a net loss. Reciprocators would disappear in such a system. Hence, the evolution of an ability to recognize and punish cheaters is necessary to maintain the stability of reciprocity (Wade & Breden, 1980).

Trivers's account resulted in a nice evolutionary story of a behavior rooted in biology and widely available across the phylogenetic tree. It provided Cosmides with the cheater detection explanation of people's behavior in the Wason card selection task. But subsequent research suggests that the story is not so simple. The data indicate that helping behavior in nonhuman animals is not reciprocal altruism. It is either kin based (i.e., the assistance is being offered to genetically related individuals) or does not meet Trivers's criteria for reciprocity (de Waal & Brosnan, 2006; Riehl

& Frederickson, 2016). In these cases, no cheater detection mechanism is required. Does such a mechanism actually exist?

Certain mammals, such as meerkats and Belding's ground squirrels, perform sentinel duty as members of a group. One would think that such activities are altruistic in that, to help the group, the individual is increasing its own chance of predation. A careful examination suggests otherwise. In the case of meerkats, once an individual is well fed, sentinel duty allows the animal to benefit from early detection of danger and actually reduces its risk of predation compared to other members of the group (Clutton-Brock, 1999). For Belding's ground squirrels, the picture is a little bit more complex. When they detect an airborne predator, such as a hawk, they "whistle." When they detect a mammalian predator (e.g., a weasel or coyote), they emit a "trill." When they emit a whistle to warn against airborne predators, they are attacked 2% of the time, compared to 28% of the time for nonwhistlers, again greatly reducing their chance of predation. So, both sentinel duty and whistle calls, despite appearances, are actually selfish acts that increase individual fitness, consistent with the Darwinian account. However, when Belding's ground squirrels emit a trill to warn against mammalian predators, they are attacked 8% of the time, compared to 4% of the time for nontrillers. In this case, the individual trilling squirrel is putting itself in harm's way to help conspecifics. But this behavior is actually driven by kinship relations (Sherman, 1985).

Even kin-based altruism was problematic for the notion of direct (or individual) fitness in the Darwinian theory of evolution, but it was nicely reconciled by several important developments in the neo-Darwinian update that incorporated knowledge of the gene and in the 1960s led to the insight that the unit of selection in evolution was not the individual, group, or species but rather the gene (Hamilton, 1963, pp. 354–355):[2] "Despite the principle of 'survival of the fittest' the ultimate criterion which determines whether [a gene] G will spread is not whether the behavior is to the benefit of the behaver, but whether it is to the benefit of the gene G."

The individual was simply the container for the gene. Based on this reformulation, William Hamilton (1964a, 1964b) redefined the notion of fitness as "inclusive fitness" and proposed kinship theory. Inclusive fitness equals direct fitness (individual reproductive success) plus indirect fitness (reproductive success of relatives the altruist has aided). Hamilton argued that inclusive fitness can be maximized by helping genetically related individuals, even at a cost to oneself, and proposed the following rule: altruistic behavior will spread if $rB > C$, where r = coefficient of relatedness, B = fitness benefit to the beneficiary, and C = fitness cost to the altruist.[3]

An example of reciprocal altruism that meets some of Trivers's criteria but fails others is provided by food sharing among vampire bats. Vampire bats feed only on blood. Approximately 8% of adults will fail to feed successfully on any given night. If a bat fails to feed on two or three consecutive nights, it will starve to death. Food sharing through regurgitation of blood meals from a successful individual to an unsuccessful individual is commonly observed in vampire bats (Wilkinson, 1984). Much of the sharing is between mothers and pups and other related female individuals (Carter & Wilkinson, 2013). However, sharing also occurs in the case of familiar but unrelated individuals (Denault & McFarlane, 1995). This practice seems to meet a number of criteria for reciprocal altruism: vampire bats live in social groups, the females are nontransient and have weak dominance hierarchies, and the benefit to the recipient bat is greater than the cost to the donor bat (because the former could otherwise die). However, there is no evidence that vampire bats can distinguish kin from associates, nor is there any evidence that they can track cooperative returns (i.e., detect cheaters) and base future donations on these returns (Carter & Wilkinson, 2013).

But what about more encephalized animals such as nonhuman primates, particularly chimpanzees, our closest living relatives? Chimpanzees live in social groups and cooperate with unrelated partners. Males cooperate in patrolling territory, hunting, sharing food, grooming, and joint mate guarding, and even form within-group coalitions for aggressive actions against other members. Despite all these prosocial behaviors, in experimental settings they will not volunteer to help another familiar but unrelated individual obtain food, even at no cost to themselves (Silk et al., 2005). In one study, they show a small increase (5.7%) in sharing food with a familiar individual who has groomed them within the past two hours compared to an individual who has not groomed them within this time period (de Waal, 1997). In other studies, they fail to show that their altruism is conditional on reciprocity.

Sarah Brosnan and her colleagues (2009) carried out an experiment on captive chimpanzees to determine whether they would more readily share food with a partner from their home group who had shared food with them on previous trials versus partners who had not shared food with them. Individuals familiar with each other were tested in pairs. One individual was offered a choice between two options: (a) deliver a food reward to themselves and another equal one to the other individual (prosocial behavior) or (b) deliver a food reward to themselves and nothing to the other individual (selfish behavior). On the next trial, the other individual was offered the same choice. The trials were repeated a number of times. Interestingly, the

choices individuals made were not affected by the choices that their partners made in previous trials. That is, any food cooperation was not contingent on previous interactions. Several leading primatologists now agree that reciprocal altruism (and hence cheater detection) is nonexistent among nonhuman animals, even including nonhuman primates (de Waal & Brosnan, 2006; Stevens & Hauser, 2004). It may be something unique to humans.

Frans de Waal and Sarah Brosnan (2006) propose three levels of reciprocity, of which the first two can be found among nonhuman animals and the third seems exclusive to humans: symmetry-based, attitudinal, and calculated or contingent reciprocity. The simplest form, symmetry-based reciprocity (i.e., "we are friends"), requires that both parties behave similarly with each other; it is based on existing relationships such as kinship, group membership, alliances, and similarity in age. It does not require scorekeeping. There is a very low degree of contingency. The altruistic behavior of meerkats, Belding's ground squirrels, and vampire bats would fall into this category. By contrast, attitudinal reciprocity requires that an individual's willingness to cooperate covary with the recent attitude of the partner ("if you are nice, I will be nice"). Both parties may not benefit simultaneously, but the requirement of scorekeeping is minimal. The contingency is immediate. The exchange is based on "general social disposition rather than specific costs and benefits" (Brosnan, de Waal, & Proctor, 2014, p. 24). The altruistic behavior of chimpanzees reported here would fall in this category. Finally, in calculated or contingent reciprocity (Trivers's reciprocal altruism), individuals expect reciprocation of at least equal value, though allow for significant time lags. If expectations are violated, cheaters will be punished.

In addition to the prerequisites identified by Trivers—the benefit to the recipient greater than the cost to the donor, opportunity for repeated interaction, reasonably flat dominance hierarchies, and the cheater detection mechanism—full-fledged contingent or calculated reciprocity requires quite sophisticated cognitive abilities, such as recognition of individuals, memory of previous events, scorekeeping, numerical discrimination, and even temporal discounting. The only robust examples of it occur in humans.

The task used to test for calculated reciprocity in chimpanzees can also be used on young children. It is reported that children 3–4 years of age choose like the chimpanzees (House, Henrich, Sarnecka, & Silk, 2013). They usually choose not to share, and this choice is independent of whether the other child shared with them on previous trials. Children 5–7 years of age, on the other hand, make the prosocial choice (70% of the time) when their partners have made the prosocial choice on previous trials. Children of this age are beginning to show signs of contingent or calculated reciprocity.

According to these data, any claims that calculated reciprocity and cheater detection are evolutionary traits widely dispersed along the phylogenetic tree are false. Contingent reciprocity does have simpler precursors (i.e., symmetry-based and attitudinal reciprocity)—that do not require cheater detection—but fails to present itself in unambiguous form even in our closest living relatives, where we might well expect it. It is possible that it arose only on the hominina or even homo branch of the phylogenetic tree, piggybacking on increased cognitive capacities, perhaps even propositional attitudes. Indeed, Trivers's original description is replete with appeals to propositional attitudes and other sophisticated cognitive and emotional systems largely confined to humans. The failure to find robust calculated reciprocity in nonhuman animals does not preclude it as a candidate for an instinct. However, the fact that it does not arise in humans until five years of age suggests a period of maturation and/or socialization. There may also be some more basic instinctual systems that feed into it.

Reciprocity in Humans: Self-Maximization, Fairness, Cheating, and Punishment

Any Darwinian model of human behavior must begin with the selfish maximization of resources. However, resources can sometimes be multiplied exponentially through mutual cooperation. A single individual may be able to hunt a rabbit or build a hut. A cooperating group of individuals can bring down a mammoth and build Chartres Cathedral; the group result may be greater than the sum of individual efforts. The starting point for any model of human cooperation needs to be based on sharing of effort and rewards. This assumes and requires a sense of *fairness*. While *self-maximization* of resources is widely present along the phylogenetic tree, fairness may be unique to the hominina or even homo branch. Human cooperative behavior is an interplay between fairness and self-maximization. Unchecked self-maximization will lead to a violation of fairness (i.e., cheating). Unabated cheating would result in a breakdown of cooperation. Fairness (hence cooperation) is maintained by not only *detecting* cheaters but also actively *punishing* them (even at a cost to self). Computational models suggest that such interacting systems can result in stable cooperation (Axelrod & Hamilton, 1981; Boyd, Gintis, Bowles, & Richerson, 2003; Fowler, 2005).

Are these traits learned or are they instincts? It has long been a tenet of Western society that these concepts, particularly fairness, are cultural and social, even religious, constructs. If this is the case, they must be learned, will emerge late with socialization, and will correlate with individual and

societal variations in beliefs. The instinct-based view advocated by evolutionary psychologists is that these notions are innate constructs and a common heritage of at least *Homo sapiens*. There is indeed empirical evidence to suggest that self-maximization, fairness, cheating, and punishment are all adaptations or instincts. They have not been arrived at through socialization (i.e., learning and reasoning).

The most interesting data in support of the innateness view of fairness are emerging from the study of young infants. Infants as young as 19 months old expect resources and rewards to be divided equally between two individuals. In a game-playing scenario with pairs of infants, 21-month-olds expect the experimenter to distribute rewards equally when both infants worked to complete the task but not when only one worked at the task and the other played another game (Sloane, Baillargeon, & Premack, 2012). It is very difficult to argue that cultural and social influences are driving the behavior at this early stage.

Three-year-olds share more equally with a collaborating partner than with a freeloader (Melis, Altrichter, & Tomasello, 2013). Three- and four-year-old children engaged in collaborative tasks objected to inequitable reward distribution, even when it favored themselves, and in such cases equalized rewards by transferring some of theirs to the partner. Chimpanzees performing the same task are insensitive to the inequity and are only concerned with maximizing their own resources (Ulber, Hamann, & Tomasello, 2017). These data indicate that the concept of fairness emerges very early, prior to extensive cultural socialization, and is thus best considered innate or instinctive. Furthermore, if the concept of fairness emerges so early, and fair-minded individuals can be taken advantage of by cheaters, emergence of cheating should follow. This is indeed the case.

While children seem to understand the concept of fairness at a very early age, they do not always follow the principle when their own resources are at stake (i.e., they often cheat). When children three to eight years old were given stickers and asked to share with children who did not receive any, they all said giving them half the stickers would be fair. When it actually came time to share, the seven-to-eight-year-olds did share half their stickers, but the younger children gave fewer than half their stickers (Blake, McAuliffe, & Warneken, 2014). The result with the older, more socialized children shows that instinctive biases can in certain situations be modulated (to varying extents) by social, belief-based factors.

In another study, children ranging in age from 5 to 15 years were asked to toss a fair coin and privately record the results (black or white). The children were to be rewarded based on the number of white trials they reported.

They were told that the experimenters would not check the actual tosses but rather just take their word for it. Statistically, one would expect approximately 50% white trials. The children reported on average 85% white trials, well above the expected 50% but also below 100%. There were no age or sex differences. The experiment was then repeated with a prior admonition not to cheat. Here the overall white responses were reduced by 13% in boys and 36% in girls (Bucciol & Piovesan, 2011). Interestingly, the admonition dampened but did not eliminate the cheating. This speaks to a role for socialization (learning and reason-based beliefs) in shaping cheating behavior and also to the limits of socialization. The fact that cheating behavior cannot be eliminated by socialization speaks to some innate components.

Given that the notion of fairness and the propensity to cheat develop very early and universally, and seem to be only modestly affected by socialization, the development of cheater detection and punishment should not be far behind. Consistent with this expectation, it is reported that children two to three years of age can understand normative rules, as in the structure of simple games, and detect violations (i.e., detect cheaters) (Rakoczy, Warneken, & Tomasello, 2008). In the context of moral transgressions by a third party, three-year-old children can detect such violations and even intervene by tattling on the transgressor (punishment) and behaving more prosocially toward the injured party (Vaish, Missana, & Tomasello, 2011). These data are consistent with the evolutionary psychologists' claim that these traits are instincts. The evolutionary psychologists would further argue that we should be able to explain human cooperative decision-making as the interaction of these various traits or instincts. To evaluate this claim, we now consider how these traits actually come into play in cooperative decision-making. That is, is the massive modularity model—of just interacting instincts—sufficient to account for the cooperative decision-making data, or do we need to introduce learning and reason to explain the data?

Tethered Rationality: Blend of Instincts and Reason in Cooperative Economic Decision-Making

A small number of economists and mathematical biologists have recently traded in the exclusively self-maximizing *Homo economicus* model of decision-making for models based on the actual study of human nature. Prominent among this group are Martin Nowak of Harvard University, and Ernst Fehr of the University of Zürich, and their colleagues. They accept the innateness of the mechanisms of self-maximization, fairness, cheater detection, and punishment, with some even including the notion of trust

(Berg, Dickhaut, & McCabe, 1995; Ortmann, Fitzgerald, & Boeing, 2000), and explore their interaction in cooperative monetary decision-making. In fact, some of these economists expand Trivers's notion of contingent or calculated reciprocity, where we reward or punish if it is in our long-term self-interest, to a notion of "strong reciprocity," where we will bear the cost of rewarding and punishing even in the absence of any long-term benefits to ourselves (Fehr & Fischbacher, 2003). While it remains unclear how this extension reconciles with the theory of evolution, it is an interesting conjecture.[4]

The question of whether human reciprocity is adequately characterized as contingent reciprocity or strong reciprocity is interesting but not particularly germane for our purpose. Either way, there are in place a set of adaptations or instincts guiding our cooperative behavior. Numerous studies characterizing human cooperative monetary decisions as a function of these traits have been undertaken. I will suggest that the valuable data and insights that they have generated are best accommodated by a model of tethered rationality where human cooperative choices are a blend of beliefs, coherence relations, and instincts.

Human decision-making, specifically choice in cooperative resource-allocation situations, is studied by economists through the use of a handful of simple games. Four such games are the Dictator Game, the Ultimatum Game, the Trust Game, and Social Dilemma Games. In the Dictator Game, there are two players. One player (called the donor) is endowed with a sum of money and instructed that he can keep it all or share a portion with the other player (called the recipient), who has received nothing from the experimenter. The recipient must accept whatever (if anything) is offered by the donor. Because the donor did nothing to earn the reward, fairness would dictate that he or she share half the money, whereas the self-maximizing choice would be to give the recipient player nothing. When the game is played as a single shot (i.e., no repetitions) and with actual money, there are significant individual differences among players: 40% of donors will choose to keep all the funds and only 20% will share equally with the recipient player, with others sharing smaller amounts (Forsythe, Horowitz, Savin, & Sefton, 1994). These results show similar trade-offs between fairness and self-maximization, with a preference for the latter, as noted in the children's data.

The results can be shifted dramatically and reliably under certain conditions. When the game is played with imaginary money, 80% of the donors will share 40%–50% of the funds with the recipient (it costs them nothing) (Forsythe et al., 1994). If there are repeat trials of the game, with the donor

and recipient alternating, then donors become even more "generous," because they know they will be at the receiving end in the next trial and will have to deal with the consequences of their reputation for defecting or cooperating (Berg et al., 1995). *Reputation* is a critically important *belief-based* factor, discussed shortly. Generosity can also be manipulated by instilling certain beliefs in the donor about the recipient (for example, the recipient is dying of cancer or has recently insulted the donor) (Eimontaite, Nicolle, Schindler, & Goel, 2013; Eimontaite et al., 2019). In these manipulations the instilled beliefs—albeit with important emotional components—modulated the outcome of the choice. Beliefs do not even have to be explicitly instilled. A manipulation whereby the two players spend a few minutes silently looking at each other increases generosity in single-shot games compared to totally anonymous interactions (Bohnet & Frey, 1999). This process allows not only for the humanization of the other player but also for identification for future interactions (i.e., it raises concerns about reputation).

The Ultimatum Game also involves two players (donor and recipient), with only one (the donor) receiving an initial sum of money. However, there is an interesting twist. The donor must offer some of the money to the nonreceiving (recipient) player. If the recipient accepts the offer, then they both get to keep the allocated funds. If the recipient rejects the offer, both walk away with nothing. The self-maximizing choice for the donor is to offer as little as possible to the recipient (as in the Dictator Game). The self-maximizing choice for the recipient is to accept any nonzero amount offered, because even if only 1¢ is offered from $10, that 1¢ is greater than the alternative of 0¢.

In actuality, any offer less than 25% of the original amount is roundly rebuffed. Only offers around the 40%–50% mark are routinely accepted. This is an instance of fairness being enforced by cheater detection and punishment. The recipient detects a violation of fairness (i.e., cheating) in low offers and punishes the donor player at a cost to themselves. They would rather have nothing—and have the donor receive nothing—than accept an unfair offer. In anticipation of this response, donors usually offer something in the 40%–50% range. Interestingly, in the case of the Ultimatum Game, it seems to make no difference whether the game is played with real money or imaginary money, presumably because of the presence of the real threat of punishment by the recipient (Forsythe et al., 1994). This is an example of a self-maximizing choice being rejected in favor of punishing the cheating behavior.

Punishment is costly. In the preceding example, in order to punish the donor, the recipient has to forgo whatever amount the donor offers. What

is even more interesting is that we will expend resources to punish not just those who cheat or harm us (or could potentially do so in the future) but also those who cheat others. This speaks to our strong sense of fairness. This can be illustrated by a modified version of the Dictator Game that includes three players: a donor endowed with the money, a potential recipient, and a third party. The donor is endowed with $10, the potential recipient with nothing, and the third party with $5. The donor may give whatever he wishes to the recipient player. Once the transfer is made, the third party is informed that they can spend money to punish the donor if they so wish. Every dollar spent by the third party reduces the income of the donor by $3. Where the donor has violated fairness norms, the third party will use some of their own funds to punish them (Fehr & Fischbacher, 2004). Again, this is not immediately self-maximizing for the third party but speaks to the important role of cheater punishment in our behavior. Empirical data (Dreber, Rand, Fudenberg, & Nowak, 2008; Fehr & Gächter, 2000) and computational models (Boyd et al., 2003; Fowler, 2005) suggest that punishment (even at a cost to oneself) is critical for maintaining cooperation based on reciprocity. But there may also be a reason-based calculation involved, specifically that long-term punishment may lead to formation of a reputation as a "punisher," thereby reducing the probability of being cheated in future interactions (Hilbe & Traulsen, 2012). Other evidence of reason-based modulation of punishment behavior includes calculation of cost-benefit trade-offs such that people are more likely to punish if the cost of the punishment is less than its consequences (Egas & Riedl, 2008).

Cheating and punishment instincts emerge very early and are ubiquitous in human social affairs. In every newspaper around the world, we will find stories such as the following (Ingalls, 2011): "Woman in Washington State living in a million dollar waterfront mansion with her Jaguar driving chiropractor husband, receives monthly welfare assistance of $1272 for housing, federal and state payments for a disability, and food stamps." We pay attention to such stories. We become outraged. We demand punishment for the cheaters. This is all in line with the adaptation story. When the cheaters belong to an out-group, such as immigrants, we are extra incensed, demand greater punishment, and generalize more broadly (e.g., "all immigrants are welfare cheats"). This may result from interaction of the cheater detection adaptation and the out-group aversion adaptation (chapter 13). Reactions to instances of cheating may be instinctive but can clearly be modulated by beliefs, as predicted by the model of tethered rationality.

Beliefs about the trustworthiness of other members—known as their *reputation*—are a critical factor in cooperation and punishment (Milinski,

2016). An individual's reputation is established by their history of choices in previous interactions. It is transmitted either directly (through firsthand knowledge of previous choices) or indirectly, via language or some other system of symbols. Either way, potential donors are less likely to help (i.e., more likely to punish) those who have previously violated fairness norms. In repeated game interactions, where both individuals know that the other will have knowledge of their past interactions, cooperation rates rise dramatically. In fact, individuals are aware of the value of a good reputation and will expend resources to gain one. In an experimental situation where an individual has the possibility of developing a positive reputation, the cooperation rate rises from 37% to 74%. Reciprocity and a good reputation reinforce each other (Gächter & Falk, 2002).

Reputation also modulates punishment. Consider the differences in how corporate tax avoidance is viewed and punished in the United States compared to single mothers collecting housing vouchers and food stamps while holding an unreported secondary job. A recent study by Oxfam America (2020) reports that between 2008 and 2014, the 50 biggest US companies received $27 in federal loans, loan guarantees, and bailouts for every dollar they paid in taxes. Each dollar that the biggest companies spend on lobbying is associated with $130 in tax breaks and more than $4,000 in federal loans, loan guarantees, and bailouts. Another study by economists estimates that tax avoidance by major corporations costs the US Treasury up to $111 billion a year (Clausing, 2016). Interestingly, these facts rarely make the front page of most newspapers or incense most of us.

Corporations are run by humans. *Homo sapiens* will cheat, to some extent, if they can get away with it, irrespective of whether they are corporate CEOs or welfare recipients. The question is, why doesn't corporate cheating activate our cheater detection and punishment modules to the same extent as a single mother welfare recipient working an unreported side job? Americans have been raised to have different beliefs about corporations and single mothers on welfare, which either attenuate or accentuate the triggering of the relevant instincts. For example, we are taught that corporations are the backbone of society. They provide jobs. They grow food, build cars, and provide health insurance, among other things. Corporations are good. They spend billions of dollars shaping our beliefs, and it works (Wu, Balliet, & Van Lange, 2016). Single welfare mothers do not have the lobbyists to explain why it might be necessary to hold down a couple of side jobs while claiming welfare assistance to pay the rent and buy food (Kohler-Hausmann, 2007), so they are stuck with the following reputation (Feagin, 1972): "About welfare? What do I think about welfare? It ought to

be cut back. The goddamn people sit around when they should be working and then they're having illegitimate kids to get more money. You know, their morals are different. They don't give a damn."

In both cases, whether it is corporate CEOs lying to get around regulations in order to increase profits or single mothers on welfare lying about holding a second job, it is cheating and should activate our detection and punishment systems equally. However, because these systems are modulated by reputations, which are often in the form of beliefs—and corporate CEOs have much better reputations than welfare mothers—welfare cheats are much more likely than corporate CEOs to go to jail.[5] These modulations can both amplify and attenuate cheater punishment, *but they cannot eliminate it*.

We will not only punish cheaters but also reward those who play fair. Consider the Trust Game. There are again two players. In this case, both players are given an equal amount of money. The first player has the option of transferring some arbitrary portion of his money to the second player, with the understanding that the experimenter will triple any amount that is transferred. The second player can then decide whether to keep all the funds or send a portion back to the first player. If the first player decides not to transfer any funds to the second player, each player keeps the initial funds. However, given the tripling rule, the self-maximizing choice for the first player is to transfer all their funds to the second player, as long as the second player then transfers half the tripled amount (or at least more than they received) back to the first player. This way, both players come out ahead. But there is a danger. What if the second player violates fairness and keeps everything for himself? If funds are transferred, it is immediately self-maximizing for the second player to keep all the proceeds and not send anything back. In this situation, the self-maximizing outcome is distant for the first player and relies on fairness, while the self-maximizing outcome for the second player is immediate and relies on cheating. Despite this, even in single-shot games, most players choose to make a substantial transfer, and the transfer back (i.e., reward) made by the second player correlates with the amount of the initial transfer (Eimontaite et al., 2013; Fehr & Fischbacher, 2003). Any knowledge, beliefs, and perceptions about the trustworthiness of the other player (i.e., their reputation) modulates the initial transfer amount as well as the returned amount (Berg et al., 1995). Repeat trial games automatically generate such knowledge and, of course, affect trust.

Tragedy of the Commons

Some of the most intractable societal problems involve allocation of public goods and take the form of a social dilemma that Garrett Hardin (1968)

famously labeled the "tragedy of the commons." These are all problems of exploiting common resources for selfish ends. The problem of initiating action to combat global warming is a classic example. The problem of maintaining universal healthcare schemes is another example. Both these problems were introduced in chapter 1. However, I will undertake the discussion of the tragedy of the commons with an analogous historical example where we have the advantage of 20/20 hindsight: the collapse of the Canadian Maritimes fisheries.

One of the greatest fish resources in the world was found on the Grand Banks, off the coast of Newfoundland, Canada. For 500 years, European vessels plied these waters to exploit the resource. It seemed endless. Based on self-maximization, each fisherman should maximize his take. Every extra fish means extra income. According to the logic of Adam Smith's "invisible hand" doctrine, individual self-maximization would be "led by an invisible hand to promote . . . the public interest" (quoted in Hardin, 1968, p. 1244). What could go wrong?

In a world of infinite resources (or where the amount removed from the resource is always less than or equal to the replenished amount), this might be reasonable advice. In the actual world, every resource is finite. As technological advances in fishing dramatically increased the catch of individual fishermen, individual trawlers, and individual corporations, they all did become dramatically wealthier, as did the Canadian Maritimes as a whole (along with communities in Iceland and Portugal). In 1968, the cod catch from the Grand Banks was 810,000 tons. In 1974, it dropped precipitously to 34,000 tons! Seemingly overnight, the cod population totally collapsed because of overfishing. This left 40,000 fishermen and related workers unemployed and financially decimated the Canadian Maritime Provinces (Pilkey & Pilkey-Jarvis, 2007). If every individual had limited their catch in line with the available resource, they and their children would still be fishing and prospering today, as would the Maritimes as a whole.

This is the tragedy of the commons. It constitutes a social dilemma where an individual receives a higher (self-maximizing) benefit for the socially noncooperating choice (e.g., overfishing, using excess energy, polluting, accessing healthcare without paying enough into the system), irrespective of what others do; however, everyone is better off if everyone cooperates. If not enough people cooperate (i.e., they take out too much or don't pay enough into the system), the resource will be depleted. In this case, everyone loses. Individual interest is at odds with the group interest.[6] The dilemma requires that the payoff matrix be as follows:

1. Payoff to the defectors > payoff to cooperators
2. Payoff for universal defection < payoff for universal cooperation

The situation can be modeled (poorly) in simple experiments such as the following. Take 10 players and give each $10. They are then given the opportunity to invest some or all their funds and are told that the return on investment will be distributed equally among the group. They privately place their contribution in an envelope. The experimenter collects the envelopes, multiplies the total of all the envelopes by five, and distributes the new total equally among all the players. If everyone contributes their $10, for a total of $100, everyone will take home $50 (after the $100 is multiplied by five and the $500 is distributed equally among the 10 players). However, if one player contributes nothing and the others contribute $10, for a total of $90, which is then multiplied by five to become $450 and redistributed, the noncontributing player will take home $55 ($10 plus $45) and the others will take home $45 each. Therefore, *Homo economicus* should contribute nothing in this situation. In every scenario in which the multiplier is less than the number of participants, the noncontributing player will come out ahead by contributing nothing or less than others. But if no one contributes, everyone will lose, as did the individual fishermen and the Maritime Provinces when the cod fisheries collapsed.

These experiments demonstrate a great deal of individual variation. More people than might be expected by the self-maximizing principle actually cooperate, but approximately 30% start as freeloaders, and this percentage increases to 80% or 90% by the tenth round of the game, leading to a rapid decline in cooperation and a depletion of the common resource (Isaac & Walker, 1988). Unsurprisingly, rates of cooperation increase with the introduction of punishment (Fehr & Gächter, 2000). Reputation also helps (Milinski, 2016). Our knowledge and beliefs regarding what others are contributing increases our own contribution, perhaps by triggering the fairness instinct. Interestingly, it does not affect the number of free riders (approximately 30%), but 50% of participants match their contributions to what they believe others are contributing, while 14% match contributions up to a certain point and then decline. The net overall result is a positive contribution, and the common resource is sustained (Dawes, 1980; Fischbacher, Gächter, & Fehr, 2001). Another important factor is the determination of the payoff matrix. One needs to understand the situation to understand the payoff matrix. Is it really advantageous to defect? How severe is the cost of group failure? Is it really the case that more is being taken out than put

back in? For instance, if the multiplier in the preceding example is greater than 10 (i.e., greater than the number of participants), it is more advantageous to cooperate. These modulating factors speak to the contribution of knowledge and the rational mind in decision-making.

If cooperation can be sustained by these reason-based modulations (at least in artificial scenarios), how did a wealthy, technologically advanced country like Canada succumb to the tragedy of the commons? Reason was employed. When questions of sustainability of the harvest arose, steps were taken to eliminate foreign participants from the fishing grounds by extending the coastal boundary line from 3 miles to 200 miles, and the best available science and technology was used to find the sustainable limit; that is, to find the actual payoff matrix. Fisheries experts from the Department of Fisheries and Oceans were consulted. They used sampling techniques to estimate the current cod population and used mathematical models to project future population levels. Based on these models, the Department of Fisheries and Oceans advised the government that imposition of proper catch limits would allow a recovery of the population within 10 years, and thereafter it would be sustainable to annually harvest 16% of the population (estimated at 500,000 tons). For 1989, they recommended a total catch of 125,000 tons. That is, the models indicated a very small individual gain and very high individual and group costs for defecting (noncooperating).

In this situation, it is irrational to continue unrestricted fishing for immediate marginal individual gains, given the inevitable dire consequences (destruction of livelihood). A rational choice would be to sustain the resource so it can continue to provide current and future benefits, albeit at a more moderate level. This choice could be implemented by cooperating with the government to initiate steps known to avert the tragedy, specifically (1) coercion or punishment of noncooperators and rewarding of cooperative behavior and (2) making sure people understand the long-term consequences of cooperating versus defecting (Dawes, 1980). These strategies essentially change the payoff matrix so it becomes less of a dilemma.

What did the fishermen actually do? The imposition of catch limits resulted in an outcry from fishermen, corporations, their communities, and the Maritime Provinces. They claimed, without any direct evidence to the contrary, that the population estimates of the Department of Fisheries and Oceans were inaccurately *low*. In response to the outcry and political pressure, the Ministry of Fisheries arbitrarily increased the quota to 235,000 tons. This saga played out annually for several years. In actuality, because of inaccuracies in sampling and the little-understood complexities of ecological systems, the Department of Fisheries and Oceans' estimates were much

too *high*, resulting in an annual harvest of 60% of the total population in the last few years instead of the predicted sustainable rate of 16%! In January 1992, the Department of Fisheries and Oceans recommended a harvest of 185,000 tons, but by June 1992 they had revised their recommendation to a complete halt to cod fishing. The fisheries were gone.

What can be learned about human decision-making from this example? Two obvious points can be highlighted for current purposes. Self-maximizing is a powerful force in every aspect of life. Like all evolutionary adaptations, it is local and shortsighted. Its concerns are immediate. But beliefs and reason were also an integral part of the tragedy. The dispute between the Department of Fisheries and Oceans and the fishing community played out as a disagreement about the payoff matrix. The fishing community did not overtly state, "We want to be selfish rather than altruistic." They did not state, "We do not care if the fisheries collapse and we all lose our livelihood a few years down the road." The fishing community argued (without evidence) that the department's methods for estimating fish populations were inaccurately low. The fish were still plentiful. *The fishing community refused to believe that they were harvesting more than was being replenished*. Why?

It is reasonable to question the accuracy of any model, based on reasons and evidence. If a model is incorrect, it can err in either direction (underestimating and overestimating). In questioning the department's estimates and mathematical models, the fishing communities were not privy to any special or additional information. Nonetheless, the fishermen argued that the model was underestimating the number of remaining fish. They refused to entertain the possibility that it might be overestimating the number of fish. They had few evidence-based reasons for their belief. Why didn't they reason as follows? "Even if it is underestimating the number of fish and recovery rates, and we nonetheless curtail our harvest, we will still benefit by having a greater future yield, whereas if it is overestimating, then curtailing the harvest is essential for our survival." Isn't "better safe than sorry" the rational choice here? One plausible explanation for the failure of reason in this case is the predominance of the principle of immediate self-maximization. Self-maximization would be one factor that biased the reasoning system, leading to a faulty conclusion, self-harmful behavior, and the tragic destruction of the fishermen's livelihood.[7]

But there was another important factor in play: the failure to consider, acknowledge, and accept the severity of the consequences of being wrong, of refusing to believe the facts as presented by the best available science at the time. We will consider two types of explanations for this refusal

in chapters 13 and 14. The first will involve the introduction of another instinct, in-group/out-group bias, and the second will involve the conjecture that worldviews are very difficult to revise once neural systems have matured. While this discussion has been undertaken with the historical example of the Canadian Maritimes fisheries, the same scenario is tragically playing out in the climate change debate, where the stakes are even higher.

This is also the appropriate time to revisit my American friend from chapter 1 and evaluate his aversion to the concept of universal healthcare. Maintaining a universal healthcare system is also a classic social dilemma situation, but the concern of my friend was not with maintaining the system but rather not wanting to opt into it. Many Americans have good health insurance coverage through their job, and if they are over 65, have subsidized Medicare coverage through the government. For them, universal healthcare offers no personal benefits. In fact, it may be in their self-interest to oppose it if they believe that the expansion of coverage will add freeloaders to the system and dilute care for them. (Ironically, the people on Medicare themselves are not fully paying into the system—and may or may not have contributed a fair share during their working years—but are equally concerned about *other* freeloaders.) Any sense of fairness or altruism is strongly subdued by self-maximization and self-righteously reinforced by reason fueled by beliefs about the "other." Remember the single welfare mothers? They are not deserving like us: "You know, their morals are different. They don't give a damn." While this may begin as a belief, it quickly activates the in-group/out-group system that will be discussed in chapter 13.

There is also a surprisingly large percentage of Americans who do not have good health insurance but also object to universal coverage. They would actually be better off with universal healthcare, despite the existence of freeloaders. It would be their rational choice. This is the category my friend falls into, but his cheater punishment instinct is so powerful that he is unable to dampen it and tolerate some cheaters in order to be personally better off. But there may also be another strand to the explanation. The universal healthcare plans being proposed are all by the *other* political party (Democrats), the un-American socialist party. They involve death panels (Gonyea, 2017). When my friend's anointed political representatives ascend to power, they are going to deliver a patriotic, American solution that will "cost much, much less and deliver much, much more" (Costa & Goldstein, 2017). We will also return to complete this discussion in chapter 13.

The data reviewed so far relate to variability in individual choices. If this variability is at least in part a function of beliefs and reason-based

modulations, one would expect to see societal-level variations where there are large differences in beliefs. This is indeed the case. For example, all industrialized societies provide some social assistance programs to their citizens. However, the variability in the amount of assistance as a percentage of GDP (for 2018) ranges from 11.1% in South Korea, to 18.7% in the United States, to 31.2% in France (OECD, 2020). Consistent with these differences, studies of 15 small-scale preindustrialized societies from around the world, including societies that engage in foraging, slash and burn horticulture, nomadic herding, and small-scale agriculture, revealed that the cooperative economic decisions of all groups were a function of self-maximization, fairness, cheating, and cheater punishment, but with considerable group variation attributable to societal factors (Buchan, Croson, & Dawes, 2002; Henrich et al., 2001). For example, in the Ultimatum Game, the mean offers in Western societies (as represented by undergraduate students) are approximately 44%. In the 15 societies in this study, they ranged from 26% to 58%. Rejection rates also varied widely. Western undergraduate students reject offers below 25% with high probability. In some of the preindustrialized societies, low offers were rarely rejected. In others, offers in the vicinity of 50% are frequently rejected. In a Social Dilemma Game, there's a 30% freeloading rate among Western undergraduate students. In one of the preindustrialized societies studied, not a single subject cooperated fully. As all these differences emerge across societal groups, it is plausible that they are a function of learned social norms or beliefs rather than instincts. This again indicates some modulation of instincts by learning and belief systems.

* * *

This discussion of reciprocity has highlighted several interesting features of instincts. Some instincts, such as the suckling response in mammals, are widely available; others, such as fairness and cheating, are largely restricted to the hominina or even homo branch of the tree. This has two obvious consequences. First, just because an instinct appears in a common ancestor does not necessarily mean that humans will (or will not) also possess it. Second, traits that are unique to humans need not be lesser candidates for instincts. Data and details matter.

The discussion has also highlighted the possibility of complex interactions between instincts. For example, when I approach a robin's nest containing chicks, the mother robin takes flight. She soon turns around to fly back to protect the chicks, but my presence near the nest again frightens her such that she stops in midflight and retreats, only to approach and

retreat again and again. The behavior of the bird from moment to moment is a function of the relative strengths of the different signals from fear and maternal instincts. Similarly, in accounting for contingent reciprocity, the "simpler" traits of self-maximization and cheating are consistent, but fairness tugs in the opposite direction. This type of relationship among instincts will complicate the prediction of behavior of individuals (as one will need knowledge of individual differences in the strength or intensity of various instincts) but will not affect the nature of the causal relationship between stimulus and response. This is the type of interaction envisioned by the massive modularity model.

But the data also clearly show the modulation of economic decision-making choice by beliefs and coherence relations (i.e., by reason).[8] Reason is a double-edged sword. It can be used to either attenuate or accentuate instincts. Conversely, instincts can also either reinforce or overcome reason (to a certain extent). In the face of this overwhelming evidence (to say nothing of common sense) for the interaction of instincts, beliefs, and coherence relations, any model restricted to instincts will be insufficient. Trying to explain these data without the postulation of reason is as futile as trying to explain them without the postulation of instincts.

What is the appropriate model to accommodate the data on cooperative decision-making that provide evidence for the involvement of both instincts and reason in economic decision-making? Part of the story is undoubtedly the existence and interaction of instincts as envisioned by the massive modularity model. Individual differences in the "setting" of these instincts lead to individual differences in choice, but the other, equally important part of the story is reason. The proposed model of tethered rationality—characterized by different kinds of minds, ranging from autonomic, instinctive, associative, and reasoning minds, with evolutionarily newer levels tethered to evolutionarily older levels—acknowledges all these critical components. Chapter 10 delves into the comparative neuroanatomy literature and makes the case for the evolution of hierarchically organized neural infrastructure to support tethered rationality.

Appendix: A Conceptual Critique of Massive Modularity

I find great value in the basic insights of evolutionary psychology that reiterate the importance of instincts in human behavior, but I reject the claim that all human behavior is to be explained in terms of instincts. This rejection is based on common sense and the empirical data reviewed in this chapter,

but there are also some deep conceptual reasons to reject the specific massive modularity instantiation of the insights of evolutionary psychology (Buller, 2006; Fodor, 2000). I register my objections here and conclude by reiterating the difference between reason and instincts. The reader more concerned about my positive account of tethered rationality can skip this appendix without loss of continuity.

The case for massive modularity is typically made at the level of computational architecture. Massive modularity is associated with a specific type of computational architecture very different from the physical symbol system architecture we met in chapter 6. Allen Newell and Herbert Simon celebrated the fact that their single GPS computer program could solve problems from different domains simply by switching data sets and could also solve the same problem in different ways by switching the algorithm, demonstrating both generality and flexibility. Cosmides and Tooby see a host of problems with this approach. They celebrate the fact that their computational architecture consists of a collection of independent, specialized programs that each solve very simple specific problems. As the number of behaviors an organism is capable of increases, so will the number of necessary modules. There could be hundreds, thousands, perhaps even hundreds of thousands of these modules, depending on the complexity of the organism. That is why this theory is referred to as "massive modularity." The human brain consists of a large number of these independent, task-specific computer programs or modules. As Cosmides and Tooby (1994a, p. 330) note:

> The human mind is powerful and intelligent not because it contains general-purpose rational methods (although it may include some), but primarily because it comes equipped with a large array of what we might call reasoning instincts. Although instincts are often thought of as the polar opposite of reasoning, a growing body of evidence indicates that humans have many reasoning, learning, and preference circuits that (i) are complexly specialized for solving the specific adaptive problems our hominoid ancestors regularly encountered; (ii) reliably develop in all normal humans; (iii) develop without any conscious effort; (iv) develop without any formal instruction; (v) are applied without any awareness of their underlying logic; (vi) are distinct from more general abilities to process information or behave intelligently. In other words, these reasoning, learning, and preference circuits have all the hallmarks of what people usually think of as "instincts."

In this passage, Cosmides and Tooby are unimpressed with the idea of a general-purpose reasoning system (i.e., the reasoning mind introduced in chapter 6). They suggest that reason and instincts are not polar opposites—even referring to "reasoning instincts"—and that we can understand the

former in terms of the latter. Let's take up their concerns about generality and then revisit the relationship between reason and instincts.

Objections to General-Purpose Reasoning Systems

Cosmides and Tooby (1994a, 1994b) raise three main objections to a general-purpose reasoning system. They argue that generality is (1) overrated and unnecessary, (2) inconsistent with evolutionary theory, and (3) leads to certain intractable computational problems. I worry that much of this discussion in the literature is conflating computational and conceptual issues. My own characterization of the reasoning mind was at the conceptual level, with the computational instantiation as a means of capturing the conceptual machinery. The computational issues only come into play after the conceptual issues have been sorted out. Accordingly, I will address the conceptual issues surrounding generality.

Objection 1: Generality is overrated The first objection is essentially that specialists (specific modules for solving specific problems) will do a better job of solving any given problem than a generalist (i.e., a general-purpose program for solving arbitrary problems). This may be true, but it misses the mark on the need for generality. Let's review how and why generality and flexibility enter into the reasoning mind. Conceptually, the reasoning mind is committed to a system whereby any specific stimulus is neither necessary nor sufficient for a specific response. That is, given any specific input, a reasoning mind is not predisposed to any specific response. This was the "gap" between stimulus and response in reasoning systems identified by Ernst Cassirer. It was discussed in the context of Hamlet killing his uncle Claudius. As noted in chapter 6, the various reasons proposed were all capable of justifying the act, but none was necessary or sufficient to cause the act. Furthermore, it is widely believed that the key to realizing such a system is the conceptual machinery of propositional attitudes and coherence relations. This same apparatus allows us to find novel ways of getting to work in the mornings and allows us to land a rover on Mars. Additionally, it has been proposed—with numerous caveats—that such a system can be mechanistically realized using a particular general-purpose computer architecture (Fodor, 1975; Fodor & Pylyshyn, 1988; Newell, 1980). Cosmides and Tooby seem to focus largely on criticizing this particular computational architecture rather than the conceptual system the architecture is meant to realize. If the criticism is that a particular computational architecture may not be the best way to realize this conceptual system, that is fine, but it fails to address why generality and flexibility at the conceptual level are unimportant.

Objection 2: Generality could not have evolved The claim that generality is inconsistent with evolutionary theory is predicated on a very narrow view of evolution that emphasizes the stability of the evolutionary environment of our Pleistocene ancestors, natural selection, gradualism, specific adaptations, and an increase in the complexity of organisms through a linear, uniform addition of adaptations. This leads to the conclusion that only situation-specific adaptations are possible and that general-purpose reasoning systems could not have evolved. I want to make the obvious suggestion that what evolved were mental representations with propositional content responsive to coherence relations. The problem they solved was that of maintaining *veridicality* between mental representations and the world and *consistency* of mental representations (chapters 6 and 11). Generality is just a consequence of this system. If a particular formulation of the theory of evolution cannot account for propositional attitudes, then based on the same rationale used to reject behaviorism in chapter 5, I would say so much the worse for that theory; it needs to be updated and enriched. One cannot pretend that the phenomenon does not exist. Chapter 11 illustrates how an evolutionary account based on comparative neuroanatomy does have the potential to naturally accommodate both reason and instincts.

Objection 3: Generality leads to the intractable Goodman relevance problem; massive modularity solves it The main objection that Cosmides and Tooby raise about general-purpose reasoning systems is that they are susceptible to the intractable "frame problem." This is actually a loose collection of problems, and it is not clear that they are all identical (Shanahan, 2016), but discussions with my evolutionary psychology colleagues suggest that the problem they are referring to is the Goodman relevance problem of selecting properties suitable for generalization (i.e., projectable predicates).[9] This problem was introduced and discussed with the dinosaur and grue examples in chapter 8. Recall that on finding one *Borealopelta* with fossilized stomach contents, we happily generalized the dietary habits of all *Borealopelta* dinosaurs, but finding the *identical* evidence for broken and missing scales, we were unwilling to generalize that all *Borealopelta* had broken and missing scales. The former property was relevant for generalization, the latter not. The issue was formulated more precisely by Goodman with the grue example.

Cosmides and Tooby are correct in noting the seriousness of this problem and the fact that the conceptual model of the reasoning mind presented in chapter 6 has no solution for it. Without a solution to this problem, there is no science-based psychology. Any candidate solution needs to be

considered carefully and, if it solves the problem, embraced. If I understand correctly, the solution massive modularity is offering for the relevance problem is to replace the large general-purpose database and single reasoning engine of physical symbol systems with many specialized smaller databases, each with its own "reasoning" engine. This reduces the problem search space that any particular module needs to traverse for a solution. By reducing the search space, any potential combinatorial explosions are supposedly avoided. By avoiding combinatorial explosions, Goodman's relevance problem is solved.

It is possible that I have misunderstood both the problem they're trying to solve and the solution they are offering, but if they are dealing with the Goodman problem of projectable predicates, then the size of the database that needs to be searched is irrelevant for determining the relevance of any particular predicate. Whether one has three predicates to consider or three billion, it is equally difficult to determine relevance. In the dinosaur example, there are only two predicates ("dietary contents" and "broken and missing scales"); one is generalizable, the other not. But the evidence provides no basis for differentiating between them. Goodman made the same point more rigorously with the predicates "green" and "grue."

The example that is often given in the literature is that of learning a grammar for natural language. Any given fragment of a natural language can be trivially described by an infinite number of grammars, so one might think reducing this infinite number to 20 or even 2 is a big step forward. Not as far as Goodman's selection problem is concerned. Even if there are only two possible grammars that the module has access to and the sampled data are equally consistent with *both* of them, how does the system decide? Notice that the relevance problem is not a computational problem; it is a conceptual problem (of specifying necessary and sufficient conditions for relevance). It needs a conceptual solution. The size of the database may become an interesting computational factor once the conceptual problem is solved, but it is not a factor in the solution to the problem itself. It is a red herring.

It is important to understand that not all minds have to confront the relevance problem. It is a problem specific to minds reasoning with propositional contents and coherence relations. Instinctual minds like Lorenz's (figure 4.1) don't have to deal with it. There is a causal connection between a specific stimulus and the animal's fixed action pattern (response), as in the example where the swollen abdomen and the posture of the female stickleback unlock the innate release mechanism of the male's mating behavior. The stimulus is causally necessary and sufficient for the response (with

the noted degrees of freedom allowed by the model). The male stickleback is not confronted with Goodman's problem. Neither are our homeostatic systems regulating various bodily functions nor our low-level visual system confronted with it. In the latter case, there is a topographic mapping from the retina, to the lateral geniculate nucleus (LGN), to the primary visual cortex. There are specific mechanisms for detecting edges, light and dark areas, line orientation, and other features. Hold up a certain pattern of lines at certain angles, and certain specific neurons in the primary visual cortex will respond. Hold up a pattern of lines in a different orientation, and another set of neurons will be activated. This is a causal story. No propositions are involved until the very end, when you formulate the belief that the sun is setting on the horizon. The Goodman relevance problem emerges with the emergence of propositional attitudes and conceptual coherence relations.

"Solving" the Goodman problem the way the stickleback does it—sidestepping it by replacing conceptual relations with direct causal relations—is a genuine workaround. But it is important to appreciate that it restricts you to a certain kind of mind, the kind that the stickleback has. If massive modularity is signing up for the stickleback solution, then it indeed sidesteps Goodman's relevance problem. But the proposed "cure" may be worse than the proverbial disease itself. It means doing without the "gap" between stimulus and response, which is the *conditio sine qua non* of the reasoning mind. My own view is that this is too high a price to pay. I want to keep my reasoning mind (though tethered to the stickleback's mind), with the understanding that at some future date the relevance problem must be discharged.

There are occasions on which Cosmides and Tooby seem to recognize the shortcomings of the stickleback's mind, and the need for the types of behaviors made possible by propositional attitudes, but are unsure how to get there (Cosmides & Tooby, 2013, p. 182):

> Large amounts of knowledge are embodied in intelligent, domain-specific inference systems, but these systems were designed to be triggered by stimuli in the world. This knowledge could be unlocked and used for many purposes, however, if a way could be found to activate these systems in the absence of the triggering stimuli; that is, if the inference system could be activated by *imagining* a stimulus situation that is not actually occurring: a counterfactual.

When they write "if a way could be found to activate these systems in the absence of the triggering stimuli," I'm assuming they mean a causal trigger. If this is the case, then they seem aware of the dilemma that massive modularity presents: either stick with the stickleback's mind and ignore

such situations or accept the solution provided by propositional attitudes and coherence relations (and put up with Goodman's relevance problem). I have chosen the latter.

Reiterating the Distinction between Reason and Instincts

Finally, Cosmides and Tooby insist that reason and instincts are not polar opposites. By contrast, I have claimed that instinct and reason constitute different kinds of minds. The reader is encouraged to return to table 6.1 and review each of the five dimensions along which kinds of minds were differentiated. Reason and instinct differ on each dimension. At the expense of some repetition, I will summarize: instincts are the preferred solution to guiding behavior that is essential, does not need to change across generations, and may be needed prior to any opportunity for learning. Where within-generation environmental fluctuations are in play, instincts on their own will not be sufficient. The least expensive and most widespread solution for this is a mechanism that learns through associations by tracking co-occurrence relations. An even more complex interaction with the environment involves the ability to track stimuli and changes that may not even be present at the time, to consider counterfactual scenarios, and to make flexible individualized responses. For example, I can easily imagine the consequences of rising oceans on coastal cities before it actually happens and choose to respond very differently than my neighbor. This requires still more sophisticated machinery.

In the cognitive account, this more sophisticated machinery consists of psychological intentional states, such as beliefs and desires with proposition-like representational contents, referred to as propositional attitudes. Propositional attitudes possess the properties of productivity, compositionality, systematicity, and inferential coherence; they relate to the world via a reference relation and to each other via semantic, logical, and conceptual coherence relations (chapter 6). Along with this machinery comes Goodman's relevance problem. As far as I can see, the massive modularity solution for doing away with the relevance problem entails doing away with this machinery. So, even if Cosmides and Tooby are correct about instincts and reason not being polar opposites, it should be clear that they are solutions to different types of problems and postulate different machinery. One cannot be successfully substituted for the other. Both are necessary to explain human behavior.

Reason and Less

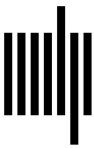
Pursuing Food, Sex, and Politics

By: Vinod Goel

Citation:

Reason and Less: Pursuing Food, Sex, and Politics
By: Vinod Goel
DOI: 10.7551/mitpress/12811.001.0001
ISBN (electronic): 9780262369701
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Goel, Vinod, author.
Title: Reason and less : pursuing food, sex, and politics / Vinod Goel.
Description: Cambridge, Massachusetts : The MIT Press, [2022] |
Includes bibliographical references and index.
Identifiers: LCCN 2021017752 | ISBN 9780262045476 (paperback)
Subjects: LCSH: Decision making. | Reasoning. | Logic. | Cognitive neuroscience.
Classification: LCC BF448 .G64 2022 | DDC 153.4/3—dc23
LC record available at https://lccn.loc.gov/2021017752