Behavioral and Brain Sciences (forthcoming)

This Target Article has been accepted for publication and has not yet been copyedited and proofread. The article may be cited using its doi (About doi), but it must be made clear that it is not the final version.

Resource-rational contractualism:
A triple theory of moral cognition

Sydney Levine

Department of Brain and Cognitive Sciences, Massachusetts
Institute of Technology (USA)
Department of Psychology, Harvard University (USA)
Allen Institute for Artificial Intelligence (USA) smlevine@mit.edu
sites.google.com/site/sydneymlevine

Nick Chater

Warwick Business School (UK)

Joshua B. Tenenbaum*

Department of Brain and Cognitive Sciences, Massachusetts
Institute of Technology (USA)
Center for Brains, Minds, and Machines (USA)

Fiery Cushman*

Department of Psychology, Harvard University (USA)

* Joint senior author

Short Abstract

We present a novel theory of moral cognition organized around resource-rational contractualism. From a contractualist perspective, ideal moral judgments are those that would be agreed to by rational bargaining agents—an idea with widespread support in philosophy, psychology, economics, biology, and cultural evolution. As a practical matter, however, investing time and effort in negotiating every interpersonal interaction is unfeasible. Instead, we propose, people use abstractions and heuristics to efficiently identify mutually beneficial arrangements. We argue that many well-studied elements of our moral minds, such as reasoning about others' utilities ("consequentialist" reasoning) or evaluating intrinsic ethical properties of certain actions ("deontological" reasoning), can be naturally understood as resource-rational approximations of a contractualist ideal.

Long Abstract

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

It is widely agreed upon that morality guides people with conflicting interests towards agreements of mutual benefit. We therefore might expect numerous proposals for organizing human moral cognition around the logic of bargaining, negotiation, and agreement. Yet, while "contractualist" ideas play an important role in moral philosophy, they are starkly underrepresented in the field of moral psychology. From a contractualist perspective, ideal moral judgments are those that would be agreed to by rational bargaining agents—an idea with wide-spread support in philosophy, psychology, economics, biology, and cultural evolution. As a practical matter, however, investing time and effort in negotiating every interpersonal interaction is unfeasible. Instead, we propose, people use abstractions and heuristics to efficiently identify mutually beneficial arrangements. We argue that many well-studied elements of our moral minds, such as reasoning about others' utilities ("consequentialist" reasoning) or evaluating intrinsic ethical properties of certain actions ("deontological" reasoning), can be naturally understood as resource-rational approximations of a contractualist ideal. Moreover, this view explains the flexibility of our moral minds—how our moral rules and standards get created, updated and overridden and how we deal with novel cases we have never seen before. Thus, the apparently fragmentary nature of our moral psychology—commonly described in terms of systems in conflict—can be largely unified around the principle of finding mutually beneficial agreements under resource constraint. Our resulting "triple theory" of moral cognition naturally integrates contractualist, consequentialist and deontological concerns.

Key Words

contractualism, ethics, moral psychology, resource-rationality, social cognition

Word Counts

Short Abstract: 112; Long Abstract: 236; Main Text: 13,997; References: 3,478; Entire Text: 17,823.

1Introduction

Scientists make their mark by disagreeing with one another, so when they agree about something, it's worth taking note. Across the many disciplines that study human morality—psychology, anthropology, economics, biology—there is widespread consensus about the function of morality. Humans have diverse interests and goals, and each person's success at achieving their goals depends on decisions made by others. We must, then, figure out individually and collectively how to adjudicate our competing interests. Morality helps us solve this problem (Baumard et al., 2013; Bicchieri, 2005; Binmore, 2005; Curry, 2016; Gauthier, 1987; J. Greene, 2014; Haidt, 2007; Rawls, 1971; Tomasello, 2009).

Considerable effort has been devoted to defining, from a rational perspective, how people should negotiate such "interdependent choices". A key, recurrent idea is that people should settle into the kinds of arrangements for mutual benefit that they would have agreed to in an idealized bargaining process (e.g. Binmore, 2005; Gauthier, 1987; Nash, 1950). And, indeed, the *agreements* we make with one another are central to our moral lives. We make promises, draw up contracts, and "talk it out" when our interests conflict. These agreements create obligations, and breaking them is morally problematic. Thus, one might naturally assume that our moral psychology would be structured around finding and adhering to mutually beneficial agreements.

Yet, while there is a large literature on the psychological mechanisms that people use to make moral judgments, relatively little of it is framed in terms of finding mutually beneficial agreements

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

(but for important exceptions see: Baumard et al., 2013; Everett et al., 2016; Kohlberg, 1969; Levine, Kleiman-Weiner, et al., 2020; Piaget, 1932; Tomasello, 2020). The field has focused instead on prohibited actions (i.e. “deontological” processes; whether established by emotions, heuristics, or rules (e.g. Baron & Ritov, 2004; Cushman et al., 2006; Haidt et al., 1993; Nichols, 2021; Nichols & Mallon, 2006)), utility-based evaluations of outcomes and welfare (i.e. “consequentialist” processes)(e.g. FeldmanHall et al., 2016; Hsu et al., 2008; Lockwood et al., 2020; Williams, 1968), and the psychological conflicts that arise when prohibitions and welfare concerns clash (e.g. Crockett, 2013; Cushman, 2013; J. Greene, 2014).

We aim to fill this gap, offering a theory of psychological contractualism. Starting with the premise that the *ultimate* function of morality is to guide interdependent choice, we argue that the *proximate* mechanisms of our moral psychology can be usefully organized around the concept of agreement—the creation and application of “contracts”, formal or informal, between people.

What would contractualist psychological mechanisms look like? The ideal way of striking mutually beneficial agreements might be to get everyone in a room together to talk through the best way to proceed—the kind of setting sometimes envisioned by game theory or moral philosophy. But explicit negotiation is usually impractical. When we lack the time, information, computational resources, and social capital for *ad hoc* negotiation, our view posits that we fall back on cognitively efficient shortcuts that approximate this ideal. These heuristics are the heart of our account. We propose a *resource-rational* contractualist psychology—one designed to trade off the cost of cognitive and social effort against the gains of mutually beneficial agreement.

Specifically, our view describes two axes of heuristic approximation of the contractualist ideal. The first involves the *terms* of the agreement. Rather than negotiating over what should be done in a single instance, we instead consider *abstractions*, or standards, that can be reused in future circumstances. The second axis of approximation involves the *process* by which agreement is reached. Rather than actually getting the affected parties together to bargain, we could *anticipate* how the negotiation would proceed using mental simulation.

This resource-rational approach has two important payoffs. First, we show how two well-understood parts of our moral psychology are, in fact, particularly useful ways of approximating *ad hoc* negotiation. One implements a form of utility maximization over weighted outcomes. Another generalizes broad constraints on allowable actions. Thus, two of the best studied elements of our moral psychology—those often called “utilitarian” and “deontological” in the moral psychology literature (Conway & Gawronski, 2013; Cushman et al., 2010; Everett & Kahane, 2020; J. D. Greene et al., 2001; Holyoak & Powell, 2016)—can be explained, in part, as resource-rational approximations to mutually beneficial agreements.

The second payoff of this view is that it explains the flexibility of our moral minds. Previous theories of moral psychology have largely been silent on how moral standards get established, updated, overridden, and revised. A resourcerational contractualist account suggests that these phenomena can be explained as renegotiations prompted by changes in the world, or by the discovery of new and better ways of achieving mutual benefit.

We call this a psychological “triple theory” of moral cognition, with a nod to Derek Parfit’s normative theory of the same name (Parfit, 2011). Like Parfit, we propose that rule-based, utility-based, and agreement-based approaches to moral decision-making are much more intimately related than they appear at first. Our theory departs from Parfit’s, however, in many respects. Chief among them is that we offer resource-rational contractualism as a powerful framework that helps explain both the connections and clashes between the three philosophical approaches, but not as an all-encompassing theory reconciling *every* facet of our moral minds. In this paper we define the predictions and potential of the resource-rational contractualist approach—and argue that many rules and welfare-based processes of moral judgment can be encompassed within it—but we also define its limitations and borders.

In sum, psychological contractualism offers a computational framework that is grounded in a rational analysis of the function of morality, that relates diverse mechanistic elements of our moral mind, and that offers clues about how they might interact.

2The Problem of Interdependent Choice

In order to understand the structure of any cognitive system it helps to identify its function—the problem that it is designed to solve (Marr, 1982; Tinbergen, 1963). Across a broad range of fields—biology, anthropology, economics, political theory, and psychology—there is broad agreement that a principle function of morality is to help guide the choices we make that affect each other, given our diverse interests (André et al., 2022; Baumard et al., 2013; Binmore, 2005; Curry, 2016; Gauthier, 1987; J. Greene, 2014; Ostrom, 1990; Rawls, 2004). After all, as humans, we are interested in pursuing our own goals based on our subjective interests.[1] Meanwhile, everyone else around us is trying to do the same. People's interests often conflict—what one person prefers may be dispreferred by others. Yet, often, each person could benefit from the help of others—there are things we can achieve together that no one can achieve alone.

An extensive formal literature in game theory attempts to characterize this problem of "interdependent choice" from a rational perspective, asking how idealized agents would act when the outcomes of their actions are partially determined by others' actions (Bacharach, 2018; Binmore, 1994; Gintis, 2014). This work establishes a normative standard for interdependent rational choice, and it also provides a template for organizing descriptive theories of human thought and behavior. Although humans are clearly not fully rational creatures, idealized rational models of choice can be a useful way to organize a scientific inquiry into the actual psychological mechanisms used by ordinary people (Chater & Oaksford, 1999; Marr, 1982).

Two ideas arising from this literature are central to our argument: That interdependent rational agents will choose arrangements of "mutual benefit", and that they can discover these by processes of bargaining and agreement. After first describing the role of these ideas in idealized models of interdependent choice (addressing the *function* of moral judgment), we then use them to motivate our key claim (addressing its *structure*): Much of our actual moral psychology arises from resource-rational cognitive mechanisms for reaching and adjusting agreements – including many situations where the agreements or agreement processes may be not be visible or only implicit.

2.1 Mutual benefit: What would rational agents agree to?

People engage in a wide variety of relationships (friendships, couples, corporations, etc.), make many types of deals, promises, and agreements, and insist on many and varied moral standards regarding how they are treated and how they must treat others. We argue that these can all be viewed as ways to achieve mutual benefit: Arrangements that tend to make each party better off than they would be otherwise. But, can we be more specific about what "mutual benefit" amounts to, and which particular arrangements we should expect?

[1] These interests need not be purely selfish. Indeed, our interests typically include concern for the well-being of others, especially our friends and family. What is crucial here is that interests are person-specific rather than shared, so that the interests of different people are likely to clash.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

2.1.1 The bargaining problem

Our strategy is to begin with a simple, well-studied, and analytically tractable case, and then to consider how its core lessons can be applied to the more complex settings which are our primary concern.

The *bargaining problem* is an economic game that formalizes the “mutually beneficial” arrangement that would be reached when two agents can create something of value together, and then must decide how to divide their profit (Nash, 1950). For instance, maybe two friends need to share a cab and divide the bill; two spouses need to divide childcare and housework; an employer and an employee need to divide company profits; or, a seller and a buyer need to agree on the price of a car. In each case, there are many different specific arrangements that would leave both parties better off than they would be if they failed to come to an agreement. But which one should they choose?

In cases like these, a widely used criterion is that rational agents will agree on a division that maximizes the product of the utility gains between the players, known as the Nash bargaining solution (Nash, 1950)[2]. When two players are identically situated, and the goods are equally valuable to them both, the Nash bargaining solution corresponds to an equal division of resources. But the players might be asymmetrically situated such that a deadlock between them is much more harmful to one of them than the other—for instance, one might go hungry while the other scarcely notices. A player with poor “outside options” if the bargain fails is said to have weak bargaining power, and the Nash Bargaining Solution predicts that they will obtain less. This is a familiar concept in labor disputes: management will obtain a better offer when they can easily hire new workers; workers will obtain a better offer if they can easily find other jobs. Meanwhile, the players might also benefit asymmetrically from the goods (e.g., a smaller person can be fed with fewer calories, while a larger person requires more to achieve the same benefit), and here the Nash bargaining solution allocates more resources to the person who requires more to achieve the same gain.

This simple, well-studied example captures the dynamics at the core of many moral situations. It also furnishes an simple example of what we mean by “mutual benefit”: The specific agreement that rational agents would come to, leaving both better off. Indeed, the logic of bargaining theory has already been used to explain the structure of human morality in simple two-person situations. Experimental work shows that when people work together to divide resources they come up with arrangements well predicted by the Nash bargaining solution, and they consider these arrangements to be fair (Binmore et al., 1993; Le Pargneux & Cushman, 2023; Mallucci et al., 2019; Rustichini & Villeval, 2014).

Obviously, most of the interdependent choices that we face deviate from the stylized assumptions of this simple case and therefore go beyond the simplified assumptions within which the Nash Bargaining Solution applies. Real bargaining is a dynamic, time-consuming process, and this can affect the bargaining solution (Binmore et al., 1986; Rubinstein, 1982). Real bargaining often involves more than two players (Harsanyi, 1963). And, we often encounter interdependent choice problems in which literal bargaining and enforceable agreements cannot be taken for granted (Misyak & Chater, 2014). As these various simplifying assumptions are relaxed, the specific agreements that people reach are sometimes predicted to deviate from the precise Nash bargaining solution. Nevertheless, what is perhaps more remarkable is how often its animating principles are preserved—that is, that we can understand each party’s behavior as an attempt to achieve a specific division of resources that both parties would agree to, leaving both parties better off than they would be otherwise, and taking into account the relative advantages and disadvantages of their negotiating positions.

[2] Other rational bargaining solutions have also been proposed (Muthoo, 1999)

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

To illustrate this in a bit more detail, consider one particular extension of the Nash bargaining solution of special relevance to our proposal. In the classic Nash bargaining problem the parties are *dividing goods*, but this logic has been generalized to the case of bargaining over the *norms or rules* that govern conduct (André et al., 2022; Baumard et al., 2013; Binmore, 2005). Rules and norms are powerful tools for shaping behavior—indeed, when enforced, they can stabilize an astonishing variety of different resolutions to conflicts of interest, including both cooperative and non-cooperative ones (Bicchieri, 2005; Boyd & Richerson, 1992; Ostrom, 1990). We know from both field and laboratory studies that people often establish rules and norms through a process of discussion, negotiation, and agreement (Ostrom et al., 1992). A classic example are the informal norms and agreements that fishermen create to govern the valuable, limited resource of their fisheries (Acheson, 1988; Ostrom, 1990). In fisheries—and across human societies more generally—we tend to observe rules and norms that stabilize cooperative behavior leading to mutually beneficial outcomes (Bowles & Gintis, 2011; Curry, 2016; Haidt, 2007; Tomasello, 2009), although not always (Herrmann et al., 2008). This is not surprising because, when bargaining over the contents of rules and norms, the key concepts of the bargaining game apply: Rational agents are expected mutually agree upon norms that generate resources distributions of mutual benefit—in some cases, ones that conform quite precisely to the Nash bargaining solution (Binmore, 2014).

2.1.2 Mutual benefit versus aggregate benefit

It is important to point out that mutually beneficial outcomes predicted by the theory of bargaining diverge from the “welfare maximizing” outcome favored by utilitarian moral theories, such as maximum aggregate benefit.

Indeed, rational agents would often *not* agree to divisions that maximize aggregate benefit. For instance, suppose that 10 tokens are to be divided between two people, and each token is worth $1 to the first person and $2 to the second person. Aggregate benefit is maximized by giving all tokens to the second player (this results in aggregate benefit of $20, the maximum *sum* possible)[3]. But we would not expect the first player to agree to this, since it provides them with no benefit at all. Bargaining theory instead predicts a mutually-beneficial division—the Nash Bargaining Solution, for instance, dictates that the players receive 5 tokens each (resulting in $5 for the first player and $10 for the second, which yields the maximum *product* possible). Although aggregate welfare is lower (the sum: $15), this is likely a better match to ordinary peoples’ moral intuitions about a fair split.

To choose another example, suppose a rich baker has money they don’t need and bread they wouldn’t enjoy much, while a poor customer has money they need dearly and would benefit a lot from the bread. The agreement they would likely strike—i.e., the predicted bargaining solution—is not for the baker to give away the bread for free (and perhaps some money, too, for good measure!), even though this would maximize aggregate benefit (the raw sum of their utilities). Rather, the agreement would likely involve the customer paying at least something for the bread—a solution that achieves mutual benefit.[4]

2.2 Finding mutual benefit

How do people arrive at mutually beneficial arrangements? While bargaining theory can describe, explain, and predict solutions to the problem of interdependent choice, it is largely silent on the actual processes that bring those solutions about.

[3] Assuming, as a rational theory must (Rabin, 2000), that utility increases roughly linearly with money for small stakes

[4] Note that in this hypothetical example, the baker still values money to some extent— otherwise, he wouldn’t be in the baking business, or indeed any business at all. That being the case, the pauper maintains some bargaining power (via whatever money he has), which enables an exchange. However, there are those in society who have little or no bargaining power, yet many people have t he moral intuition that they are nonetheless owed basic rights.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Existing work tends to focus on the evolutionary dynamics that lead to mutual benefit. Evolutionary models show that adaptive dynamics favor (although do not always guarantee) cooperative Nash bargaining solutions in many cases of conflicting interest (J. M. Alexander, 2000; André et al., 2022; Bruner, 2021; Skyrms, 2014), and competitive equilibria predicted by game theory in others (Nowak, 2006; Smith & Price, 1973). In these models, this arises as the outcome of a typical "blind" evolutionary process (whether biological or cultural), not due to any reasoning or explicit bargaining on the part of the agents themselves. Similarly, there is considerable theorizing about how selection among candidate human norms might occur by "blind" biological or (more often) cultural evolutionary processes (Binmore, 2014; Boyd & Richerson, 1990; Henrich, 2004). These sorts of evolutionary accounts are aligned with our emphasis on the logic of contractualism, but are beyond the scope of the psychological account we provide.

Our focus is different. Blind evolutionary processes can be powerful, but they take a long time and their results—the policies, structures or mechanisms they construct—are rigid. But the human moral world changes all the time—there is new information, new opportunities, and new interdependent choice structures to navigate and we have to figure out how to act in light of these constantly changing circumstances. Cognitive processes—in contrast to evolutionarily-endowed intuitions—can adapt quickly and flexibly to a changing world. The human mind is designed to solve an open-ended space of new problems and tasks. We argue that the cognitive processes of human morality respond to and exhibit this flexibility, which derives from agreement-based processes; people approximate bargaining solutions by explicitly or implicitly negotiating with each other, both over specific resource distributions but also over rules and norms. It is the cognitive processes of negotiation that allow us to design agreements that "make sense" to both parties (rather than being the product of blind variation and selection).

Our main purpose is this article, then, is twofold. First, we argue that agreement-based methods are in fact a cornerstone of human moral cognition, and we chart a framework for identifying and characterizing them at a mechanistic level. Second, we argue that many of the ways we rely on agreementbased methods involve cognitively efficient heuristics. These "resource-rational" versions of contractualism explain large and important parts of our moral psychology that have not previously been understood as grounded in the logic of agreement and mutual benefit.

This is a challenge for contractualist theories broadly (both normative and descriptive). Our view does not claim that *all* moral judgments can be explained by agreement-based processes and this is an example of a set of judgments that is beyond the scope of our theory in its present form (though see our discussion of this point in §6).

2.3 Limitations and bounds of the bargaining framework

We argue that a wide range of moral judgments and behaviors can be organized around resource-rational contractualist principles. But not all of them can, and it is important to understand the limits of the theory.

First, the moral judgments we attempt to explain all fall within the domain of interdependent choice—cases that involve the interacting utilities of multiple agents. However, there are some moral judgments that fall outside this scope. For instance, sometimes people condemn "victimless" crimes (such as consensual incest (Haidt et al., 1993) or the eating of certain foods (Levine, Rottman, et al., 2020)) as morally wrong, which our account does not explain. Moreover, a virtue-based perspective on moral judgment posits that it is morally good to pursue certain "self-regarding" virtues like fortitude, prudence, and courage (Taylor & Wolfram, 1968), which need not have anything to do with other people. Our account likewise does not explain these virtues, unless they are framed as having down-stream impacts on other agents (Taylor & Wolfram, 1968, as indeed, they sometimes are) (though see §6).

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Even some cases of interdependent choice are outside the scope of our proposal because they bypass the cognitive processes of negotiation, such as the sense of moral duty to provide for one's children. This intuition, and others like it, likely originates in "blind" adaptive processes rather than agreement-based reasoning. As discussed in § 2.2, evolutionary processes can lead to behaviors or instincts with an underlying contractualist logic, but are nevertheless outside the scope of our view because they bypass the cognitive processes underlying individual learning and reasoning.

Conversely, bargaining theory makes some predictions that many people find to be unjust or immoral. Consider a case where a person agrees to their own exploitation, such as a worker who accepts a very low wage and very poor working conditions because they truly have no better option. The most straightforward application of an agreement-based view would hold that there is nothing morally wrong with this arrangement—after all, everyone is better off than they would be otherwise. Indeed, an agreement favoring the advantaged party is specifically predicted by the Nash bargaining solution. Yet, many people hold that there is something morally wrong with this arrangement. The origin of our preference for equal, universal human rights, regardless of a person's bargaining position and power, is something that contractualist theories often regard as outside the scope of what emerges from agreement(Gauthier, 1986).[5]

Meanwhile, many people probably also hold a roughly "libertarian" view of the exploitation case—namely, that any social arrangement is morally acceptable as long as it is mutually agreed upon (or feel the force of the viewpoint even if balanced against competing egalitarian viewpoints). This part of our moral psychology is well explained by our theory. Finally, while there is also an extensive literature purporting to show the pervasiveness of egalitarian preferences around resource distributions (Starmans et al., 2017, for a review, see), Starmans and colleagues note that these typically match recipients on a variety of important attributes such as effort, ability, and desert (Starmans et al., 2017). Indeed, when a bargaining situation is perfectly symmetric, the Nash bargaining solution is egalitarian (as pointed out by André et al., 2022). When effort, ability, and desert are manipulated or observed, however, people tend to prefer *unequal* distributions —just as bargaining theory would predict if those attributes were to be indicative of bargaining power.

Following past work (André et al., 2022; Melkonyan et al., 2017), we use the Nash Bargaining Solution to illustrate the broad principles arising from agreement-based methods. But many other bargaining solutions have been proposed in the literature in game theory; and the general process of bargaining and negotiation is likely to depend on many factors not typically considered in formal models of bargaining (e.g., including reputation, historical precedent, salience, and many more). Irrespective of the detailed account, the bargaining process has two important features: (1) that parties can walk away freely from an interaction without interference, and (2) that once an agreement is made, neither party can unilaterally back out. These conditions are common, for instance, in economic markets regulated by states—nobody can be compelled to enter a contract but, once a contract is entered, it is backed by the force of law. But there are cases in which a person may be compelled to participate in social interactions that they would not have chosen to enter into in the first place (e.g., cases of coercion); and/or there is no credible enforcement of agreements (either by the parties themselves or an external agent). It remains for future work to determine whether, and how, principles of agreement in moral psychology apply in such cases (see, for example, (Hoffman et al., 2016)).

Finally, in this paper we largely restrict our analysis to the cognitive mechanisms people use to making decisions involving a few others, or small groups. We provide little discussion of inter-group

[5] Nevertheless, as we describe below, it is possible that globally-applied cached welfare and action standards—the products of *resource rational* contractualism, may be psychological origin of these intuitions. See also, our discussion on this point in §6, "The boundaries of a bargain".

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

relations, or of society's formal, largescale institutions (though see the discussion of this point in § 6). Particularly in the case of some institutions, it is natural to think that agreement-based principles might have substantial explanatory power. After all, institutions often have formal mechanisms for finding and enforcing agreements. What is less clear, however, is whether these mechanisms are meaningfully grounded in a contractualist moral psychology—one which, to the extent that it is evolved, presumably evolved in smaller and less institutionally formalized settings. These psychological mechanisms are our primary focus.

3Mechanisms of Moral Cognition

Much prior work in moral psychology takes inspiration from philosophical theories, but builds almost exclusively on two different approaches in moral philosophy: deontology (focusing on restrictions on actions) and consequentialism (focusing on calculating outcomes). In contrast, our proposal draws in part on the contractualist tradition in moral philosophy. Contrasting these three approaches provides a helpful framework for organizing the current state of the field.

3.1 Contractualism and its psychological counterparts

Philosophical contractualism is a normative theory: It attempts to explain how we *should* think and act.[6] At the broadest level, philosophical contractualism posits that human agreements, institutions, and rules ought to be modeled on what rational agents should agree to in order to achieve mutual benefit. This takes many different particular forms: an idealized dialog between people attempting to adjudicate their differences (Habermas, 1996); the identification of behaviors that could not be "reasonably rejected" by anyone (Scanlon, 1998); the behaviors of rational agents acting in concert (Gauthier, 1986); the choices we would make behind a "veil of ignorance" (Rawls, 1971); or, those we could rationally will as a "universal law" governing the conduct of all (Kant, 1785; O'Neill, 2012). What is common among all is the idea that we must find arrangements to which all relevant parties would agree.

As we have already seen, actual human moral judgments, norms, institutions, and rules often show broad alignment with these models. It is natural to assume, then, that our moral psychology would be organized at least in part around agreement-based methods. Yet, surprisingly little psychological research has characterized the precise mechanisms by which people agree on what is right and wrong.

Piaget's (Piaget, 1932) and Kohlberg's (Kohlberg, 1969) theories of moral development are early and important exceptions, having proposed that certain stages of a child's moral understanding were organized around convention and agreement. One important line of work in child development has elaborated on these ideas quite directly (Killen, 1995; Killen & Turiel, 1991). Others share its more general focus on the role of mutual agreement in children's understanding of how moral norms are established or updated (Tomasello, 2020; Zhao & Kushnir, 2018).

Outside of the developmental literature, agreement-based descriptions of the moral mind have received less attention, although with some notable exceptions. Everett and colleagues have suggested that contractualist notions (in particular, a "respect for persons and the honoring of social contracts") can explain our responses in certain moral dilemmas (Everett et al., 2016), and Levine

[6] Some scholars draw a distinction between contractualism and contractarianism. When defined narrowly, contractualism is typically associated with Scanlon and his view, inspired by Kant, that an act's moral permissibility is based on whether the policy gui ding the act could be reasonably rejected by anyone affected. In contrast, contractarianism finds its roots in Hobbes' writings, and views contracts as agreements between self-interested actors. In this paper, we use the term "contractualist" in the broad sense, covering both views, and referring to the general class of theories that derives moral permissibility from some form of agreement.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

and colleagues have described mechanisms through which this may occur (Levine, Kleiman-Weiner, et al., 2020; Levine et al., 2024). Sell and colleagues' recalibrational theory of anger views anger as a mechanism that people use to bargain for better treatment with those they interact with (Sell et al., 2017). Most ambitiously, André and colleagues have argued that many of our moral judgments can be explained via an evolutionary contractualist approach (André et al., 2022). Specifically, they propose that evolutionary processes naturally lead our moral judgments to approximate generalized Nash bargaining solutions, providing an adaptive rationale for our intuitions about distributive justice, ownership, authority, our responses to moral dilemmas, special obligations towards kin, and even the moralization of supposedly harmless wrongs. This description of the ultimate function of our moral psychology leaves open questions about its proximate-level psychological implementation.

Nonetheless, the most notable thing about the psychological literature on contractualism is its sparsity. This stands in stark contrast to its two main philosophical rival: consequentialism and deontology.

3.2 Consequentialism and its psychological counterparts

Consequentialist theories posit that the moral properties of an action depends exclusively on the "state of affairs" (consequences) that it brings about. Often, the relevant consequence is something like welfare[7], and the claim is often that it should be impartially maximized in aggregate. "Welfare maximization" also plays a central role in many theories of moral psychology. For instance, many computational accounts of moral decision-making are structured around a backbone of expected value maximization, where value is often defined largely in terms of one's own and others' welfare. These models have been frequently and successfully applied in studies of the neural basis of moral decision-making (FeldmanHall et al., 2016; Hsu et al., 2008; Lockwood et al., 2020; Williams, 1968), which often show striking overlap with the neural basis of self-interested expected value maximization in ordinary, non-social decision making (KleimanWeiner et al., 2015; Shenhav & Greene, 2010). Also, a rich and productive literature shows that when people face welfare tradeoff dilemmas, such as the trolley problem, a major (although not exclusive) contributor to their moral judgments is a mechanism that makes characteristically consequentialist choices (Cushman et al., 2006; J. Greene, 2014). In other words, it decides whether to sacrifice one life in order to save five by "doing the math" and favoring the welfare of five over one. We refer to this mechanism of moral judgment as "welfare-based:" it determines what to do by considering the welfare consequences to various people given a model linking actions to outcomes. Unlike many versions of philosophical consequentialism, however, most psychological models of this kind assume that different peoples' welfares are weighted differently—i.e., that we care more about some people (ourselves, family, friends, etc.) than others.

3.3 Deontology and its psychological counterparts

Deontological theories are concerned not with outcomes but with whether actions conform to moral rules, rights, and duties (L. Alexander & Moore, 2021). Deontological theories stand in contrast to consequentialiast theories in positing that some actions are morally impermissible regardless of the consequences that they bring about (L. Alexander & Moore, 2021). A wealth of psychological evidence shows that people's moral judgments are indeed responsive to intrinsic properties of actions, and not just their consequences. For example, judgments of moral cases seem to be affected by whether the case involves a commission or an omission (Baron & Ritov, 2004; Cushman et al.,

[7] Or otherwise, utility, which is sometimes assumed to be substantive and measurable. For example, it might correspond to amounts of pleasures or pains, or self-reported wellbeing. Since the "revealed preference" revolution in economics (Samuelson, 1938), a popular alternative viewpoint is that utility is a derived notion, which can be inferred from any pattern of coherent rational choices (Broome, 2017)

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

2006; Petrinovich & O'Neill, 1996), the relationship of the victim to the actor (Kleiman-Weiner et al., 2015; Petrinovich & O'Neill, 1996), whether or not the harm was intended or accidental (Barrett & Saxe, 2021; Cushman et al., 2013; Piaget, 1932), whether the harm was a means or a side-effect (Cushman et al., 2006; J. Greene, 2014; Levine & Leslie, 2020; Mikhail, 2011), whether the harm was brought about through bodily contact (Cushman et al., 2006; J. Greene, 2014; Pellizzoni et al., 2010), whether the harm is caused "directly" or indirectly (Royzman & Baron, 2002), to name just a few. Some theories argue that this arises from explicit representations of abstract categories such as harm, knowledge, and intention, although perhaps outside of conscious awareness (Cushman, 2015; Darley & Shultz, 1990; Mikhail, 2011; Nichols, 2021). Others accounts emphasize the role for moral emotions (J. D. Greene et al., 2001; Haidt, 2001)) or other mechanisms of value representation (Crockett, 2013; Cushman, 2013) in triggering the condemnation of certain types and properties of actions. What these views share is that moral judgment does not depend on consideration of the welfare consequences of actions, but on properties of actions themselves. Whether or not these properties are explicitly represented in rule-like form, they dictate which categories of action are permissible, obligatory, or forbidden. We refer to this broad and rather diverse class of moral mechanisms—those concerned with properties of action, rather than the outcomes of those actions—as "actionbased" mechanisms.

3.4 Conflict and consilience between mechanisms of moral judgment

How can we best understand the relationship between these parts of our moral minds: the consequentialist (welfare based), deontological (action based), and contractualist (agreement based)?

Generally, philosophers think of these normative theories as rivals. Much philosophical work is organized around finding cases that reveal the conflicts between them—where different philosophical views render different judgments— to help adjudicate which theory should be preferred. Similarly, psychologists have seen different theories as standing in opposition, and have exhaustively explored how moral dilemmas reveal cognitive conflict between welfare-based and action-based mechanisms(Conway & Gawronski, 2013; Crockett et al., 2010; J. D. Greene et al., 2004; Koenigs et al., 2007).

However, there has also been some effort to understand how welfare-based and action-based mechanisms of moral judgment are *related*. It has been influentially argued that action-based rules are cognitively cheap approximations of welfare maximization (Baron, 1994; Crockett, 2013; Cushman, 2013; J. Greene, 2014). But two things are striking about this attempt to relate welfare-based and action-based mechanisms. First, it proposes that the ultimate function of our moral psychology is to maximize aggregate welfare. As we have already argued, however, rational models of interdependent choice predict not welfare maximization, but instead bargaining solutions. Second, this proposal provides no natural account of the relationship between agreement-based methods and action- or welfare-based ones.

We draw inspiration from a proposal in the philosophy literature that unifies all three. Derek Parfit famously pointed out that contractualism, deontology, and consequentialism agree on moral judgments far more often than they disagree (Parfit, 2011). This suggests that they may be quite intimately related.

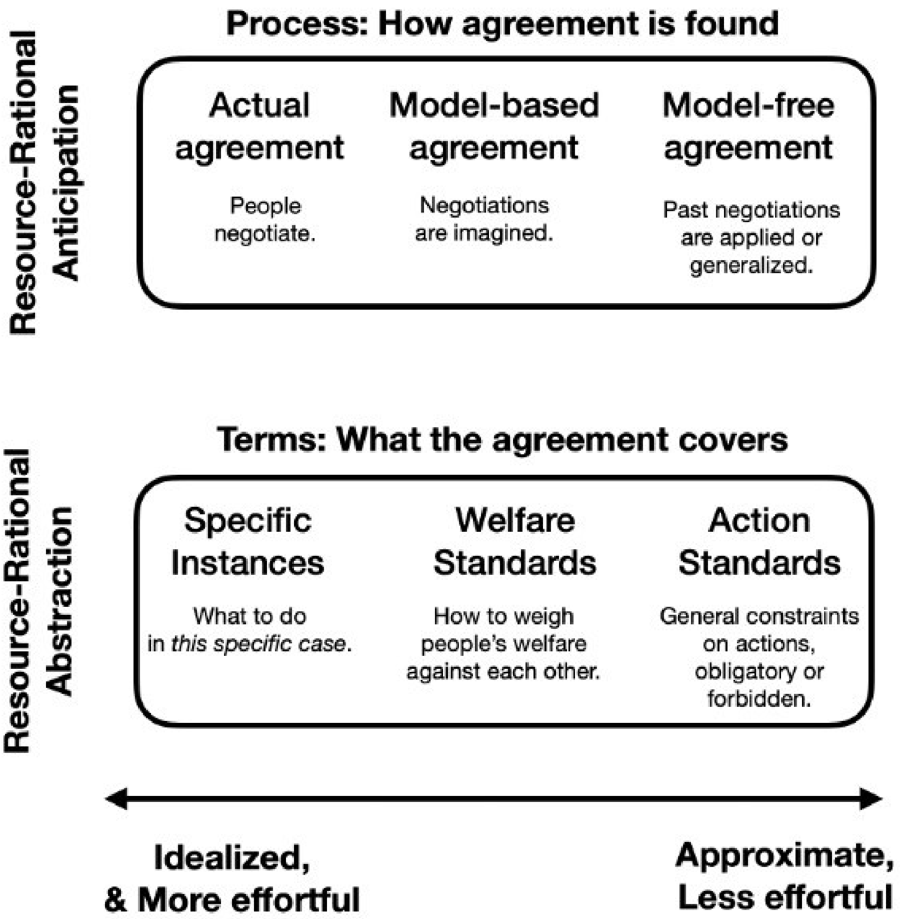
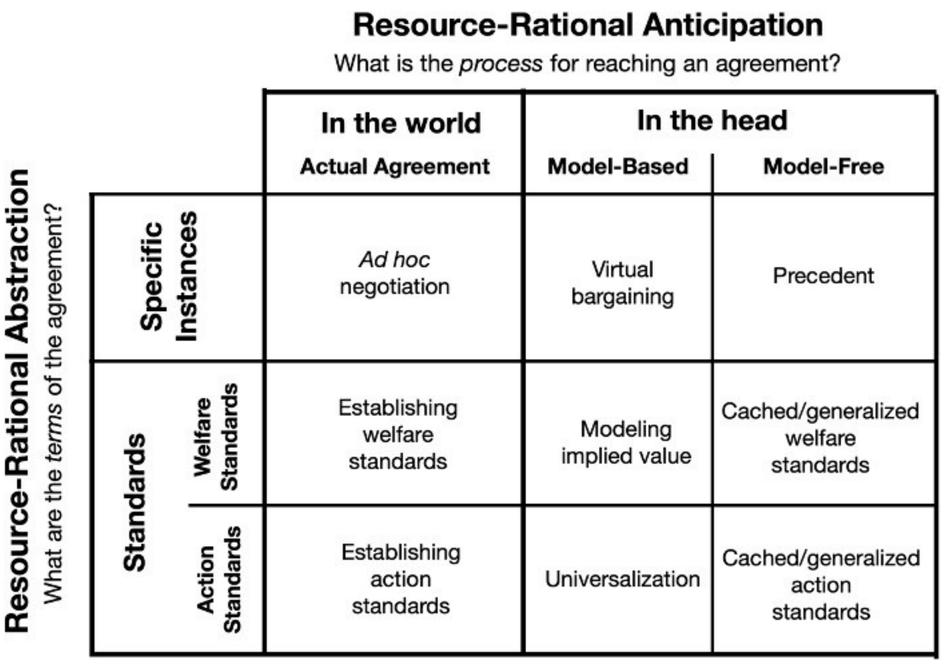

Process: How agreement is found
Resource-Rational Anticipation
Actual agreement
People negotiate.
Model-based agreement
Negotiations are imagined.
Model-free agreement
Past negotiations are applied or generalized.
Terms: What the agreement covers
Resource-Rational Abstraction
Specific Instances
What to do in *this specific case*.
Welfare Standards
How to weigh people's welfare against each other.
Action Standards
General constraints on actions, obligatory or forbidden.
Idealized, & More effortful
Approximate, Less effortful

Figure 1: The idealized form of contractualist moral decision-making involves a particular process (getting everyone together to negotiate) and particular terms (an agreed upon solution to the specific instance under discussion). Both the process and the terms have resource-rational approximations that trade off accuracy against effort. We can replace the *process* with an *anticipation* of what an in-person negotiation would yield via either model-based or model-free strategies. Likewise, we can replace the *terms* with those that are more *abstract* than those dealing with particular instances, such as welfare-based and action standards.

He proposes that contractualism—i.e., agreement between people—provides a basis for accepting or rejecting rules that govern action, and does so on the basis of their welfare consequences. His key insight—one which holds great promise for descriptive theories of moral psychology—is that contractualism provides the ultimate *justification* for moral standards, but that these standards may be framed in deontological or consequential terms. By analogy, we propose that many action-based or welfare-based judgments can be explained as resourcerational approximations of agreement-based ones. We turn next to consider how this works.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press



Resource-Rational Anticipation

What is the *process* for reaching an agreement?

Resource-Rational Abstraction

What are the *terms* of the agreement?

		In the world	In the head	
		Actual Agreement	Model-Based	Model-Free
Specific Instances		*Ad hoc* negotiation	Virtual bargaining	Precedent
Standards	Welfare Standards	Establishing welfare standards	Modeling implied value	Cached/generalized welfare standards
	Action Standards	Establishing action standards	Universalization	Cached/generalized action standards

Figure 2: Two principle axes of resource-rationality: abstraction and anticipation. The ideal is to actually negotiate for mutual benefit with all relevant parties over each specific instance (top left cell). Given that this is resourceintensive, we can instead bargain over more *abstract* welfare or action standards (rows) or *anticipate* bargaining solutions through model-based or model-free methods (columns).

4Resource-rational Contractualism

How would a contractualist moral psychology work, in practice? Agreementseeking in its idealized form is often infeasible in the real world. Thus, while we might sometimes observe thoughts and behaviors that are close to that ideal, what we will find more often are cognitively and socially expedient heuristics designed to approximate it. Our strategy is to divide the idealized version of contractualism into two elements: the process of agreement (*how* we find it), and the terms of the agreement (*what* it covers). We then consider potential heuristic approximations for each.

4.1 The process of agreement: Resource-rational anticipation

The contractualist ideal is actual bargaining. But this imposes both cognitive and social demands, requiring multiple people to sit down and think hard together.

As a heuristic, rather than engaging in actual bargaining, we can instead *anticipate* the likely outcome of the bargaining process. One way to do this is “model-based”: We mentally simulate the bargaining process to guess its hypothetical outcome. Put simply, we imagine what another person would agree to if we were talking things over with them. Another way is “model-free”: We generalize from past agreements, as relevant precedent. Our use of the terms “model-based” and “model-free” in this context draws an intentional analogy to theories of reinforcement learning and decision-making (i.e., sequential valueguided decision-making), where these terms first emerged. In that context, as here, model-based methods rely on inference (via planning, simulation, or other means) in the moment using a *generative causal model*—in this case, a model of bargaining. Model-free methods draw up previously cached (or "amortized") model-based computations, by generalising from them, or by other methods such as statistical generalization, analogy, or other means.

In summary, we propose a spectrum of possibilities from the costly but exact process of actual bargaining, through a less costly mental model of bargaining, ultimately to the least costly possibility of generalizing from past agreements.

4.2 The terms of agreement: Resource-rational abstraction

In the idealized case, the terms of any agreement will be precise and narrow: Agents agree to the specific behaviors that attain mutual benefit, but the agreement would narrowly cover their conduct only in the present circumstances, since no future circumstances will ever be completely identical. As a less costly alternative, however, people may strike more *abstract* bargains—ones that apply to many cases, preventing the need for constant re-negotiation.

We consider two possible forms of abstraction, motivated by the kinds of decision-making mechanisms with which we know the human mind is wellequipped for purposes of individual decision making (Dolan & Dayan, 2013).

First, people often solve decision-problems by the logic of expected value maximization. When planning a day, for instance, one considers a range of possible actions, along with one's goals and preferences, and then chooses the specific actions that maximize expected value. Since humans are good at this kind of computation, we might expect that they would use it to solve the problem of interdependent choice by defining a single utility function over all the agents concerned and then maximizing it. In this case, the nature of the utility function—which utilities, and especially whose utilities, it favors—would constitute the terms of the bargain.

This is an approximation of the ideal computation of mutual benefit because it substitutes a weighted welfare maximization problem for a bargaining problem. Nevertheless, it preserves the essential characteristics of agreementbased methods because the *weightings* applied during welfare maximization are subject to negotiation. We call the result of negotiating those weightings, a "welfare-based" standard.

Moreover, as a general rule for choice, maximizing overall choice will typically benefit most people, most of the time; indeed, when aggregating over a sequence of many choices, the results of maximizing overall benefit, and ignoring distributional concerns, is likely to yield benefit to most people. Overall welfare maximization provides a "rising tide" that will raise most, if not all, boats.

A second common and even more computationally efficient method of making decisions is to define standards directly over *actions* in context—that is, to assign actions intrinsic value, or an intrinsic probability of selection, contingent on a class of situations (Konda & Tsitsiklis, 1999; Watkins & Dayan, 1992). Representations of this sort—which we sometimes call rules, habits, or customs—can often make individual decision-making very straightforward. While each situation is unique, they can often usefully be organized into broad classes, so that once we have figured out the best way to act in one situation (whether by learning from experience, or by model-based planning to maximize utility), we can then generalize that guidance to all instances of the class without the hard computational work of learning or thinking through each new situation from scratch. After one encounter with a hot stove, for example, we may simply encode and operate a rule to keep our distance in the future.

The same is true in the context of interdependent choice: multiple agents or groups can strike agreements about how to act jointly that are designed to achieve mutual benefit across broad classes of situations. Since people are good at implementing rules like this in the case of individual

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

decision-making, we might expect them to characterize an important part of psychological contractualism as well. We call agreements of this sort "action standards".

We next consider each of these two forms of abstraction in more detail, describing the specific predictions rendered when construing welfare-based and action-based standards as resource rational approximations of agreement.

4.2.1 Welfare-based standards

Sometimes people bargain over the relative weighting of their welfare —a standard governing how each values the other (Adams, 1965; Delton & Robertson, 2016; Tooby et al., 2008). In certain contemporary institutionalized settings, negotiation over welfare standards is explicit; one example is the distribution of limited resources through public health agencies such as the World Health Organization or the UK's National Institute for Health and Care Excellence (Hirth et al., 2000; Owen & Fischer, 2019). For the most part, however, people don't have debates along the lines of "You seem to value me at half of the value you place on yourself, but I feel like it should be more like three quarters." Nevertheless, real-world social processes involve implicit negotiation over welfare trade-offs. For instance, there is a rich research program which proposes that certain kinds of social emotions—anger, jealousy, resentment, gratitude, remorse, etc.—are actually a vehicle for just this kind of welfare-based bargaining to occur (Adams, 1965). Anger can be viewed as an attempt for a wronged person to express that someone else is not taking their welfare into account as much as they would like. And, remorse can be viewed as the "recalibration" of a welfare standard in response to justifiable anger (Chang et al., 2011; Sell et al., 2017). For instance, when a child fusses about whether they got as much juice as their sister, or an employee bristles when the boss celebrates a colleague's birthday but ignores his, their anger may principally reflect not the literal resources at question, but the implied *relative valuation*—and, indeed, that anger may act as a bid for *revaluation*. Because we know that people often make decisions by taking others' welfare interests into account—but to varying degrees—it makes good sense that we would have implicit yet powerful mechanisms for negotiating and renegotiating how each others' welfares ought to be weighted.

Once welfare standards are established, people can then make interdependent decisions by selecting the action that maximizes the agreed *weighted* aggregate welfare for some set of relevant parties. For instance, given that we value ourselves greatly, our immediate family highly, certain friends and acquaintances slightly less, even strangers to some degree, etc., we can then decide how big a gift or favor to give, how scarce resources should be divided, and on whom various burdens can be imposed, etc. (Marshall et al., 2020, 2022; McManus et al., 2020); and we anticipate that others will judge our behaviour using similar weightings. A wealth of research demonstrates that people do, in fact, often navigate interdependent choice problems in precisely this way (FeldmanHall et al., 2016; Hsu et al., 2008; Lockwood et al., 2020; Williams, 1968, e.g.).

It is useful to bargain over welfare standards when we will interact with the same person many times, but in varied situations. For instance, perhaps we are in a long-term relationship with a friend, coworker, or romantic partner. In any of these relationships we might encounter unusual circumstances that involve interdependent choice. For instance, will I pick up their dry cleaning across town, despite my urgent work deadline, knowing that they are on the phone dealing with the fallout from their sister getting arrested?[8] It would be possible to engage in *ad hoc* negotiation in a case like this, but is likely to be more efficient to fall back on a general sense of the relative costs and benefits and the weighting of our interpersonal interests. Thus, when we participate in enduring relationships that elicit very diverse interdependent choices, it may be rational to agree (presumably implicitly) on welfare standards.

[8] Hypothetically.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Any plausible theory of welfare-based moral standards requires certain situational and role-specific constraints on their application. When a kindergarten teacher is at work, for instance, he will value each of his students quite highly, while thinking relatively little about the welfare of his elderly parents. While visiting his parents over the holidays the reverse is true. In each context his behavior can be partially understood as the product of welfare-based reasoning—a balancing of the interests and needs of children against each other, or parents against himself, etc. But one cannot perfectly capture his behavior *crosssituationally* according to a consistent set of welfare standards: A sacrifice he makes for a needy student in the classroom does not imply an equivalent sacrifice for the same student over holidays. This situational sensitivity is implicit in the tug-and-pull of interpersonal negotiation over welfare standards. An employee who complains that her boss should value her more just means to say she should be more valued *at work*, not that the boss should devote more attention to her home life.

These variations make sense if we see welfare as approximating the results of agreements: the teacher has "signed up" to care for his class in a specific context; a boss is primarily responsible for the welfare of her employees at work, and so on. In contrast, these variations are relatively harder to capture on a standard consequentialist view of welfare standards, according to which the valuation we place on another's welfare should apply equally across situations. This becomes particularly apparent if we consider the norm of evenly splitting windfall resources: the "fair" 50/50 split. This norm is shared across many cultures (Henrich et al., 2005, 2010, although not all, see), and dictates how people behave even in one-shot anonymous interactions. It is obviously not the case, however, that in these cultures everybody weights others' welfare equally with their own. Rather, the norm applies to a narrow class of situations in which windfall profits are divided.

This analysis invites a new perspective on welfare-based moral judgment. It has been argued that one can view such welfare-based mechanism as basically consequentialist; and that consequentialism is also the rational standard that other mechanisms of moral judgment (such as action-based constraints) are designed to approximate. According to the present proposal, however, welfarebased mechanisms of moral judgment are best understood as a heuristic approach to approximating the *ad hoc agreements* that we would reach under idealized circumstances: the psychologically relevant rational standard is not utilitarian but consequentialist.

4.2.2 Action-based standards

People can also bargain over *action-based standards* such as rules and norms. This is obviously true at an institutional level, where explicit bargaining over rules, norms, laws, and policies often happens. Debate and negotiation in legislative bodies, for instance, concerns what rules should be adopted to govern a community such as a school, a place of work, organization, city, region or an entire country. But, action standards are also negotiated informally as well. Extensive fieldwork from the past half century—specifically looking at how groups self-organize to manage "common pool resources"—reveals that bargaining over rules happens via a predictable processes (Acheson, 1988; Ostrom, 2000). These are generally most successful when the directly affected parties negotiate their own solutions (rather than solutions being imposed by outside authorities), when their solutions impose enforced constraints on behavior, when resources are shared or distributed across all those subject to regulation, and when each person's near-term costs are outweighed by their understanding of the long-term benefits of the system (Ostrom et al., 1999). In other words, the standards established by the group help regulate each individual's behavior in order to achieve an outcome that leaves nobody worse off than they would have been in the absence of an agreement (Ostrom, 2000). Similarly, children use the agreement and negotiation to establish the rules of games

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

and other social interactions by middle childhood (Piaget, 1932), drawing in part on their ability to form joint intentions and plans (Tomasello, 2020).

Once established, rules and norms tell us how to act, or introduces constraints on action, across a wide range of relevantly similar circumstances. Thus, we have: "No non-resident street parking after 8pm on weekdays"; "I'll pick up the kids for gymnastics every Wednesday afternoon this fall"; "Good restaurant service should be rewarded with a 20% tip". Naturally, bargaining over action standards will be efficient when sufficiently similar circumstances recur. Rather than bargaining anew, and repeatedly arriving at very similar bargaining solutions, we can establish a standard—a sort of moral heuristic—that can be quickly retrieved and applied, and which efficiently gets us close enough to the optimal answer on each new occasion.

Action standards need to strike a delicate balance between accuracy and generality. Consider the rule: "Don't steal money from the woman wearing a floral dress, sitting in a cafe on the corner of Kirkland and Washington Streets on May 7". Such an over-specific rule may always be accurate, but no efficiency is gained by establishing this rule because the circumstance is unlikely to arise more than once. On the other hand, the (wildly) over-general rule "Don't do *anything*!" is universally applicable, but errs in every case where something really *should* be done. Rules are useful when the costs of their inevitable errors are outweighed by the benefits of frequent, cognitively cheap application (Hare et al., 1981; Sunstein & Ullmann-Margalit, 1999). Finding this balance is, in essence, a problem of resource rationality.

Explicit action standards, such as rules, also have benefits for communication and coordination. Communication is enhanced when rules are formulated in simple terms over observable features. They also help groups coordinate actions for mutual benefit, such as choosing between multiple equilibria (e.g. driving on one side of the road or backing the same candidate) (Anderson, 2001; Gauthier, 1986; Harsanyi, 1977). Action standards also enable coordinated action because whether a rule is being followed is typically more easily observed than, for instance, how much welfare various individuals obtain, or how an actor weights the welfare of others. Thus, action based standards help us determine who to trust, praise, or blame with greater certainty than welfare based standards—and also afford greater confidence that others will arrive at the same conclusion (DeScioli & Kurzban, 2013).

Not only can action-based moral standards be efficiently reused, they can also be culturally transmitted in the form of rules, norms, laws, etc. Collectively, this set of standards is an important part of our cultural inheritance—one that each new generation of children imbibes partly through explicit guidance and instruction from others and partially through observation and inference (Bandura & Walters, 1977; Cialdini & Trost, 1998; Henrich & Muthukrishna, 2021; Nichols, 2021; Stegall et al., 2023).

Thus, just as with welfare standards, action standards can be best understood as a resource-rational approximation of the *ad hoc* agreements we would reach if we bargained over every unique situation that arises.

4.3 The mechanisms of resource rational cognition

Having described the different *abstractions* that can be employed during bargaining, we now turn to the specific heuristic cognitive mechanism we can use to *anticipate* the outcome of bargaining processes (Figure 3).

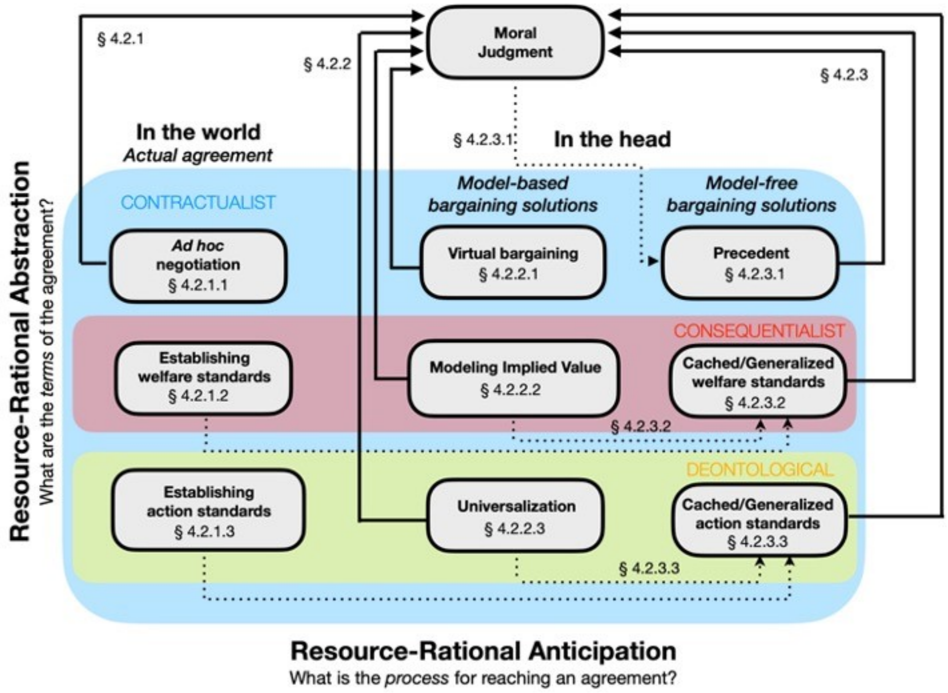
4.3.1 Model-based contractualist cognition

One way to anticipate bargaining outcomes through private mental processes is to employ a cognitive model of bargaining. We consider several accounts of this kind, which focus on model-

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

based anticipation of bargains over *ad hoc* circumstances, over welfare-based standards, and over action-based standards.

4.3.1.1 Model-based bargaining over individual instances: Virtual bargaining It is quite intuitive to think that in some situations we decide whether an action is OK by asking whether the party affected by that action would agree to it. For instance, if you are hungry and you know your co-worker keeps granola bars in her desk, can you simply take one? We might answer this question by asking, "what *would* she answer if I could ask her?" This same idea plays a much weightier role in the law surrounding emergency medical care. A doctor who encounters an unconscious victim of a car collision, for instance, is permitted to perform life-saving procedures because of the presumption that the patient would consent to that treatment if he were able to (Easton et al.,



Moral Judgment
§ 4.2.1
§ 4.2.2
§ 4.2.3
In the world
Actual agreement
§ 4.2.3.1
In the head
Resource-Rational Abstraction
What are the terms of the agreement?
CONTRACTUALIST
Model-based bargaining solutions
Model-free bargaining solutions
Ad hoc negotiation § 4.2.1.1
Virtual bargaining § 4.2.2.1
Precedent § 4.2.3.1
CONSEQUENTIALIST
Establishing welfare standards § 4.2.1.2
Modeling Implied Value § 4.2.2.2
Cached/Generalized welfare standards § 4.2.3.2
§ 4.2.3.2
DEONTOLOGICAL
Establishing action standards § 4.2.1.3
Universalization § 4.2.2.3
Cached/Generalized action standards § 4.2.3.3
§ 4.2.3.3
Resource-Rational Anticipation
What is the process for reaching an agreement?

Figure 3: Relationship of each mechanism to each other and to moral judgment. Contractualist methods of moral judgment (all those pictured) are highlighted in blue. Processes that have been described as "consequentialist" (i.e., concerned with welfare outcomes) are highlighted in red and those described as "deontological" (i.e., concerned with action constraints) in yellow. Yellow and red boxes sit on top of the blue one, indicating that "deontological" and "consequentialist" processes are versions of contractualist ones. The arrows indicate how the processes relate to one another and to moral judgment. Moral judgment can be rendered directly by *ad hoc* negotiation, by any of the model-based bargaining processes, or by the application of model-free bargaining approximations (solid lines). Bargaining over standards either in person (left-most column) or through mental simulation (middle column) can lead to those standards being cached for future re-use (right-most column) (dotted lines). The section numbers in the figure refer to the discussion of each process in the text.

2007). In each case, if you were to act (taking their food, or performing a medical intervention) in a way that you would *not* have expected the affected person to consent to, this would be a gross moral violation.

This method of making moral judgments can be addressed using a process of "virtual bargaining"—-using a cognitive model of bargaining processes to coordinate behavior and resolve *ad hoc* any conflicts of interest.[9] Virtual bargaining has been developed primarily outside the moral domain, to understand the reasoning that underpins how people are able flexibly to coordinate on communicative and social conventions "in the moment" (Chater et al., 2022; Misyak & Chater, 2014; Misyak et al., 2014). But the same mechanism applies naturally to moral judgments.

The virtual bargaining approach contrasts with standard game-theoretic reasoning, in which agents do not think about possible *bargains*, but attempt to choose between Nash equilibria: strategically stable combinations of actions such that no person can benefit by unilaterally changing their choice of action. Virtual bargaining allows people to find mutually beneficial agreements that are *not* Nash equilibria; and also provides a mechanism for choosing between the multiple possible bargains (including many Nash equilibria) that arise in any but the very simplest interactions.

In order to distinguish these hypotheses, social interactions can be formalized as a multiplayer game. Consistent with the predictions of virtual bargaining, in game-based experiments people often choose non-Nash equilibrium bargains, where these are most mutually beneficial; and they also select between the many possible Nash equilibria (and other options) by coordinating on the option with greatest mutual benefits.[10]. That is, people choose actions characterized by the equilibrium that would be mutually agreed upon, according to the theory of bargaining—even in the absence of any actual negotiation. Related work shows that this model predicts not only their behavior, but also their moral judgments (Le Pargneux et al., 2023).

The virtual bargaining view makes predictions not only about peoples' behaviors, but also about their moral judgments. Confirming this prediction, Levine and colleagues (Levine et al., 2024) presented participants with a hypothetical situation in which somebody is able to earn a large sum of money by painting their neighbor's house blue while their neighbor cannot be contacted. Participants' moral judgments are best modeled by an account of virtual bargaining—which assumes that accepting the money is permissible if a sizeable enough portion is given to the neighbor (such that the neighbor would have agreed to it in a negotiation)—but not by rule-based or welfare-based accounts.

The theory of virtual bargaining also makes the clear developmental prediction that children's behavior and moral judgments will change as they gain the ability to reason about bargaining and agreement. Some current data, while circumstantial, shows that signatures of virtual bargaining emerges by the preschool years. For instance, young children find it morally permissible to cause a small harm to someone to prevent a greater one to that same person (Jambon & Smetana, 2014; Levine & Leslie, 2020)—perhaps because the wouldbe victim would agree to it. Moreover, when a child collaborates on a task with another person and then receives a greater reward than their partner, they will often share the excess reward (Hamann et al., 2011)—potentially because they realize that the collaborator's contribution was contingent on a tacit agreement that the reward would be shared. Moreover, there are contexts in which children treat implicit commitments as having the binding power of explicit ones(Kachel & Tomasello, 2019). This suggests that, by a young age, children are able to infer the agreements that would be made if they could discuss the situation

[9] Here we reserve "virtual bargaining" for imagined negotiation regarding specific situations. However,the term can also be used more broadly, to refer to implicit negotiation of any kind, including over welfare- and action-standards.

[10] Rather than, for example, choosing at random among Nash equilibria, or choose the equilibrium that would result in the greatest personal payoff, the equilibrium that lead to the outcomes with the greatest summed payoffs (Misyak & Chater, 2014, see) or other ingenious criteria that have been proposed for choosing between Nash equilibria

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

at hand, and feel that they are bound by those agreements. It remains to be seen whether these patterns of judgment emerge only as the child becomes able to reason about agreement, and this stands out as an important area for further study.

4.3.1.2 Model-based bargaining over welfare standards: Implied valuation When trying to anticipate others' reactions to our behavior, we often consider not only whether they would agree to the particular act in question, but also whether they would accept the *implied value* we assign to their welfare (Adams, 1965). Put another way, when we make choices, we often consider whether those choices would make others feel as if we value them sufficiently. And this make sense, given that people do in fact interpret social behavior not just in light of its immediate, local consequences, but also in terms of the way it signals a more global valuation—how much they are cared for in general (Adams, 1965; McManus et al., 2020; Radkani & Saxe, 2023; Shaw, 2013; Uhlmann et al., 2015). This, for instance, is a natural way to understand "recalibrational" emotions like anger, gratitude, and so on (Sell et al., 2017).

This view predicts that people have and use cognitive models of how others would infer their welfare-based standards and react to them. By contrast, it might be conjectured instead that even if people *use* welfare-based standards to choose their actions and make moral judgments, they may fail to anticipate how others would respond to or attempt to renegotiate those standards.

There is, though, already some evidence that people explicitly model the anticipated (i.e. model-based) reactions of a social partner (Houlihan et al., 2023; Railton, 2017). This is consistent with a large literature on impression management (Leary & Kowalski, 1990). It is also consistent with experimental evidence that when people allocate resources between themselves and another, they are attuned to their social partner's perception of how the resources are divided (Houlihan et al., 2023), even when those perceptions are false (van Baar et al., 2019). (Such attunement to false beliefs is a hallmark of model-based reasoning about another's mental states). This same literature suggests that in anticipating those reactions, we often focus on the implied degree to which we generally care about the other person's welfare versus our own (rather than the specific bargain we would strike over a single act).

4.3.1.3 Model-based bargaining over norms and rules: Universalization Finally, there are circumstances in which we use model-based reasoning to determine what norms or rules people would agree to. When considering whether to litter, vote, speed, recycle, etc., we often feel as if we should adopt the pattern of behavior that the relevant community would prefer to be universally adopted. This is often called universalization. Put simply, it asks : "what if everybody did that?"

Universalization appears, in one form or another, in many contractualist views in moral philosophy (Gauthier, 1986; Habermas, 1990; O'Neill, 2012; Rawls, 1971; Scanlon, 1998). A variety of theoretical traditions point out that suitably enforced norms play a crucial role in allowing people to solve social dilemmas, such as the protection and use of public goods (Bicchieri, 2005; Henrich, 2004; Ostrom, 2000; Tomasello, 2009). Norms allow people to transform *any* pattern of behavior into a stable equilibrium via enforcement (Boyd & Richerson, 1992). A key question, then, is which norm will be chosen. The logic of universalization allows people to anticipate the answer to this question, finding equilibria that generate mutual benefit, and to set their own behavioral standards accordingly, through a cognitive model of collective agreement over rules (Binmore, 2014).

Consistent with these predictions, many people spontaneously render moral judgments consistent with the logic of universalization (Kwon et al., 2022, 2023; Levine, Kleiman-Weiner, et al., 2020). Existing work shows how universalization guides the judgment of individual acts, however. Our view suggests that universalization should naturally be used not only as a method of telling us which *acts* to perform by considering what would happen if rule were in force, but also as a method

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

of determining which *rules* to adopt[11]. After all, its very logic asks us to consider the question: What if this action were adopted as a universal rule?

Testing this prediction is an important area for future work. One of the important implications of a rule-based model of universalization is that it may help us to understand the application of egalitarian rules to even cases where one party holds a bargaining advantage. Viewed through the lens of *ad hoc* bargaining, the egalitarian solution is not predicted by a contractualist account. Viewed through the lens of a society-wide bargain over suitable standards to cover a wide variety of circumstances—ones where individuals are sometimes advantaged and other times disadvantaged, or may not know what position they will hold—egalitarian rules may be easier to explain (see also §6).

4.3.2 Model-free contractualist cognition

A still more frugal approach to anticipating bargaining outcomes is to extrapolate from previous bargains. This can be described as a "model-free" method of moral judgment in the sense that it does not rely on a cognitive model of bargaining, but instead retrieves a stored solution. It offers a natural account of how people make moral judgments based on precedent, welfare maximization, and action-based moral standards such as rules and norms.

4.3.2.1 Precedent: Judgment by generalization Perhaps the simplest model-free approach to moral decision-making is to apply precedents from previous agreements to govern sufficiently similar novel cases. For instance, suppose that a student shows up to class 5 minutes late on the first day and the professor says nothing, but when they do the same on the second day the professor gives them no credit for class participation that day. Here, the student might feel as if the professor's initial reaction established an implicit precedent, and that the departure from this precedent is unfair.

Reasoning from precedent is one example of moral "generalization": A process by which one moves from established moral agreements of a narrower scope to assumed moral agreements of a wider scope (Stegall et al., 2023, see also). Its cognitive efficiency is obtained by relying on a simple, intuitive sense that a past situation is sufficiently similar to the present one.

Consistent with the predictions of this view, precedent plays an important role in moral judgment. Indeed, there is evidence of this both in informal moral commitments (Chen & Saxe, 2023; Graeber, 2012; Theriault et al., 2021) and in formal structures like the law, where judicial decisions are strongly guided by precedent (Daston, 2022; Schauer, 1987).

4.3.2.2 Model-free welfare-based judgment When people have agreed on welfare standards—how much a person should care for themselves versus others—this supports a particularly efficient form of decision-making in future situations. One can simply apply these without any need for actual or imagined renegotiation, by engaging in a weighted cost-benefit analysis. Thus, for instance, a family might have previously come to the understanding that desserts are divided equally among children. When they are presented with a new conundrum—a fancy cake, where one child want lots of frosting, another the corner piece, and the third a piece with frosting rose—the parents need not conduct a messy negotiation, or even imagine one, but instead merely determine which gerrymandered cuts maximize each child's preferences maximally and equally.

[11] Indeed, the logic of universalization might also be applicable to setting welfare standards. It has been argued, for instance, that adopting a general standard of assigning at least minimal weight to the welfare of all strangers (Ullmann-Margalit, 2011) allows for mutually beneficial comity, and that adopting the general standard of sharing resources "fairly" (i.e., dividing them equally) in certain economic exchanges allows for the maintenance of mutually beneficial market exchange (Henrich et al., 2010). Possibly, then, the logic of universalization could be used in order to determine which universally applicable welfare standards would be agreed upon by a moral community. This stands out as a fertile area for further study.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Or, imagine a doctor who witnesses a car accident and has the opportunity to care for its victims. What degree of concern for the victims' welfare is incumbent upon the doctor? They might analogize from other situations where the duties of care have already been established. Is this like their obligation to a patient in the emergency room? Or more like their obligation (or lack there of) to any stranger they meet? Both of these possibilities embody generalization from existing welfare standards to a novel, unanticipated case.

Crucially, in each of these cases we imagine that the decision-maker does not mentally *renegotiate* the welfare-based standard (a "model-based" approach), but instead *generalizes* from relevantly similar ones already negotiated (a "modelfree" approach). This makes it less computationally demanding than modelbased approaches.

Still, engaging in welfare-based reasoning will usually be more cognitively demanding than to simply retrieve a simple rule or heuristic (J. Greene, 2014). This is because it does require model-based planning of a different sort: Deriving the welfare consequences of ones behavior from a generative model linking actions to outcomes. In other words, this method avoids having to model *negotiation*, but retains modeling *the effects of one's action on others*. Concretely, one can say "Adopting my standard approach, I ought to satisfy each child's wishes equally" and proceed to considering how to slice the cake, without first asking, "if they bargained over welfare standards, what standard would these children agree to?"

Since many other theories posit that people engage in welfare-based reasoning during moral judgment, what are the distinctive predictions of the contractualist view? The key question is how the "weights" on welfare are determined: From past agreements (real, or imagined), or exclusively from other sources, such as a rational commitment to broadly utilitarian preferences (J. Greene, 2008; Kahane et al., 2015), or from blind adaptive processes of biological or cultural evolution (Kleiman-Weiner et al., 2017)?

The contractualist view helps to explain why those weightings are highly parochial (i.e., showing strong preferences for some individuals' welfare over others), because of the disparate social benefits and bargaining power of different individuals we interact with (Chen & Saxe, 2023; Le Pargneux & Cushman, 2023; O'Connor, 2019). Thus, for instance, we might feel stronger moral obligations towards those with whom relationships are especially beneficial to us, or those who have more available alternative social relationships rather than fewer. These are features that a standard analysis of utilitarian reasoning, in terms of something approximating philosophical consequentialism (J. Greene, 2008), struggles to explain (although parochialism depending on biologial relatedness is much easier to explain from the standpoint of blind adaptation). The contractualist framework also explains why our welfare-based reasoning is highly *situational* (i.e., the way we value others' welfare depends greatly on context and social role, such as a doctor who cares for her patients on shift but not off), since the agreements we reach are often situationally constrained, and generalizations from past to present agreements will depend on the apparent situational similarity. This feature of welfare-based reasoning is harder to explain either in terms of folk-utilitarianism or biological adaptive processes, although it could arise from precedent-based cultural evolution.

4.3.2.3 Model-free action-based judgment People also often make moral judgments based on action standards: by attending to rules, norms, etc., that render certain classes of actions permissible or impermissible. In the present framework, this approach achieves maximum computational savings, since it is neither necessary to model the bargaining process, nor is it necessary to model the consequences of one's actions on others' welfare.

We have already raised a few of the ways that rules can be established. One is actual negotiation, whether formal or informal. A second is through modelbased methods, such as virtual bargaining or universalization. (These methods are not model-free, of course, but once they establish a rule it can

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

be deployed in a model-free manner subsequently.) Action standards can also be established by generalizing from sufficiently similar precedents. For instance, if you are aware of the rules that apply at one swimming pool, you might readily generalize the same rules to other swimming pools.

When rules take the form of laws or norms, they can also be socially learned. Many formal rules, such as laws, are explicitly taught. Informal moral norms can also be inferred by observational social learning—i.e., seeing how people around you act (Goldstein & Cialdini, 2011; Nichols, 2021). In this case the bargaining processes that establishes the standard occurs in one set of individuals, and then the resulting rule is acquired and deployed in another.

Many prior discussions of moral rules agree that that they are cognitively efficient heuristics, but propose that they are designed to approximate a utilitarian ideal (Baron, 1994; J. Greene, 2014; Sunstein & Ullmann-Margalit, 1999). In contrast, we propose that moral rules are often heuristics designed to approximate a contractualist ideal of instance-based decision-making—i.e., rules are supposed to guide us towards the acts we would have agreed to in negotiation. This aligns with the utilitarian ideal in one specific way: its concern for outcomes. After all, the idealized bargain derives the quality of acts from the quality of the outcomes that they will tend to produce. However, it diverges from the utilitarian ideal in that, rather than maximizing aggregate benefit, bargaining solutions aim for mutual benefit—the outcomes predicted by gametheoretic bargaining solutions.

Once we look closely at actual moral rules, we see the fingerprints of the contractualist ideal everywhere. Rules governing property rights, for instance, can be elegantly explained through the lens of attempting to bring about a bargaining solution (Smith & Price, 1973). Formal analysis shows that self-interested agents bargaining over resources will tend to settle into an equilibrium where each fights hard to maintain one's current resources, while acting deferentially towards others' comparable claims. It is harder to explain how strict adherence to absolute property rights bring about overall utility maximization, however; on the contrary, it is often apparent that taking things from one person and giving them to another would maximize welfare, at least in certain specific cases. The institution of promising also finds a natural fit with a contractualist function. Promising someone something creates a moral obligation to carry through. But why? A welfare-maximization model would suggest that we should ignore a promise if the harm to one party is offset by the benefit to another. But, as fits our intuitions, a contractualist approach holds that a promise, like other agreements, must be kept unless *both* parties agree to break it (or would do so, where actual negotiation is not possible).

5Moral flexibility

It is tempting to suppose that people trust and rely on moral standards because of their sturdy and stable appearance. But upon closer inspection our moral standards are neither sturdy nor stable. People constantly revise moral rules, craft exceptions to them, and generalize them to new situations. (In this sense our morals are like airplanes—what naive passengers call a jet, seasoned pilots joke, is just a collection of spare parts flying in close formation). Situations constantly arise in which the moral standards that we've assembled conflict, cannot be clearly applied, or manifestly fail to achieve their original purpose. In these cases we are forced to flexibly reassemble new morals from spare parts on the fly.

Current theories of moral psychology fall short of explaining how, when, and why this occurs. The resource-rational contractualist account, however, makes clear predictions about how, when, and why people will exhibit moral flexibility. Specifically, it predicts that moral flexibility arises from *renegotiation*, whether actual or virtual, and that it will occur when the benefits of finding a better

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

solution outweigh the efficiency gains of sticking with a heuristic approach (such as a moral rule) which produces results we would be unlikely to agree to if negotiating explicitly.

5.1 The flexibility of moral standards

Moral standards are seldom absolute and unchanging. Consider the months following the emergence of the COVID virus. Early on, for many people, the normative landscape was characterized by relatively broad standards: “Socially isolate completely”; “Wear a mask at all times”, etc. Over time, however, people began to confront situations that seemed to warrant limited exceptions to these standards: Perhaps it would be acceptable to have social interactions with a limited, defined “pod” of close others; perhaps it would be acceptable to go maskless outside, when maintaining sufficient distance; perhaps one could remove a mask briefly to take a sip of water on an airplane; etc.

In doing so, we suggest, people likely drew on the entire set of cognitive mechanisms described above. After all, even though each mechanism approximates the ideal standard of achieving mutual benefit in instance-specific ways, each approximation operates differently, leading to different patterns of judgment. We propose that people revise their moral standards when they become dissatisfied with one pattern and seek another. The key question, then, is how we determine when one set of standards is failing, and how to find a better alternative.

5.1.1 When do we negotiate and renegotiate?

Broadly speaking, renegotiation can be divided into two types. First, we may need to renegotiate agreements because the world has changed, so that our past agreements and standards are out of date. The early days of the COVID pandemic are an excellent example. Just as occurred during the pandemic, we might often expect that our initial attempts at renegotiation will depend on especially wide-scope abstractions—a highly heuristic, rough “first-pass” at achieving mutual benefit.

During moments of relative stability, however, renegotiation can serve a distinct purpose: To refine the precise contours of moral judgment to better achieve mutual benefit in the cases that our broad agreements do not handle well. The later days of the COVID pandemic, and the myriad particular rules, standards and exceptions that occurred, exemplify this variety of moral flexibility.

A contractualist framework predicts that several factors will tend to encourage more effortful approaches to moral judgment.

First—as we have already suggested—the more unique a situation is, the greater the likelihood that generalizing from past agreements (whether based on precedent, action standards, or welfare standards) will not adequately approximate the outcome of *ad hoc* negotiation. Unique situations might also introduce sufficient uncertainty about others’ welfare interests and beliefs that mentally simulating those interests will be unreliable. Unique situations, therefore, will tend to justify more cognitively or socially costly methods.

Second, the higher the stakes of a moral decision, the greater the cost of a sub-optimal choice. Thus, like unique situations, high-stakes situations should favor those mechanisms of moral judgment that achieve greater accuracy at the cost of effort.

Third, people may be motivated to question existing standards when they suspect that they, personally, would profit from instance-based renegotiation. For instance, suppose that two colleagues are vying for a new and desirable office space that has become vacant. If either suspects that they would have superior leverage in *ad hoc* negotiation (but would be disfavored under existing standards, such as a rule based on seniority), they might favor the cognitive and social costs of renegotiation.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

5.2 Inverting a model of agreement

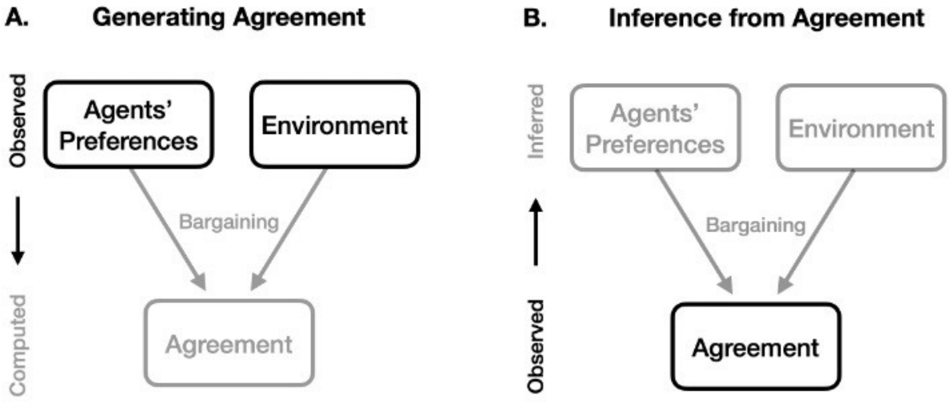
So far, we have discussed how people can use a generative model of agreement in a "forward" direction—i.e., moving *from* an understanding of others' utilities *towards* a conclusion about the standards that they would agree to. A key insight of contemporary cognitive science, however, is that we can use generative models to support inference on unobserved variables(Chater et al., 2010; Griffiths et al., forthcoming; Tenenbaum et al., 2006, 2011). For instance, we can use a generative model of others' minds (a "theory of mind") to infer their unobserved mental states, or we can use a generative model of physics to infer unobserved forces such as gravity or the wind (Baker et al., 2017; Battaglia et al., 2013; Kleiman-Weiner et al., 2015; Ullman et al., 2009).

Similarly, our view predicts that people should be able to use a generative model of agreement to infer unobserved variables that are relevant to moral decision-making: The utilities of the agents we are interacting with, and the functions of the standards that currently govern our conduct. In other words, we can ask, "what motives best explain the agreements we see around us" in cases where the motives themselves are not perfectly known.

5.2.1 Inferences about the environment and the utilities of agents

Imagine, for a moment, that you are a traveller to a small foreign village. You learn that in this village people are allowed to take as much water as they want from a nearby well, but each person is allowed to take only one bucket of water each week from a distant well, unless they are pregnant or sick, in which case they can take a bucket each day. From this information alone, you can reasonably infer certain things about the environment and peoples' preferences in the village. It would be reasonable to infer that (1) the water from nearby well is plentiful; (2) the water from the distant well is a limited resource; and (3) the water from the distant well is more valuable than the water from the nearby well, and perhaps (4) it is less prone to pathogens, in particular. These sorts of inferences come naturally to us, but it is worth reflecting on why. Presumably it is because we have an understanding of the *generative process* by which standards are constructed: Namely, they are constructed when people come to agreement on binding constraints in order to achieve mutual benefit. Thus, we can ask ourselves, "which facts about the environment and people's preferences would best explain why people agreed to these standards?"

Although its details are fanciful, the basic structure of this case is ubiquitous. We acquire many moral standards by social learning: We are either taught the standards, or we infer them by observation (Nichols, 2021). In either case, the



https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Figure 4: In addition to (A) using one's knowledge of agents' preferences and environmental constraints to model the mutually beneficial agreements they would come to, one can also (B) invert this model to infer agents' preferences and environmental constraints from the agreements they have reached, including the welfare and action standards that they employ.

standards become apparent to us often without any discussion of the underlying principles that establish or justify them. A young child might observe, for instance, that she is permitted to slurp soup in front of her disgusted sister, but she is absolutely forbidden from doing so in front of her impassive grandmother. From this she can infer that, despite her grandmothers impassivity, she really doesn't like slurping noises. She can also infer that, in her family, grandma's preferences about social niceties count for much more than her sister's. Testing this prediction of our view is a critical avenue for future work.

Our view also predicts that performing these inferences should help us to better exercise moral flexibility in the future. Having inferred "most people don't like slurping noises" and "grandma's preferences count for a lot" one can draw better conclusions about whether to slurp soup in front of the rabbi, or prank grandma with a whoopee cushion. It can also help you identify permissible exceptions to the standard—for instance, it informs you that, when at grandma's house, *she* can tell you, "it's alright to slurp your soup", but your *sister* cannot.

5.2.2 Inferences about the function of standards

Many standards have specific functions that can be stated with greater precision than just "achieve mutual benefit". These specific functions allow us to substitute local, tractable problems for global, intractable ones. For instance, in many cultures there is a strong norm that one must wait in line for goods and services, such as coffee in a shop. Its function is to achieve a fair and predictable distribution of costs (waiting times) and benefits (coffee) across customers. Optimizing the rules of a line to balance costs and benefits is much easier than trying to figure out what arrangement leads to mutual benefit in a global, all-things-considered sense (Awad et al., 2022; Kwon et al., 2022).

Crucially for our present purposes, our theory predicts that people should be able to infer those functions and, when they do, use them to guide their reasoning about exceptions and other forms of moral flexibility. Examples of this are everywhere in daily life. Returning to the example of lines in coffee shops, people are willing to consider certain exceptions to this norm. For instance, if you are given the wrong coffee, perhaps you can cut in line to replace it. There are other exceptions people will not grant, however: Just being "in a big hurry" does not permit you to skip to the head of the line. Why? This discrepancy is hard to explain in terms of "mutual benefit", since skipping ahead might benefit both the hurried customer and the customer with the wrong order equally. It is easier to explain, however, in terms of the *specific* function of the line: To distribute costs (waiting) and benefits (coffee) equally. The hurried customer seeks to obtain a benefit without paying any cost, creating inequity, while the customer with the wrong coffee seeks to obtain the proper benefit for a cost they have already paid.

How do we discover the specific function of a moral standard? We propose that this can be accomplished by inverting a generative model of the process by which the standard was adopted. For instance, we know that it is wrong to charge customers twice as much for water on a hot day, scoot down the empty shoulder of a highway in a traffic jam, or scoop all the pennies out of the change jar by the checkout counter; yet, many people who know these things have never been explicitly told what function each standard serves. Still, we can make educated guesses by asking ourselves how they are designed to achieve mutual benefit in some local sense. Store-owners and clients mutually profit when prices are predictable; one person can scoot down a highway, but if

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

everyone tried it nobody would be better off and emergency vehicles wouldn't be able to get through, etc.

5.3 Summary

We have proposed that moral judgments can be rendered by a variety of processes that occupy distinct points on a continuum governed by principles of resource rationality. Although the representations at each level are distinct, nevertheless information can be exchanged among them. This can occur in two dimensions. One can use the bargaining process, or a cognitive model of it, to move "forwards" from a representation of specific agents, their environment, and their utilities towards cached standards. Or, one can invert a cognitive model of the bargaining process in order to infer properties of agents, their environment, their utilities, and so forth, based on the cached standards and specific judgments that one observes. This is an application of the general principle of "representational exchange" (Cushman, 2020; Vélez et al., 2022) to the moral domain.

6Contractualist frontiers

What, in principle, can we bargain over? The trichotomy of instance-based, welfare-based and action standards provides an appealing organizational scheme because it aligns with major areas of philosophical and psychological inquiry: contractualism, consequentialism, and deontology. It is by no means, however, an exhaustive set. Several other potential targets of negotiation suggest promising avenues for further study.

Bargaining over virtues People do not just evaluate acts, consequences, and agreements, but also other *people*. In philosophy this is studied under the rubric of virtue ethics (see (Hursthouse & Pettigrove, 2022) for an overview), and in psychology there is also a rich tradition of thinking about person- or characterbased moral judgment (Merritt et al., 2010; Miller, 2014). A potential topic of negotiation, then, is which virtues count and how much. For instance, what counts as being a good cooperator? Should a good cooperator be fiercely loyal to those in their in-group, or generous to all regardless of group status (Enke, 2019; Graham et al., 2011)? Similarly, how should a person (e.g., in a public position) balance the virtues of impartiality and integrity against family loyalty or kindness to a deserving individual (Dungan et al., 2015; McManus et al., 2020)? The definition and weighting of various virtues may be negotiated and renegotiated across cultures.

Bargaining over beliefs People disagree not just about values, but also about facts. In cases of factual disagreement, we use argument and persuasion to attempt to change each others' minds. Now, of course, we can't trade beliefs as we can goods—e.g., one person cannot give up one belief if the other agrees to relinquish another. But we often do negotiate a common understanding of a situation *for the purposes of future negotiation*. This is one function of courts: they decide whether a person will be treated as guilty or innocent henceforth by the state and society at large, largely overriding any personal beliefs. That is, we collectively agree that we should treat the person as if innocent or guilty, irrespective our own opinions (of course, we might challenge the court's decision). Similarly, when a person is "declared the winner" of a literary competition, after intense negotiation in a judging panel, there is a collective agreement concerning the result. To a limited extent, companies, professions, and governments also have "received views" that dictate how their members must act and decide, independent of their personal opinions.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

There are limits to this analogy. One may not be able to directly apply gametheoretic concepts like a bargaining solution or equilibrium analysis—concepts defined over the payoff matrices of multiplayer games—to the dialectical process (Rubinstein, 2000, but see). And, whereas there is no particular challenge in agreeing to any *ad hoc* exchange of goods or services, there is a distinctive challenge to agreeing to *believe* something. Thus, one may not be able to simply say, "Fine, since you insist, I will simply go ahead and believe that the sky is green", and then make good on the promise. But one might be able to say, when negotiating an insurance claim "OK, let's work on the basis that all crops were wiped out by the flood" even if one doubts this is quite correct; or in a police investigation we might implicitly agree to work on the assumption that a victim of theft legitimately owned the now-stolen artworks, whatever one's secret suspicions. The potential for "epistemic" rather than "deontic" negotiation deserves further investigation.

Bargaining in social groups and institutions This paper has largely been concerned with how contractualism explains the moral cognition of dyadic and small-group interactions. However, there are several reasons to think that contractualism can potentially also explain aspects of how people navigate larger social groups and institutions and how those groups interact with one another. First, formal agreement-based processes are ubiquitous in democratic governance, from legislatures determining the policies that should guide conduct in a society to juries arguing to reach verdicts in trials. Second, private institutions also implement formal agreement-based methods for self-governance (Hadfield & Weingast, 2014). Third, intergroup relations may be characterized by similar bargaining principles as are interpersonal relations. Recent work emphasizes that conflict and cooperation between social groups are dynamic, with shifting alliances over time (Cikara, 2021). Theories of cooperative games offer a powerful tool to predict and explain coalitional behavior (Ray, 2007). Yet, while it seems likely that the logic of agreement can contribute much to any analysis of institutions and social groups, it remains an open question when, and to what extent, its value here arises from a contractualist moral psychology operating in individual minds. By analogy, some aspects of economic activity are very usefully analyzed through the lens of individual cognitive mechanisms for value-guided decision-making (e.g., consumer behavior), while other aspects are less so (e.g., the monetary policy of the Federal Reserve). Presumably there are also some cases in which inter-group and institutional decision-making is usefully understood through the lens of our individual cognitive mechanisms for resource-rational contractualism, and other cases in which it is not.

The boundaries of a bargain A contractualist approach to moral judgment immediately raises the question "Who is in on the bargain?" It is difficult to specify exactly how this boundary should be drawn. For instance, when parents decide who will pick up the kids, are they the only ones who get a say—or do the kids? When one country sets its carbon emissions policies, do other countries get a say? And so forth. It remains to future work to explore how people resolve this problem.

Moreover, we sometimes feel moral commitments to people beyond what a contractualist approach would seem to demand. We might feel obligations to strangers who we will never interact with or need to reach agreement with— distant victims of a crisis, for instance. And, we might feel obligations to provide people with more than they would have bargained for—to share resources equitably, for instance, even when we hold an advantaged position. How should we make sense of these cases?

Some cases of this kind may simply lay beyond the scope of our account. There are clearly some elements of our moral psychology that are best explained by principles other than contractualism. One example is our moral commitment to our own children. The most obvious explanation is

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

grounded in the logic of kin selection, not rational agreement. After all, who has less "bargaining power" than an infant? (See also, §2.3.)

On the other hand, there are other ways our account might be extended to cover certain cases of this kind. First, if people generalize agreements (whether *ad hoc*, welfare-based, or rule-based) to novel contexts as a resource-rational heuristic, a natural consequence is that they may overextend moral concern to others where it is not warranted by the logic of bargaining. (Other times, of course, they may undershoot.) For example, a shop owner may treat a onetime out-of-town patron with the same courtesy he extends to repeat customers because he generalizes from the welfare standard that typically applies to a usual circumstance. Second, there may be important reputational benefits to overextending favorable treatment to those who cannot bargain for it, in order to signal one's favorable qualities as a social partner to others who can. This logic has been very successfully developed elsewhere (André et al., 2022; Barclay & Willer, 2007). It may help to explain why contemporary large-scale societies, in which there is a robust market for productive positive-sum social relationships via market transactions, are characterized by especially high levels of moral concern for socially distant others (Enke, 2019; Henrich et al., 2010).

Finally, following the contractualist tradition in political philosophy, it is possible that such feelings of obligation can partially be explained as emerging from an agreement that is made under certain normative conditions. Rawls, for instance, famously suggested that moral principles arise from negotiations that take place behind "a veil of ignorance", a state that obscures each person's specific circumstances from themselves (Rawls, 1971). Future work should investigate the potential role of veil-of-ignorance reasoning in a contractualist moral psychology (Huang et al., 2019, see also).

7Conclusion

Tucked into the pages of his magnum opus, and penned in twilight of his life, Parfit offers an arresting image (Parfit, 2011). Like mountaineers, moral philosophers labor to ever higher positions of insight. Each asserts that their own path is best. But it is hard to see any standard by which they could judge one path better than another. From this discord, however, Parfit offers a hopeful view. Perhaps the climbers' paths share one essential feature: They all point up the same mountain. As each path is pursued to its logical conclusion, Parfit suggests, it will be discovered that they all converge at a common summit. From this he draws a lesson for moral philosophy: That the surest way forward for each moral theory is to discover its points of convergence with the others, rather than to dispute their discrepancies.

Parfit's vision was inspired by the fragmentary and conflicting state of moral philosophy, but the discord he describes applies equally to the prevailing picture of our moral psychology. According to this prevailing view, the moral mind is composed of fragmented and warring mechanisms vying for control over our actions and decisions. Perhaps, though, there a more abstract vantage point from which the many distinct and discordant capacities for moral judgment can be viewed as paths towards the same goal. In this paper we have suggested that there is.

There is wide-spread agreement across numerous fields that the goal of morality is to help us get along with each other. Humans face the challenge of interdependent choice: each person must figure out how to act given all the ways that their actions are intertwined with those of others. We aim for solutions to these problems that achieve mutual benefit. The most accurate way to find mutual benefit will often be actual negotiation between the affected parties, where all relevant information is discussed on a case-by-case basis. But humans have limitations—our time, effort, and

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

computational power needs to be used wisely. We therefore draw on our species' unique strengths to offset these constraints to make rational use of our limited resources.

Human moral judgment is governed by the logic of resource rationality along two principle axes. First, we not only bargain over one case at a time, but also over general standards that will apply in the future. Bargains can therefore concern how to manage welfare trade-offs, or which actions are permissible or impermissible. Second, bargaining is not limited to actual face-to-face negotiation. Instead, we can imagine and extend the outcomes of negotiations in a model-based or model-free way.

We can now see how the seemingly fragmentary moral mind is, in fact, elegantly unified. Distinct moral mechanisms can be seen as resource-rational approximations of bargaining solutions to multi-agent decision problems. Thus, morality reflects the ways in which humans are both sophisticated and limited. We have tremendous powers to understand others' minds, make good decisions, build world models, infer latent structure from sparse data, and craft useful abstractions and rules. Yet our computational resources are limited. Thus, the restrictions we accept on our behavior in order to navigate the social world reflect not only the kind of world that we live, but also the kind of minds that we have.

Acknowledgements

The authors gratefully acknowledge the feedback they received on the ideas presented in this paper from the following individuals: Jean-Baptist André, Nicolas Baumard, Yejin Choi, Joshua Cohen, Molly Crockett, Léo Fitouchi, Ori Friedman, Josh Greene, Gillian Hadfield, Max Kleiman-Weiner, Chris Kutz, Arthur Le Pargneux, John Mikhail, Jonathan Phillips, Iyad Rahwan, Rebecca Saxe, Walter Sinnott-Armstrong, Tan Zhi-Xuan. Likewise, the following groups provided invaluable feedback over the years in which these ideas were developed: Evolution and Social Cognition group in the Department of Cognitive Sciences at ENS - Paris, the Behavioural Science Group at the Warwick Business School, the Moral Psychology Research Lab at Harvard, the Computational Cognitive Science Lab at MIT, and the Social Heckling and Manuscript Editing (SHaME) writing group (East-coast chapter). Murphy was a constant source of encouragement during the writing of this manuscript. Finally, the authors thank Gottlob Frege for his continual inspiration.

Funding

This project was supported by grants from the Templeton World Charity Foundation and a US Army Research Office MURI grant (W911NF-19-1-0217) to J.B.T and by grant N00014-22-1-2205 from the Office of Naval Research and grant 61061 from the John Templeton Foundation to F.C. N.C. gratefully acknowledge support from the European Union/UKRI under Horizon Europe Programme Grant Agreement no. 101120763 – TANGO. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. For the purpose of open access, the author has applied a Creative Commons Attribution (CC-BY) licence to any Author Accepted Manuscript version arising from this submission.

Competing Interests

The authors declare none.

References

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Acheson, J. M. (1988). *The lobster gangs of maine*. Upne.

Adams, J. S. (1965). Inequity in social exchange. In *Advances in experimental social psychology* (pp. 267–299, Vol. 2). Elsevier.

Alexander, J. M. (2000). Evolutionary explanations of distributive justice. *Philosophy of Science*, *67*(3), 490–516.

Alexander, L., & Moore, M. (2021). Deontological Ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University.

Anderson, E. (2001). Symposium on amartya sen's philosophy: 2 unstrapping the straitjacket of'preference': A comment on amartya sen's contributions to philosophy and economics. *Economics and Philosophy*, *17*(1), 21.

André, J.-B., Debove, S., Fitouchi, L., & Baumard, N. (2022, May). Moral cognition as a nash product maximizer: An evolutionary contractualist account of morality. https://doi.org/10.31234/osf.io/2hxgu

Awad, E., Levine, S., Loreggia, A., Mattei, N., Rahwan, I., Rossi, F., Talamadupula, K., Tenenbaum, J. B., & Kleiman-Weiner, M. (2022). When is it acceptable to break the rules? knowledge representation of moral judgement based on empirical data. *CoRR*, *abs/2201.07763*. https:// arxiv.org/abs/2201.07763

Bacharach, M. (2018). Beyond individual choice. In *Beyond individual choice*. Princeton University Press.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Bandura, A., & Walters, R. H. (1977). *Social learning theory* (Vol. 1). Englewood cliffs Prentice Hall.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1610), 749–753.

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences*, *17*(1), 1–10.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational behavior and human decision processes*, *94*(2), 74–85.

Barrett, H. C., & Saxe, R. R. (2021). Are some cultures more mind-minded in their moral judgements than others? *Philosophical Transactions of the Royal Society B*, *376*(1838), 20200288.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59–78.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Binmore, K., Rubinstein, A., & Wolinsky, A. (1986). The nash bargaining solution in economic modelling. *The RAND Journal of Economics*, 176– 188.

Binmore, K., Swierzbinski, J., Hsu, S., & Proulx, C. (1993). Focal points and bargaining. *International Journal of Game Theory*, *22*, 381–409.

Binmore, K. (1994). *Game theory and the social contract: Just playing* (Vol. 2). MIT press.

Binmore, K. (2005). *Natural justice*. Oxford university press.

Binmore, K. (2014). Bargaining and fairness. *Proceedings of the National Academy of Sciences*, *111*(supplement_3), 10785–10788.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Bowles, S., & Gintis, H. (2011). A cooperative species. In *A cooperative species*. Princeton University Press.

Boyd, R., & Richerson, P. J. (1990). Group selection among alternative evolutionarily stable strategies. *Journal of Theoretical Biology*, *145*(3), 331– 342.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology*, *13*(3), 171–195.

Broome, J. (2017). *Weighing goods: Equality, uncertainty and time*. John Wiley & Sons.

Bruner, J. P. (2021). Nash, bargaining and evolution. *Philosophy of Science*, *88*(5), 1185–1198.

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, *70*(3), 560–572.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in cognitive sciences*, *3*(2), 57–65.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811– 823.

Chater, N., Zeitoun, H., & Melkonyan, T. (2022). The paradox of social interaction: Shared intentionality, we-reasoning, and virtual bargaining. *Psychological Review*, *129*(3), 415–437.

Chen, A. M., & Saxe, R. (2023). People have systematic expectations linking social relationships to patterns of reciprocal altruism.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.

Cikara, M. (2021). Causes and consequences of coalitional cognition. In *Advances in experimental social psychology* (pp. 65–128, Vol. 64). Elsevier.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of personality and social psychology*, *104*(2), 216.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363–366.

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences*, *107*(40), 17433–17438.

Curry, O. S. (2016). Morality as cooperation: A problem-centred approach. In *The evolution of morality* (pp. 27–51). Springer.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273–292.

Cushman, F. (2015). From moral concern to moral constraint. *Current opinion in behavioral sciences*, *3*, 58–62.

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, *43*, e28.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21.

Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. In J. M. Doris (Ed.), *Moral psychology handbook*. Oxford University Press.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, *17*(12), 1082–1089.

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual review of psychology*, *41*(1), 525–556.

Daston, L. (2022). *Rules: A short history of what we live by*. Princeton University Press.

Delton, A. W., & Robertson, T. E. (2016). How the mind makes welfare tradeoffs: Evolution, computation, and emotion. *Current Opinion in Psychology*, *7*, 12–16.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological bulletin*, *139*(2), 477.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Dungan, J., Waytz, A., & Young, L. (2015). The psychology of whistleblowing. *Current Opinion in Psychology*, *6*, 129–133.

Easton, R. B., Graber, M. A., Monnahan, J., & Hughes, J. (2007). Defining the scope of implied consent in the emergency department. *The American Journal of Bioethics*, *7*(12), 35–38.

Enke, B. (2019). Kinship, cooperation, and the evolution of moral systems. *The Quarterly Journal of Economics*, *134*(2), 953–1019.

Everett, J. A., & Kahane, G. (2020). Switching tracks? towards a multidimensional model of utilitarian psychology. *Trends in Cognitive Sciences*, *24*(2), 124–134.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772.

FeldmanHall, O., Dalgleish, T., Evans, D., Navrady, L., Tedeschi, E., & Mobbs, D. (2016). Moral chivalry: Gender and harm sensitivity predict costly altruism. *Social psychological and personality science*, *7*(6), 542–551.

Gauthier, D. (1986). *Morals by agreement*. Oxford University Press on Demand. Gauthier, D. (1987). *Morals by agreement*. clarendon Press.

Gintis, H. (2014). *The bounds of reason: Game theory and the unification of the behavioral sciences-revised edition*. Princeton University Press.

Goldstein, N. J., & Cialdini, R. B. (2011). Using social norms as a lever of social influence. In *The science of social influence* (pp. 167–191). Psychology Press.

Graeber, D. (2012). *Debt: The first 5000 years*. Penguin UK.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Greene, J. (2008). The secret joke of kant's soul. *Moral Psychology Handbook*, *3*, 35–79.

Greene, J. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Griffiths, T., Chater, N., & Tenenbaum, J. (forthcoming). Bayesian models of cognition.

Habermas, J. (1990). *Moral consciousness and communicative action*. MIT press.

Habermas, J. (1996). Between facts and norms, trans. william rehg. *Polity, Oxford*, 274–328.

Hadfield, G. K., & Weingast, B. R. (2014). Microfoundations of the rule of law. *Annual Review of Political Science*, *17*(1), 21–42.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J. (2007). The new synthesis in moral psychology. *science*, *316*(5827), 998–1002.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of personality and social psychology*, *65*(4), 613.

Hamann, K., Warneken, F., Greenberg, J. R., & Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature*, *476*(7360), 328–331.

Hare, R. M., Hare, R. M., Hare, R. M. H., & Hare, R. M. (1981). *Moral thinking: Its levels, method, and point*. Oxford: Clarendon Press; New York: Oxford University Press.

Harsanyi, J. C. (1963). A simplified bargaining model for the n-person cooperative game. *International Economic Review*, *4*(2), 194–220.

Harsanyi, J. C. (1977). Morality and the theory of rational behavior. *Social research*, 623–656.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Henrich, J. (2004). Cultural group selection, coevolutionary processes and largescale cooperation. *Journal of Economic Behavior & Organization*, *53*(1), 3–35.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., et al. (2005). "economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and brain sciences*, *28*(6), 795–815.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., et al. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *science*, *327*(5972), 1480–1484.

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, *72*, 207–240.

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367.

Hirth, R. A., Chernew, M. E., Miller, E., Fendrick, A. M., & Weissert, W. G. (2000). Willingness to pay for a quality-adjusted life year: In search of a standard. *Medical decision making*, *20*(3), 332–342.

Hoffman, M., Yoeli, E., & Navarrete, C. D. (2016). Game theory and morality. In *The evolution of morality* (pp. 289–316). Springer.

Holyoak, K. J., & Powell, D. (2016). Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin*, *142*(11), 1179.

Houlihan, S. D., Kleiman-Weiner, M., Hewitt, L. B., Tenenbaum, J. B., & Saxe, R. (2023). Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A*, *381*(2251), 20220047.

Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *science*, *320*(5879), 1092–1095.

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good. *Proceedings of the national academy of sciences*, *116*(48), 23989–23995.

Hursthouse, R., & Pettigrove, G. (2022). Virtue Ethics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2022). Metaphysics Research Lab, Stanford University.

Jambon, M., & Smetana, J. G. (2014). Moral complexity in middle childhood: Children's evaluations of necessary harm. *Developmental Psychology*, *50*(1), 22.

Kachel, U., & Tomasello, M. (2019). 3-and 5-year-old children's adherence to explicit and implicit joint commitments. *Developmental Psychology*, *55*(1), 80.

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209.

Kant, I. (1785). *Groundwork for the metaphysics of morals*.

Killen, M. (1995). Preface to the special issue: Conflict resolution in early social development. *Early Education and Development*, *6*(4), 297–302.

Killen, M., & Turiel, E. (1991). Conflict resolution in preschool social interactions. *Early education and development*, *2*(3), 240–255.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. *CogSci*.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908–911.

Kohlberg, L. (1969). *Stage and sequence: The cognitive-developmental approach to socialization*. Rand McNally.

Konda, V., & Tsitsiklis, J. (1999). Actor-critic algorithms. *Advances in neural information processing systems*, *12*.

Kwon, J., Tan, Z.-X., Tenenbaum, J., & Levine, S. (2023). When it's not out of line to get out of line: The rules of rule-breaking. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*.

Kwon, J., Tenenbaum, J., & Levine, S. (2022). Flexibility in moral cognition: When is it okay to break the rules? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Le Pargneux, A., Chater, N., & Zeitoun, H. (2023). Contractualist reasoning influences moral judgment and decision making.

Le Pargneux, A., & Cushman, F. (2023). Bargaining power, outside options, and moral judgment.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological bulletin*, *107*(1), 34.

Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F., & Tenenbaum, J. (2024). When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*.

Levine, S., & Leslie, A. (2020). Preschoolers use the means-ends structure of intention to make moral judgments. *PsyArXiv*.

Levine, S., Rottman, J., Davis, T., O'Neill, E., Stich, S., & Machery, E. (2020). Religious affiliation and conceptions of the moral domain. *Social Cognition*.

Lockwood, P. L., Klein-Flügge, M. C., Abdurahman, A., & Crockett, M. J. (2020). Model-free decision making is prioritized when learning to avoid harming others. *Proceedings of the National Academy of Sciences*, *117*(44), 27719–27730.

Mallucci, P., Wu, D. Y., & Cui, T. H. (2019). Social motives in bilateral bargaining games: How power changes perceptions of fairness. *Journal of Economic Behavior & Organization*, *166*, 138–152.

Marr, D. (1982). *Vision: The philosophy and the approach*. W.H. Freeman; Company.

Marshall, J., Gollwitzer, A., Mermin-Bunnell, K., Shinomiya, M., Retelsdorf, J., & Bloom, P. (2022). How development and culture shape intuitions about prosocial obligations. *Journal of Experimental Psychology: General*.

Marshall, J., Wynn, K., & Bloom, P. (2020). Do children and adults take social relationship into account when evaluating people's actions? *Child Development*, *91*(5), e1082–e1100.

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science*, *31*(3), 227–242.

Melkonyan, T., Zeitoun, H., & Chater, N. (2017). Collusion in bertrand vs.

cournot competition: A virtual bargaining approach. *Management Science*.

Merritt, M. W., Doris, J. M., & Harman, G. (2010). Character. In *The moral psychology handbook*. Oxford University Press.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

Miller, C. B. (2014). *Character and moral psychology*. OUP Oxford.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Misyak, J. B., & Chater, N. (2014). Virtual bargaining: A theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130487.

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: Virtual bargaining underpins social interaction, culture, and society. *Trends in cognitive sciences*, *18*(10), 512–519.

Muthoo, A. (1999). *Bargaining theory with applications*. Cambridge University Press.

Nash, J. (1950). The bargaining problem. *Econometrica: Journal of the econometric society*, 155–162.

Nichols, S. (2021). *Rational rules: Towards a theory of moral learning*. Oxford University Press.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530–542.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *science*, *314*(5805), 1560–1563.

O'Connor, C. (2019). *The origins of unfairness: Social categories and cultural evolution*. Oxford University Press, USA.

O'Neill, O. (2012). Kant and the social contract tradition. *Kant's Political Theory: Interpretations and Applications, E. Ellis (ed.), The Pennsylvania Press, Pennysylvania*, 25–41.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of economic perspectives*, *14*(3), 137–158.

Ostrom, E., Burger, J., Field, C. B., Norgaard, R. B., & Policansky, D. (1999). Revisiting the commons: Local lessons, global challenges. *science*, *284*(5412), 278–282.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American political science Review*, *86*(2), 404–417.

Owen, L., & Fischer, A. (2019). The cost-effectiveness of public health interventions examined by the national institute for health and care excellence from 2005 to 2018. *Public health*, *169*, 151–162.

Parfit, D. (2011). *On what matters: Volume one* (Vol. 1). Oxford University Press.

Pellizzoni, S., Siegal, M., & Surian, L. (2010). The contact principle and utilitarian moral judgments in young children. *Developmental science*, *13*(2), 265–270.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, *17*(3), 145–171.

Piaget, J. (1932). The moral judgement of the child.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, *68*, 1281–1292.

Radkani, S., & Saxe, R. (2023). What people learn from punishment: Joint inference of wrongness and punisher's motivations from observation of punitive choices.

Railton, P. (2017). Moral learning: Conceptual foundations and normative relevance. *Cognition*, *167*, 172–190.

Rawls, J. (1971). *A theory of justice*. Harvard university press.

Rawls, J. (2004). A theory of justice. In *Ethics* (pp. 229–234). Routledge.

Ray, D. (2007). *A game-theoretic perspective on coalition formation*. Oxford University Press.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social justice research*, *15*(2), 165–184.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, 97–109.

Rubinstein, A. (2000). *Economics and language: Five essays*. Cambridge University Press.

Rustichini, A., & Villeval, M. C. (2014). Moral hypocrisy, power and social preferences. *Journal of Economic Behavior & Organization*, *107*, 10–24.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

Scanlon, T. (1998). *What we owe to each other*. Harvard University Press.

Schauer, F. (1987). Precedent. *Stanford Law Review*, 571–605.

Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., Rascanu, R., Sugiyama, L., Cosmides, L., & Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition*, *168*, 110–128.

Shaw, A. (2013). Beyond "to share or not to share" the impartiality account of fairness. *Current Directions in Psychological Science*, *22*(5), 413–417.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667–677.

Skyrms, B. (2014). *Evolution of the social contract*. Cambridge University Press.

Smith, J. M., & Price, G. R. (1973). The logic of animal conflict. *Nature*, *246*(5427), 15–18.

Starmans, C., Sheskin, M., & Bloom, P. (2017). Why people prefer unequal societies. *Nature Human Behaviour*, *1*(4), 1–7.

Stegall, J., Mikhail, J., Nichols, S., & Kushnir, T. (2023). Underdetermination and obligation rules: Adult and children's use of closure principles in moral learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Sunstein, C. R., & Ullmann-Margalit, E. (1999). Second-order decisions. *Ethics*, *110*(1), 5–31.

Taylor, G., & Wolfram, S. (1968). The self-regarding and other-regarding virtues. *The Philosophical Quarterly (1950-)*, *18*(72), 238–248.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309–318.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Theriault, J. E., Young, L., & Barrett, L. F. (2021). The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews*, *36*, 100–136.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie*, *20*(4), 410–433.

Tomasello, M. (2009). *Why we cooperate*. MIT press.

Tomasello, M. (2020). The moral psychology of obligation. *Behavioral and Brain Sciences*, *43*, e56.

Tooby, J., Cosmides, L., Sell, A., Lieberman, D., & Sznycer, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. *Handbook of approach and avoidance motivation*, *15*, 251.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, *22*.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press

Ullmann-Margalit, E. (2011). Considerateness. *Iyyun: The Jerusalem Philosophical Quarterly/:* 205–244.

van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature communications*, *10*(1), 1483.

Vélez, N., Wu, C. M., & Cushman, F. A. (2022). Representational exchange in social learning: Blurring the lines between the ritual and instrumental. *Behavioral and Brain Sciences*, *45*.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*, 279–292.

Williams, T. C. (1968). *The concept of the categorical imperative: A study of the place of the categorical imperative in kant's ethical theory*.

Zhao, X., & Kushnir, T. (2018). Young children consider individual authority and collective agreement when deciding who can change rules. *Journal of Experimental Child Psychology*, *165*, 101–116.

https://doi.org/10.1017/S0140525X24001067 Published online by Cambridge University Press