Fall 10-1-2021

Relational Morality

Brian D. Earp
Yale University Graduate School of Arts and Sciences, brian.earp@gmail.com

Abstract

Relational Morality

Brian D. Earp

2021

Contemporary work in moral psychology has focused primarily on judgments concerning interactions between strangers. However, it increasingly is recognized that much of human moral judgment takes place in the context of -- and is shaped by -- multiple dyadic social relationships, such as parent-child, teacher-student, close friends, long-term romantic partners, neighbors, teammates, and so on. In this dissertation, I show how such relationships are associated with distinctive patterns of socially prescribed cooperative functions (such as care, hierarchy, or mating), which can be used to predict out-of-sample moral judgments of both blame and praise regarding various actions in relational context. I then proceed to focus on the long-term romantic partner relationship, showing how the ordinary concept of "true love" is likewise imbued with normative expectations. In the final part of the dissertation, I discuss how descriptive findings about people's moral judgments in relational context may be used to inform substantive moral questions about relationships in philosophy and bioethics.

Relational Morality

A Dissertation

Presented to the Faculty of the Graduate School

Of

Yale University

In Candidacy for the Degree of

Doctor of Philosophy

By

Brian D. Earp

Dissertation Directors:

Molly J. Crockett and Joshua Knobe

Committee Chair: Paul Bloom

December 2021

Table of Contents

List of Figures

List of Tables

Source Material and Author Contributions

Chapter 1. Introduction. A brief overview of philosophical and psychological theories of human morality that center relational context. This chapter is an updated version of my Theme Essay and does not include any previously published material.

Chapter 2. How social relationships shape moral wrongness judgments. This chapter is based on Earp, B. D., McLoughlin, K. L., Monrad, J. T., Clark, M. S., & Crockett, M.J. (2021). How social relationships shape moral wrongness judgments. *Nature Communications*, 12(5776): 1-13.

Chapter 3. Praise and blame in relational context. This chapter is based on Earp, B. D.,* McLoughlin, K. L.,* Owen, A., Caraccio, M., Calcott, R., Monrad, J. T., Clark, M. S., & Crockett, M.J. (in prep). Praise and blame in relational context: predicting moral judgments for strengthening or weakening cooperative functions of care, hierarchy, mating, and transaction across multiple social relationships, *working manuscript*. *Co-first authors.

Chapter 4. True love: a normative relational concept. This chapter is based on Earp, B. D., Do, D., & Knobe, J. (2021). The ordinary concept of true love. In C. Grau & A. Smuts (eds.), *Oxford Handbook of Philosophy of Love*. Oxford: Oxford University Press, doi: 10.1093/oxfordhb/9780199395729.013.38.

Chapter 5. Conclusion. Summary, next steps, and normative implications. Please note that, in addition to some material from the original discussion section of "The ordinary concept of true love" (on which Chapter 4 of the dissertation is based), this chapter incorporates ideas and material from Earp, B. D.,* Lewis, J.,* Dranseika, V., & Hannikainen, I. (in press). Experimental philosophical bioethics and normative inference. *Theoretical Medicine & Bioethics*, in press. *Co-first authors. Finally, some paragraphs are adapted from a grant application regarding this work co-authored with Molly J. Crockett, Margaret S. Clark, and Maria Gendron.

Dedication

This dissertation was only possible because of some very important relationships: my close friends and family members, mentors, colleagues, classmates, students, and a particular doctor-patient relationship (*gracias* for seeing me through).

Chapter 1

Introduction

Abstract

Much recent work moral psychology has framed participant judgments in terms of their apparent adherence to abstract moral principles, such as the utilitarian requirement to impartially maximize welfare or Kant's categorical imperative. Moreover, participants often are asked to make judgments about cases involving hypothetical interactions between strangers. But most of our moral judgments in the real-world concern neither abstract principles nor strangers; rather, we tend to make concrete judgments about the behavior of those with whom we regularly interact – people we know, and with whom we stand in particular social relationships. Recognizing this, a growing number of moral psychologists are shifting their focus to the study of moral judgment in social-relational context. In this chapter, I give a brief overview of recent work in this vein, highlighting both strengths and weaknesses, and setting up a new 'relational norms' model of moral judgment which my co-authors and I have developed and tested over the past few years. To situate this discussion, I also draw on influential philosophical theories of human morality that foreground relational context. I suggest that these theories, in addition to those such as utilitarianism that have shaped so much of the literature on moral psychology, should perhaps be given more attention by moral psychologists than has so far been the case.

Introduction

An out-of-control trolley is barreling down the track. Five workers will be killed if it continues on its way. You are standing by a switch that, if flipped, will redirect the trolley to a side-track, where a single worker would be killed instead. Should you flip the switch? (adapted from Foot, 1967).

In one sense, this is a familiar scenario. Readers of the contemporary moral psychology literature will have encountered countless papers focused on 'trolley dilemmas' like the one just described. Starting with a classic line of studies by Greene and colleagues (Greene et al., 2001), participant responses to such hypothetical dilemmas conventionally have been divided into two main categories, *consequentialist* and *deontological*. When participants choose the option that is stipulated to maximize expected welfare—e.g., by saving the most lives—they are said to have made a 'consequentialist' judgment. When they fail to choose that option if it requires causing instrumental harm to an innocent person, especially through up-close-and-personal physical contact, they are often said to have made a 'deontological' judgment (Greene, 2015; Greene et al., 2004).

The basis for this division and the extent to which the two categories meaningfully correspond to their namesake theories in moral philosophy is controversial (Berker, 2009; Kahane, 2015). Nevertheless, more than a decade of research and hundreds of studies employing this paradigm have converged on the view that consequentialist judgments (so conceived) are characteristically associated with deliberative mental processes driven by such considerations as explicit cost-benefit analyses, whereas deontological judgments (so conceived) are characteristically associated with more reflexive mental processes driven by emotion

or intuition (Crockett, 2013; Cushman, 2013; Greene, 2008, 2015; Greene et al., 2004, 2004; Patil et al., 2019).

On this basis, it has been claimed that such dual mental processes should be understood as *psychological natural kinds*: dissociable patterns or modes of moral thinking that are deeply rooted in our cognitive architecture and thus part of the fundamental structure of our moral minds (Greene, 2008). Alongside other leading theories in moral psychology, including the Universal Moral Grammar theory (Dwyer et al., 2010; Harman, 2008; Hauser et al., 2008; Mikhail, 2007) and the Social Intuitionist/ Moral Foundations theory (Haidt, 2001, 2007)—each of which will be discussed in later sections—this Dual Process theory has become massively influential in the cognitive science of moral judgment and behavior (Demaree-Cotton & Kahane, 2018)

In recent years, critics have raised various concerns about the methodology underlying the Dual Process model. For example, Kahane and colleagues (Kahane, 2015; Kahane et al., 2015, 2018; Kahane & Shackel, 2010) have argued that participant responses to 'sacrificial' moral dilemmas of the kind exemplified by the famous trolley problem—sacrificial because at least one innocent life typically must be sacrificed to save a greater number—cannot reliably be used to infer either consequentialist or deontological motivations on the part of participants (for additional methodological critiques, see Berker, 2009). But even if the dilemmas can in fact be used to draw such inferences (Conway et al., 2018), or perhaps to reliably assess participants' moral intuitions in some other way, there still would be good reason to expand our methodological toolkit beyond the use of such dilemmas.

One reason goes like this: however familiar the trolley dilemmas have become in one sense—the sense noted previously, based on their widespread use—there is

another, perhaps more important sense in which they are unfamiliar, and hence unlikely to elicit common moral intuitions (Bauman et al., 2014; Fried, 2012; Gold et al., 2014). Apart from emergency first-responders, police officers, soldiers in wartime, or similar personnel, relatively few people in real life ever will face such a stark decision to sacrifice the life of a stranger to save a greater number. Indeed, even during the height of the COVID-19 global pandemic, when hospital staff, for instance, faced very real 'triage dilemmas' about how best to allocate life-saving medical care (Kneer & Hannikainen, 2020), it was not in the capacity of 'strangers' that these decisions were made. Rather, upon entering the hospital in need of care, the stranger becomes a *patient*, and the hospital staff their *doctor* – thus, a doctor-patient relationship is created, with all the special obligations this relationship entails (Gillon, 1986)

People do sometimes encounter strangers in need of help. But the way they respond to such strangers might often differ from how they would respond if the same individuals were ones with whom they stood in a more personal relationship (Lee & Holyoak, 2018). If one's parent, child, or romantic life-partner were on the side-track in the opening scenario, for instance, one presumably would be much less inclined to divert the trolley, even if more lives would be saved. Although study participants do seem to behave in a moderately generous way toward strangers under experimental conditions—for example, while playing the Dictator Game (Engel, 2011)—it has been proposed that unknown individuals with whom one has had no prior contact, and with whom one does not expect to interact in the future, more typically will fail to register as particularly important, or even relevant to one's moral intuitions, under the more ordinary conditions of real life (Bloom, 2011).

Thus, while we may have considerable evidence concerning the various factors that bear on hypothetical decision-making in uncommon or unrealistic scenarios with ostensibly moral connotations, we do not have comparably strong evidence—from the mainstream moral psychology literature, at least—of the factors underlying more common moral decisions that people make in real life. If our aim is to understand how the 'moral mind' works, therefore, we need to shift our focus away from one-shot interactions with anonymous individuals taking place under unusual conditions (Gray & Keeney, 2015), and toward interactions with particular individuals in relational context under the conditions of daily life (Clark & Boothby, 2013; Hofmann et al., 2014).

Taking a Relational Turn

Most of our moral decisions in everyday life have little to do with strangers. Instead, they concern people with whom we stand in some kind of prior, often personal relationship; with whom we have had multiple previous interactions; with whom we expect to continue interacting in the future; and with respect to whom the moral issues at stake do not typically have life-or-death significance (Clark et al., 2015). Indeed, within the literature on prosocial behavior, there are many studies showing that relational context matters a great deal for shaping moral responses, and that relational context shapes whether a given decision is even considered to be a moral one in the first place (for an overview, see Clark, Boothby, Clark-Polner, & Reis, 2015). To see this in an intuitive way, consider a very different scenario from the trolley problem at the start of this chapter. The scenario first was introduced by Gopnik (2009), and later adapted by Bloom (2011). Bloom's version is as follows:

> A young woman meets a much younger [male] and takes him into her home. He suffers from terrible limitations. He cannot walk or talk or even sit up; he cannot be left alone and must be carefully fed. He often needs attention at night, and she spends the first years with him in a sleep-deprived fog. Still, this is the most important relationship of her life. She would die for him. She spends many years nursing him as he gradually becomes able to walk, to toilet himself, and to express and understand speech. After they have been together for a decade, he becomes interested in other women and begins to date, and eventually he leaves her home and marries someone else. The woman continues to love and support him, helping to raise the children that he has with his new wife. (p. 26).

If we imagine that the much younger male in this story is an adult stranger whom the woman recently met on the street, her actions would seem incomprehensible. But if we imagine instead that he is her son, to whom she has just given birth—and thus 'met' as a baby—our moral intuitions change completely. Far from seeing her caretaking behavior as befitting only a saint or a madwoman, we now see it as perfectly normal and even expected. Indeed, if she *failed* to exhibit such care and sacrifice for her son, treating him as she would an actual stranger, her behavior would be judged to be deeply immoral (Bloom, 2011, p. 27).

This example highlights the importance of taking relational context into account if we want to make sense of moral motivations, behavior, and judgment in the real world (Bloom, 2011; Clark et al., 2015; Korsgaard, 1993; Lee & Holyoak, 2018; Mason, 2014; McGraw & Tetlock, 2005; Rai & Fiske, 2011; Shabo, 2012; Simpson, Laham, & Fiske, 2016; Strawson, 1962). Everyday experience, bolstered by the very large empirical literature on the psychology of close relationships (for an overview, see Clark, Lemay, & Reis, 2018), reveals that one and the same act, or sequence of actions, may be regarded as morally appropriate or inappropriate depending upon—

among other things—the identity of the actor (or the social role the actor is occupying in a given context or society), the identity or social role of the person with whom the actor is interacting, and, crucially, the nature of the relationship between them, whether existing or desired (Clark et al., 2015, p. 330; for a related discussion, see Earp, Douglas, & Savulescu, 2017). Indeed, across societies, a given act will be judged as good, fair, virtuous, honorable, morally correct, and so forth, when it takes place in a certain social-relational context, and as bad, unfair, lacking in virtue, dishonorable, or morally wrong (etc.) when it occurs in certain other social-relational contexts (Rai and Fiske, 2011, p. 57).

In practical terms, this means that one should be able to hold an act constant and manipulate the social role or relationship between individuals, and dramatically alter how one sees the moral status of the act in question (as well as the moral character of the actor) (Marshall et al., 2020; McManus et al., 2021). For a simple illustration of this phenomenon, consider the following scenarios from Clark et al. (2015, p. 229):

> 1. Jim has a demanding job and very limited vacation time. Nonetheless, he takes a full day off from work to help a total stranger move across town.
>
> 2. Tasha and Elaine have been best friends since childhood, and they often share their most private thoughts and feelings. When Tasha calls Elaine in distress late one evening with an urgent need to talk to her friend, Elaine says, "I'm happy to talk to you Tasha, but could you call back tomorrow, during normal business hours?"
>
> 3. Anne runs out of gas on her drive home from work. She calls her sister, who promptly picks her up and brings her to a nearby gas station to buy a canister of gas. Once they've ensured that the car is running again, Anne's sister hands her a bill for her time and labor.

These cases are supposed to strike the reader as strange. Although each of the scenarios involves prosocial—and thus ostensibly moral—activity of one kind or another (helping someone move, comforting someone in distress, rescuing someone who is stranded), some of the activity seems puzzling or even inappropriate. Intuitively, there is mismatch between the actor's behavior and the *type of relationship* the actor is described as having with the beneficiary of that behavior. Why would Jim miss work to help a complete stranger move across town? Why would Elaine make her close friend call back during business hours? Why would Anne's sister expect to be compensated for helping out with the gas?

If one changes the relationship context, however, such puzzles are immediately resolved. Instead of a stranger, suppose that Jim is helping his *fiancé* make the move across town; instead of a lifelong friend, suppose that Elaine is Tasha's *therapist*; and instead of her sister, suppose that the person helping Anne is an *employee* from a roadside assistance company (Clark et al., 2015, p. 330). Now their behavior does not seem quite so strange.

What these examples suggest is that we cannot fully understand, or even *begin* to understand, the moral status of an action in a given social context unless we know who is performing the action, who is affected by it, and the nature of the relationship between them. By ignoring relational context, therefore, we remove the factor with perhaps the greatest explanatory power from our models of moral judgment and behavior (Clark et al., 2015; Rai & Fiske, 2011). And yet *most* current research in moral psychology does not take this context into account. Instead, relationship-oriented considerations are often treated as 'noise'—distorting factors that must be filtered out with a focus on one-time interactions between strangers (Hester & Gray, 2020).

Such approaches appear to rest on the assumption that there must be an underlying source of moral judgment whose realization in a particular context might be biased by social or relational factors (Rai and Fiske, 2011, p. 58). But in reality, the opposite is true: social-relational factors must be taken into account to accurately predict moral judgment in a given situation. Thus, we suggest, they should be treated as the *data* or *signal*, rather than experimental noise. In summary, the time is ripe for moral psychologists to focus more heavily on theories and studies that do not filter out, but rather are rooted in, common, real-life behaviors and activities taking place between individuals in interpersonal relationships (of which the relationship between strangers is but one particular type) (Clark & Boothby, 2013; Hofmann et al., 2014; Rai & Fiske, 2011).

Recognizing this, some psychologists who study moral judgement and behavior have indeed begun to steer their research programs in a more relationship-oriented direction: the beginnings, perhaps, of a broader 'relational turn' in moral psychology. In a later section, we will survey some of the major work that has been done in this vein so far, calling attention both to strengths and weaknesses. First however, we will lay out a theoretical roadmap to give shape to this relational turn.

Philosophical Foundations

Casual readers of the moral psychology literature could be forgiven for having the impression that there are two, and only two, explicit theories of morality discussed by philosophers that might be relevant to their empirical work. These would be the consequentialist or utilitarian theories most closely associated with Bentham (1789) or Mill (1863), and the deontological theories most closely associated with Kant and his interpreters (Kant 1785/1997). Despite often being characterized as competing, or

even contradictory visions of right and wrong, the two theory-types have much in common.

One shared feature is a dominant focus on universalizing, abstract moral principles, designed to apply every person in any situation at all times (Randall, 2019). Classical act utilitarianism, for example, holds that the morally correct action is the one that reasonably is expected to bring about the greatest happiness for the greatest number of people, without regard to one's personal relationship to any of the people in question. And Kant's categorical imperative (for a classic discussion, see Foot, 1972) famously holds that one should always act in accordance with moral maxims that one rationally can will to become a universal law. Neither of these formulations rules out taking relational context into account in deciding what to do in a specific situation: for example, what it is that maximizes welfare often depends on who will be affected the decision or action and their relationship(s) to the actor. Nevertheless, there are other important philosophical theories of morality that take relational factors more centrally into account. Moreover, in addition to determining what is typically *judged* to be morally required in a given situation, as we touched on in the previous section, these theories hold that the social and relational context determines what is *in fact* morally required—that is, from a substantive or "normative" perspective (Noddings, 2013; Randall, 2019). According to Tan and Snell (2002, p. 362), in predominately Chinese societies, for instance,

> morality is both role and act dependent. Morality deriving from Confucian teachings emphasizes virtuous personal qualities (e.g., loyalty, honesty, obedience, sincerity, etc.) required in performing roles, and regards rights, privileges, claims, immunities, expectations, obligations, and responsibilities as particularistic, relationship-based and role-related. In this moral tradition, required moral behavior varies according to a person's role, position and

relationship with the other role-players in a highly differentiated and hierarchical social nexus.

Within the Western tradition, Strawson (1962), for example, explicitly grounded his moral theory in relational or intersubjective terms. Specifically, he grounds them in the "non-detached attitudes and reactions of people directly involved in transactions with each other; of the attitudes and reactions of offended parties and beneficiaries; of such things as gratitude, resentment, forgiveness, love, and hurt feelings" (p. 5). Strawson argues that intimate relationships of the kind almost everyone values are not psychologically possible[1] without such 'reactive' attitudes, and that such attitudes imply certain genuinely normative, rather than merely descriptive, moral commitments or practices: for example, practices of holding one another morally accountable.

Building on Strawson's insights, Darwall has argued in an extensive body of work that utilitarian-style considerations of benefit and harm cannot, even indirectly, ground the existence of moral rights or duties; and that, among other things, the very concept of moral blame cannot be understood from outside what he calls the 'second-person standpoint' – the standpoint that two parties to a relationship (a 'you' and an 'I') necessarily take toward each other when engaging in such practices as assigning responsibility, making claims on one another, and holding each other accountable (Darwall, 2009, 2010, 2018; Dill & Darwall, 2014).

Similarly, Korsgaard (1993), argues that basic values—by which she means moral values with real or binding normative force—are ultimately grounded in, and supervenient upon, the structure of interpersonal relationships. She calls her view

[1] See Sommers (2007) for an objection to this view. See Mason (2014) for a response. See also Kelley et al. (2003) for a psychological 'atlas' of interpersonal relationships and their structure and significance.

intersubjectivism. On this view, the subject matter of morality is not *what* we should aim to do or bring about in the abstract, based on some invariant maxim or principle, for example, but rather it is *how* we should relate to one another (p. 25). Korsgaard argues that individual, subjective interests or values become intersubjective values when we take an attitude toward each other that impels us to recognize—and share—each other's ends in a given context (see also Kelley, 1979, for a psychological perspective).

Scanlon (2008) also grounds his theory—known as *contractualism*—in relationships. Contractualism holds, most basically, that "an act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject as a basis for informed, unforced, general agreement" (Scanlon, 1998, p. 153). Thus, like utilitarianism and Kant's categorical imperative, there is a universalized statement of the view. But Scanlon's contractualist account is not primarily focused on determining, on such abstract grounds, which actions are right and wrong. Rather, it is focused on the *reasons* for acting, and on the types of reasons that can be *justified to others*. Thus, as with Korsgaard (1993), there is a central role for reasons (or ends) that are shared between oneself and others with whom one stands in some kind of relationship. It follows, then, that the reasons for acting that will in fact be justifiable to those others will depend on who they are—and on the nature of their relationship to the agent.

According to this way of thinking, to make a sensible moral judgment, it is not enough to give an abstract analysis of an agent's intentions, actions, and/or the causes or consequences of those actions. Rather, the permissibility of a given act depends, in large part, on the reasons for which it was performed within the context of a given

relationship: on what Scanlon (2008) calls its "meaning." This refers to the significance of the action, for the agent as well as others affected by it (whether directly or indirectly), of the agent's willingness to perform the action on behalf of particular reasons. To say that an act is blameworthy, therefore, is to make a claim about its meaning: specifically, it is to claim that the act indicates something about the agent's attitudes that impairs her relations with others, such that to blame someone "is to understand one's relations with that person as modified in the way that such a judgment holds to be appropriate" (p. 6). As a consequence of this, contractualist blame "only makes sense within a relationship, such as friendship or family relationships. Impairment must be judged against the standard of what is appropriate *within that relationship*" (Ashford & Mulgan, 2018, emphasis in original). Our aim in the present work is, in large part, to give a theory supported by empirical results that can explain what is and is not appropriate (in this sense) in different relationships.

None of this focus on personal relationships is to say that we do not have moral obligations toward strangers, nor they toward us. Although the nature of the relationship between strangers often will differ in important ways from the nature of the relationship between, for example, close friends or family members, it is nevertheless the case that all individuals stand in *some* relationship to one another, characterized by such minimal values as mutual recognition of (and giving weight to) each other's interests. Such a bare-bones moral relationship can be thought of as "a normative ideal, like a normative ideal of friendship that specifies attitudes and expectations that we should have regarding one another [such that] morality requires that we hold certain attitudes toward one another simply in virtue of the fact that we stand in the relation of 'fellow rational beings'" (Scanlon 2008, pp. 139–40). Thus,

there are certain fundamental attitudes that we rightly expect even from strangers, and we may be justified in blaming them if their behavior appears to indicate a lack of such attitudes, including a basic level of care or respect (Ashford & Mulgan, 2018).

How can moral psychologists contribute to this discussion? If the source of normativity is in fact second-personal accountability, as Strawson (1964) and Darwall (2009) suggest, and this normativity supervenes on the structure of interpersonal relationships, as Korsgaard (1993) suggests—what exactly is that structure? What is its underlying logic? How might it have come about? Why did it evolve? What predictions does this structure make for real-life moral judgments and behavior? And if Scanlon (2008) is right that moral judgements such as blame only make sense in terms of what is appropriate for a given relationship—how is appropriateness determined? Why are some things seen as appropriate in certain relationships, but not in others?

In his classic work, Strawson (1962) anticipated such questions. In order to develop a more robust moral theory on the basis of interpersonal relationships, "we should think of the many different *kinds* of relationship which we can have with other people – as sharers of a common interest; as members of the same family; as colleagues; as friends; as lovers; as chance parties to an enormous range of transactions and encounters" (p. 6, emphasis added). Although we typically expect at least some degree of cooperation or goodwill on the part of those with whom we stand in such relationships, the specific antecedents, content, and consequences of such cooperation will vary significantly from one relationship to another. In our work, we offer a model of relationship 'kinds' that we think can explain such systematic variation, based on a set of *cooperative functions* that different relationships within a society must serve in order to solve recurrent coordination problems of our species.

Evolutionary Beginnings

If the philosophical views just surveyed are on the right track, then to understand human morality, whether descriptively or normatively, we should be very interested in exploring the 'structure' of interpersonal relationships (Kelley, 1979; Kelley et al., 2003). Moreover, we should expect that structure, and the reasons for acting that will be justifiable to others, to vary from relationship to relationship—leading to different 'reactive attitudes' depending on the nature of the relationship between each of us and those with whom we are interacting.

Evolutionary theory may shed some light here. Presumably, our moral psychology did not evolve primarily to guide our thoughts and behaviors in interactions with anonymous strangers. Rather, it likely evolved to help us avoid common causes of interpersonal conflict with familiar others that would threaten our evolutionary fitness, or otherwise help us solve coordination problems among those with whom we interacted on a regular basis (Bloom, 2011). In the environment of evolutionary adaptation (EEA; Bowlby, 1982; Volk & Atkinson, 2013), such coordination problems almost certainly would have involved—first and foremost—our immediate family and other close kin, but also friends, mating partners, and other members of our social group with whom we might seek to form an alliance, whether temporary or more enduring.

There are several plausible accounts of how this evolutionary process may have played out (Cosmides & Tooby, 2006; Haidt & Kesebir, 2010; Kitcher, 2011; Machery & Mallon, 2010; Pettit, 2015; Isern-Mas & Gomila, 2020); but they all share a similar basic thrust: those ancestors who were motivated to strategically cooperate with others in their immediate environment, find alternatives to violence in situations

of competing interests, punish defectors, and so forth, would have had greater reproductive success than those who did not have such motivations, and that is why we have such motivations today. According to Baumard and colleagues (Baumard et al., 2013), our moral motives are—more specifically—an adaptation to an environment in which individuals regularly competed to be chosen for mutually beneficial interactions with those around them.

On this 'partner choice' model, an instinct toward cooperative behavior would evolve for a very specific reason. Those with a reputation for having a reliable motive to cooperate in a mutually advantageous way—neither exploitative nor subject to exploitation—would tend to be preferred as social partners in contexts where cooperation, rather than defection, would enable them to reap the greatest reproductive benefit over the long run. And the most efficient, secure way to gain a reputation for having such a reliable motive, Baumard et al. (2013) argue, is simply to *have* the motive—that is, to sincerely believe that the 'right' thing to do is to behave in a cooperative (or otherwise moral) manner, and to feel negative emotions such as guilt for failing to behave in such a way. This, then, would explain how our moral instincts could have evolved over evolutionary time.

This theory has been critiqued on various grounds, including for purporting to give an account of our 'moral' sense while giving instead an account of our 'fairness' sense, which is but one part of our moral psychology. And it has been critiqued for ignoring relational context (Clark & Boothby, 2013). Thus, although the core of the theory is plausible, it could be improved by incorporating a recognition that cooperative motives rarely are impartial, in the sense of being indifferent to the type of relationship one has, or desires to have, with a potential cooperation partner. In the EEA, there would have been a strong reproductive advantage to showing favoritism in

cooperation (and other forms of moral behavior) toward friends, family, romantic partners, and others in one's close emotional sphere (Bloom, 2011). And even within that sphere, it would not have been adaptive to follow a single, consistent, or uniform moral norm in each relational context; rather, it would make adaptive sense to follow different norms in different contexts depending on the specific nature of the relationship—whether existing or desired—as well as its developmental stage, with these factors shaped, in turn, by the underlying evolutionary functions the relationship-type would be expected to serve in the given environment (Clark et al., 2015).

Rather than a generalized 'moral' sense, therefore, a friends-and-family cooperation bias, with more specific norms applied to particular relationships and even relationship-stages, must have played a large role in shaping our intuitive moral psychology. This lesson applies more generally. Consider the Universal Moral Grammar (UMG) theory championed by Mikhail and others (Dwyer et al., 2010; Harman, 2008; Hauser et al., 2008; Mikhail, 2007). This theory holds that our moral emotions reflect the evolutionary design and output of an innate moral faculty or 'organ'—an unconscious, computational system that maps input about agents, patients, intention, causation, and action onto underlying representations of moral principles—akin to a grammar with set rules or parameters—to produce an intuition about whether the action is permissible, obligatory, or forbidden.

Again, the core idea is plausible. Yet as with the Dual Process model discussed in the introduction, much of the empirical support for this theory comes from testing it against trolley problems in a dizzying array of permutations (Mikhail, 2007). To explain people's moral intuitions about more common scenarios, it will be necessary to be much more specific about the 'agents' or 'patients' involved, and to

build in information about the nature of the relationship between them (Hester & Gray, 2020; Schein, 2020). This will require theoretically grounded studies into moral judgment and behavior that systematically vary relational context.

Prior Significant Empirical Work Taking Relational Context into Account

One of the earlier studies to manipulate relational context while holding moral behavior constant was conducted by McGraw and Tetlock (2005). These researchers found that "moral outrage" and "cognitive confusion" could be invoked by asking participants to engage in what they call a *taboo trade-off*. This refers to any mental comparison or social interaction that violates a deeply-held intuition about the nature of given relationship and the values that undergird it (McGraw & Tetlock, 2005, p. 3). For example, they found that a substantial proportion of participants refused to name an acceptable selling price for a gift that was said to have been acquired in the context of a close relationship. Why might it be taboo to name such a price?

In a large body of work (beginning with Clark & Mills, 1979), Clark and her collaborators have shown that different relationships operate on different underlying values. One such value difference is between what she calls *communal* versus *exchange* relationships. In communal relationships, the giving of a benefit in response to a need for that benefit is seen as appropriate, whereas in exchange relationships, the giving of a benefit in response to past (or future, anticipated) receipt of a similarly valued benefit is seen as appropriate. One way to create a taboo trade-off, then, would be to ask someone to perform an action within the context of one such relationship that is governed by the underlying values of the other.

For example, one might apply a market-based heuristic to the transfer of goods or services within a relationship characterized by a norm of communal sharing (Clark

& Mills, 2012). Indeed, this is exactly what we saw with Anne—the stranded motorist—and her sister in an earlier section, and it explains why her sister's behavior seemed so strange. As Fiske and Tetlock (1997) argue, people in general tend to view such trade-offs as morally impermissible, and will respond with indignation when forced to assess the value of an action or entity that is normally governed by the rules and expectations of one relational model by the rules or expectations of a disparate or otherwise conflicting relational model (p. 256).

The relational models employed by McGraw and Tetlock (2005) in their experiments were the ones proposed by Fiske in a seminal paper (1992), building in part on the earlier work by Clark (e.g., Clark & Mills, 1979). According to Fiske, there are four fundamental relationship structures or schemas that can be combined in various ways to describe and make sense of the socially meaningful features of nearly any social situation across cultures. These are: communal sharing, authority ranking, equality matching, and market pricing. As summarized by McGraw and Tetlock (2005):

> *Communal sharing* (CS) slices the social world [into classes that] differentiate in-groups and out-groups without degree of distinction. Everyone in a community—which could be as small as a romantic dyad or as large as a nation state—shares certain rights and incurs certain duties. Nonmembers may be excluded entirely. Within the relationships, people give as they can and take as they need.
>
> *Equality matching* (EM) defines socially meaningful intervals that can be added or subtracted to keep score in social interaction. The social prototype is collegial or friendship networks in which in-kind or tit-for-tat reciprocity is a dominant exchange norm [and in which members must] calibrate degrees of indebtedness and strive for balance.

> *Authority ranking* (AR) imposes an ordinal ranking on the social world that permits lexical decision rules. One's location in this ranking scheme determines one's relative status in a collective and the prevailing direction of accountability for decision making. Military ranking serves as the social prototype …
>
> *Market pricing* (MP) makes possible ratio comparisons of the values of diverse entities through the use of a single value or utility metric. This structure underlies capitalism and monetary transactions that range in sophistication from simple loans to [much more complex] financial instruments. (p. 3).

In more recent theoretical work with Rai (Rai & Fiske, 2011), Fiske explicitly ties such relational structures to a discrete set of hypothesized moral motivations, resulting in what those authors term the Relationship Regulation Theory of moral psychology (RRT). According to RRT, Communal Sharing is associated with *Unity* (described as a motive to provide care and support to members of an in-group based on need or compassion rather than desert, and to protect the integrity of the group's boundaries by neutralizing the threat of contamination); Equality Matching is associated with *Equality* (a motive for in-kind reciprocity, balance, fair treatment, and equal opportunity); Authority Ranking is associated with *Hierarchy* (a motive to respect social rank, such that superiors are entitled to deference but must also guide and protect subordinates); and Market Pricing is associated with *Proportionality* (a motive for merit-based rewards and punishments, calibration of benefits to contributions, and cost-benefit analyses on a utilitarian framework) (p. 57).

Rai and Fiske (2011) claim that all cultures base their moral judgments and behaviors on this same underlying set of motives; that these motives exist for the purpose of regulating social relationships; and that these shared relational structures

and corresponding motivations explain the existence of apparent universals in moral beliefs and behavior across cultures. By the same token, moral diversity and disagreement, they claim, stems from the deployment of different relational models in a single circumstance, or the different implementation of a given relational model in one or more circumstances.

To date, there have been few empirical tests of this theory in terms of the predictions it makes for specific patterns of moral judgment. In part, this might be due to the difficulty one may have in formulating precise predictions. Although RRT posits just four fundamental moral motives—Unity, Equality, Hierarchy, and Proportionality—the actual content or descriptions of these motives, and the examples used to illustrate them, suggest a more complicated, and somewhat less coherent picture.

In a recent collaboration with Simpson and Laham (Simpson, Laham, and Fiske 2016), for instance, Fiske states that, according to RRT, moral judgment largely depends on the "varying activation" of the four moral motivations used to coordinate various interpersonal relationships (Simpson et al., 2016, p. 595). This, in turn, leads to the broad hypothesis that "varying the relational context of moral violations should predict variability in wrongness judgments independently of other factors relevant to moral judgment" such as gender, political ideology, and religious affiliation (Simpson et al., 2016, p. 597).

In an attempt to derive more specific predictions, the authors import a set of norm violations from Moral Foundations Theory (MFT) (Haidt, 2007). This theory proposes five innate 'modules' for moral cognition that are hypothesized to have evolved to serve certain adaptive ends. These have been summarized by Suhler and Churchland (2011), in a critical piece, as follows:

Harm/Care — protect and care for young, vulnerable, or injured kin

Fairness/Reciprocity — reap benefits of dyadic cooperation with non-kin

Ingroup/Loyalty — reap benefits of group cooperation

Authority/Respect— negotiate hierarchy, defer selectively

Purity/Sanctity — avoid microbes and parasites

Simpson et. al (2016) reasoned that Harm/Care violations from MFT would be judged to be more wrong in relationships prototypically characterized by Communal Sharing (Unity motive, in particular the aspect concerned with empathy and provision of care) than any of the other relationship types; that Fairness/Reciprocity violations would be judged *least* wrong in such relationships; that Authority/Respect violations would be judged most wrong in prototypically Authority Ranking relationships (Hierarchy motive); and that Ingroup/Loyalty and Purity/Sanctity violations would be judged to be most wrong in Communal Sharing relationships (due to the postulated aspect of the Unity motive concerned with protecting the group from outside contamination).

Violations were expressed as short sentences of the form "Person A does X to B" where X violates one of the MFT foundations (e.g., "Person A borrows $20 from Person B's bag without asking" as a violation of Fairness/Reciprocity), and Person A and B are everyday relationship pairings, such as brother-sister, student-professor, or customer-salesperson. Simpson et al. (2016) found mixed support for their specific predictions. For example, relational context did not consistently or uniquely predict wrongness judgments of care violations, which they took to suggest that relational context is not especially important for judging the moral status of physically or

emotionally harmful transgressions (p. 605). There was, however, a large main effect for relationship context, such that manipulating context explained variability in moral judgment above and beyond such potent predictors as gender, political ideology, religious affiliation, and even explicit endorsement of MFT foundations. As Simpson et al. (2016) concluded, "if we want to gain a thorough understanding of why individuals vary in their moral judgments, relational context cannot be ignored" (p. 605).

There are clear limitations to the studies by Simpson et al. (2016), as they acknowledge: (1) they used a small number of relationship exemplars (e.g., Student-Professor), with each one chosen to represent a single relationship model (e.g., Authority Ranking) in an all-or-none fashion (i.e., no consideration of degrees); (2) only wrongness judgments were assessed, in response to violations of moral foundations (no corresponding assessments were made of praiseworthy judgments based on morally positive behaviors); and (3) relationship models were assessed in isolation (presumably, the combination of certain relationship models will provoke the strongest moral judgements, for example, when they are at cross-purposes within a single dyad, as Simpson et al. point out). Nevertheless, these authors have called vital attention to, and begun to explore the contours of, an enormous gap in the existing moral psychology literature—most of which does not even consider relationship context, much less manipulate it in a systematic way.

More recently, Tepe and Aydınlı-Karakulak (2018) sought to test the RRT, predicting that moral judgments would not be based on such considerations as harmfulness or impurity alone, but rather that violations of relational motivations would underlie perceptions of moral wrongness of harmful or impure behaviors. To test this, they compared the relative utility of harm, impurity, and relational motive

violation in predicting ratings of perceived immorality of various acts. Their first step was to develop scenarios that would theoretically violate one of the four RM principles (moral motivations) while also exhibiting either harmful or impure content. For example, the action "kissing passionately" was adapted to the four relational models situated in different scenarios: two siblings kissing each other passionately (Communal Sharing context), a boss and a subordinate kissing each other passionately (Authority Ranking context), two colleagues kissing each other passionately (Equality Matching context), and a salesperson and customer kissing each other passionately (Market Pricing context).

They chose sixteen such scenarios based on participant ratings of the extent to which each moral motive was violated by each scenario in a pilot study (see Table 1 for a sample of 8 of these scenarios, taken from Appendix C of Tepe and Aydınlı-Karakulak, 2018). In the main study, participants were asked to rate each scenario in response to the question, "Is this behavior morally wrong?" (1 = No, not at all wrong to 7 = Yes, definitely wrong). A regression analysis was conducted with RM violation, harmlessness violation, purity violation, and participant country of origin (Turkey or United States) entered as predictors in the first step, and all possible two-way interactions entered in the second step. They found that moral wrongness was predicted by RM violation ($\beta = 0.66$, $t(256) = 10.39$, $p < .001$) and harmfulness scores ($\beta = 0.15$, $t(256) = 2.36$, $p = .02$), [$R2 = .62$ (95% CI [.54, .68]), $F(4, 256) = 102.24$, $p < .001$], with no significant two-way interactions.

Table 1
Eight selected scenarios from Tepe and Aydınlı-Karakulak (2018)

Violations of Relational Motivation	*Harmful Scenarios*	*Impure Scenarios*
Communal Sharing/ Unity violation	An older brother comes home drunk and beats his sister.	An older brother comes home drunk and passionately kisses his sister.
Communal Sharing/ Unity violation	Someone makes fun of his/her overweight friend's appearance in public	Someone sends obscene messages to an overweight friend about his/her appearance.
Authority Ranking/ Hierarchy violation	Unlike everyone else, someone refuses to stand up when a judge enters a courtroom and throws a stone at him/her.	Unlike everyone else, someone refuses to stand up when a judge enters a courtroom and urinates in front of him/her.
Authority Ranking/ Hierarchy violation	A high school student talks with his/her teacher in an overly familiar way, stands up and pushes the teacher.	A high school student talks with his/her teacher in an overly familiar way, and pays sexual compliments to the teacher.
Equality Matching/ Equality violation	Someone is beaten up while walking outside because he/she is homosexual.	Someone receives offers for oral sex while walking outside because he/she is homosexual.
Equality Matching/ Equality violation	Women being paid less compared to men.	Women being exposed (more often) to obscene dialogue in the workplace (compared to men).
Market Pricing/ Proportionality violation	A company employee not receiving his/her monthly salary, because he / she did not make enough sales.	A company employee being made responsible for the toilets, because he / she did not make enough sales.
Market Pricing/ Proportionality violation	Cutting off the hand of someone who steals.	Making someone who steals walk around naked in public.

In a subsequent study, they developed a new set of 13 RM scenarios based on pilot testing to see whether the predictive power of harmfulness on judgments of moral wrongness would be affected by RM violation, expecting that its predictive utility would decrease when the degree of RM violation was low. A linear regression was performed with centered scores for harmfulness and RM violation entered as predictors in the first step, and their interaction entered in the second step. The interaction was significant, with a slopes test revealing that the relationship between harmfulness and wrongness was significant when RM violation was either strong (β = 0.30, 95% CI [0.10, 0.50], $t(121)= 3.02$, $p < .01$) or moderate (β = 0.20, 95% CI [0.01, 0.38], $t(121)= 2.13$, $p = .04$), but non-significant when RM violation was low (β = 0.10, 95% CI [- 0.12, 0.31], $t(121) = 0.87$, $p = .38$).

Several interesting findings emerged from their experiments. For example, they found that violations of the hierarchy motive fell into two distinct categories: top-down violations (committed by a superior) and bottom-up violations (committed by a subordinate). While both categories could be seen violations of the same relational motive, participants in their studies responded differently depending on the 'direction' of the action. This suggests that asymmetrical relationship dyads (in terms of power, for example) may need to be theorized in a more specific manner.

Another finding was that the proportionality motive could be seen as vague or technical, and hard to distinguish from other motives. Participants generally saw proportionality violations as violations of equality or as an abuse of power (i.e., violations of top-down hierarchy). For example, proportionality violations in the workplace were attributed to an exploitive management style (a top-down violation of authority), whereas disproportionate resource distributions were often seen as unequal treatment (a violation of equality). Hence, they suggest that the proportionality motive may need to be reconsidered or at least more clearly differentiated from other RMs.

Finally, like Simpson et al. (2016), each relationship dyad was chosen to represent a single relational model and associated motive violation in an all-or-none fashion, when many real-life relationships may involve different motivations to different degrees: for example, a parent-child relationship may involve both Unity or provision-of-care motivations as well as Hierarchy motivations, especially when the child is relatively young, and even Equality motivations when the parent and child are on equal footing in a given context, for example, when playing a game together (Bugental, 2000). Also consistent with Simpson et al. (2016), only judgments of moral wrongness were assessed, with no assessments made of moral praiseworthiness.

From Types to Functions

While Rai and Fiske (2011), Simpson et al. (2016), and Tepe and Aydınlı-Karakulak (2018) – among others – have rightly drawn attention to relational context in moral judgment, there are some important weaknesses in the theoretical foundation for that work that must be addressed to move this area of research forward. Of particular concern is the relatively descriptive, typological nature of RRT, which does not allow for clear explanations of moral judgments or behavior because it gives insufficient attention to the *functions* served by social relationships of various kinds. In line with this more functional approach, which we use as the basis for our own work, Bugental (2000) has theorized five domains of social life, understood as algorithms—that is, effective procedures for solving particular problems (in this case, the recurrent problems of social coordination likely faced by our ancestors over time).

Drawing together evidence from cognitive, social, biological, evolutionary, and developmental psychology, Bugental (2000) takes care to articulate the specific adaptive problems to be solved, the characteristic information an algorithm would need to solve those problems, the approximate timing of the emergence or development of associated capacities within a lifespan, the psychological processes involved in exercising those capacities (primarily if-then contingencies), and the key neurohormonal regulators of the relevant emotions and behaviors. As she notes, her model differs from the taxonomy proposed by Fiske (1992) in several ways, including in its focus on the processes by which a given algorithm is acquired and the relevant biological mechanisms mediating its function (Bugental, 2000, p. 192). In pursuing this focus, Bugental identified a different relational structure to that captured by the RRT, such that different divisions had to be drawn between postulated functions or domains in order to account for the totality of the data.

The finished Bugental (2000) model includes an *Attachment* function (for safety maintenance and the promotion of fundamental welfare needs, characterized by co-evolved careseeking and caregiving roles); a *Mating* function (for acquiring and maintaining a sexual partner, characterized by mate guarding behavior and pair-bonding); a *Reciprocity* function (for maximizing joint outcomes among functional equals, characterized by conscious or unconscious recordkeeping and tit-for-tat exchanges); a *Hierarchical* function (for optimizing welfare and the balance of control between those of unequal power or status, characterized by dominant and subordinate roles); and a *Coalitional* function (for acquiring and defending shared resources and territory especially in the face of out-group threats, characterized by norms of coordination and conformity and the establishment of a group-level identity).

This model differs from the RRT in several ways. For example, whereas the Communal Sharing component of the RRT model collapses care and attachment functions, on the one hand, and coalitional functions, on the other hand, in the Bugental model, these functions are kept distinct. Moreover, the RRT model does not include a mating function, despite the centrality of this function to reproductive success, whereas this is a core function of the Bugental model.

Unlike the RRT, Bugental's (2000) model also does not propose a unified moral motivation for each relationship domain or function. But this is only appropriate. Pre-theoretically, one should not expect that moral judgments and emotions would have a one-to-one, abstract relationship with such domains. Rather, one should expect that the relationship between these factors would depend on the particular combination of social-relational functions that are in play in a given situation, and on how these functions are distributed among the various parties to the

interaction (as in the example of a parent and child playing a game, which would contextually trigger an equality-based or reciprocal function).

This suggests a plausible first step for making predictions about people's moral intuitions as a function of relational context. Specifically, it seems necessary to build a map of ways in which particular relationships or relationship dyads (parent-child, professor-student, brother-sister, customer-seller, and so on) are intuitively seen as fulfilling or ideally serving the different underlying functions described by Bugental (2000) in a given society.

In order to build such a map, of course, it would be important not to assume that a given dyad—such as professor-student—solely exhibits, or can be used to study, a single underlying relationship type or function—much less such a function conceived of as operating in an all-or-none, or present-or-absent fashion. Instead, common intuitions about the *degree* to which *each* relevant function is ideally served by a given relationship should systematically be measured, and the resulting data used to formulate predictions. This is the approach we take in this work, described in this thesis.

Conclusion

Much recent work moral psychology has framed participant judgments in terms of their apparent adherence to abstract moral principles, such as the utilitarian requirement to impartially maximize welfare or Kant's categorical imperative. Moreover, participants often are asked to make judgments about cases involving hypothetical interactions between strangers. But most of our moral judgments in the real-world concern neither abstract principles nor strangers; rather, we tend to make

concrete judgments about the behavior of those with whom we regularly interact – people we know, and with whom we stand in particular social relationships.

Recognizing this, a growing number of moral psychologists are shifting their focus to the study of moral judgment in social-relational context. In this chapter, I have given a brief overview of recent work in this vein, highlighting both strengths and weaknesses, and setting up a new 'relational norms' model of moral judgment we have developed and tested over the past few years. To situate this discussion, I also drew on influential philosophical theories of human morality that foreground relational context. I suggest that these theories, in addition to those such as utilitarianism that have shaped so much of the literature on moral psychology, should be given more attention by moral psychologists than has so far been the case.

Chapter 2

How Social Relationships Shape Moral Wrongness Judgments

Abstract

Judgments of whether an action is morally wrong depend on who is involved and the nature of their relationship. But how, when, and why social relationships shape moral judgments is not well understood. We provide evidence to address these questions, measuring cooperative expectations and moral wrongness judgments in the context of common social relationships such as romantic partners, housemates, and siblings. In a pre-registered study of 423 U.S. participants nationally representative for age, race, and gender, we show that people normatively expect different relationships to serve cooperative functions of care, hierarchy, reciprocity, and mating to varying degrees. In a second pre-registered study of 1,320 U.S. participants, these relationship-specific cooperative expectations (i.e., relational norms) enable highly precise out-of-sample predictions about the perceived moral wrongness of actions in the context of particular relationships. In this work, we show that a 'relational norms' model better predicts patterns of moral judgments across relationships than alternative models based on genetic relatedness, social closeness, or interdependence, demonstrating how the perceived morality of actions depends not only on the actions themselves, but also on the relational context in which those actions occur.

Introduction

Moral psychology has been dominated by studies of judgments and behaviors concerning strangers: individuals who stand in no particular relationship with one another, and who may or may not interact in the future (Hester & Gray, 2020). Researchers conducting such studies commonly ask participants to make decisions that impact anonymous others (Batson et al., 1997; Bowles, 2008; Crockett et al., 2014) or to judge the moral acceptability of hypothetical actions taken by thinly-described agents, as in sacrificial dilemmas where participants must judge the permissibility of killing one person in order to save a greater number (Conway et al., 2018; Greene, 2008; Kahane et al., 2015; Mikhail, 2007). Of course, people often do encounter strangers as they go about their lives, and the interpersonal standing implied by such encounters can be seen as a bare-bones social relationship involving certain minimal obligations: for example, a "duty of easy rescue" in the case of emergencies (Sterri & Moen, 2020). The copious research on moral judgments in the context of stranger-stranger relationships thus sheds important light on at least one important aspect of our moral psychology.

However, the vast majority of our moral judgments in everyday life do not concern strangers. Rather, they concern familiar others with whom we stand in particular, often ongoing relationships (Clark et al., 2015). The stakes of such moral judgments for the maintenance of our personal social networks typically are higher than the stakes of analogous judgments pertaining to strangers. Moreover, moral judgments about interactions between strangers often will differ in systematic ways from judgments about interactions between friends, family members, or other familiar individuals in the same situation (Bloom, 2011; Clark & Boothby, 2013; Ko et al., 2020). For example, consider someone who could easily feed a hungry individual but

fails to do so. If this person is a mother failing to feed her own child, she likely will be seen as highly blameworthy. But if the person is a local restaurant owner failing to feed a non-paying customer, the same behavior likely will not be seen as blameworthy under ordinary conditions (Clark et al., 2020).

A number of theorists have highlighted relational context as likely to be important for understanding moral judgment and behavior (Bloom, 2011; Clark et al., 2015; Isern-Mas & Gomila, 2020; Rai & Fiske, 2011; Schein, 2020; Tomasello, 2020). In line with these developments, there is now a small but growing empirical literature which explores how moral judgments of particular actions vary across different types of social relationships (Everett et al., 2018; Koleva et al., 2014; Lee & Holyoak, 2020; Mammen et al., in press; Marshall et al., 2020; McGraw & Tetlock, 2005; McManus et al., 2020; Rowe et al., 2020; Selterman et al., 2018; Simpson et al., 2016; Simpson & Laham, 2015; Sunar et al., 2020; Tepe & Aydınlı-Karakulak, 2019; Waytz et al., 2013; Weidman et al., 2020). How these relationships are theorized depends on the study. For example, one recent study characterized relationships in terms of the *genetic relatedness* of the interaction partners, and showed how varying this factor affects moral judgments about helping behavior (McManus et al., 2021). Another recent study characterized relationships in terms of the authors' intuitive sense of the *social closeness* and relative *interdependence* of the interaction partners – regardless of genetic relatedness – and tested the influence of these factors on judgments about violations of care (Gilead et al., 2018). Researchers have also sought to predict moral wrongness judgments of actions in relational context from a single *cooperative function* thought to characterize a given relationship (e.g., care for a sibling relationship, hierarchy for a teacher-student relationship, and so on) (Simpson et al., 2016).

These studies demonstrate that moral judgments of one and the same action often differ across different types of relationships, depending on how relationship "type" is understood. What is missing, however, is a systematic, data-driven account of the *multiple* cooperative functions that can characterize any given social relationship (Clark et al., 2020), and an explicit comparison of how well such cooperative functions predict relationally-situated moral judgments relative to alternative models such as genetic relatedness, social closeness, and interdependence. We aim to fill that gap with the present research.

In contrast to genetic relatedness, which can be determined objectively, and the constructs of social closeness and interdependence, both of which have been carefully defined within the relationship science literature, there is no agreed-upon set of cooperative functions prescribed for different social relationships to solve characteristic coordination problems. Recognizing both the theoretical overlap and diversity among the various existing taxonomies of cooperative functions (Curry et al., 2019; Haidt & Joseph, 2007; Rai & Fiske, 2011), we build on work by Bugental (2000). This work describes a distinctive set of cooperative functions that serve to coordinate behavior in interpersonal relationships. Each function represents an efficient, socially acceptable solution to a particular type of recurrent coordination problem (Curry et al., 2019), enabling cooperation partners to mutually benefit over repeated interactions (Baumard et al., 2013; Haidt & Joseph, 2007; Sznycer & Lukaszewski, 2019). We focus here on four cooperative functions that solve dyadic or two-party coordination problems: care, reciprocity, hierarchy, and mating (Table 1).

As has been noted previously, any given relationship may serve multiple cooperative functions, either characteristically or in a specific context (e.g., Bugental, 2000, p. 192; Rai & Fiske, 2011, p. 60). We propose that within a given society, there

are prescriptive norms for the set of cooperative functions different relationships should serve ('relational norms'). In the present work, we sought to (i) describe patterns of relational norms for a large set of common dyadic relationships in a U.S. cultural context; (ii) use these patterns of relational norms to predict out-of-sample judgments of moral wrongness for actions that violate those norms across relationships; and (iii) to compare this 'relational norms' model with alternative ways of characterizing dyadic relationships, i.e., in terms of genetic relatedness, social closeness, and interdependence.

Table 1
Cooperative functions of dyadic relationships, adapted from Bugental (2000)

Cooperative function	**Coordination problem to be solved**
Care	Securing basic welfare needs through non-contingent provision (or acceptance) of help or support; maintaining safety; encouraging learning
Reciprocity	Coordinating behavior between individuals with functionally similar (or equal) status, power, authority, or claim on a resource
Hierarchy	Coordinating behavior between individuals with different (unequal) status, power, authority, or claim on a resource
Mating	Finding and maintaining sexual partners; ultimately, producing and ensuring the survival of offspring

Note: the care function is based on the work of Clark and colleagues concerning "communal" relationships (e.g., Clark & Mills, 1993); it conceptually overlaps with, and replaces, the "attachment" function in Bugental's model. Because our model is focused on dyadic interactions, we also do not include Bugental's group-level "coalition" function in this table (see Appendix 1, Supplement Section 1.4.3. for data pertaining to the coalition function).

With respect to aims (i) and (ii), we predicted that relational norms would robustly predict moral judgments about the wrongness of actions in relational context. The basis for this prediction is straightforward: the more a particular set of cooperative functions matters for a given relationship, the morally worse it should be judged to be to neglect or frustrate those same functions within that relationship. Because relationships vary in terms of the set of cooperative functions they are expected to serve, a given action may be judged to be seriously wrong in the context of one relationship but entirely acceptable in the context of another.

We further predicted that our relational norms model would better explain the variance in moral judgments across relationship dyads than genetic relatedness, social closeness, and interdependence, which we believe offer incomplete predictive accounts of moral judgments in relational context. To illustrate, imagine that Person A fails to behave in a deferential manner toward Person B. Insofar as the relationship is normatively expected to be governed by the hierarchy function (see Table 1), with Person A in the subordinate position, such behavior will likely be judged as morally wrong. By contrast, consider how genetic relatedness might explain the wrongness of this action. Some genetically close relationships, such as the parent-child relationship, may, indeed, normatively rely on the hierarchy function to coordinate behavior, and to the extent they do, the action might be judged to be wrong. However, other genetically close relationships, such as siblings of a similar age, are less likely to rely on the hierarchy function, while some genetically distant relationships, such as a typical boss-employee relationship, might be equally or even more likely to rely on the function. Thus, genetic relatedness may ultimately prove to be largely *independent* of the question of what makes certain actions liable to be judged morally wrong.

In summary, unlike most prior work in moral psychology, which has been designed to predict moral judgments from features of actions *regardless* of who performs the action or their relationship to the affected other, here we consider features of common social relationships that we predict will shape moral judgments of actions that occur in the context of *specific* relational dyads. We show that the similarity between relationship dyads in terms of their prescribed cooperative functions—or relational norms—corresponds to similarity in moral judgments between relationships. Put another way, dyads with similar relational norms within a given society are associated with similar patterns of moral judgments across actions,

whereas dyads with dissimilar relational norms are associated with divergent patterns of moral judgments across actions. Finally, we show that relational norms more strongly predict patterns of moral judgments across relationships than alternative predictors, including genetic similarity, social closeness, or interdependence.

Results.

Relational norms vary across common dyadic relationships. We first measured relationship-specific patterns for prescribed cooperative functions (i.e., relational norms) for a set of common dyadic relationships in the U.S. (study design, sampling plan, and exclusion criteria pre-registered at https://aspredicted.org/). Participants (final n = 423, U.S. nationally representative for age, race, and gender; "Sample 1") rated 20 common dyads on the extent to which each is normatively expected to serve the functions of care, reciprocity, hierarchy, mating, and coalition. Results for all 20 relationships across the four functions from Table 1 are depicted in Figure 1 (for coalition data, see Appendix 1, Supplement Section 1.4.3.).

As can be seen in Figure 1, relational norms varied markedly across dyads in several respects. The reciprocity function was generally prescribed for most dyads (M across dyads = 54.23, SD = 49.64; higher than scale midpoint with Bonferroni corrected alpha = .0125, $t(1,422) = 100.47$, $p < .001$, $d = 1.09$; note that all tests reported in the manuscript are two-sided). Meanwhile, the mating function was *negatively* prescribed (i.e., proscribed) for most dyads (M across dyads = -63.02, SD = 62.01; lower than the scale midpoint with same correction, $t(1,422) = -93.48$, $p < .001$, $d = -1.02$), with a few obvious exceptions (romantic partners, M = 95.12, SD = 12.94; friends-with-benefits, M = 58.43, SD = 51.21). Participants demonstrated higher levels of agreement about whether dyads were expected to serve the mating (SD_{mean} across relationships = 32.26) and care (SD_{mean} across dyads = 37.82)

functions, relative to the reciprocity (SD_{mean} across dyads = 42.25) and hierarchy (SD_{mean} across dyads = 53.72) functions.

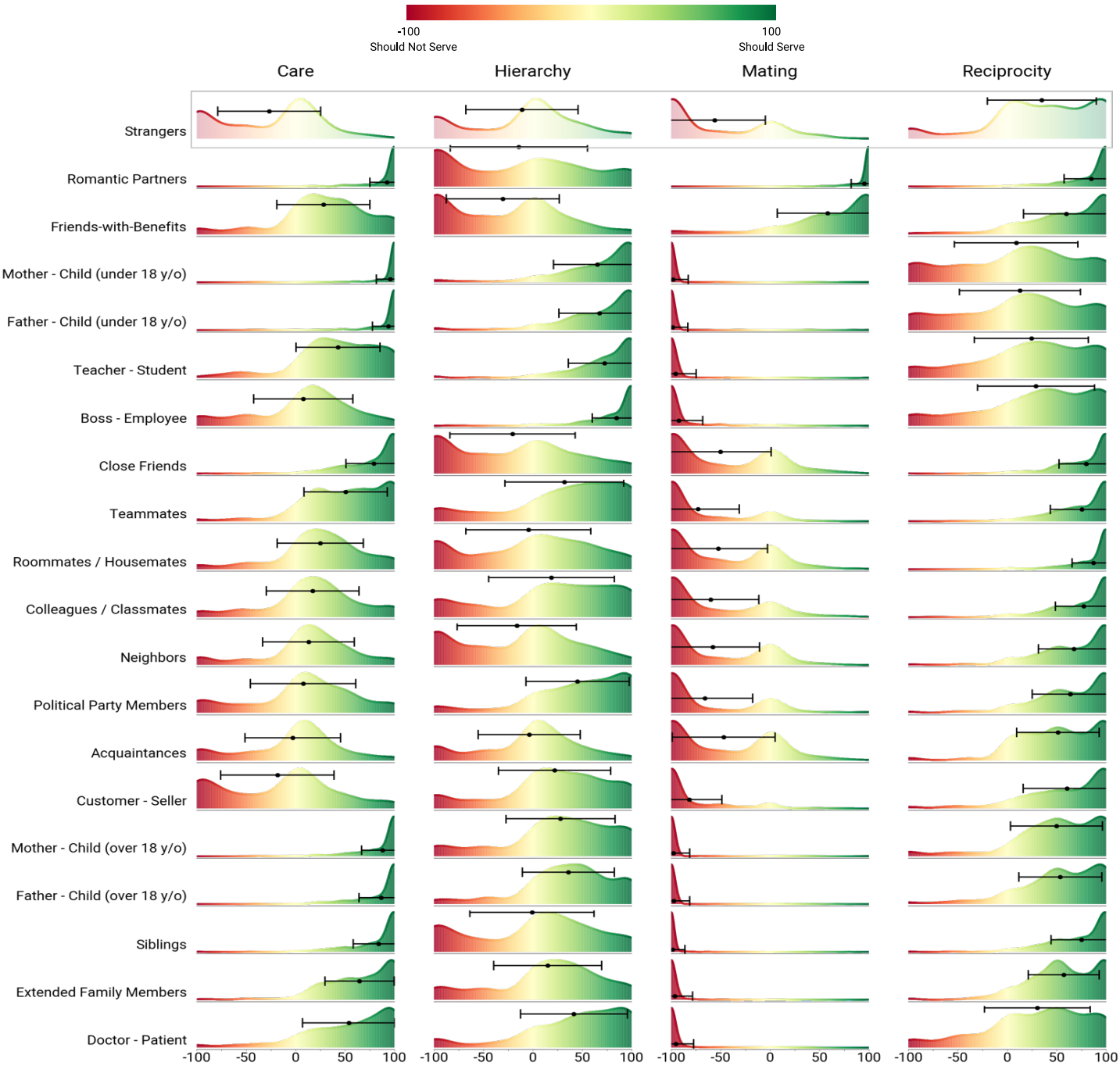


Figure 1. Prescribed cooperative functions. Kernel density plots of prescribed cooperative functions for 20 common relationship dyads. Dots represent the population mean prescription for each cooperative function within each relationship, caps represent +/- one standard deviation. The height of the curve represents density: the likely proportions of scores (relative to each function) that fall within the given range along the x-axis.

Figure 2 depicts the four-dimensional relational norm profiles (i.e., sets of prescribed cooperative functions) for a subset of relationships studied, and illustrates several additional features of our data (see Appendix 1, Supplement Section 1.4.4. for functional profile plots for all 20 relationships). First, some relationships are highly functionally "polarized," showing substantial deviation in mean prescriptions across the four cooperative functions, with at least one function anchored at an extreme end of the scale. An example is the parent and under-18 child relationship (*SD* across cooperative functions = 85.33, 85.2 for the mother-child and father-child relationship, respectively), which is characterized by a strongly positive expectation for care and a strongly negative expectation for mating. By contrast, other relationships are less functionally polarized, such as the relationship between strangers (*SD* across functions = 37.93). In these relationships, prescribed cooperative functions are relatively evenly spread across the measured spectrum.

Second, some relationships are functionally "specific," that is, they are only strongly expected to serve a single cooperative function. For example, the roommate/housemate relationship is strongly expected to serve the reciprocity function (*M* = 87.30, *SD* = 21.71), but less so the care (*M* =24.9, *SD* = 43.64), hierarchy (*M* = -4.48, *SD* = 63.00), and mating functions (*M* = -52.39, *SD* = 49.85). Similarly, the boss-employee relationship is strongly expected to serve the hierarchy function (*M* = 84.75, *SD* = 24.68), but less so reciprocity (*M* = 29.14, *SD* = 58.93) or care *(M* = 7.86, *SD* = 50.21), and not at all mating (*M* = -92.17, *SD* = 23.98). By contrast, other relationships are functionally "pluralistic," that is, they are strongly expected to serve multiple cooperative functions. A key example is the romantic partner relationship (*M* across functions = 64.61), which is strongly expected to serve three of the four cooperative functions: care (*M* = 92.43, *SD* = 17.06), mating (*M* =

95.12, *SD* = 12.92), and reciprocity (*M* = 84.95, *SD* = 27.28) but not, in this sample from the United States, the hierarchy function. See Appendix 1, Supplement Sections 1.4.1. and 1.4.2. for the rankings of all 20 relationship dyads on the dimensions of polarization and specificity.

We also find gender differences in prescribed cooperative functions across relationships. After scaling the raw scores to each participant's mean rating, we built a mixed linear effects regression model controlling for relevant demographic information (age, income, religiosity, and political orientation entered as fixed effects), including participant and relationship dyad type as random effects. With a Bonferroni correction (alpha = .0125 for each of the following effects), the model revealed that women (*M* = 0.43, *SD* = 0.75), compared to men (*M* = 0.37, *SD* = 0.79), reported stronger average expectations that relationships will serve a function of care ($p < .001$, 95% CI [.03, .10]), consistent with the existing literature (43–46). This divergence was most apparent for the roommate/housemate ($M_{female} - M_{male} = 0.17$), customer-seller ($M_f - M_m = 0.15$), teacher-student ($M_f - M_m = 0.14$), neighbor ($M_f - M_m = 0.14$), and colleague/classmate ($M_f - M_m = 0.13$) relationships. Regarding mating, the opposite pattern was found, also consistent with the existing literature (Jonason et al., 2015; Mark et al., 2015). Men (*M* = -1.12, *SD* = 0.97), compared to women (*M* = -1.2, *SD* = 0.89), reported stronger average expectations that relationships will serve a mating function ($p < .001$, 95% CI [.03, .11]). This divergence was most apparent for the friends-with-benefits ($M_m - M_f = 0.27$), roommate/housemate ($M_m - M_f = 0.25$), acquaintance ($M_{male} - M_{female} = 0.24$), close friend ($M_m - M_f = 0.24$), colleague/classmate ($M_m - M_f = 0.21$), stranger ($M_m - M_f = 0.17$), and neighbor ($M_m - M_f = 0.17$) relationships. Additional demographic analyses are reported in the Appendix 1 Supplement Section 1.4.5.

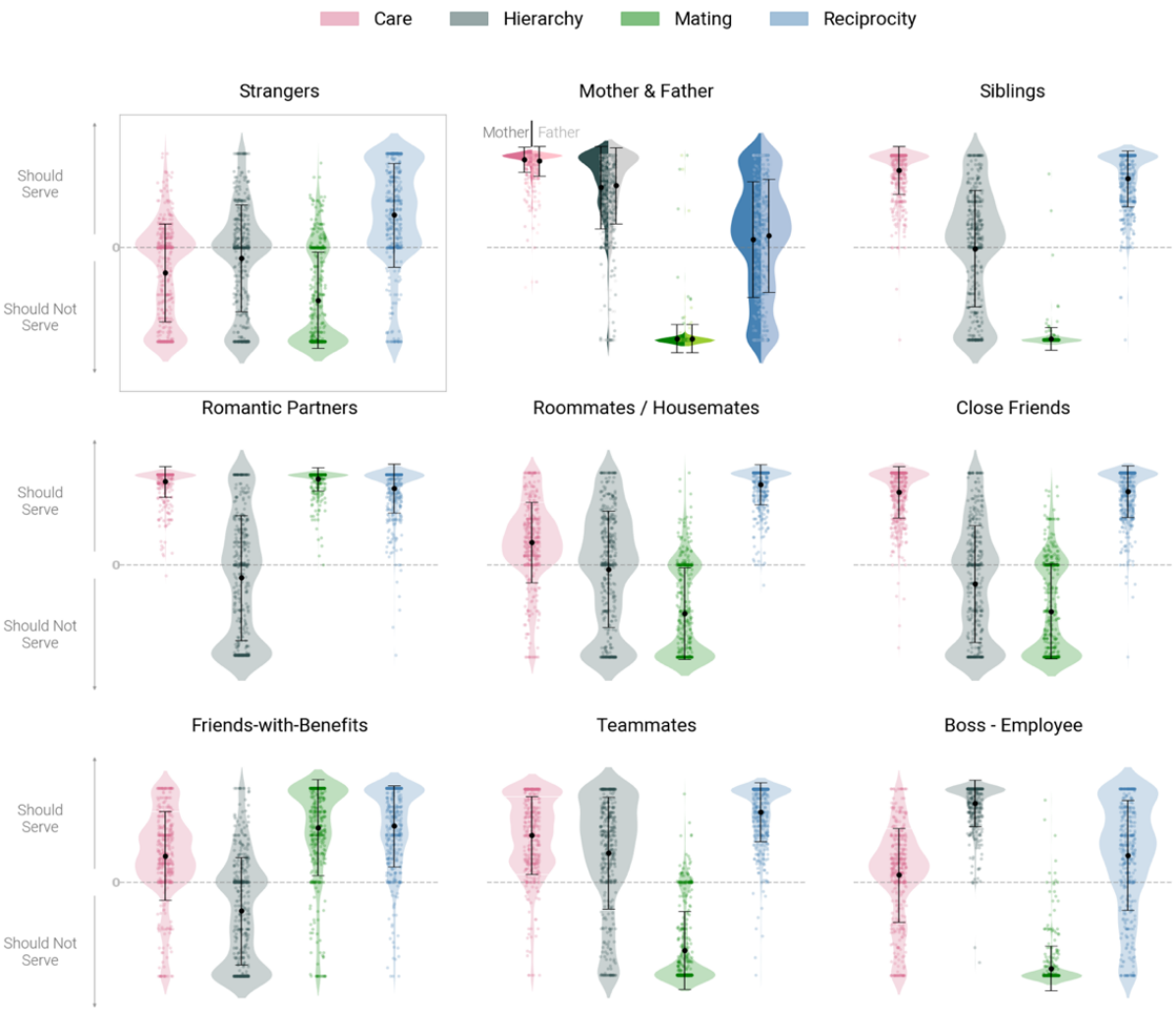


Figure 2. Relational norm profiles for a subset of 10 relationships. Pink represents care, black represents hierarchy, green represents mating, blue represents reciprocity. The raw data (n = 423 independent ratings per function per relationship; total n = 8,460) are shown in individual dots; error bars represent the mean (dot) and +/- 1 SD (caps). Note: Mother/Father and under-18 child have been combined into a single plot. Plots for all 20 relationships are in the Appendix 1, Supplement Section 1.4.4.

Common relationships are hierarchically clustered around relational norms.

Next, we sought to quantify the distinctiveness of each relationship in four-dimensional relational norm space. Because in many instances patterns of prescribed cooperative functions were not normally distributed in our study population (see Figure 1), characterizing relationship differences in terms of their average relational norm scores would sacrifice considerable information. We therefore calculated the Kolmogorov-Smirnov (K-S) distance statistic (a quantification of the difference in

overall shape between any two empirical distributions) for each cooperative function for each possible pair of relationships, and averaged across functions to calculate the overall dissimilarity in relational norms for each relationship pair. This approach is conceptually similar to representational similarity analysis (Kriegeskorte et al., 2008), but incorporates information about the shapes of the relational norm distributions in addition to distribution means.

We used the relational norm dissimilarity values to conduct a hierarchical clustering analysis using a farthest-point algorithm: $d(u,v)=\max(\text{dist}(u[i],v[j]))$ (Voorhees, 1985). This revealed four main clusters, depicted in Figures 3a and 3b, which align with intuitive relational categories. The first cluster consists of sexual relationships (romantic partners and friends-with-benefits). The second cluster consists of hierarchical relationships with highly unequal authority between individuals (parents and their minor children, teacher-student, boss-employee). The third cluster includes relationships characterized largely by reciprocal interactions between equals (e.g., customer-seller, roommates/housemates, strangers). And the fourth, final cluster includes familial or other caring relationships (e.g., siblings, extended family members, parents and their adult children).

Based on these analyses, we identified a subset of 10 relationships with relatively distinctive relational norms (see Appendix 1, Supplement Section 2.1. for the selection procedure). This subset included long-term romantic partners, friends with benefits, boss and employee, colleagues or classmates, mother/father and under-18 child, siblings, close friends, roommates or housemates, teammates, and strangers. Relational norm profiles for these relationships are depicted in Figure 2. We next sought to predict moral judgments of actions performed in the context of these relationships on the basis of their relational norm profiles.

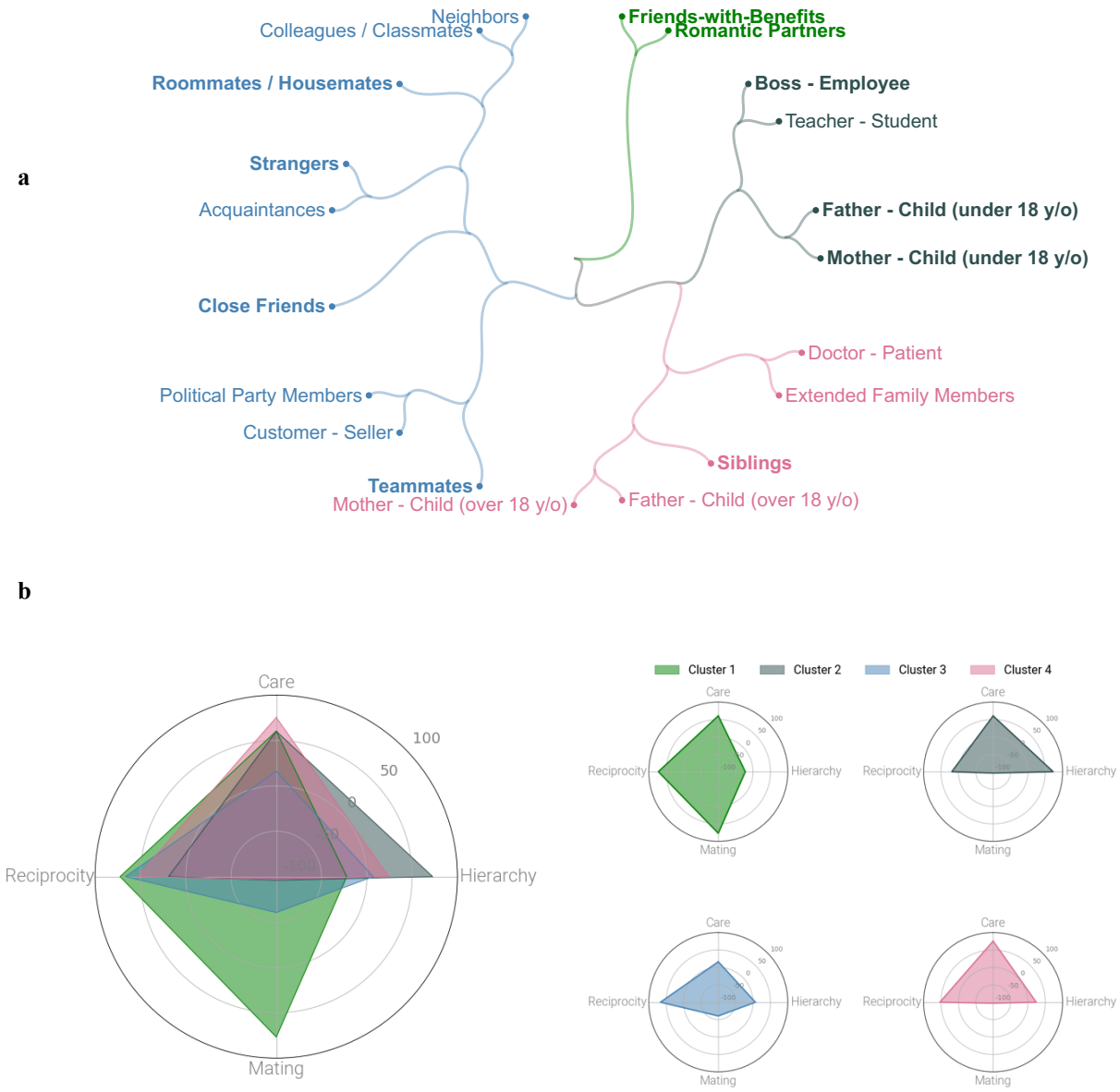


Figure 3. Hierarchical clustering of relationships. Circular dendrogram visually representing the mean Kolmogorov-Smirnov (K-S) distance between relationships in four-dimensional relational norm space, clustered hierarchically according to the Voorhees (1985) method (**a**); relationships selected for Study 2 are highlighted in a darker shade. Radar plots derived from the hierarchical cluster model are depicted in the bottom half of the figure (**b**). The left panel shows the overlapping clusters; the right panel shows each cluster on its own set of axes. Source data are provided as a Source Data file.

Relational norm profiles predict relationship-specific moral judgments out of sample. To test the hypothesis that relational norm profiles would predict patterns of moral judgments across common relationships, we first assembled a set of common behaviors that would plausibly weaken or violate one or more of the cooperative

functions. Fifteen trained judges rated 86 action statements of the form "Person A does X to Person B" on the extent to which each described action would characteristically weaken (that is, violate or impair) or strengthen each of the cooperative functions, *setting moral questions aside* (that is, they were instructed not to think about whether an action might be right or wrong in any relationship, but only whether it would weaken or strengthen each function). There was very high interrater agreement in these ratings (ICC(3, k) = .97). Using these data, we selected a final set of 12 characteristic function-violating action statements, with 3 statements for each of the 4 dyadic functions (see Methods for the algorithm used to select the final sub-set). See **Figure 4.**

As can be seen in Figure 4, each action was rated by the judges as having both a main (i.e., "target") effect on a given function, as well as "side effects" on the other cooperative functions. For example, "Person A sees Person B crying and walks away from them" was rated as most characteristic in weakening the care function (M = -87.9, SD = 15.5), but also was rated as characteristically weakening the mating function, albeit to a lesser extent (M = -40.1, SD = 35.0). The fact that one and the same action might simultaneously weaken several interpersonal functions is to be expected, depending on the "logic" of each function and the nature of the action. To account, then, for the specificity of each action as a function-violator, we computed a "target specificity" variable (i.e., main effect minus the mean of side effects) for each action for use in subsequent analyses.

	Actions judged to characteristically weaken one or more cooperative functions	Care Mean (SD)	Hierarchy Mean (SD)	Mating Mean (SD)	Reciprocity Mean (SD)
1	Person A sees Person B crying and walks away from them	-87.9 (15.5)	-15.9 (24.3)	-40.1 (35.0)	-19.0 (32.3)
2	Person A keeps checking their cellphone while Person B tells a sad personal story	-75.1 (27.8)	-22.1 (29.6)	-36.5 (39.1)	-35 (35.1)
3	Person A watches passively while Person B carries several heavy boxes up the stairs, even though they could easily help	-73.9 (28.6)	-27.9 (31.3)	-17.4 (29.4)	-29.1 (28.6)
4	Person A refuses to follow a reasonable order from Person B	-22.1 (29.5)	-89.5 (17.7)	-10.7 (13.1)	-16.9 (24.6)
5	Person A repeatedly interrupts Person B while they are speaking	-42.2 (28.5)	-71.6 27.7)	-15.9 (20.3)	-50.3 (37.0)
6	Person A decides to skip a meeting scheduled with Person B without a good excuse	-42.5 (34.7)	-69.8 (30.8)	-21.3 (31.0)	-38.4 (32.4)
7	Person A refuses to have sex with Person B	-9.3 (17.7)	-10.7 (32.6)	-95.3 (10.7)	0.1 (2.1)
8	Person A repeatedly turns down Person B's offer to go on a romantic date	-10.5 (20.2)	-15.3 (19.4)	-77.7 (23.3)	-9.9 (34.3)
9	Person A invests time and energy in a romantic relationship with someone other than Person B	-26.5 (34.8)	-4.5 (27.6)	-74.0 (28.8)	-31.9 (35.8)
10	Person A decides not to pay Person B back, hoping Person B won't remember	-37.1 (31.3)	-27.0 (33.9)	-12.4 (17.8)	-85.2 (26.7)
11	Person A decides not to return Person B's nice favor	-35.5 (29.9)	-9.2 (24.1)	-22.7 (21.6)	-82.0 (21.8)
12	Person A charges Person B $50 for an item worth $25	-34.9 (27.9)	-28.1 (31.2)	-13.4 (23.6)	-69.4 (30.7)

Figure 4. Characteristic function-weakening actions. Heatmap showing mean ratings of judges (n = 15) of the extent to which each action would characteristically neglect or weaken the care, hierarchy, mating, and reciprocity functions, respectively, between any two people (i.e., not stating whether the relationship between "Person A" and "Person B" should in fact serve any of those functions). These items were chosen as experimental stimuli from a much larger set by an algorithm using the judges' ratings, where -100 represents the most characteristic function-weakening effect (see Methods). Darker shades represent more extreme ratings. Note: when rating actions on the "hierarchy" dimension, judges were asked to imagine that Person A was in a *subordinate* role, specifically; when rating actions on the "care" dimension, judges were asked to imagine that Person A was in a *caregiving* (as opposed to care-seeking) role, specifically.

Having identified a set of actions, drawn from everyday life, that were judged to characteristically weaken or violate one or more prescribed cooperative functions,

our next step was to assess moral judgments concerning those actions in the context of specific relationships. To do this, we recruited a new group of participants (online US convenience sample, final n = 1,320; "Sample 2"), after pre-registering our hypothesis, study design, sampling plan, exclusion criteria, and analysis approach at https://aspredicted.org/blind.php?x=ta5yj5. These "naïve" Sample 2 participants were given no information about cooperative functions. Rather, each participant was assigned randomly to consider 1 of the 10 functionally distinctive relationships identified above, and was asked to rate the moral wrongness of all 12 actions listed in Figure 4 in the context of that relationship (e.g., "Imagine that an employee refuses to follow a reasonable order from their boss. How morally wrong would that be, if at all?"). We also asked participants to rate each action on how *likely* it would be to occur in real life, in order to be able to control for violations of nonmoral (i.e., social-conventional) expectations (Turiel, 2008) in a pre-registered secondary analysis (see "action likelihood" variable below). For each participant, we computed the mean moral wrongness rating for each of the four cooperative function-violation categories within their assigned dyad. See Figure 5 for distributions of moral wrongness ratings for each function-violation for each relationship (for demographic analyses, see the Appendix 1 Supplement Section 2.5.1.).

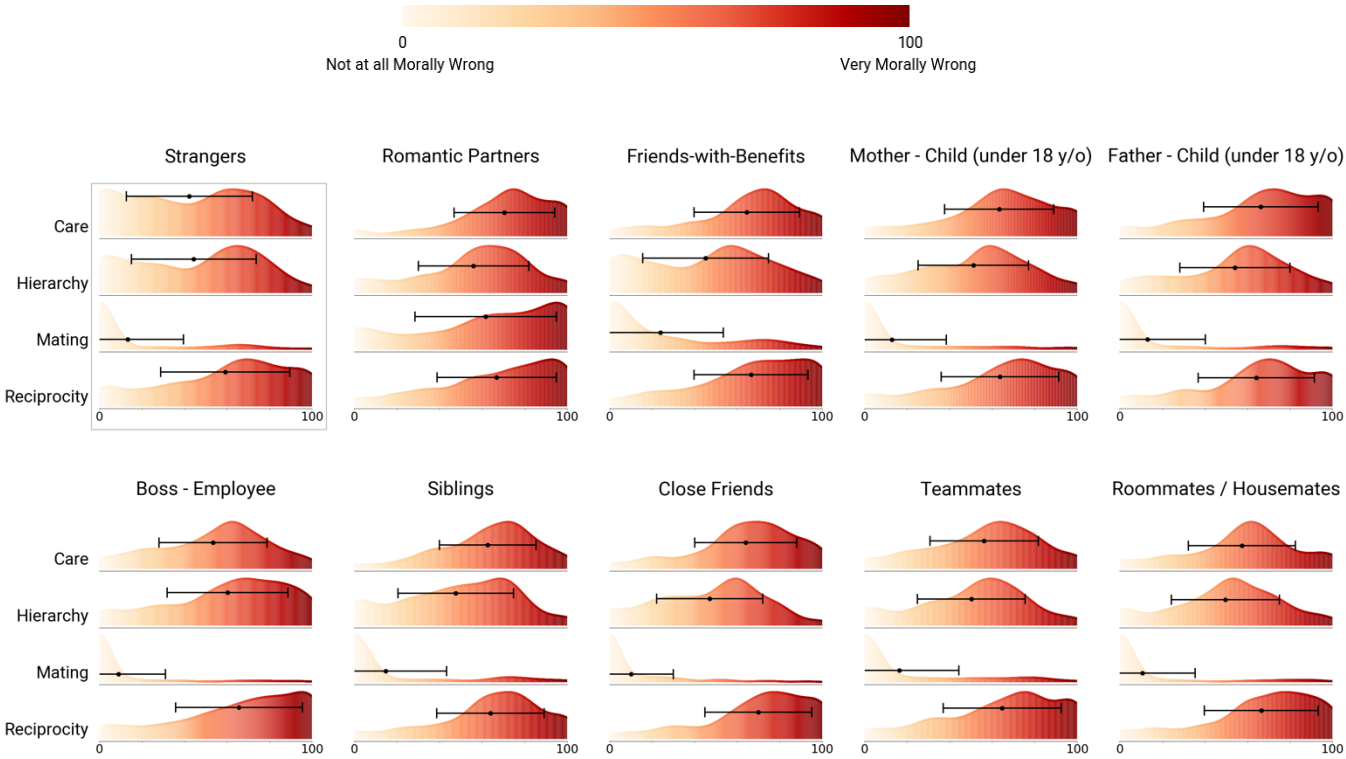


Figure 5. Moral wrongness judgments: Sample 2 moral wrongness judgments for cooperative function violations in different relationships: kernel density plot of wrongness judgments (0 = not at all morally wrong, 100 = very morally wrong) concerning characteristic function-weakening actions for each of four dyadic cooperative functions across 10 relationships. Dot represents the mean, with 95% confidence intervals. Height of the curve represents density (see **Figure 1** for explanation). This experiment was conducted once, with all data shown here. Note that actions which *weaken* the mating function (e.g., refusing to have sex with someone) were judged closer to "not at all wrong" than "very wrong" for all dyads apart from the romantic partner relationship. Otherwise, the relative lack of visually dramatic differences in the shape of the moral wrongness judgment distributions between relationships can likely be explained by the mild or 'everyday' nature of the function-weakening actions employed in this study (see **Figure 4**). Such actions were deliberately chosen to contrast with the more extreme, unusual, or bizarre actions often studied in moral psychology; thus the ability of our model to predict even subtle variance in moral wrongness judgments between relationships for common, non-extreme actions (see analysis below) can be seen as a strength.

We turn now to our main, pre-registered hypothesis. As a first approach, we sought to predict Sample 2 moral wrongness judgments (i.e., for violating each of the four cooperative functions) directly from Sample 1 relational norm profiles in a linear mixed regression model. Sample 2 participants were entered as the highest-level grouping variable, with relationship dyad and function-violation type then entered as

crossed random factors. This variance structure accounts for the fact that for each relationship a judgment was made for every function-violation type (i.e., a crossed design). The mean relational norm estimates from Sample 1 were entered alongside both "action likelihood" and "target specificity" as continuous fixed factors for the reasons given above.

The results from this model supported our hypothesis. Relational norms derived from Sample 1 significantly predicted the moral wrongness judgments of Sample 2 participants ($p < .001$, 95% CI [15.63, 16.88]), accounting for 63% of the variance in mean moral wrongness judgments according to an R^2 analysis (Nakagawa & Parker, 2015). Breaking the model down further, we find that target specificity was positively correlated with moral wrongness judgments ($p < .001$, 95% CI [.34, .40]), indicating that the more "on-target" the effect of an action in violating a given function, the more harshly that action was judged. Action likelihood was also negatively correlated with moral wrongness judgments ($p < .001$, 95% CI [-.21, -.18]), indicating that rarer actions were judged more harshly, consistent with past research (Lindström et al., 2018). These results are robust when controlling for demographic factors. For the full regression tables, see Appendix 1, Supplement Section 2.5.2.

The "action likelihood" variable serves an additional, theoretically important purpose. As we alluded to previously, it can help account for the variance in moral judgments that is due to potentially non-moral violations of social-conventional expectations (i.e., deviations from what is socially expected, whether or not the expectation tracks a perceived moral obligation) (Turiel, 2008) as opposed to violations of relational norms specifically. By comparing the R^2 effect size estimates and AIC goodness-of-fit scores (i.e., of relational norm versus action likelihood

models) we can judge the relative impact of each metric in explaining moral judgments across relationships. We find that, in a model with no information about relational norms, action likelihood alone does significantly predict moral wrongness judgments in the absence of other predictors ($p < .001$). However, this model explains much less variance, with a poorer goodness-of-fit score (marginal $R^2 = .08$, AIC = 136,496.9) than a model based only on relational norms (marginal $R^2 = .30$, AIC = 130,804). Moreover, the beta value for relational norms (16.26) is more than 80 times larger than that for the action likelihood ratings (-.20) when both are included in the same model (see Appendix 1, Supplementary Table #11f in Supplement Section 2.5.2.). This shows that relational norms explain moral judgments in this study far better than do merely conventional norms regarding what is socially expected.

Having confirmed that relational norms predict between-relationship variation in moral judgments, over and above mere uncommonness or unexpectedness of behavior, we sought to further explore the nature of this predictive relationship. Specifically, we sought to predict the distance between each pair of relationships in moral judgment space (based on Sample 2 patterns of moral judgment) from their corresponding distances in four-dimensional relational norm space (from Sample 1). To do this, we relied on the same K-S distance approach as described above, comparing the moral judgment distributions for each type of function violation for each possible pair of relationships, and averaging across functions to produce an overall moral judgment dissimilarity score for each relationship pair. We then ran a Spearman's correlation between these moral judgment dissimilarity values and the previously computed relational norm dissimilarity values, hypothesizing that the average K-S distance between every pair of relationships in relational norm space would predict the corresponding K-S distance between the same pairs of relationships

in moral judgment space. As can be seen in Figure 6, this hypothesis was confirmed (r = .43, p = .003). Looking at the same K-S distances, but on a function-by-function basis (see Appendix 1, Supplement Section 2.5.3. for the corresponding scatterplots), we find that the positive correlation between prescribed cooperative functions and moral judgment K-S distances holds for care (r = .48, p < .001), mating (r = .73, p < .001), and hierarchy (r = .31, p = .041), but not for reciprocity (r = -.13, p = .39). We will return to this unexpected result for reciprocity in the general discussion.

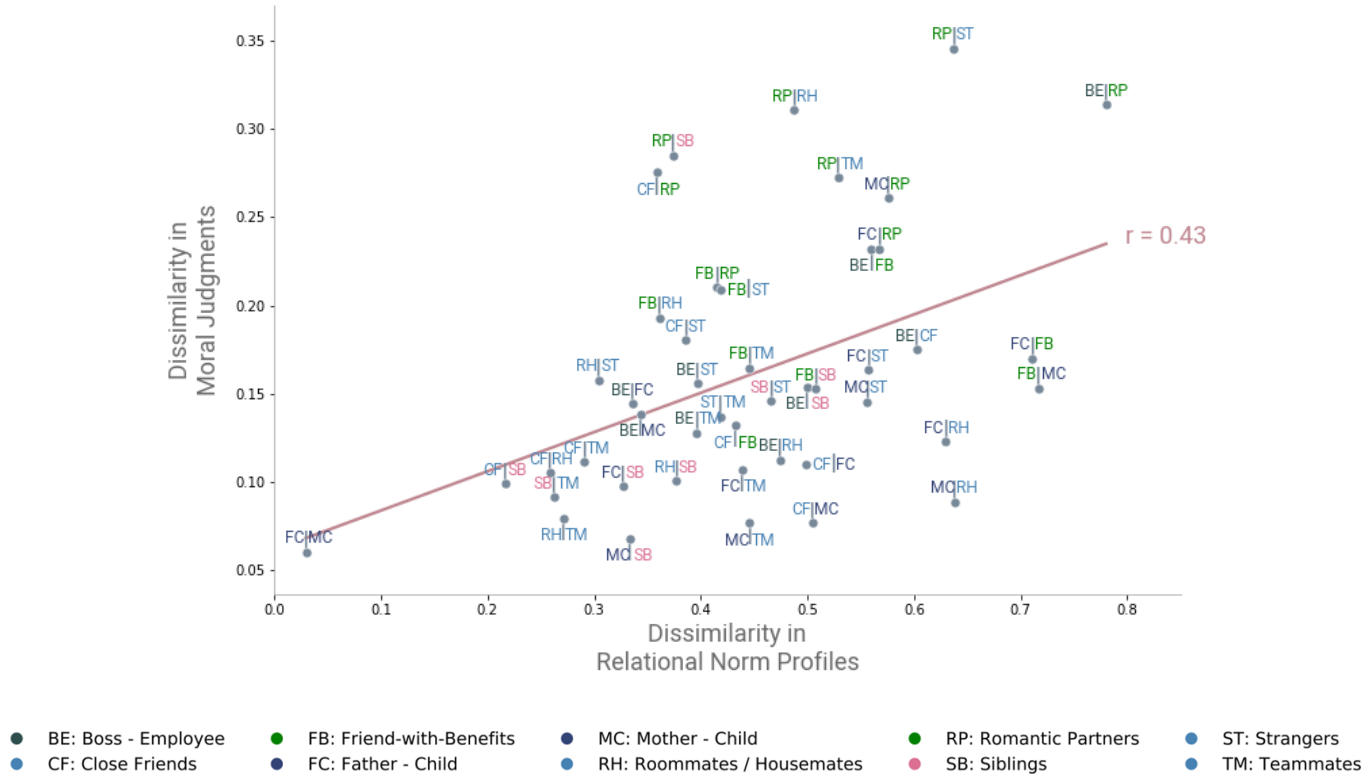


Figure 6. Relational norm and moral judgment dissimilarity. Scatterplot showing the predicted correlation in K-S distance between each pair of relationship dyads in relational norm space (x-axis) and the K-S distance between those same dyads in moral judgment space (y-axis). Spearman's r = .43, p = .003. Note that the color of each relationship reflects the cluster in which it is located from Figure 3.

Relational norms explain more variance in moral judgments than alternative models. Prior work has sought to predict relational variance in moral judgments from factors such as genetic relatedness (Burnstein et al., 1994), social closeness (Mills et al., 2004), and degree of interdependence (Berscheid et al., 1989)

of the interaction partners. How do these alternative predictors compare to relational norms in terms of explaining variance in relationally situated moral judgments?

To address this question, we asked a third sample of participants (online US convenience sample; final n = 85) to rate the extent to which a well-functioning instance of each of the 10 distinctive relationship dyads would be characterized by social closeness and interdependence. More specifically, for social closeness, we asked how much the partners would "deeply understand each other," "accept and validate each other's natures," and "strive to care for and promote each other's overall well-being" (taking the mean of these three items). For interdependence, we asked how "frequently," "strongly," and in "how many ways" each partner would affect the other's thoughts, feelings, and behaviors across different situations (ditto). Genetic relatedness for each relationship was determined objectively.

We then entered genetic relatedness, social closeness, and interdependence ratings as predictors in separate linear mixed models similar to those described previously, regressing moral judgments on relational norms. We found no relationship between mean moral wrongness judgments and any of social closeness (p = .11, 95% CI [-.04, .39]), interdependence (p = .14, 95% CI [-.05, .38]), or genetic relatedness (p = .78, 95% CI [-12.41, 9.27]). By contrast, relational norms remained significantly predictive of moral wrongness judgments (p < .001, 95% CI [.10, .15]), even controlling for the other factors. In addition, measures of model fit suggest that the relational norms model (marginal R^2 = .69, AIC = 841) performed substantially better than any of the alternative models: social closeness (marginal R^2 = .44, AIC = 908.04), interdependence (marginal R^2 = .44, AIC = 908.00), and genetic relatedness (marginal R^2 = .44, AIC = 910.33). See Appendix 1, Supplement Section 3.4. for the full regression tables.

Discussion.

Several scholars have stressed the importance of taking relational context into account in understanding our moral psychology. Yet the way that relational context has been conceptualized so far has suffered from certain limitations. Most commonly, relational context has been understood to vary in a one-dimensional way: for example in terms of the genetic relatedness of the interaction partners (McManus et al., 2021), or their social closeness or interdependence (Gilead et al., 2018). A more promising approach, we think, is to conceptualize relationships in terms of the distinctive cooperative functions they are normatively expected to rely upon for coordinating behavior in a given society (Bugental, 2000; Clark et al., 2020; Simpson et al., 2016). Although a number of authors have proposed various taxonomies of cooperative functions (Curry et al., 2019; Haidt & Joseph, 2007; Rai & Fiske, 2011) that overlap theoretically with the set employed here (Bugental, 2000), it has remained unclear how these functions actually are embedded in different types of relationships. Consequently, we undertook to measure relationship-specific patterns of prescribed cooperative functions (i.e., relational norms) in a U.S. cultural context and to demonstrate how these relational norms predict relationship-specific moral judgments.

To do this, we first measured prescribed cooperative functions for a large set of common dyadic relationships, yielding four-dimensional relational norm profiles for each relationship. Quantifying the distinctiveness of these relationships in terms of their relational norms revealed several distinct clusters of relationship types spanning the domains of care, hierarchy, mating, and reciprocity. Consistent with our predictions, such relational norms predicted out-of-sample moral judgments in relational context, and explained more relational variance in moral judgments than

genetic relatedness, social closeness, or interdependence of relationship partners. This suggests that moral judgments of actions within a given relationship are guided by the extent to which the actions thwart or neglect prescribed cooperative functions for that relationship. Moreover, relationships with more similar relational norms showed more similar patterns of moral judgments (Figure 6). These findings reveal a robust underlying structure of expected relational obligations which shape our moral judgments.

Lewin famously argued that behavior is a product of the person and the situation (Lewin, 1951). In a similar spirit, our data confirm that judgments of moral behavior cannot be understood solely with reference to a given act or actor, but rather, must be interpreted in light of the situation, which, in this case, is the type of relationship existing between two individuals. Relationships in a given society can be characterized by distinctive profiles of cooperative norms. They will, therefore, typically be one of the most important situational factors in terms of explanatory power (Clark et al., 2018). Although relationship theorists have, for decades, worked to characterize the structural elements of various close relationships (Kelley et al., 2003) and have sometimes categorized relationships in terms of cooperative functions necessary for human thriving (Bugental, 2000; Clark & Mills, 1993), here we systematically described lay perceptions of the ideal functional make-up of a wide range of relationships as identified by ordinary language. Moreover, we were able to use this information to make accurate out-of-sample predictions of moral judgments concerning various actions. We hope that our approach will inspire further research in this vein, both theoretical and empirical, at the interface of relationship science and moral psychology. Ideally, such research will help to integrate and enrich work in both domains, which has so far remained largely separate.

From a theoretical perspective, one aspect of our current account that requires further attention is the *reciprocity* function. In contrast with the other three functions considered, relationship-specific prescriptions for reciprocity did not significantly predict moral judgments for reciprocity violations. Why might this be so? One possibility is that the model we tested did not distinguish between two different types of reciprocity. In some relationships, such as those between strangers, acquaintances, or individuals doing business with one another (McGraw & Tetlock, 2005), each party tracks the specific benefits contributed to, and received from, the other (Clark et al., 1989). In these relationships, reciprocity thus takes a tit-for-tat form in which benefits are offered and accepted on a highly contingent basis. This type of reciprocity is *transactional*, in that resources are provided, not in response to a real or perceived need on the part of the other, but rather, in response to the past or expected future provision of a similarly valued resource from the cooperation partner. In this, it relies on an explicit accounting of who owes what to whom, and is thus characteristic of so-called "exchange" relationships (Clark & Mills, 1993).

In other relationships, by contrast, such as those between friends, family members, or romantic partners – so-called "communal" relationships – reciprocity takes a different form: that of mutually expected responsiveness to one another's needs. In this form of reciprocity, each party tracks the other's needs (rather than specific benefits provided) (Clark et al., 1989) and strives to meet these needs to the best of their respective abilities, in proportion to the degree of responsibility each has assumed for the other's welfare (Clark & Mills, 1993). Future work on moral judgments in relational context should distinguish between these two types of reciprocity: that is, mutual care-based reciprocity in communal relationships (when

both partners have similar needs and abilities) and tit-for-tat reciprocity between "transactional" cooperation partners who have equal standing or claim on a resource.

A further limitation of the current studies is that they only concern moral *wrongness* judgments, based on actions that weaken one or more of the expected relational functions. What about judgments of moral rightness, goodness, or praiseworthiness as these relate to actions which *strengthen* one or more of the functions (Anderson et al., 2020)? Will people be judged positively for "merely" meeting functional expectations, as when a parent-child relationship fulfills the care function, or will such judgments be reserved for so-called supererogatory behaviors, going above and beyond the call of duty (Archer, 2016)? Either way, we expect that praiseworthiness judgments for a given action will depend, among other things, on the relational context (functionally understood).

Much of the prior literature in moral psychology has focused on judgments of strangers involved in moral dilemmas that pit distinct ethical principles against one another: for example, a utilitarian imperative to maximize welfare, versus a deontological rule that forbids using individuals as a mere means to an end (Conway et al., 2018). A key tenet of utilitarianism is that welfare should be maximized *impartially*, rather than prioritizing the well-being of family members (for example) over distant strangers (Mill, 1863). Descriptive research on moral dilemmas shows that many people are not in fact impartial in this sense (Kahane et al., 2018), consistent with our observations here that people have different cooperative expectations for different relationships, leading in turn to different moral judgments depending on relational context. One intriguing possibility is that individuals who more strongly endorse impartial beneficence will have more uniform prescriptions for cooperative functions across relationships, leading to more uniform moral judgments

across relational contexts. This perspective also suggests possible antecedents of impartial beneficence. Because care is normative in close relationships (with family, friends, and romantic partners), caring for partners in these relationships does not typically elicit special approbation. Perhaps those who find a sense of purpose or belonging not in tending to close relationships, but in widely being admired (Hirsch & Clark, 2019), tend to "distribute" care across a broader set of relationships (thus showing relatively impartial beneficence).

We note that the generalizability of our findings may be limited in several ways. First, apart from relational role and gender (for mothers and fathers), we did not consider the possible impact of such target characteristics as race, religion, politics, age, and social class, on moral judgments. Each variable itself may impact moral judgments (Hester & Gray, 2020; E. F. Jones et al., 1999) and interact with relational context in systematic ways. Second, again apart from gender, we did not comprehensively evaluate how observer (i.e., participant) characteristics along those same demographic lines shape moral judgments, thereby impacting the correspondence between relational norms and moral judgments in relational context. Other individual differences among participants, for example in their relative tendency to engage in different styles of moral reasoning (Kahane et al., 2018) will be important to assess in future research. We see our work as a starting point that may launch further investigations into how both target and observer relational qualities interact with each other and with other kinds of characteristics in shaping moral cognition.

Because we studied participants in the U.S., it also will be important to investigate whether our results generalize across different cultures (Awad et al., 2018). Although we expect that humans in all cultures form (or stand in) relationships

which rely on one or more of the underlying cooperative functions we have highlighted, the patterning of relational norms likely will vary by culture. Indeed, long-standing programs of research have documented such differences using alternative theoretical frameworks. Hindu research participants from the city of Mysore in southern India, for instance, expected care from a broader array of people -- from parents, friends, and even strangers -- than did research participants from the city of New Haven in the United States (Miller et al., 1990). The same difference applied to reciprocity (Miller & Bersoff, 2016). In another study, U.S. American wives felt that husbands should care more for them than for their mothers whereas the reverse held true for Egyptian wives (Pataki et al., 2013). Future studies might also compare how "tight" (that is, lacking in variance across situations) relational norms are in each culture (Gelfand et al., 2011).

Our primary goal for this research has been simple: to investigate how relational context – in particular, the functional cooperative norms that prescriptively govern dyadic interactions of various kinds – shapes moral judgments. A secondary goal has been to push researchers studying human moral psychology to look at behaviors and associated judgments that are more characteristic of people's day-to-day lives than heretofore has been the case. Much remains to be done, including more precise and sophisticated analyses of which cooperative functions apply to which relationships, how these functions relate to one another, and how they can be used to predict praiseworthiness judgments (not just judgments of moral wrongness as we have undertaken here). As we and others pursue work that places the study of morality in both geopolitical *and* relational cultural context, we anticipate the emergence of a more nuanced literature on human morality that becomes better

integrated with broader and long-standing programs of research on relationships and prosocial behavior.

Methods.

All studies were reviewed and approved by the Yale University Institutional review board (protocol #20000022385); informed consent was obtained from participants in each instance prior to data collection. We have posted all study materials, pre-registration forms, raw data, and analysis code on the Open Science Framework (https://osf.io/zxjt6/). For complete study descriptions and supplementary findings, see the Appendix 1 Supplement.

Stage 1. For Stage 1, the design, measures, sampling plan, and exclusion criteria were pre-registered at aspredicted.org (#26400). We used an online polling software (https://www.nbrii.com/our-process/sample-size-calculator/) to determine that at least 385 participants would be needed to obtain population estimates of functional expectations with a 5% margin of error and 95% confidence level. Anticipating participant exclusions, we over-sampled by about 15% and aimed to recruit 450 U.S. participants via the Prolific Academic platform (Prolific); 493 ultimately took the survey, each of whom was paid at a rate of $7.25 per hour. Seventy (70) participants were excluded based on the pre-registered exclusion criteria, leaving us with a final sample of 423 participants ("Sample 1") who completed an online survey. Participants were given descriptions and definitions of all five cooperative functions adapted from Bugental (57): care, coalition, hierarchy, mating, and reciprocity (see Appendix 1, Supplement Section 1.2.1. for the full descriptions). To ensure that participants were thinking of the functions in the way we intended, participants were not able to advance to the main part of the study before passing multiple comprehension checks.

We then asked participants to indicate how much each of 20 common relationships should ideally serve each of the five cooperative functions, specifying: "if this kind of relationship was the best possible relationship of its kind it could be [i.e., according to general societal standards], how much should it serve each of those 5 relationship functions?" Participants rated each relationship type in random order. For each relationship type, we included a brief description (see Appendix 1, Supplement Section 1.2.2. for the descriptions). Then, for each combination of relationship and function, participants rated how much the relationship ideally should serve the given function on a sliding scale from -100 (definitely should not serve) to +100 (definitely should serve). Since every participant assessed all five functions for all 20 relationships, we obtained 100 data points per participant. Finally, we collected a battery of demographic measures (described next) as well as exploratory measures for future studies not included here.

In analyzing the demographic information, we first excluded the coalition ratings for the reasons described in the main text. We then used a linear mixed model to regress prescribed cooperative function scores on participant gender (female, male; 4 participants who marked 'other' were excluded) for each of the four remaining functions. Reported annual income ("low" = $35K or less, "high" = more than $35K; split based on U.S. median income), religiosity ("high" versus "low" based on a mean split), and both social and economic political ideology (ditto) were entered into the model as categorical covariates. Full model summaries are in Appendix 1, Supplement Section 1.4.5.

Stage 2. For Stage 2, the hypothesis, design, measures, sampling plan, exclusion criteria, and analysis approach were pre-registered at aspredicted.org (#31592). Two main steps were involved: first, selection of a subset of relationships

from Stage 1 plus the generation of "action items" to be rated for subsequent use; and second, the actual study, collecting ratings from a new sample ("Sample 2"). As before, the Sample 1 coalition data were excluded.

Using the (remaining) Sample 1 cooperative function scores for all 20 relationships, we performed an analysis that is conceptually similar to a representational similarity analysis (RSA), except that it relies on the Kolmogorov-Smirnov (K-S) distance statistic rather than distribution means. The goal of this analysis was to identify relationships with relatively dissimilar relational norm profiles, so that 10 of the least functionally redundant relationships could be used in the current Stage. For the RSA-like analysis, each relationship was compared to every other relationship on the dimensions of care, hierarchy, mating, and reciprocity. The mean of the four corresponding K-S distance statistics was used for this comparison. Next, we ranked each pair of relationships by its mean K-S distance, from least to most distant (that is, from most functionally redundant to least functionally redundant). We then dropped the relationship from each pair that had the lowest mean K-S distance from all other relationships in the set. Note: for theoretical reasons (i.e., to allow gender comparisons) we decided in advance to retain both the father-child and mother-child relationships in case they faced off. The final set of relationships identified by this procedure is shown in Figure 2 of the main text.

We then created a set of 86 "action statements" describing behaviors that would plausibly weaken or violate specific cooperative functions based on their underlying logic (i.e., how each function solves its corresponding coordination problem). To determine the extent to which certain actions would characteristically weaken (or strengthen) each of the four dyadic cooperative functions, we had 15 trained judges rate all 86 action items in our set. These judges were recruited among

lab members and colleagues and were given extensive training, either in-person or using an online video conferencing platform, to ensure high quality ratings. They were instructed to consider *only* the functional implications of each action, setting any potential moral considerations strictly aside.

Ratings were obtained via an online survey. The survey included the same formal descriptions of cooperative functions used in Stage 1. Following multiple comprehension checks, the judges were shown the 86 action statements, in random order, in the format "Person A does X to Person B." For each action and function combination, they made their judgment on a sliding scale ranging from "Would characteristically weaken [the function]" (-100) through "It depends / Would neither weaken nor strengthen" (0) to "Would characteristically strengthen" (+100).

Next, we created an algorithm to select 12 action items that were rated among the most characteristic in weakening each of the four cooperative functions (three statements per function). First, for each function, the algorithm ranked the actions, in ascending order, by their mean weakening "characteristicness" rating and randomly selected 3 out of the seven most characteristic actions. Second, it computed the mean rating across the three selected actions, yielding one mean score per function. Third, the algorithm computed the standard deviation of the four function-specific means generated in the previous step. Finally, steps one to three were repeated 10,000 times to find the combination of three action statements that yielded the lowest standard deviation of scores across functions. The second iteration of the algorithm was subjected to two further constraints so that we could ensure consistency with potential function-strengthening items planned for testing in future studies ('strengthen' set). The first constraint was that the minimum mean score in the 'weaken' set could not be lower than the minimum mean score in the 'strengthen' set. The second constraint

was that the average of the final ‘weaken’ scores could not be more than one point lower than the average of the final ‘strengthen’ scores. So that future studies can be straightforwardly compared with the present study, we selected the ‘weaken’ action items so that they would weaken the cooperative functions to a similar degree as future ‘strengthen’ items would strengthen the cooperative functions. The reason for doing this is because we wanted to make sure that we identified a set of ‘weaken’ items that were not more extreme (in the ‘weaken’ direction) than future ‘strengthen’ items (in the ‘strengthen’ direction). This process resulted in a final set of 12 function-weakening action statements, with three per function, as shown in Figure 4.

Proceeding to the second main part of Stage 2, a set of naive/lay participants (Sample 2) was recruited, this time on Amazon’s Mechanical Turk (MTurk). To power for the same confidence and margin of error as in Sample 1, but this time with a between-subjects design, it was determined that we would need ratings from 1,551 participants (see Appendix 1, Supplement Section 2.2. for the full rationale). Based on the Sample 1 exclusion rate, we over-recruited by about 10% and thus aimed to recruit 1,706 participants; 1,822 ultimately filled out at least part of the survey (not all finished), each of whom was paid $1.00. Five hundred and two (502) participants were excluded based on pre-registered exclusion criteria, leaving us with a final sample of 1,320 participants. As in Stage 1, they were shown brief descriptions of their assigned relationship. They were told that they would be asked to rate the moral wrongness of various actions within the relationship. To orient them to the rating scale, we clarified that none of the actions they would see would be extreme (e.g., murder), but would rather all be actions that might plausibly occur within the course of day-to-day life.

After passing several attention and comprehension checks, participants were shown, in random order, all 12 action items, tailored to their assigned relationship. For instance, if they were assigned the romantic partner relationship, one of their items was: “Imagine that someone keeps checking their cell phone while their romantic partner tells a sad personal story. How morally wrong would that be, if at all?” Responses were recorded on a sliding scale from “Not at all morally wrong" (0) to "Very morally wrong" (100). Finally, we collected exploratory data about how likely or unlikely it was that each of the rated actions would happen in real life, and administered the same battery of demographic measures as were used in Stage 1.

For the K-S distance analysis reported in the main text, please note that the functional ratings from Sample 1 were first z-scored to each Sample 1 participant in order to account for individual differences in scale use; for the moral wrongness ratings, no such z-scoring was performed because each Sample 2 participant made only 12 ratings (on account of the between-subjects design). For the linear mixed regression model reported in the main text, although the moral wrongness variable was not normally distributed, Q-Q plots indicated that this did not violate the normality assumption of the model. See Appendix 1, Supplement Section 2.5.1 for details.

Stage 3. For Stage 3, we powered to have as many observations per distribution as were obtained in Stage 1. Given design differences between the studies, we determined that we would need ratings from 150 participants for the current study (see Appendix 1, Supplement Section 3.1. for the full rationale). This is the number we recruited on MTurk; ultimately, 149 participants completed the survey, each of whom was paid $1.00 for their time. Sixty-four (64) participants were excluded based on the predetermined exclusion criteria, leaving us with a final sample

of 85 participants. Participants were shown the same descriptions of relationships used in Stage 2 and asked to rate them along three dimensions each of social closeness and interdependence (see Appendix 1, Supplement Section 3.2.1. for the precise wording). Responses -- regarding the extent to which a well-functioning instance of each relationship would be characterized by each dimension of both constructs -- were recorded on a sliding scale from 0 to 100, labelled appropriately for each dimension. Similar demographic measures to those used in the previous studies were administered. Please note that, given the unexpectedly large proportion of excluded participants in this study, we performed a sensitivity/robustness analysis with no exclusions (see Appendix 1, Supplement Section 3.4.2. for details), and the results remain substantively the same.

Data availability. All original data (anonymized) and study materials are available on the Open Science Framework at https://osf.io/zxjt6/. Source data are provided with this paper.

Acknowledgments. Thank you to Billy Brady, Elena Khusainova, Henry Glick, Kevin Anderson, Clara Colombatto, and Hongbo Yu for statistical advice and coding assistance on this project. Thank you to Rachel Calcott, Aden Goolsbee, Nell Mermin-Bunnell, Yuri Munir, Lillian Yuan, Vivan Fung, Vanessa Copeland, Heeral McGhee, Alan Presburger, Samar Allibhoy, Elena DeBre, Isobel Munday, Gargi Singh, Stephanie Brown, Ryan Cox, Brian Bink, Qihe Sun, and Daniel Do, for help in drafting stimuli or serving as judges to rate items used in the study. Finally, thank you to the members of the Crockett, Clark/Bargh, Bloom, and Knobe labs at Yale University for feedback on this research and/or comments on earlier drafts of this manuscript.

Chapter 3

Praise and Blame in Relational Context

Abstract

Contemporary work in moral psychology has focused primarily on judgments concerning interactions between strangers. However, it is increasingly recognized that much of human moral judgment takes place in the context of -- and is shaped by -- multiple dyadic social relationships, such as parent-child, teacher-student, close friends, long-term romantic partners, neighbors, teammates, and so on. In the recent literature, relationship 'type' has been understood in various ways, including in terms of the genetic relatedness or social closeness of the interaction partners. An alternative approach, based on distinctive patterns of cooperative functions (such as care, hierarchy, or mating) that different relationships are normatively expected to serve, has recently shown promise for predicting moral wrongness judgments for a wide range of behaviors taking place in relational context (see Chapter 2). However, this work has focused solely on actions that characteristically impair or neglect ("weaken") cooperative functions, and on associated moral wrongness judgments. Actions that characteristically promote or fulfill ("strengthen") cooperative functions, as well as corresponding judgments of moral goodness or praiseworthiness, remain understudied.

In the present work we begin to address this gap. In Study 1 (N = 388, U.S. nationally representative for age, race, and gender) we measured normative cooperative expectations ("relational norms") regarding care, hierarchy, mating, and transaction for 20 common social relationships. We then used these relational norms to predict out-of-sample moral judgments of blameworthiness and praiseworthiness for actions that characteristically weaken (Study 2, N = 1,660) or strengthen (Study 3, N = 1,431) one or more cooperative functions. Implications and future directions are discussed.

Introduction

Until recently, the psychology of obligation -- our sense of what we owe to one another, morally speaking -- has not received much attention from moral psychologists (Tomasello, 2020). Nor has the way in which our sense of obligation is shaped by social-relational context received much attention (Clark et al., 2015, 2020). Instead, within moral psychology, the focus of research largely has been on judgments about interactions between strangers, as exemplified by so-called trolley problems and other similar paradigms (Bostyn et al., 2018; Conway et al., 2018; Goldstein-Greenwood et al., 2020; Greene, 2015; Greene et al., 2001; Kahane et al., 2015, 2018). However, it is increasingly recognized that much of human moral judgment takes place in the context of -- and is shaped by the nature of -- multiple social relationships, such as those between parents and children, bosses and employees, customers and sellers, and so on; thus, a single morally relevant action may be judged very differently depending on the social relationship within which it occurs (Bloom, 2011; Clark & Boothby, 2013; Simpson et al., 2016). A growing empirical literature is bearing this out (Berg et al., 2021; Earp et al., 2020; Haidt & Baron, 1996; Lee & Holyoak, 2020; Mammen et al., 2021; Marshall, Mermin-Bunnell, et al., 2020; Marshall, Wynn, et al., 2020; McGraw & Tetlock, 2005; McManus et al., 2020, 2021; Rowe et al., 2020; Simpson et al., 2016; Sunar et al., 2021; Tepe & Aydınlı-Karakulak, 2019; Waytz et al., 2013; Yudkin et al., 2021).

For example, in a recent study, an agent who helped a stranger was judged as more morally admirable than an agent who helped a close relative under similar conditions, whereas an agent who helped a stranger *instead* of helping a relative was judged to be morally worse (McManus et al., 2021). This pattern of results can be

explained by the fact that kin, compared to strangers, are widely seen as having a stronger moral obligation to help one another (Crimston et al., 2016; Ko et al., 2020). In other words, while the agent who helped a stranger in the first scenario might be seen as going above and beyond the call of duty, thereby eliciting a positive moral judgment, the agent who helped a stranger in the second scenario could do so only by failing to meet an even weightier obligation, thereby eliciting a negative moral judgment (Law et al., 2021).

Along the same lines, but regarding friendship rather than family ties, another recent study found that, among older children and adults, an unhelpful friend was seen as "meaner" than an unhelpful stranger (Marshall, Wynn, et al., 2020). This is consistent with the fact that, when compared to strangers, friends -- like family -- are typically regarded as having a stronger mutual obligation of care. A similar pattern applies to a larger set of relationships: parent-child, boss-employee, strangers, siblings, long-term romantic partners, friends-with-benefits, close friends, teammates, and roommates/housemates. In this study, the more a relationship was normatively expected to be governed by a norm of care -- based on the ratings of one group of participants -- the more morally wrong a second group of participants judged it to be for someone within that relationship to fail to be caring (Earp et al., 2020). Clearly, in many cases, judgments about the moral status of an act (or actor) will depend on who performs the action, who is affected by it, and the type of relationship between them.

Types of Relationships

But how should relationship 'type' be understood? Social relationships -- as identified by lay language terms such as friend, sibling, or teammate -- vary along

multiple dimensions, any number of which might be relevant for understanding the perceived moral obligations embedded within them. Relative expectations of care are only the tip of the iceberg. Thus, in the recent empirical literature, relationship 'type' has been understood -- and experimentally manipulated -- in different ways. In the study by McManus et al. (2021), relationships were defined in terms of the degree of genetic relatedness of the interaction partners (e.g., strangers, distant relatives, close relatives). Other recent studies have categorized relationships in terms of the social closeness of the interaction partners (Berg et al., 2021; Gilead et al., 2018; Yudkin et al., 2021). And still others have conceived of relationships in terms of various cooperative functions -- including, but not limited to, care -- that each might normatively be expected to serve within a given society (Earp et al., 2020; Simpson et al., 2016; Tepe & Aydınlı-Karakulak, 2019).

Let us consider each classification in turn. The link between genetic relatedness and a perceived obligation of care is straightforward. Classic work on kin altruism predicts that individuals who are more closely genetically related to each other should feel more strongly inclined to unconditionally promote each other's welfare (i.e., care for them) than should individuals who are more distantly genetically related, all else being equal (Burnstein et al., 1994; Ko et al., 2020; Foster et al., 2006). This inclination has undoubtedly become codified within many human cultures, serving as a normative expectation or moral norm (Isern-Mas & Gomila, 2020). The widespread value of being loyal to one's family above most or all others is a reflection of this (Lee & Holyoak, 2020; Waytz et al., 2013).

However, in complex modern societies, family isn't everything. For example, a person might have a tighter affiliation or cooperative bond with a close friend than with a distant cousin, say -- and therefore feel a stronger obligation of care toward the

friend -- even though the cousin is more closely genetically related (McManus et al., 2021). Depending on the society, this stronger perceived obligation to the friend may be more or less culturally sanctioned. And to the extent that it is culturally sanctioned, it might seem that it is the degree of *social closeness* of the interaction partners -- a quality that tends to overlap with genetic relatedness, but which also often comes apart from it -- that more directly explains the perceived obligation. Indeed, within the relationship science literature, the construct of social closeness is, itself, partly defined in terms of the strength of one's caring disposition toward another (i.e., motivation to promote their well-being) (Berscheid et al., 1989; Mills et al., 2004), lending some support to this approach.

But just as there is more to care than family loyalty, there is more to human morality than care. Instead, as several theorists have emphasized, our sense of moral obligation goes far beyond a concern for helping (or avoiding harming) certain people; it also covers concerns about fairness or reciprocity, respect for authority, and sexual propriety among other considerations (Bugental, 2000; Haidt & Hersh, 2001; Haidt & Joseph, 2007; Rai & Fiske, 2011).

One way to understand these concerns is that they reflect certain *cooperative functions* that different relationships and different combinations of relationships within a society must serve to solve recurrent coordination problems of our species (see Figure 1). For example, if interaction partners have highly unequal knowledge or skill with respect to a joint task that needs doing, this could create a coordination problem if the less able partner insisted on being in charge. The cooperative function of *hierarchy* provides a solution to this kind of problem: in such a situation, the more able partner should have final say about how to proceed, while also, ideally, providing leadership in helping the less able partner to grow in their knowledge and skill. The

less able partner should cooperatively follow the leader. If the partners successfully cooperate on this basis, they will be more likely to achieve their joint goal.

Other cooperative functions -- including care, mating, and transaction (a specific form of reciprocity in which benefits are exchanged on a tit for tat basis and records are kept informally or formally) -- are discussed below. The point for now is that different social relationships within a society may typically or chronically face different kinds of coordination problems, and so will characteristically be expected to employ these different functions to different degrees (Earp et al., 2020; Rai & Fiske, 2011). In each case, the moral obligations that are widely perceived to be associated with the relationship can be expected to track the relationship's characteristic cooperative functions (albeit with certain of these functions coming to the fore or receding into the background depending on the situation) (Yudkin et al., 2021).

There is now preliminary evidence that this last, cooperative-functional account -- described in more detail in the following sections -- can successfully predict, and explain more variance in, moral wrongness judgments for a range of potentially objectionable behaviors in relational context than can accounts based on genetic relatedness or social closeness alone (Earp et al., 2020). However, whether such an account can predict or explain judgments of moral *goodness* or *praiseworthiness* for positive behaviors is not yet known. We discuss this issue in the following section.

Cooperative function	Coordination problem to be solved
Care	Securing overall welfare through non-contingent provision (or acceptance) of benefits or resources in response to need
Transaction	Balancing contingent provision and acceptance of benefits for mutual gain over repeated interactions; avoiding exploitation
Hierarchy	Coordinating behavior between individuals who have unequal authority over one another
Mating	Finding and maintaining sexual partners; ultimately, producing and ensuring the survival of offspring

Figure 1. *Cooperative functions of dyadic relationships,* adapted from Earp et al. (2020), building on Bugental (2000) and Clark (e.g., Clark & Mills, 2012). The care and transaction functions, in particular, are based on the work of Clark and colleagues concerning "communal" and "exchange" relationships, respectively (Clark & Mills, 1993, 1979; Clark & Taraban, 1991). Note that in our original model, transaction was defined more broadly in terms of generalized reciprocity (coordinating behavior between functional equals) rather than, as here, in terms of the tit-for-tat logic of exchange-based cooperation. Hierarchy was also defined more broadly, in terms of unequal power, status, or responsibility rather than, as in the present model, more specifically in terms of unequal authority of the interaction partners over one another. See the section entitled "Model Updates" for a discussion of the reasoning behind these changes.

Predicting Positive Moral Judgments

The aforementioned lack of knowledge regarding judgments of moral goodness or praiseworthiness in relational context is consistent with a wider 'blind spot' in contemporary moral psychology research. This research has, overwhelmingly, focused on agents and behaviors deemed to be morally bad, and thus on judgments of moral wrongness, assignment of blame, and associated motivations or decisions to punish perceived offenders (Anderson et al., 2020; Guglielmo & Malle, 2019). Although it is understandable to be especially concerned with the worst of human nature -- our ability to hurt, harm, and offend one another -- the sphere of human morality is much broader than this. To work toward a better world, or simply to understand ourselves in a more comprehensive way, we also need to make sense of our ability to do good: to help, support, and show respect to one another or more

generally to cooperate and promote each other's flourishing (Curry, Mullins, et al., 2019; Law et al., 2021).

How do moral judgments factor into this? In a theoretical paper drawing together findings from social, cognitive, developmental, and consumer psychology, Anderson, Crockett, and Pizarro (2020) propose that blame and praise, while conceptually opposed, are not in practice simple opposites or mirror-images of one another. Rather, they suggest, moral praise is "a fundamentally unique form of moral attribution" that serves a different social function in our lives: "while blame is primarily for punishment and signaling one's moral character, praise is primarily for relationship building" (p. 694). For example, praise may be used to strengthen an affiliative bond or alliance to secure or promote more effective cooperation. It is striking, then, relatively little empirical work has looked at moral judgments of praise in the context of different kinds of relationships (Marshall, Wynn, et al., 2020; McManus et al., 2021).

In this paper we begin to address this gap. In the following sections, we introduce a "relational norms" model of moral judgment based on one particular set of cooperative functions -- care, hierarchy, mating, and transaction -- that different relationships and different combinations of relationships within a society might serve. We then explain how behaviors that characteristically weaken or strengthen these functions might tend to elicit judgments of praise or blame depending on the relationship within which they occur. We then test these predictions in two main studies: one focused on actions that characteristically weaken one or more of these cooperative functions, and a second one focused on actions that characteristically strengthen one or more cooperative functions.

Broadly speaking, we predict that weakening a cooperative function that is strongly normatively expected -- i.e., prescribed -- within a given relationship (for example, a parent failing to care for their child) will be judged more harshly than weakening the same function in relationships for which the function is less strongly prescribed (for example, a roommate failing to care for another roommate) (consistent with Earp et al., 2020). We also anticipate that weakening a function that is negatively expected -- i.e., proscribed -- within a relationship (for example, a doctor refusing to engage in 'mating' behavior initiated by a patient) will be judged positively, perhaps even being seen as praiseworthy.

When it comes to strengthening cooperative functions, a similar logic applies: actions which serve to strengthen a cooperative function that is prescribed within a relationship should be seen as neutral or praiseworthy, whereas strengthening a function that is proscribed within a relationship should be seen as blameworthy. However, particular patterns of moral judgment will differ from function to function, as we emphasize below in laying out our more specific hypotheses. First, we describe our relational norms model.

Cooperative Functions and Relational Norms

We humans are fundamentally social creatures, who must cooperate with one another to survive (Curry, Jones Chesters, et al., 2019; Curry, Mullins, et al., 2019; Gellner et al., 2020). As such, almost all of us are situated within complex networks of relationships -- some entered into voluntarily; others given by nature or circumstance (Bayer et al., 2020; Brinberg et al., 2021; Clark et al., 2020; Clark & Boothby, 2013; Fei, 1992). To understand and coordinate our behavior within these relationships, we often make use of recognized social categories or relational roles,

such as parent-child, boss-employee, neighbor, teammate, long-term romantic partner, or sibling. Such roles typically are not merely descriptive, but rather, carry with them certain normative expectations for how individuals who occupy the roles should, or should not, behave toward one another under various conditions (Haidt & Baron, 1996; Rowe et al., 2020). So, for example, the role of 'parent' is both a descriptive concept, referring to someone who contributes genetic material to an offspring or adopts a child, but also a normative concept pointing to a particular set of relational obligations, including a special duty to care for the child, discipline them, and support their learning and development (Shweder, 1992).

Some of the characteristic prescriptions and proscriptions embedded within such relational roles are social-conventional in nature (Turiel, 2008), having to do with matters of politeness or etiquette, for example. However, others reflect more basic cooperative functions -- such as care, hierarchy, mating, or transaction -- that different relationships within a society must serve in order to solve recurrent coordination problems of the species (Bugental, 2000; Curry, Mullins, et al., 2019). Fulfilling or violating these more basic cooperative expectations, then, will tend to elicit not only social-conventional judgments of approval or disapproval, as might be the case for certain errors of etiquette, but rather, moral judgments of praise or blame -- often as a precursor to rewarding or punishing the actor for their perceived cooperativeness or lack thereof.

To make sense of moral judgments in these cases, a necessary first step is to identify a core set of cooperative functions that societies use to solve recurrent coordination problems (see Figure 1). Various sets of functions have been proposed along these lines, often with overlapping theoretical content (Bugental, 2000; Clark & Mills, 1993; O. S. Curry, Jones Chesters, et al., 2019; Haidt & Joseph, 2007; Rai &

Fiske, 2011b; Shweder et al., 1997). We have adapted a set from Bugental (2000) with modifications based on Clark (Clark & Mills, 1993, 2012), as we discuss in greater detail below. For a summary, see Table 1.

Table 1
Descriptions of relationship functions (as shown to Study 1 participants)

Function	Description
Care	**Full version:** This function coordinates behavior between people to ensure that their overall well-being is secure, without any strings attached to the giving or receiving of help (like expecting direct compensation or favors in return, or feeling an explicit debt). In other words, it ensures that people have someone in their corner they can truly count on for aid and support, in good times and bad. When relationships serve this function, each person pays attention to the genuine needs of the other, and strives to ensure that those needs are met to the best of their respective abilities. When needs and abilities are equal, the individuals may take turns helping each other, or otherwise share burdens equally. When needs and/or abilities are unequal, one person may have to do more to help the other, but this does not create a specific debt to be repaid. **Brief version:** the function of giving or receiving support based on need, without creating a debt
Hierarchy	**Full version:** This function coordinates behavior between people in situations where they have different authority over -- and hence different responsibility for -- one another. In many situations, it is most effective for one person to be in charge or have final say about what happens. So the hierarchy function involves assigning people roles based on their respective authority and responsibility in a given situation, in order to coordinate behavior in a more efficient manner and help accomplish goals. When relationships serve the hierarchy function, there are two main roles: the person who is ultimately in charge (or who has more say about what happens) and the person who is not in charge (or who has less say about what happens). **Brief version:** the function of coordinating behavior between people with unequal authority over one another
Mating	**Full version:** This function coordinates behavior between people to allow them to find and maintain a sexual partner. For our ancestors, the ultimate point of mating was to produce healthy offspring, so that we could pass on our genes and continue as a species. Of course, today we have birth control, and people often have sexual relationships without consciously planning to have children. But the underlying "logic" of the mating function --in terms of the feelings and motivations it tends to inspire remains the same: to attract and secure a mate and stay with that person long enough to at least potentially have children together. **Brief version:** the function of establishing and maintaining a sexual partnership

Transaction	**Full version:** This function coordinates behavior between people in situations where they do not have a special responsibility for helping each other unconditionally, but when they can still mutually benefit through cooperation or exchange. When relationships serve this function, each person asks "What's in it for me?" Accordingly, each individual pays close attention to who has contributed what to a joint activity or agreement, and strives to ensure that the respective benefits are proportional to what each has contributed. "You get back what you put in" -- or "you get what you paid for" -- is the logic of this function. When contributions or claims on a benefit are equal, the fairest solution will be to share equally, take turns enjoying the benefit, or trade comparable benefits with one another. In any case, it's important to keep track of who owes what to whom, so the scales don't get out of balance. **Brief version:** the function of balancing mutual benefits by keeping track of who owes what to whom

Given their usefulness for solving cross-culturally common interpersonal coordination problems, cooperative functions such as these[2] are ubiquitous in human societies. However, to predict moral judgments in relational context, it is not enough to identify an abstract set of functions. Recent research suggests that, over the course of development, children learn that particular relationships within their social environment are normatively associated with different cooperative functions -- or concomitant moral obligations -- to different degrees (Chalik & Dunham, 2020; Mammen et al., 2021; Marshall, Mermin-Bunnell, et al., 2020; Marshall, Wynn, et al., 2020). By adulthood, they will need to understand, at least implicitly, that one and the same action might be judged quite differently depending on (a) the relationship within which it occurs, (b) the specific pattern of cooperative functions that is normatively

[2] Note: we do not suggest that these are the *only* cooperative functions that humans use to solve common (dyadic) coordination problems. For example, Curry, Mullins and Whitehouse (2019) suggest that *possession* (recognition of prior resource possession or property rights) might also be a culturally universal cooperative function, although they do not propose how this or any other function might be normatively embedded within any particular social roles. We do not take a stand on that issue here; our claim is simply that the cooperative functions in our model are likely to be among the most important functions for resolving recurrent two-party coordination problems across a wide range of human cultures, while staying open to the possibility that there may be other important functions as well.

expected for that relationship in their society, and (c) how the action bears on the 'logic' of each embedded cooperative function, either working to strengthen it -- for example, by facilitating the successful execution of the function in the appropriate social-relational context -- or to weaken it, for example, by failing to execute the function and/or executing the 'wrong' function given the context.

One reason why children need to learn these associations, is that relationship-specific patterns of normatively expected cooperative functions -- what we call "relational norms" -- often vary between societies (Argyle et al., 1986; Atari et al., 2020; Deng et al., 2021; Kaspar et al., 2016; Miller et al., 1990; Miller & Bersoff, 2016; Rai & Fiske, 2011); and they may change over time in response to sociopolitical reforms, economic developments, or (other) processes of cultural evolution. To predict and explain moral judgments of praise or blame within the context of any given relationship or society, therefore, the currently-prevailing relational norms of the relevant culture or subculture must be taken into account (Earp et al., 2020).

Consider an example. In many modern societies, an ideal boss-employee relationship is characterized by a particular profile of relational norms, involving, roughly, a strong expectation of both hierarchy and transaction, a weaker expectation of care, and a strongly negative expectation of mating (Chuang, 1998; Earp et al., 2020; Judge & Piccolo, 2004; Williams et al., 1999). As such, the employee should usually follow the boss's instructions, not the other way around (hierarchy); the employee should provide their labor only if fairly compensated, while the boss should pay the employee only insofar as the latter does the agreed-upon work (transaction); the boss should be sympathetic to the personal needs of the employee, but generally should not go out of their way to provide unconditional support (care), and bosses

should not make sexual passes at their employees, nor should employees try to date their bosses, at least while under contract (mating).

By contrast, an ideal relationship between long-term romantic partners within the same societies typically will be characterized by a very different relational norm profile: roughly, negative expectations of both hierarchy[3] and transaction and strongly positive expectations of mating and care (Earp et al., 2020). It follows from these contrasts (among many others that could be used for illustration) that one and the same kind of behavior taking place within these different relationships could elicit very different moral judgments from observers. For example, if a boss in contemporary U.S. society offers to give their employee a sensual massage at the office, this is likely to be seen as objectionable and even downright blameworthy; whereas, if someone in the same society offers to give a sensual massage to their long-term romantic partner, this is likely to be regarded much more positively (all else being equal).

Or suppose that an employee declines to spend the weekend helping their boss with an important personal project, unless the boss is willing to pay them. This is unlikely to be seen as morally objectionable, since, as noted, a typical boss-employee relationship is expected to be governed by a norm of transaction, more so than by a norm of care. Whereas, declining to help a long-term romantic partner with an important personal project unless one is directly compensated is likely to be seen as much more problematic.

[3] In mainstream contemporary U.S. culture, at least, such a relationship is not expected to be strongly hierarchical (Earp et al., 2020). In some other (sub)cultural or historical contexts, by contrast, greater asymmetry in authority between romantic partners may indeed be normatively expected, often in spousal relationships along the lines of gender, in line with patriarchal marriage norms (Bartkowski, 1997; James-Hawkins et al., 2017; Siraj, 2010).

From a theoretical perspective, such contrasting moral judgments for identical behavior suggests an advantage to thinking about relationship 'type' in terms of relational norms based on cooperative functions -- rather than, say, genetic relatedness. After all, members of a boss-employee relationship vs. a long-term romantic partnership are typically equally genetically (un)related. Yet they characteristically serve very different cooperative functions, plausibly giving rise to different relationship-specific obligations. A similar point can be made about social closeness -- another relational dimension that has been manipulated in recent studies, as mentioned previously (Berg et al., 2021; Gilead et al., 2018; Yudkin et al., 2021). One aspect of social closeness is a motive of benevolence. Given this aspect, such a construct may be useful for predicting moral judgments concerning actions that strengthen or weaken the care function. But it is less clear how it could be used to predict judgments concerning hierarchy, mating, or transaction.

Consider hierarchy as an example. Some socially close relationships, such as a typical parent-child relationship, are also characterized by asymmetrical authority. Children are thus expected to behave in a relatively subordinate or deferential manner toward their parents, and failure to do so (without adequate excuse) is liable to earn them a rebuke. But other socially close relationships, such as siblings of a similar age, are not normatively expected to be significantly hierarchical in many societies. Accordingly, the failure of one sibling to behave in a consistently subordinate manner toward the other is less likely, in the relevant societies, to be judged as morally wrong. And yet this likely difference in judgments regarding similar behavior cannot be explained by appealing to social closeness, as this quality is typically roughly comparable between the two relationships (Earp et al., 2020, Supplement Section 3.4.3). It can, however, be explained by appealing to one or more cooperative

functions, such as hierarchy, that are differently embedded (i.e., with different strengths) within their respective relational norm profiles.[4]

To summarize, if we want to answer the question, "How good or bad is it for Person A to do X to Person B," we need to know, not only what X -- the behavior -- is, but also, who Person A and Person B are. In particular, we need to know the nature of the relationship between them, including the social roles they respectively occupy as well as the underlying cooperative functions that are normatively associated with those roles (i.e., relational norms) in the relevant culture.

Model Updates

Before turning to our specific hypotheses for this paper, we should highlight some key updates to our relational norms (RN) model as described in our earlier publications (Clark et al., 2020; Earp et al., 2020). The first update concerns the hierarchy function. In our previous model (call it RN 1.0), hierarchy was defined more broadly as a function for coordinating behavior between people of different status or power (Magee & Galinsky, 2008). However, we came to realize that two people merely having different status or power – i.e., in general, as opposed to over one another -- does not necessarily create a coordination problem between them. To see this, suppose that Person A is directing a theater production, while Person B is serving as one of the actors. Now suppose that Person B (the actor) is both wealthier and more popular than Person A (the director) and thus generally more able to exert influence over others so as to achieve various desired outcomes.

[4] An additional potential advantage of this latter approach is that moral judgments concerning actions that bear on the logic of multiple functions -- for example, highly transactional behavior that (also) undermines care -- can in principle be explained within this framework.

In the context of the rehearsal room, none of that should matter. Rather, what matters for coordinating behavior effectively with respect to the cooperative task at hand -- namely, putting on a high-quality play -- is that the person occupying the social role of director requires greater authority, in the relevant context, than the person occupying the social role of actor, if they are to have the best chance of achieving their shared goal. Accordingly, it is the assignment of interaction partners to a dominant versus subordinate decision-making role that is central to the hierarchy function, rather than simply coordinating behavior between people who happen to have different status or power (Rai & Fiske, 2011). This is now reflected in the definition of hierarchy in the current version of our model (call it RN 2.0).

The second update is more substantial. The cooperative function termed "reciprocity" in RN 1.0 has been replaced in RN 2.0 with a function we call "transaction." In RN 1.0, reciprocity was conceived broadly as a cooperative function used to solve coordination problems "between individuals with functionally similar (or equal) status, power, authority, or claim on a resource" (Earp et al., 2020, p. 3). However, in contrast to the other cooperative functions we tested -- care, hierarchy, and mating -- we were not able to predict moral wrongness judgments for actions that had previously been rated as characteristically weakening the reciprocity function. Based on these null results, we reasoned that our model had failed to distinguish between two different types of reciprocity. As we noted (Earp et al., 2020, p. 16):

> In some relationships, such as those between strangers, acquaintances, or individuals doing business with one another (McGraw & Tetlock, 2005), each party tracks the specific benefits contributed to, and received from, the other (Clark et al., 1989). In these relationships, reciprocity thus takes a tit-for-tat form in which benefits are offered and accepted on a highly contingent basis. This type of reciprocity is *transactional*, in that resources are provided, not in

> response to a real or perceived need on the part of the other, but rather, in response to the past or expected future provision of a similarly valued resource from the cooperation partner. In this, it relies on an explicit accounting of who owes what to whom, and is thus characteristic of so-called "exchange" relationships (Clark & Mills, 1993).
>
> In other relationships, by contrast, such as those between friends, family members, or romantic partners – so-called "communal" relationships – reciprocity takes a different form: that of mutually expected responsiveness to one another's needs. In this form of reciprocity, each party tracks the other's needs (rather than specific benefits provided) (Clark et al., 1989), and strives to meet these needs to the best of their respective abilities, in proportion to the degree of responsibility each has assumed for the other's welfare (Clark & Mills, 1993)

We concluded that subsequent work on moral judgments in relational context should distinguish between these two types of reciprocity: "that is, mutual care-based reciprocity in communal relationships (when both partners have similar needs and abilities) and tit-for-tat reciprocity between 'transactional' cooperation partners who have equal standing or claim on a resource" (Earp et al., 2020, p. 16).

As indicated above, our new transaction function now captures the distinctive tit-for-tat contingencies of the latter type of reciprocity, whereas the former, care-base type of reciprocity is -- as that description suggests -- already reflected in our care function, whose description we have slightly modified to more explicitly focus on needs and abilities (see Table 1). Moreover, the "functional equality" aspect of reciprocity as conceived of in RN 1.0 is, in effect, represented by our hierarchy function, which we now characterize as coordinating behavior between people who have unequal authority over one another. As such, a low normative expectation for hierarchy already implies that the interaction partners should have similar or equal

'say' in the situation, which suggests that the latter quality does not need to be measured separately.

A further advantage of this approach is that, rather than having hierarchy and reciprocity defined, somewhat redundantly, in opposition to each other (as in RN 1.0), hierarchy and transaction are now conceptually orthogonal (similar to Hamilton & Sanders, 1981; Chuang, 1998). Thus, it should now be possible more accurately to capture, empirically, the relational norm structure of social role pairings that are both hierarchical *and* transactional (such as a typical boss-employee relationship) as well as those that are transactional without being hierarchical (such as a typical roommate or housemate relationship). It should now also be possible to capture, empirically, the relational norm structure of social role pairings that are both hierarchical and care-based (such as a typical parent-child relationships) as well as those that are care based without being hierarchical (such as relationships between friends). So too should it be possible more clearly to distinguish between relationships that are ideally characterized by care-based reciprocity (such as a typical long-term romantic partnership) and those that are ideally characterized by transactional reciprocity (such as a typical customer-seller relationship), even when comparing cases -- such as those just mentioned -- in which the parties to either kind of relationship are normatively expected to have similar or equal 'say' in most cooperative situations.

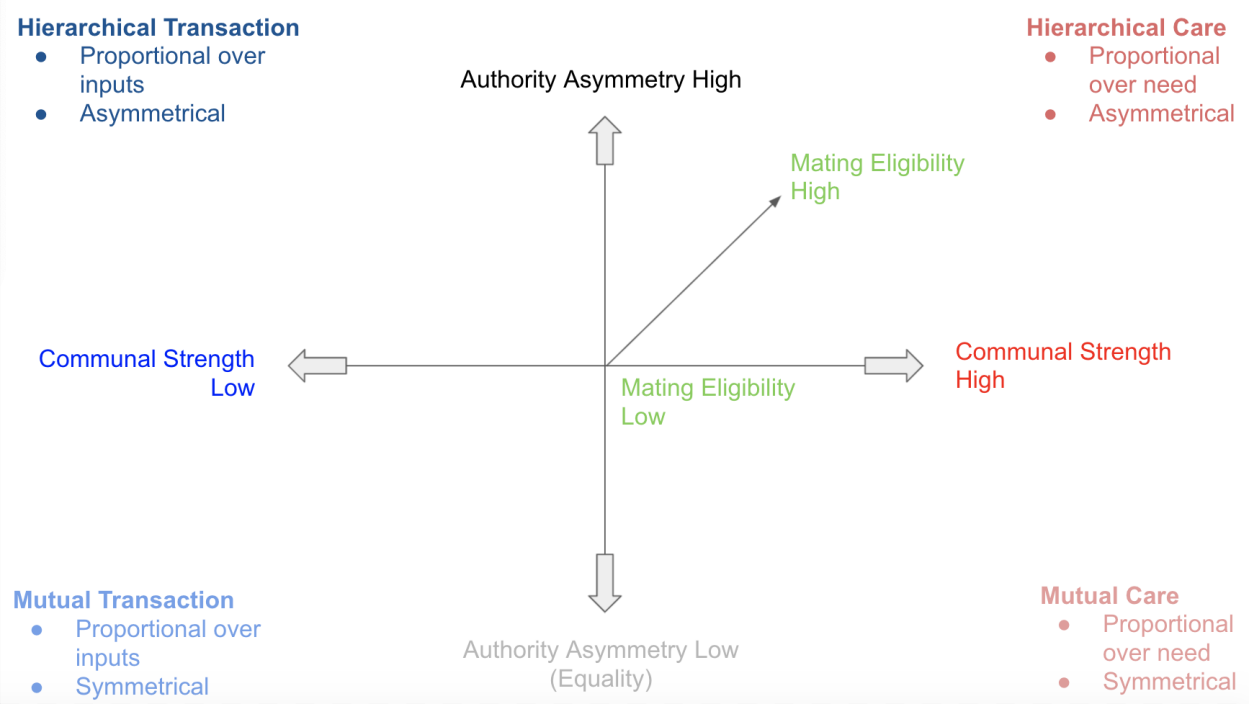


Figure 2. A new theoretical model. Model of proposed conceptual relationships among the four cooperative functions to be tested in the current research. There are three main dimensions along which relationships might vary: their degree of communal strength (i.e., how motivated the partners are to respond to one another's needs) (Mills et al., 2004), their degree of asymmetrical authority over one another (Hamilton & Sanders, 1981; Chuang, 1998), and their degree of eligibility for forming a socially acceptable mating relationship (Frayser, 1989). In principle, any social dyad could be normatively located within this three-dimensional space, raising characteristic coordination problems and -- in well-functioning relationships -- activating the associated cooperative-function schemas (anchored at the corners of the figure) for guiding behavior between the partners, inclining them toward mutually beneficial solutions to the coordination problems in question. For example, a typical relationship between a parent and young child might normatively be located near the top-right corner of the figure (with mating eligibility set to zero along the "Z" axis), suggesting that both the hierarchy and care functions should characteristically be strongly activated, but not the mating or transaction functions.

The Present Studies

The goal of the present studies is to improve, validate, and extend the theoretical scope of our initial relational norms model (RN 1.0) for predicting moral judgments in the context of various social relationships.[5] In RN 1.0, we used

[5] Note that whereas for purposes of this dissertation I focus on RN 2.0 and its links to judgments of morality, RN 2.0 should also prove useful for understanding other types of judgments and behaviors. For instance, we are currently examining its applicability to understanding when various emotions will be experienced and expressed as well as when the expression of various emotions will be deemed appropriate.

population estimates of relational norms for 10 such relationships to predict the moral wrongness judgments of a separate sample of participants, regarding a set of behaviors previously rated as characteristically weakening the cooperative functions of care, hierarchy, mating, and reciprocity. On a function-by-function basis, we were able to make highly precise predictions for care, hierarchy, and mating, but not for reciprocity (Earp et al., 2020). In the present studies, in addition to testing an updated model with a new transaction function replacing reciprocity (RN 2.0), we aim to predict not only negative moral judgments as before, but also positive moral judgments of praiseworthiness. Moreover, we aim to do so not only with regard to actions that characteristically weaken one or more cooperative functions, but also with regard to actions that characteristically strengthen one or more cooperative functions.

With respect to weakening functions, we hope to replicate our previous results (but for all four functions, including the new transaction function): roughly, by replicating our previous results, we mean that we expect to find that the more a given relationship is normatively expected to serve one or more cooperative functions, the morally worse it should be judged for someone within that relationship to perform an action that characteristically weakens those very functions. However, whereas in our previous study participants registered their judgments on a unipolar scale ranging from "not at all morally wrong" to "very morally wrong," in the present study, they will register judgments on a fully bipolar scale ranging from "very blameworthy" to "very praiseworthy." This should allow us to capture more variance in participant appraisals of actions that weaken certain functions -- such as mating -- that are negatively expected (i.e., proscribed) in various relationships. In our previous study, the most participants could say about a doctor, for example, refusing to have sex with

their patient -- thus weakening the mating function between them -- is that such a refusal is “not at all morally wrong.” However, given the opportunity, they might have wanted to express that such behavior is a positive moral good or even praiseworthy; in the present study, they will be able to do so.

With respect to strengthening functions -- not evaluated in our previous work -- our pre-registered hypothesis is as follows: “the more a function is normatively expected within a cluster of relationships, the higher the moral judgment rating for strengthening that function [within the relevant relationships] will be on the blame-to-praise scale.” By implication, the less a function is normatively expected within a cluster of relationships -- including negative expectations or proscriptions -- the lower the moral judgment for strengthening the function should be on the same bipolar scale, potentially crossing over the midpoint (thus representing judgments of blameworthiness).

This prediction is most straightforward for the mating function: in relationships for which mating is strongly proscribed, such as close kin relationships, performing an action that characteristically strengthens the mating function (such as sexually propositioning the other person) will likely be seen as highly blameworthy. Whereas, in relationships for which a function is less strongly proscribed, or perhaps only weakly prescribed -- as might be the case for transaction or hierarchy in various relationships -- actions that strengthen those functions might more likely be seen as mildly blameworthy, or at best ‘relatively less praiseworthy’ (i.e., when compared to relationships for which the functions are strongly prescribed).

Finally, the care function may present an anomaly: whereas mating behavior is straightforwardly inappropriate in normatively platonic relationships; and whereas hierarchical behavior (such as subordinating oneself to an interaction partner or,

conversely, ordering them around) may be inappropriate in normatively egalitarian relationships; and whereas transactional behavior (such as requesting direct compensation for helping with a task) may be inappropriate in normatively communal relationships (Clark & Mills, 1979; McGraw & Tetlock, 2005); caring behavior (definitionally: promoting another's welfare by striving non-contingently to meet their needs) might seem to break the mold. Unlike mating, hierarchy, and transaction, care is almost always welcomed irrespective of relational context.[6] Moreover, it may be precisely those relationships in which care is *least* normatively expected -- such as the relationship between strangers or acquaintances -- wherein acts of care are liable to be seen as supererogatory (going above and beyond the call of duty) and thus as especially praiseworthy (consistent with McManus et al., 2021).

We will return to these predictions in due time. First, we need to measure the currently-prevailing U.S. relational norms for a set of common social relationships with respect to the theorized cooperative functions of care, hierarchy, mating, and transaction. That is the goal of Study 1.

Study 1: Measuring Relational Norms

Method.

Ethics review and open science. All studies in this paper were reviewed and approved by the Yale University Institutional review board (protocol #2000022385); informed consent was obtained from participants in each instance prior to data

[6] A possible exception to this is suggested by the work of Clark and colleagues (Clark & Mills, 1979; Clark & Waddell, 1985), who find that the desirability of transactional vs. caring behavior among strangers or acquaintances depends, crucially, on whether the parties want or expect the relationship to become more communal (for example, by moving toward friendship) or whether one or both of them prefers or expects to keep at a distance (for example, by maintaining a strictly transactional relationship). However, testing the influence of desired or undesired relational development goes beyond the scope of the present work.

collection. We have posted all study materials, pre-registration forms, raw data, and analysis code on the Open Science Framework at https://osf.io/zxjt6/. The design, measures, sampling plan, and exclusion criteria for Study 1 were pre-registered at aspredicted.org (#53912).

Participants. We used an online polling software (https://www.surveysystem.com/sscalce.htm) to determine that at least 385 participants would be needed to obtain U.S. population estimates of normative cooperative functional expectations (i.e., relational norms), nationally representative for age, race, and gender, with a 5% margin of error and 95% confidence level. Anticipating participant exclusions, we over-sampled and aimed to recruit 450 U.S. participants via the Prolific Academic platform (Prolific); 451 ultimately completed the survey, each of whom was paid at a rate of $7.25 per hour. Data from 63 participants were excluded prior to data analysis based on the pre-registered exclusion criteria (see Appendix 2, Supplementary Table 1 for details), leaving us with a final sample of 388 participants (see Table 2 for key demographics).

Table 2
Key demographics of Study 1 participants

Age	N (%)	Race	N (%)	Gender	N (%)
18 - 27	60 (15.46%)	White	272 (70.10%)	Woman	203 (52.32%)
28 - 37	85 (21.91%)	Black/ African American	53 (13.66%)	Man	175 (45.10%)
38 - 47	54 (13.92%)	Asian	23 (5.93%)	Other	10 (2.58%)
48 - 57	65 (16.75%)	Multiracial	17 (4.38%)		

58 +	124 (31.96%)	Hispanic/Latinx	16 (4.12%)
Missing	0 (0.00%)	American Indian/ Alaska Native	2 (0.52%)
		Other	3 (0.77%)
		Prefer not to say	2 (0.52%)

Procedure. Participants were first given descriptions and definitions of all four relationship functions used in this study: care, hierarchy, mating, and transaction (see Table 1 above for full descriptions). To ensure that participants were thinking of the functions in the way we intended, participants were not able to advance to the main part of the study before passing multiple comprehension checks. We then asked participants to indicate how much each of 20 common relationships (from Earp et al., 2020) would, if well-functioning, serve each of the four relationship functions, specifying: "We just want your best, honest judgment about how much a good relationship of each type would serve 1 or more of the 4 relationship functions, if the relationship were working as well as it could for the kind of relationship it is." The relationships included in this study were: long-term romantic partners, mother and under-18 child, father and under-18 child, mother and over-18 child, father and over-18 child, siblings, strangers, close friends, boss-employee, acquaintances, extended family members, roommates/housemates, teacher-student, work colleagues/classmates, political party members, friends-with-benefits (Bisson & Levine, 2009), doctor-patient, teammates, neighbors, and customer-seller.

Participants rated each relationship type in random order. For each relationship type, to reduce ambiguity, we included a brief, specific description of

what we meant by the relationship. For example, for the acquaintance relationship, we specified: “This refers to people who know each other, and interact now and then, but don’t consider one another to be friends. It includes people one might know from work, school, or the neighborhood” (see Appendix 2, Supplementary Table 2 for full descriptions). Then, for each combination of relationship and cooperative function, participants rated how much a maximally well-functioning instance of each relationship would characteristically rely on the given function on a sliding scale from -100 (“definitely WOULD NOT”) to +100 (“definitely WOULD”). Finally, we collected a battery of demographic measures, including age, gender, ethnicity, relationship status, parental status, education, self-ascribed socioeconomic status, social and economic political ideology, and religiosity, as well as exploratory individual difference measures for future studies not included here (see Appendix 2, Supplementary Section 1.1.2. for details).

Results.

Relational norms vary across common dyadic relationships. As can be seen in Figure 3, relational norms varied markedly across dyads in several respects. The care function was generally prescribed, at least to some extent, for almost all dyads apart from the customer-seller relationship (M = -25.99, SD = 63.47) (M across dyads = 50.16, SD = 55.49; this is higher than the scale midpoint with a Bonferroni corrected alpha of .0125), $t(7,759) = 178.22$, $p < .001$, $d = 2.02$ (all tests reported in the manuscript are two-sided). In our previous model, RN 1.0, the reciprocity function was also generally prescribed for almost all relationships, leading to a relatively high mean expectation (RN 1.0 reciprocity M across dyads = 54.23, SD = 49.64; higher than the scale midpoint with a Bonferroni corrected alpha of .0125), $t(8,459) = 100.47$, $p < .001$, $d = 1.09$. In RN 2.0, however, the transaction function -- which

explicitly codifies a tit-for-tat dynamic rather than being solely about the equal standing of the interaction partners (see "Model Updates" for details) -- is now negatively prescribed (i.e. proscribed) for several relationships, including the mother and under-18 child relationship, (M = -53.72, SD = 59.99), father and under-18 child relationship (M = -49.72, SD = 61.189), the long-term romantic partner relationship (M = -28.53, SD = 63.49), the mother and over-18 child relationship (M = -26.73, SD = 64.18), the father and over-18 child relationship (M = -21.11, SD = 61.85), and the teacher-student relationship (M = -21.1, SD = 64.41), leading to a much lower mean expectation across relationships: M across dyads = 9.19, SD = 66.69; this is lower than the mean expectation for reciprocity from the previous model (reported above), $t(16{,}218) = 14.55$, $p < .001$, $d = .23$.

Similar to RN 1.0, in RN 2.0, the mating function was also proscribed for most dyads (M across dyads = -54.63, SD = 66.47; lower than the scale midpoint with the same Bonferroni correction), $t(7{,}759) = 217.3$, $p < .001$, $d = 1.70$, with a few obvious exceptions (long-term romantic partners, M = 91.88, SD = 20.87; friends-with-benefits, M = 61.31, SD = 50.85). With respect to hierarchy, there was wide variation between relationships, with some relationships rated as distinctly hierarchical in expectation, for example, the boss-employee (M = 83.93, SD = 29.20), teacher-student (M = 76.15, SD = 34.84), father and under-18 child (M =66.48, SD = 42.13), and mother and under-18 child (M = 62.01, SD = 48.50) relationships, while others were normatively expected not to be hierarchical, for example, the friends-with-benefits (M = -44.99, SD = 54.99), close friend (M = -44.38, SD = 59.34), neighbor (M = -42.80, SD = 58.79), and long-term romantic partner (M = -42.23, SD = 62.83) relationships.

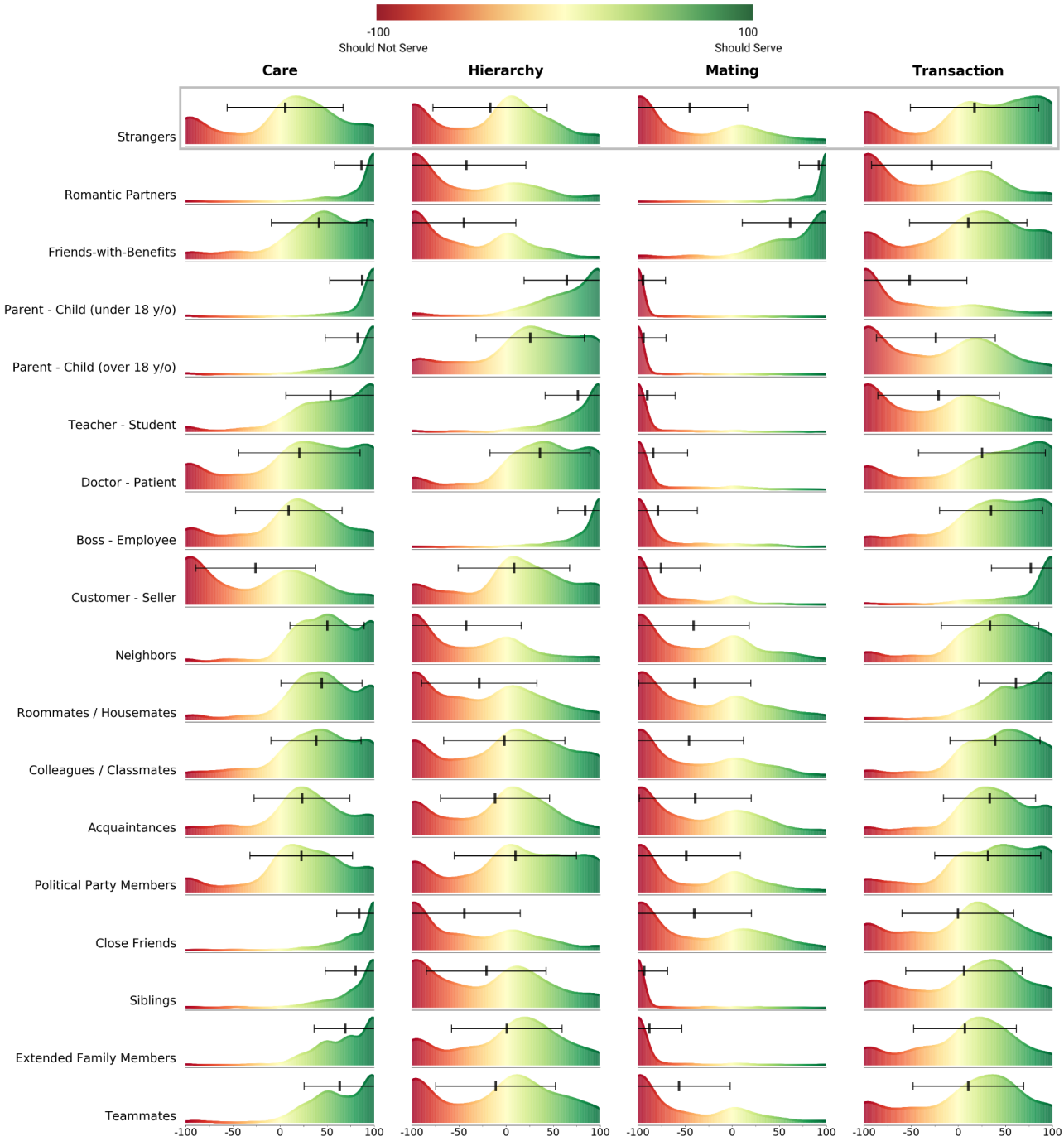


Figure 3. Prescribed cooperative functions. Kernel density plots of prescribed cooperative functions for 20 common relationship dyads. Dots represent the population mean prescription for each cooperative function within each relationship, caps represent +/- one standard deviation. The height of the curve represents density: the likely proportions of scores (relative to each function) that fall within the given range along the x-axis. Figure caption adapted from Earp et al. (in press).

Compared to RN 1.0, wherein hierarchy was defined solely in terms of unequal power or status, in RN 2.0, where hierarchy is defined more specifically in terms of unequal authority between the interaction partners (see "Model Updates"), we find that hierarchy in this latter sense is more strongly proscribed in several

relationships, leading to a lower overall mean for hierarchy in the new model (*M* across dyads = 6.37 in the current model vs. 21.42 in RN 1.0, *SDs* = 69.28 vs. 63.94 respectively), $t(16{,}218) = 3.92$, $p < .001$, $d = .06$). For demographic results, see Appendix 2, Supplementary Section 1.2.

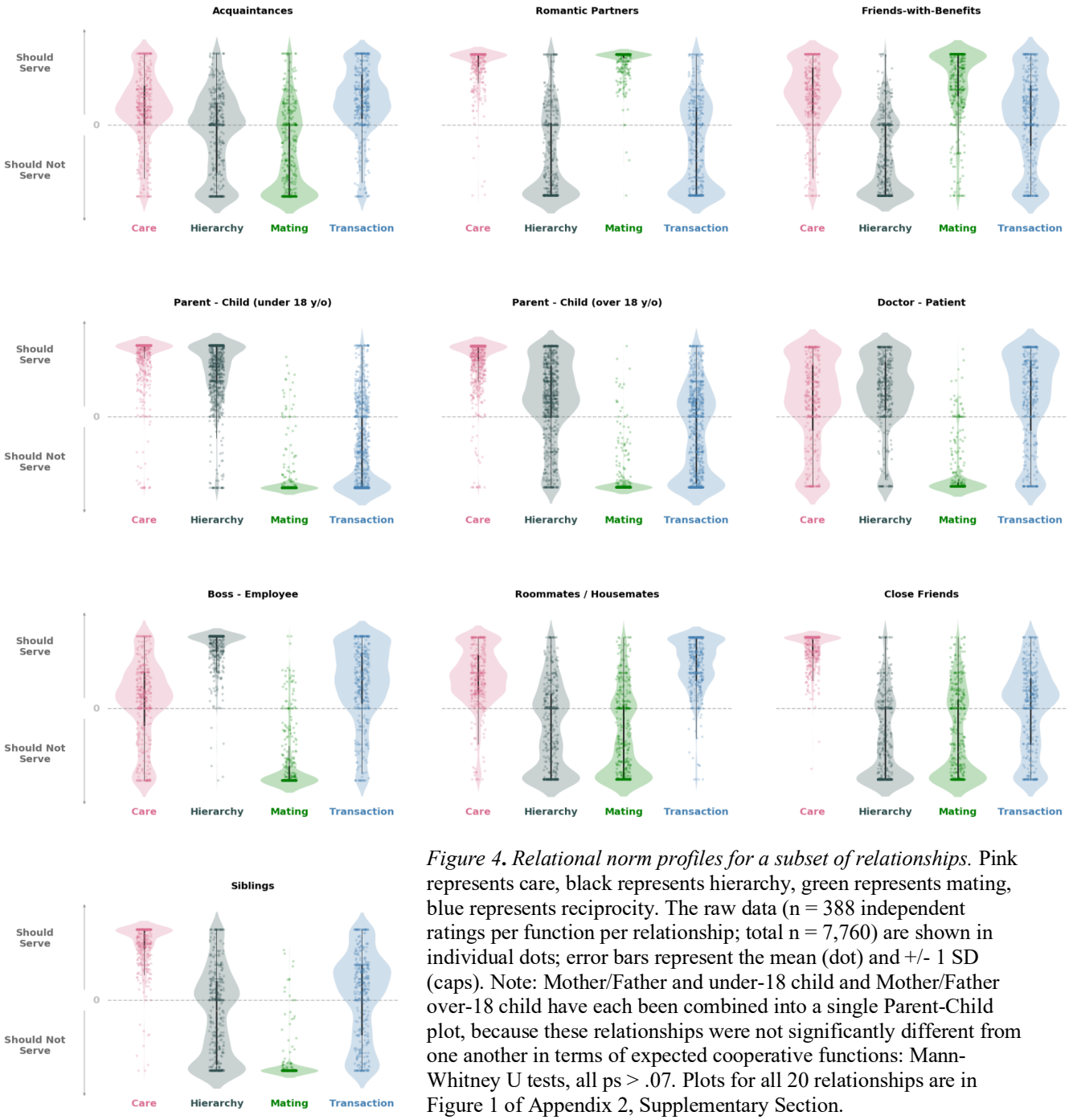


Figure 4. Relational norm profiles for a subset of relationships. Pink represents care, black represents hierarchy, green represents mating, blue represents reciprocity. The raw data (n = 388 independent ratings per function per relationship; total n = 7,760) are shown in individual dots; error bars represent the mean (dot) and +/- 1 SD (caps). Note: Mother/Father and under-18 child and Mother/Father over-18 child have each been combined into a single Parent-Child plot, because these relationships were not significantly different from one another in terms of expected cooperative functions: Mann-Whitney U tests, all ps > .07. Plots for all 20 relationships are in Figure 1 of Appendix 2, Supplementary Section.

Figure 4 is derived from the same ratings as Figure 3, but with the data organized to highlight the four-dimensional relational norm profiles (i.e., sets of prescribed/proscribed cooperative functions) for a subset of 10 relationships selected for predictive modeling in Studies 2 and 3 (see below for details).

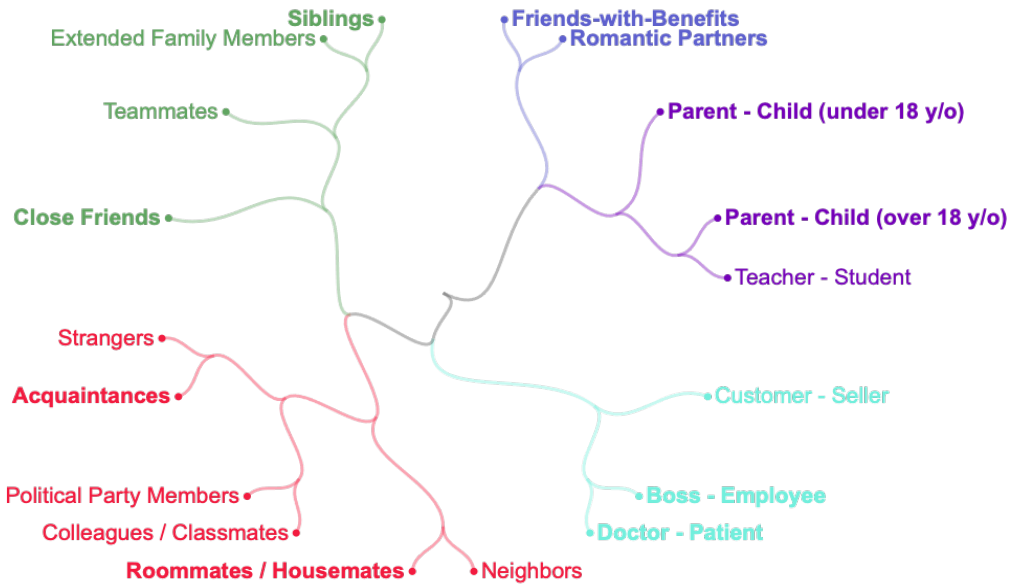
Common relationships are hierarchically clustered around relational norms. Next, we sought to quantify the distinctiveness of each relationship in four-dimensional relational norm space. Because in many instances the ratings were not normally distributed (see Figure 3), characterizing between-relationship differences in terms of their mean relational norm scores would be misleading and lose considerable information. We therefore calculated the Kolmogorov-Smirnov (K-S) distance statistic (a quantification of the difference in overall shape between any two empirical distributions) (Fabbri & De León, 2017) for each cooperative function for each possible pair of relationships. We then averaged across functions to quantify the overall dissimilarity in relational norms for each relationship pair. This approach is conceptually similar to representational similarity analysis (Kriegeskorte et al., 2008), but incorporates information about the shapes of the relational norm distributions in addition to distribution means.

We used the relational norm dissimilarity values to conduct a hierarchical clustering analysis using a farthest-point algorithm: $d(u,v)=\max(\text{dist}(u[i],v[j]))$ (Voorhees, 1985). This revealed five main clusters, depicted in Figures 5a and 5b, which aligned with our theoretical expectations (see Figure 2). The first cluster consists of sexual relationships (long-term romantic partners and friends-with-benefits). The second cluster consists primarily of relationships that are both hierarchical and caring (e.g., teacher-student, parents and children). The third cluster includes relationships that are both hierarchical and transactional (e.g., boss-

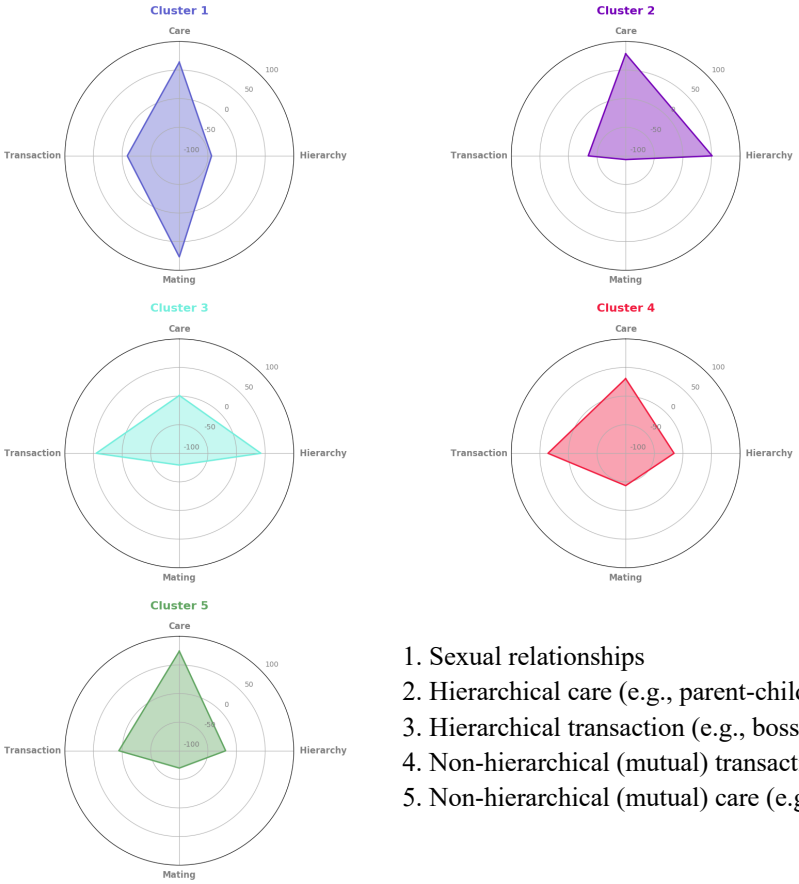
employee, doctor-patient). The fourth cluster includes relationships that are transactional without being hierarchical (e.g., neighbors, roommates/housemates). And the fifth, final cluster is characterized by relationships that are caring without being hierarchical (e.g., siblings, close friends, teammates).

Based on these analyses, we identified a subset of 10 relationships with relatively distinctive relational norms, taking care to ensure that each main cluster 'contributed' a comparable number of relationships (see Appendix 2, Supplementary Section 1.2.2. for the selection procedure). Note that there were no significant differences between participant expectations for the mother versus father relationships (see Figure 4 caption for details); these were therefore collapsed for the subsequent studies into parent-child relationships. The final subset of relationships selected for use in Studies 2 and 3 consisted of roommates, siblings, parent and over-18 child, parent and under-18 child, long-term romantic partners, friends, boss-employee, doctor-patient, friends-with-benefits, and acquaintances. Relational norm profiles for these relationships are depicted in Figure 4 above.

a.



b.



1. Sexual relationships
2. Hierarchical care (e.g., parent-child)
3. Hierarchical transaction (e.g., boss-employee)
4. Non-hierarchical (mutual) transaction (e.g., roommates/housemates)
5. Non-hierarchical (mutual) care (e.g., adult siblings)

Figures 5a-5b. Hierarchical relationship clustering. Circular dendrogram visually representing the mean Kolmogorov-Smirnov (K-S) distance between relationships in four-dimensional relational norm space, clustered hierarchically according to the Voorhees method (Voorhees, 1985) (**a**); relationships selected for Study 2 are highlighted in a darker shade. Radar plots derived from the hierarchical cluster model are depicted in the bottom half of the figure (**b**).

Discussion.

In this first study, we established population-level estimates of current relational norms in the United States for the cooperative functions of care, hierarchy, mating, and transaction. Our inclusion of the last of these functions -- transaction -- represents a significant departure from, and improvement to, our previously established estimates (RN 1.0, Earp et al., 2020), which were based in part on participant ratings of the extent to which each of 20 relationships should ideally serve a "reciprocity" function, broadly conceived. As transaction is defined more narrowly to capture the contingent provision (or acceptance) of benefits within exchange-based relationships -- rather than simply the relatively equal status or power of the

interaction partners, as in RN 1.0 -- a larger number of relationships in RN 2.0 are rated as proscribing transaction than was true of reciprocity. Similarly, the mean normative expectation of hierarchy across relationships is now lower than in our previous model, as hierarchy also has been more strictly defined to refer to the unequal authority of interaction partners over one another, rather than simply the relatively unequal status or power of the interaction partners in general.

An additional consequence of these updated definitions is that, in RN 2.0, hierarchy and transaction are conceptually independent of one another, whereas in RN 1.0, hierarchy and reciprocity were defined in opposition to each other and were thus inversely correlated. In the current model, it is now care and transaction that are defined in opposition to each other, consistent with the existing literature on "communal" vs. "exchange" relationships (Clark & Mills, 1993). Roughly, this literature finds that the stronger the communal bond between relationship partners, the more the partners tend to rely on a care-based cooperative framework. In this framework, the partners strive to promote each other's overall well-being by practicing mutual responsiveness to one another's needs, to the best of their respective abilities and in proportion to the degree of responsibility each one has -- or has taken on -- for securing the other's welfare. Whereas, the weaker the communal bond, the more the partners will tend to rely on an exchange-based (i.e., transactional) cooperative framework or, at times, simply distance themselves from the other. In this framework, each partner provides benefits, not in response to the real or perceived needs of the other partner, but rather, in response to a previously received benefit of comparable value -- or in anticipation of the future receipt of such a benefit -- from the other partner. Thus, to avoid exploitation, each party keeps a more explicit tracking of who owes what to whom (Clark et al., 1989).

As a result of these changes, we find a different configuration of relationship clusters in relational norm space. Whereas, in RN 1.0, there were four clusters (roughly: sexual relationships, hierarchical relationships, egalitarian relationships, and caring relationships), in RN 2.0, there are five clusters, with each set of relationships -- apart from the sexual ones -- grouped around at least two main cooperative functions. This is consistent with our theoretical framework laid out in Figure 2, which allows for both hierarchical and non-hierarchical caring relationships, and hierarchical and non-hierarchical transactional relationships -- roughly corresponding to the new relationship groupings identified by our clustering model (for similar results drawn from a Taiwanese sample, see Chuang, 1998).

Having used this model to determine which of the 20 overall relationships are closest to one another in four-dimensional relational norm space (i.e., most functionally redundant, based on the updated cooperative function definitions), we opted to select a subset of 10 relationships that are comparatively functionally distinct for purposes of testing our predictions regarding moral judgment in Studies 2 and 3 (see Appendix 2, Supplementary Section 2.1. for the selection procedure). We turn to the first of those studies next.

Study 2: Moral Judgments of Function-Weakening Actions

The purpose of Study 2 is to replicate our previous results: predicting moral judgments of behaviors that characteristically weaken one or more cooperative functions in relational context, out-of-sample, from a set of previously measured U.S. nationally-representative relational norms (Earp et al., 2020). Here, however, we do so with the updated cooperative functions, and with a new bipolar dependent measure

of moral judgment, ranging from “very blameworthy” to “very praiseworthy.” For this study, we pre-registered two specific predictions:

First, that blameworthiness/praiseworthiness judgments from the present study will be predicted directly from Study 1 relational norms in a linear mixed regression model (described below). Second, that relationship dyads that are relatively similar in their respective relational norm profiles will also be associated with a similar pattern of moral judgments regarding actions occurring within them. In particular, we predict that, for every pairing of relationship dyads (e.g., siblings paired with roommates), there will be a positive correlation between their distance in relational norm space, as measured by the Kolmogorov-Smirnov coefficient, and their distance in moral judgment space.

Method.

Open science. Hypotheses, design, measures, sampling strategy, analysis plan, and exclusion criteria were pre-registered at aspredicted.org (#63736). All materials, data, and code are available at https://osf.io/zxjt6/.

Stimulus development. To test the hypothesis that relational norms (from Study 1) would predict patterns of moral judgments across multiple common relationships (in Study 2), we first assembled a set of action items describing behaviors that would plausibly weaken or strengthen one or more of the cooperative functions. Most of these items were drawn from our previous study (Earp et al., 2020), however, we also edited or added candidate items to more closely track the logic of the cooperative functions based on their updated definitions. For example, for the care function, we made sure to include candidate items that explicitly referenced non-contingent responsiveness to need; for the new transaction function, we made

sure to include candidate items that exhibited a tit-for-tat logic, rather than simply a more general concern for equality or fairness; for hierarchy, we made sure to include candidate items that implied an asymmetrical authority between the interaction partners, as opposed to a mere difference in power or status.

Twelve trained judges (6 women, 5 men, 1 non-binary, M_{age} = 31.92, SD_{age} = 11.51) rated 78 action items with the form "Person A does X to Person B" on the extent to which each described behavior would characteristically weaken or strengthen each of the cooperative functions, *setting moral questions aside:* that is, the judges were instructed not to consider whether an action might be morally good or bad in any particular relationship -- or in relationships in general -- but only whether the action would characteristically weaken or strengthen each function based on its own operational logic, independent of relational context. Ratings were given on a sliding scale ranging from -100 ("Would characteristically weaken") to + 100 ("Would characteristically strengthen"), with the middle of the scale marked 0 ("Neither/It depends").

There was very strong interrater agreement in these ratings (ICC(3,k) = .95). Following a similar procedure as in Earp et al. (2020), we entered these data into an algorithm to select a final set of 12 of the most highly characteristic function-weakening action statements, with 3 statements for each of the 4 dyadic functions (see Appendix 2, Supplement Section 2.1.1 for details). See Figure 6 for the selected items.

Participants. Having selected our stimuli for the present study, participants were then recruited through the CloudResearch platform running on Amazon's Mechanical Turk (MTurk). To power for the same confidence interval and margin of error as in Study 1, but this time using a between-subjects design and a convenience

sample, it was determined that we would need ratings from 1,796 participants, which we rounded up to 1,800 (see Appendix 2, Supplementary Section 2.1.2. for the full rationale). Since we used a quality control measure in the CloudResearch platform ("Exclude low-quality participants"), we did not over-recruit to account for potential exclusions. Ultimately, 1,824 participants completed the survey, each of whom was paid $1.30. Data from 164 of these participants were excluded prior to data analysis based on our pre-registered exclusion criteria, leaving us with a final sample of 1,660 participants (see Table 3 below for key demographics; see Appendix 2, Supplementary Table 4 for exclusion details).

Table 3
Key demographics of Study 2 participants

Age	N (%)	Race	N (%)	Gender	N (%)
18 - 27	305 (18.7%)	White	1,175 (72.04%)	Female	847 (51.93%)
28 - 37	579 (35.50%)	Black/ African-American	177 (10.85%)	Male	778 (47.70%)
38 - 47	360 (22.07%)	Asian	140 (8.58%)	Other	6 (0.37%)
48 - 57	215 (13.28%)	Hispanic/ Latinx	103 (6.32%)		
58 +	172 (10.55%)	Other	26 (1.59%)		
Missing	0 (0.00%)	American Indian/ Alaska Native	7 (0.43%)		
		Hawaiian/ Pacific Islander	3 (0.18%)		

Procedure. Participants were assigned randomly 1 of the 10 relationships previously selected for their functional distinctiveness -- shown in Figure 4 -- and given brief descriptions of their assigned relationship (the same as in Study 1). They were told that they would be asked to rate the blameworthiness/praiseworthiness of various actions within the relationship. To orient them to the rating scale, we clarified that none of the actions they would see would be extreme (e.g., murder), but would rather all be actions that might plausibly occur within the course of day-to-day life.

After passing several attention and comprehension checks, participants were shown, in random order, all 12 function-weakening action items shown in Figure 6 below, tailored to their assigned relationship. For instance, if they were assigned the long-term romantic partner relationship, one of their items was: “Imagine that someone chooses not to make time for their long-term romantic partner when their partner is in need of some company. How blameworthy or praiseworthy would that be, if at all?” Responses were recorded on a sliding scale from “Very blameworthy" (-100) to "Very praiseworthy" (+100). We also asked participants to rate each action on how *likely* it would be to occur in real life, in order to be able to control for violations of nonmoral (i.e., social-conventional) expectations (Turiel, 2008) (see “action likelihood” variable below). For each participant, we computed the mean moral judgment rating for each of the 4 cooperative function-weakening categories within their assigned dyad. We also administered a battery of demographic questions, asking participants to self-report their gender, age, race/ethnicity, English language fluency, income, education, social and economic political ideology, and religiosity.

Results.

Stimulus selection: action item characteristicness. As can be seen in Figure 6, each action was rated by the judges as having both a main (i.e., "target") effect on a given function, as well as "side effects" on the other cooperative functions. For example, "Person A doesn't invest any time or energy in attending to Person B's needs" was rated as most characteristic in weakening the care function (M = -95.42, SD = 8.85), but also the mating function to a lesser extent (M = -54.25, SD = 37.4). As another example, "Person A refuses to accept payment or compensation in return for a valuable resource they provided to Person B" was rated as most characteristic in weakening the transaction function (M = -59.92, SD = 47.71), but -- consistent with the oppositional logic of care and transaction, discussed above -- was rated as *strengthening* the care function (M = 58.50, SD = 34.83). Of course, the fact that one and the same action might simultaneously affect multiple cooperative functions is to be expected, depending on the operational logic of each function and the nature of the action. To account for the function-affecting specificity of each action, then, we computed a "target specificity" variable (i.e., main effect minus mean of side effects) for each action for use in subsequent pre-registered analyses.

Moral judgment ratings. As noted, each participant was asked to rate all 12 action items shown in Figure 6 on their blameworthiness/praiseworthiness in the context of their assigned relationship. We calculated the mean moral judgment rating of each subset of 3 actions (resulting in one mean rating per function per participant for their assigned relationship), with the resulting distributions of moral judgment ratings for all assigned relationships depicted in Figure 7.

	Actions judged to characteristically weaken one or more cooperative functions	Care	Hierarchy	Mating	Transaction
		Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
1	Person A loudly complains about their own troubles after Person B reveals a sad personal story.	-57.25 (51.15)	-14.92 (23.10)	-32.42 (34.17)	10.58 (16.92)
2	Person A doesn't invest any time or energy in attending to Person B's needs.	-95.42 (8.85)	-20.92 (23.77)	-54.25 (37.4)	-3.17 (32.45)
3	Person A chooses not to make time for Person B when person B is in need of some company	-73.83 (27.33)	-14.33 (17.54)	-43.75 (37.26)	0.17 (44.95)
4	Person B openly expresses their disagreement with Person A's judgement in front of others.	-24.25 (21.25)	-63.08 (40.81)	-18.83 (23.46)	-5.17 (18.7)
5	Person B refuses to carry out a task in the manner Person A instructs	-25.25 (30.54)	-83.67 (32.65)	-12.08 (13.25)	-24.33 (28.89)
6	Person B pursues their own plan of action, despite contradictory instructions from Person A.	-26.33 (29.10)	-80.75 (26.94)	-18.17 (19.53)	-18.67 (29.29)
7	Person A intentionally avoids any kind of sexual interaction with Person B.	8.92 (5.13)	24.08 (34.21)	-87.58 (23.36)	12.33 (28.88)
8	Person A goes on romantic dates with people other than Person B without seeking Person B's permission.	-35.17 (35.98)	-18.17 (33.11)	-86.75 (15.8)	-4.58 (11.3)
9	Person A invests time and energy in a romantic relationship with someone other than Person B	-20.92 (29.89)	4.00 (16.26)	-55.92 (50.47)	0 (0)
10	Person A decides not to share a reward with Person B even though Person B did an equal amount of work.	-36.33 (33.24)	-41.92 (27.74)	-26.08 (32.65)	-81.50 (25.25)
11	Person A refuses to accept payment or compensation in return for a valuable resource they provided to Person B.	58.50 (34.83)	1.33 (27.43)	21.33 (26.39)	-59.92 (47.71)
12	Person A doesn't offer to compensate Person B for the time Person B spent helping them with a task.	2.17 (41.21)	-11.08 (43.95)	-4.50 (6.36)	-79.67 (24.53)

Figure 6. Characteristic function-weakening actions. Heatmap showing mean ratings of judges (n = 12) of the extent to which each action would characteristically weaken or strengthen the care, hierarchy, mating, and transaction functions, respectively, between any two people. These items were chosen as experimental stimuli from a much larger set by an algorithm (see Appendix 2, Supplement Section 2.1.1. for details) using the judges' ratings, where -100 represents the most characteristic function-weakening effect. Darker shades represent more extreme ratings in the weakening direction. Note: when rating actions on the "hierarchy" dimension, judges were asked to imagine that Person A was in a *subordinate* role, specifically; when rating actions on the "care" dimension, judges were asked to imagine that Person A was in a *caregiving* (as opposed to care-seeking) role, specifically. The text for this Figure legend is adapted from a similar legend from Earp et al. (2020).

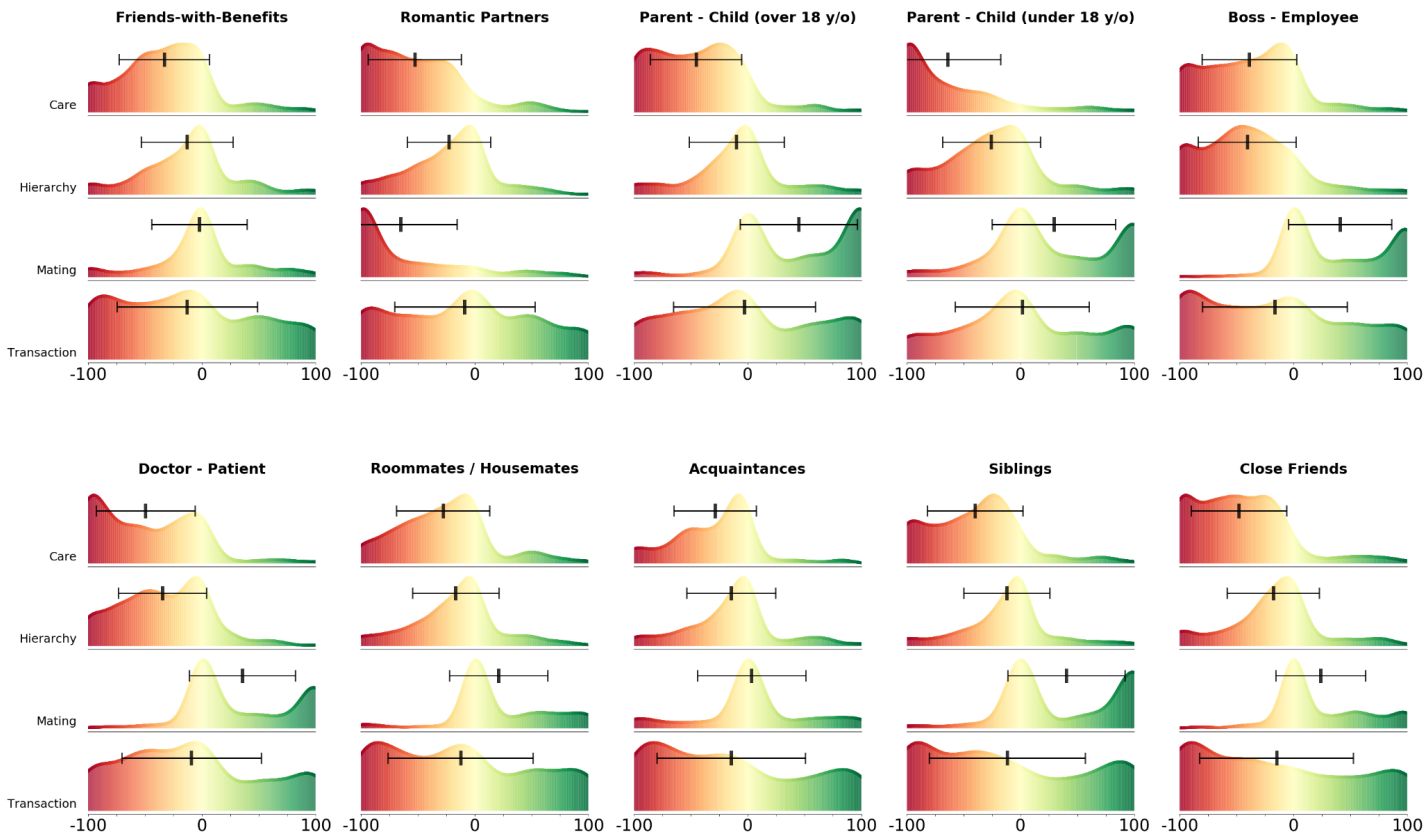


Figure 7. Moral judgments of praise and blame for function-weakening actions. Study 2 moral judgments of actions rated as characteristically weakening one or more cooperative functions in different relationships: kernel density plot of moral judgments (-100 = very blameworthy, +100 = very praiseworthy). Dot represents the mean, with 95% confidence intervals. Height of the curve represents density (see Figure 3 for explanation). This experiment was conducted once, with all data shown here.

Predicting moral judgments: linear regression approach. We turn now to our main, pre-registered hypothesis. As a first approach, we sought to predict Study 2 moral judgments (i.e., for weakening one or more cooperative functions) directly from Study 1 relational norms in a linear mixed regression model. Accordingly, Study 2 participants were entered as the highest-level grouping variable, with assigned relationship dyad entered as a nested random intercept. The mean relational norm estimates from Study 1 were entered alongside both "action likelihood" and "target specificity" as continuous fixed factors for the reasons given above.

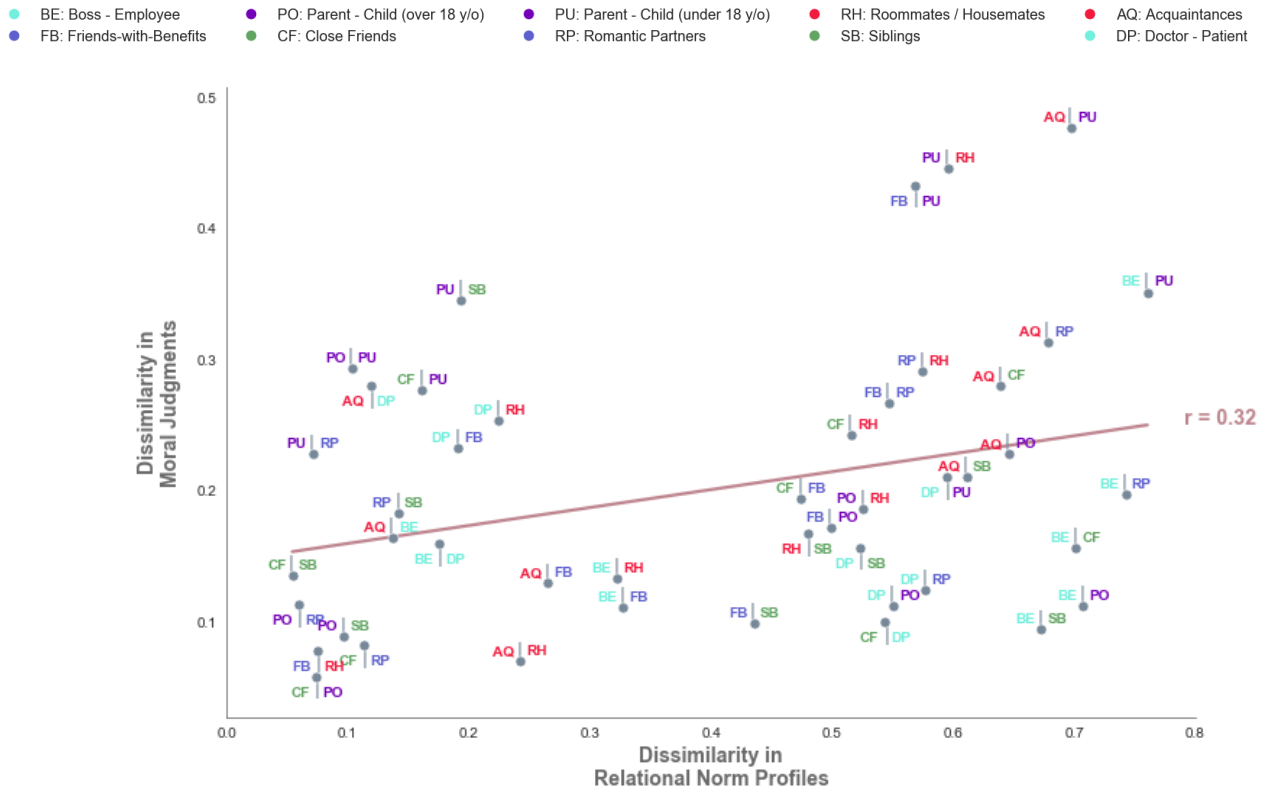
The results from this model supported our hypothesis for each function. Study 1 relational norms significantly predicted Study 2 moral judgments for care ($p < .001$, 95% CI [-.27, -.14]), hierarchy ($p < .001$, 95% CI [-.14, -.07]), mating ($p < .001$, 95%

CI [-.36, -.30]), and transaction ($p < .001$, 95% CI [-.11, -.05]). These results are also robust when controlling for the following demographic factors: gender, education level, income, social and economic ideology, and religiosity (see Appendix 2, Supplementary Section 2.2 for full regression tables).

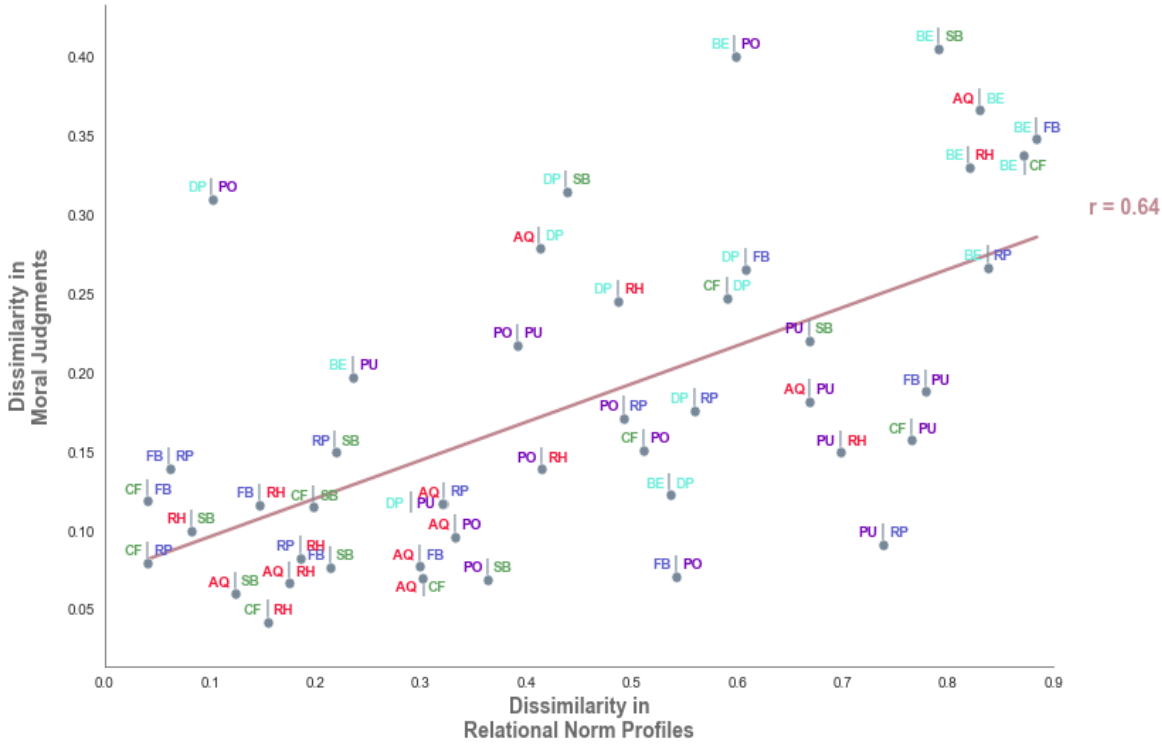
Predicting moral judgments: K-S distance correlation approach. Having confirmed that Study 1 relational norms predict between-relationship variation in moral judgments in the current study, over and above mere uncommonness or unexpectedness of behavior (see "action likelihood" discussion above), we sought to further explore this finding with a second pre-registered analysis. Our hypothesis was that "dyads with similar relational norms within a given society [will be] associated with similar patterns of moral judgments across actions, whereas dyads with dissimilar relational norms [will be] associated with divergent patterns of moral judgments across actions" (Earp et al., 2020, p. 4). Accordingly, we sought to predict the K-S distance between each pair of relationships in moral judgment space (based on current ratings) from their corresponding K-S distances in relational norm space (from Study 1).

To do this, we relied on the same K-S distance approach described earlier, comparing the moral judgment distributions for each type of function violation for each possible pair of relationships. We then computed a Spearman's correlation between these moral judgment dissimilarity values and the previously computed relational norm dissimilarity values. As can be seen in Figures 8a-8d, our above-stated hypothesis was confirmed for each function. Thus, we find a positive correlation between the extent of dissimilarity in relational norms (as measured by the K-S distance coefficient) and the extent of dissimilarity in associated patterns of moral judgment for care ($r = .34$, $p = .02$; note, however, that this is not statistically

significant when a Bonferroni correction is applied), mating (r = .86, p < .001), hierarchy (r = .63, p < .001), and -- in contrast to our inability to predict this relationship with the reciprocity function in RN 1.0 -- also transaction (r = .41, p = .006). See Figures 8a-8d.

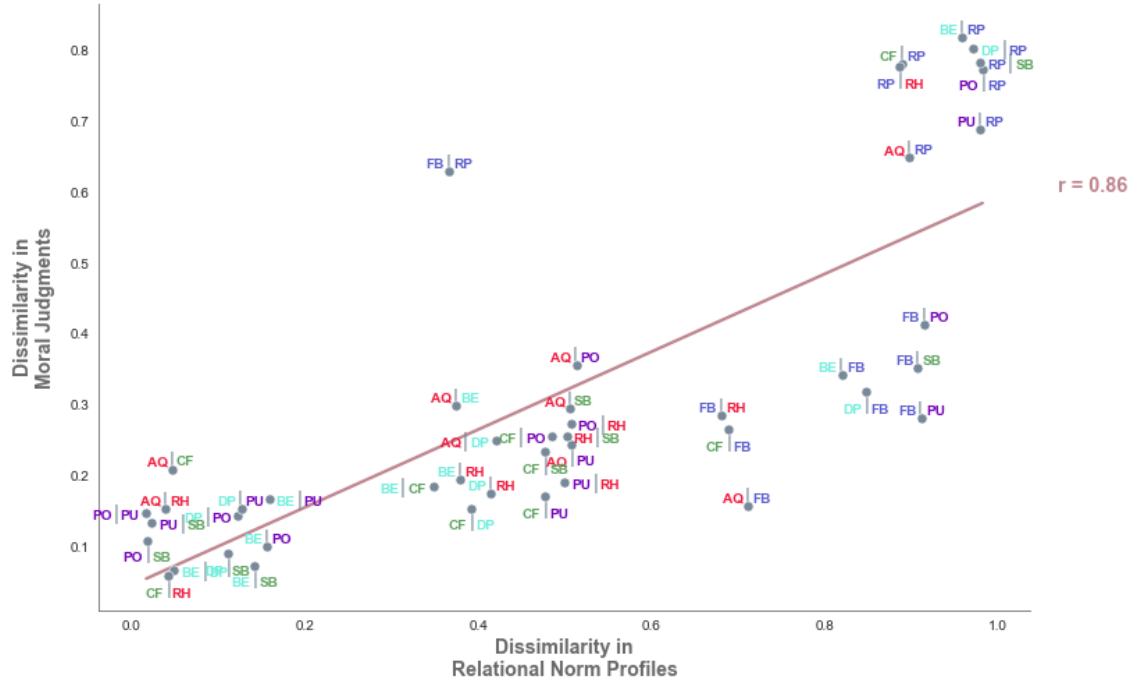


a. Care

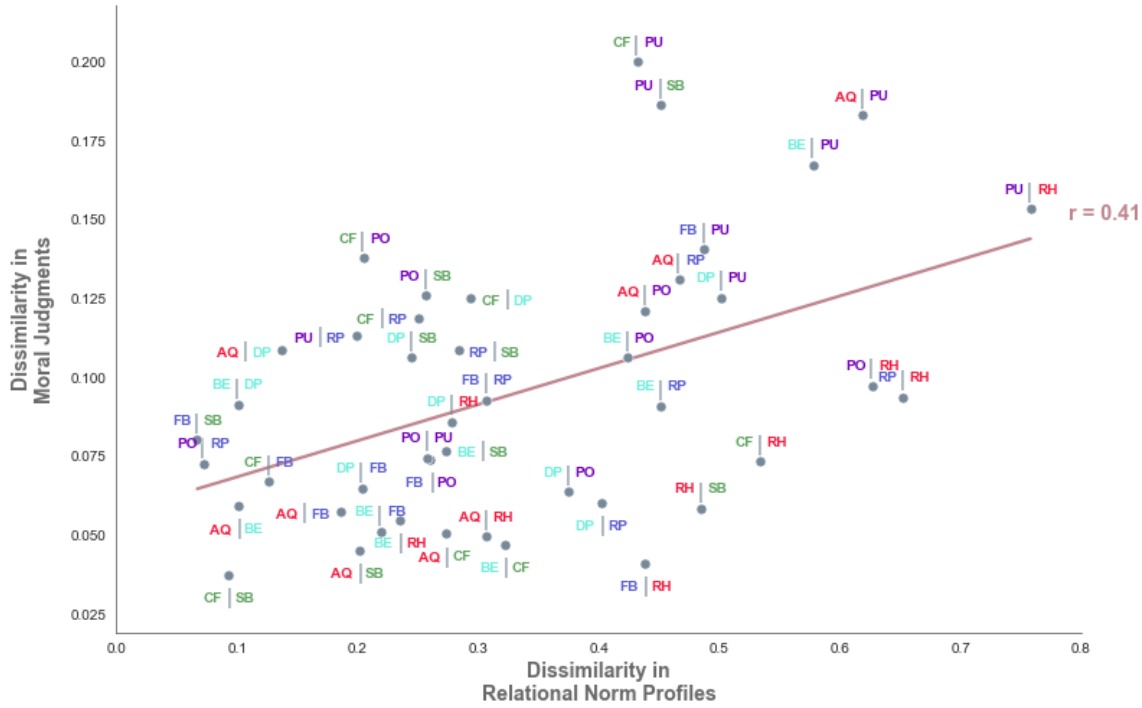


b. Hierarchy

c. Mating



d. Transaction



Figures 8a - 8d. Relational norm and moral judgment dissimilarity for function-weakening actions. Scatterplots showing the correlation in K-S distance between each pair of relationship dyads in relational norm space (x-axis) and the K-S distance between those same dyads in moral judgment space (y-axis) for each cooperative function. Spearman's *r* is reported in each case. Note that the color of each relationship reflects the cluster within which it is located from Figures 5a-5b.

Discussion.

In this study, we replicated our previously reported findings (Earp et al., 2020), while also going beyond them in two key ways. First, whereas in our RN 1.0 study, we were only able to successfully predict moral judgments from relational norms (using the K-S distance correlation approach) for three of the four functions (care, hierarchy, and mating, but not reciprocity), in the current study, our predictions were successful for all four functions (care, hierarchy, mating, and the new transaction function, replacing reciprocity). Second, whereas in our previous study, the predicted moral judgments were moral wrongness judgments specifically, in the current study, we were able to predict moral judgments that ranged from "very blameworthy" through to "very praiseworthy." However, consistent with our previous study, these judgments only pertained to actions that had been rated as characteristically *weakening* one or more cooperative functions. It is not yet known whether our relational norms model can also successfully predict moral judgments pertaining to actions that characteristically *strengthen* one or more cooperative functions. We address this issue next.

Study 3: Moral Judgments of Function-Strengthening Actions

The purpose of Study 3 is to determine whether it is possible to predict moral judgments of both blame and praise for actions rated as characteristically strengthening one or more cooperative functions using the same relational norm approach as in Study 2. As this is an exploratory study, although we pre-registered the same linear regression and K-S distance analyses used in Study 2, we did not make a specific prediction about the anticipated results. However, we did pre-register a

directional hypothesis for a novel analysis, described in more detail below, as follows: "We predict that the more a function is normatively expected within a cluster of relationships, the higher the moral judgment rating for strengthening that function will be on the blame-to-praise scale."

Method.

Open science. Hypothesis, design, measures, sampling strategy, analysis plan, and exclusion criteria were pre-registered at aspredicted.org (#69821). All materials, data, and analysis code are available at https://osf.io/zxjt6/.

Stimulus development. Using the same algorithm described for selecting function-weakening items in Study 2, but this time operating over candidate function-strengthening items, we selected 12 such items for use in the present study. See Figure 9.

Participants. Participant recruitment was the same as in Study 2, with the same target sample of 1,800 participants. Ultimately, 1,901 participants completed the survey, each of whom was paid $1.30. Four hundred and seventy (470) participants were excluded prior to data analysis based on pre-registered exclusion criteria (see Appendix 2, Supplementary Section 3.1.1 for details), leaving us with a final sample of 1,431 participants. See Table 4 for key demographics.

Table 4

Key demographics of Study 3 participants

Age	N (%)	Race	N (%)	Gender	N (%)
18 - 27	252 (17.82%)	White	1,017 (71.92%)	Female	712 (50.35%)
28 - 37	542 (38.33%)	Black/ African-American	187 (13.15%)	Male	696 (49.22%)
38 - 47	316 (22.35%)	Asian	120 (8.49%)	Other	6 (0.42%)

48 - 57	185 (13.08%)	Hispanic/ Latinx	48 (3.39%)
58+	119 (8.42%)	Other	25 (1.77%)
		American Indian / Alaska Native	16 (1.13%).
		Hawaiian / Pacific Islander	2 (0.14%)

Procedure. The procedure was identical to Study 2.

Results.

Stimulus selection: action item characteristicness. Results are displayed in Figure 9.

	Actions judged to characteristically strengthen one or more cooperative functions	Care	Hierarchy	Mating	Transaction
		Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
1	Person A lends Person B an important item that Person B needs, telling them to use it as long as they require it, with no questions asked.	80.58 (21.85)	13.92 (27.04)	19.41 (20.37)	-11.42 (63.78)
2	Person A goes out of their way to find a thoughtful gift for Person B to cheer them up when they are feeling depressed.	85.5 (10.84)	12.33 (29.68)	38.75 (32.28)	-9.75 (40.91)
3	Person A provides cold-weather clothing for Person B when Person B can't acquire it for themselves.	82.83 (20.27)	12.42 (24.03)	36.08 (35.31)	3.17 (35.29)
4	Person A takes care to follow Person B's instructions as exactly as they can.	30.08 (35.04)	82.08 (30.26)	14.33 (21.03)	16.17 (24.72)
5	Person A agrees to respect Person B's authority.	2.75 (19.37)	86.92 (22.96)	-1.67 (16.35)	10.08 (27.92)
6	Person A accepts that Person B should have "final say" about what happens in a given context.	-3 (18.66)	73.33 (24.35)	-5.83 (5.43)	-4.67 (9.02)
7	Person A expresses romantic feelings for Person B.	9.08 (27.99)	-20.58 (32.76)	71.58 (29.27)	-6.92 (17.69)
8	Person A tells Person B about a sexual fantasy they had involving Person B.	-3.33 (5.88)	-16.67 (37.01)	80 (25.49)	2.92 (10.32)
9	Person A offers to give Person B a sensual massage.	7.42 (26.77)	-7 (39.22)	82.83 (19.11)	2.67 (19.04)
10	Person A makes a point of returning a favor that Person B performed the month before.	7.41 (29.74)	23.58 (28.21)	16.42 (15.21)	88.58 (16.18)
11	Person A agrees to help Person B if Person B compensates them for their time and energy.	-53.67 (43.41)	24.83 (49.29)	-36.25 (37.28)	84.83 (18.37)
12	Person A asks Person B for compensation after teaching Person B a skill.	-47.75 (31.87)	18.25 (37.81)	-24.5 (28.17)	71 (39.42)

Figure 9. Characteristic function-strengthening actions. Heatmap showing mean ratings of judges (n = 12) of the extent to which each action would characteristically strengthen the care, hierarchy, mating, and transaction functions, respectively, between any two people. These items were chosen using the same method as used in Study 2, where +100 represents the most characteristic function-strengthening effect. Darker shades represent more extreme ratings in the strengthening direction.

Moral judgment ratings. Results are displayed in Figure 10.

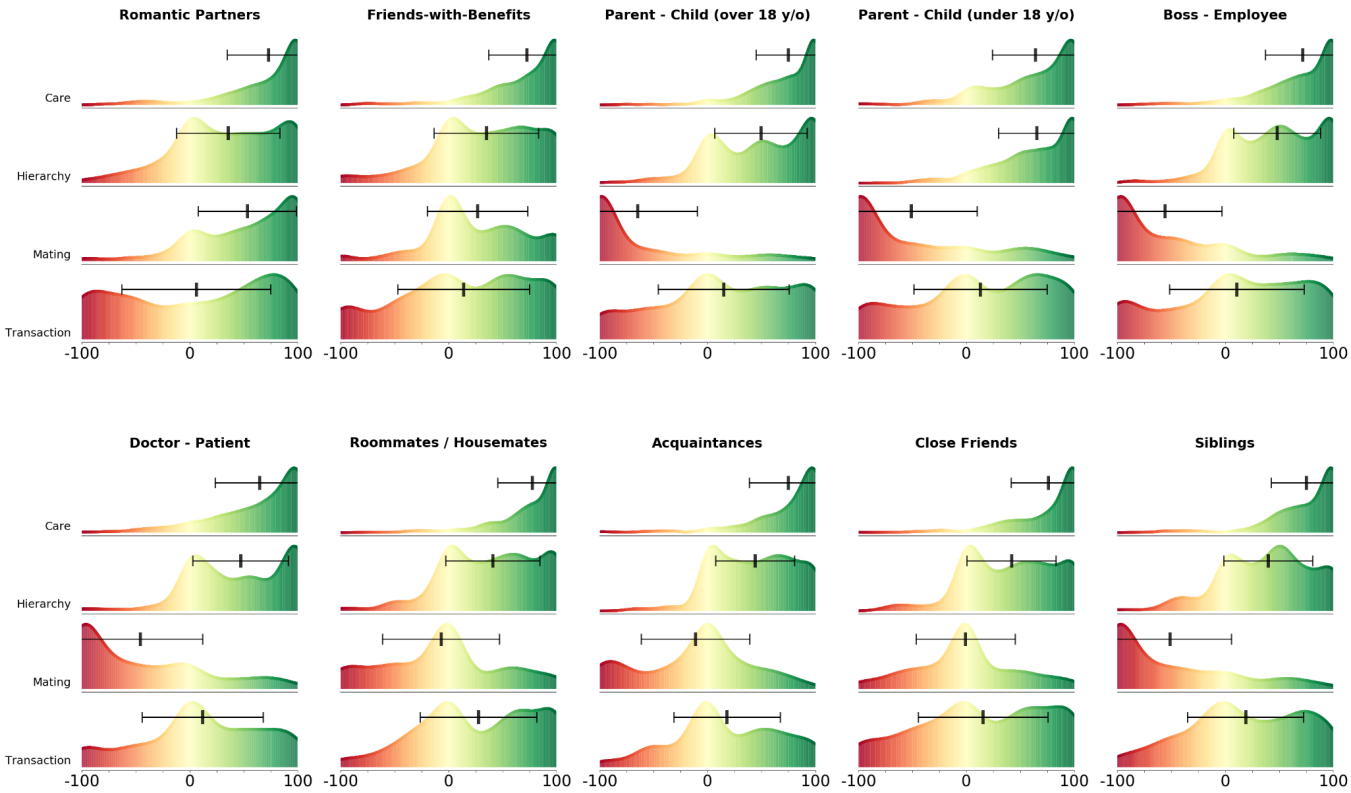
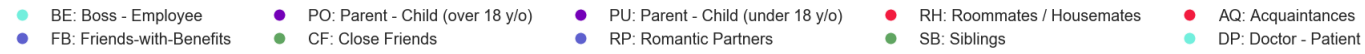


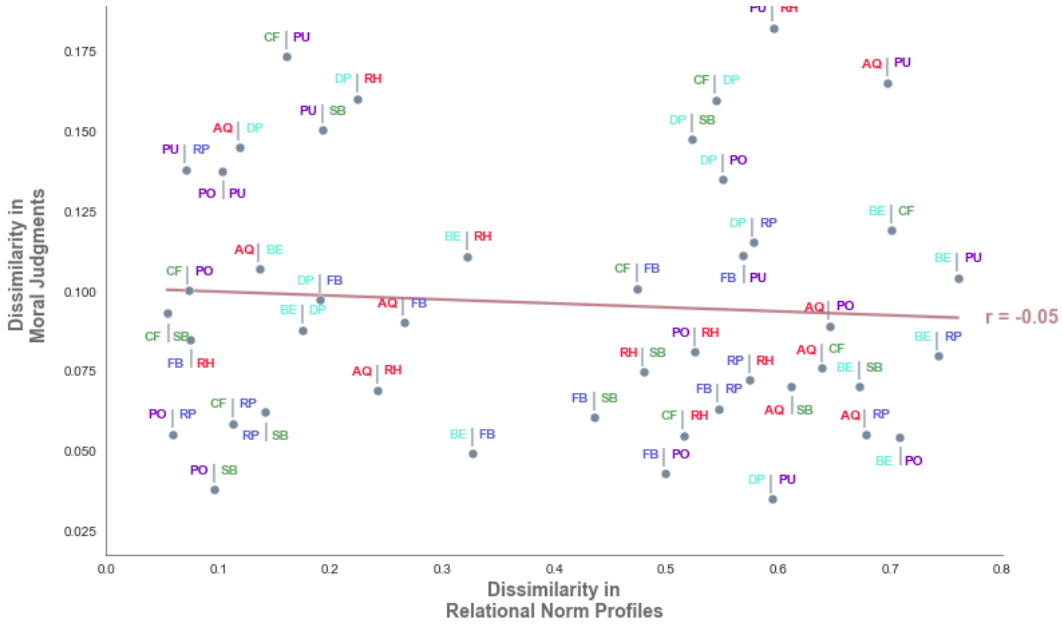
Figure 10. Moral judgments of praise and blame for function-strengthening actions. Study 3 moral judgments of actions rated as characteristically strengthening one or more cooperative functions in different relationships: kernel density plot of moral judgments (-100 = very blameworthy, +100 = very praiseworthy). Dot represents the mean, with 95% confidence intervals. Height of the curve represents density (see Figure 3 for explanation). This experiment was conducted once, with all data shown here.

Predicting moral judgments: linear regression approach. As with Study 2, we sought to predict Study 3 moral judgments (here, for function-strengthening actions) directly from Study 1 relational norms in a linear mixed regression model (same as described above). The results from this model supported a predictive relationship for some of the functions, but not others. Specifically, relational norms from Study 1 significantly predicted the moral judgments of Study 3 participants for care ($p < .001$, 95% CI [-.14, -.04]), hierarchy ($p = .02$, 95% CI [.007, .07]; however, note that this result does not survive a Bonferroni correction), and mating ($p < .001$, 95% CI [.20, .26]), but not for transaction ($p = .13$, 95% CI [-.007, .06]). The same basic pattern of results is observed when controlling for the same demographic factors as above (see Appendix 2, Supplementary Section 3.2 for the full regression tables).

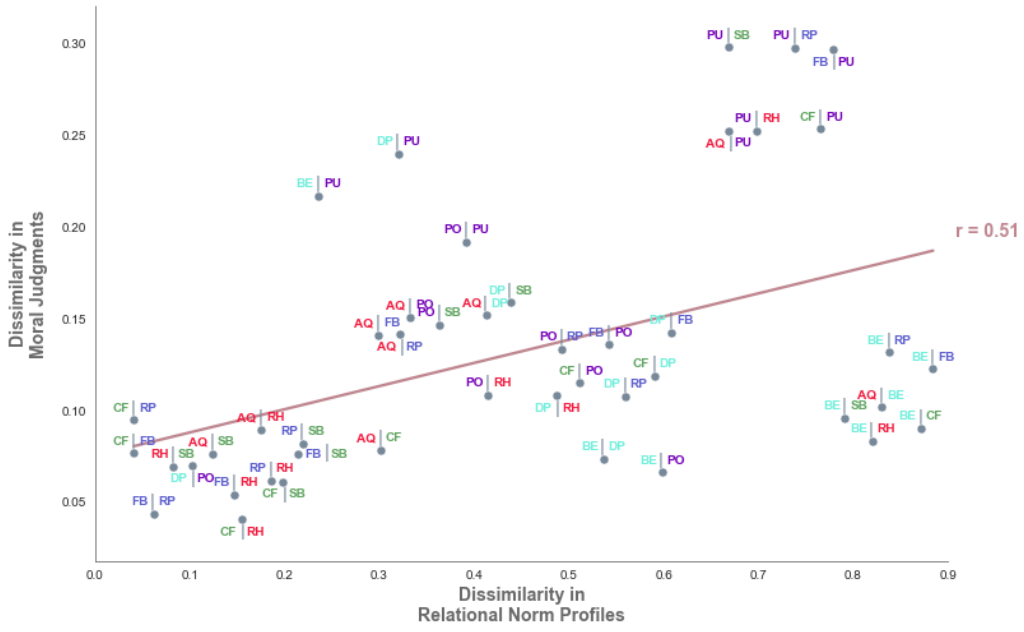
Predicting moral judgments: K-S distance correlation approach. As before, we also sought to predict the K-S distance between each pair of relationships in moral judgment space (based on current ratings) from their corresponding K-S distances in relational norm space (from Study 1). As can be seen in Figures 11a-11d, this predictive relationship holds for some, but not others, of the functions. In particular, it holds for mating (r = .86, p < .001), and hierarchy (r = .51, p < .001), but not for care (r = -.05 p = .80) nor transaction (r = .18, p = .24).



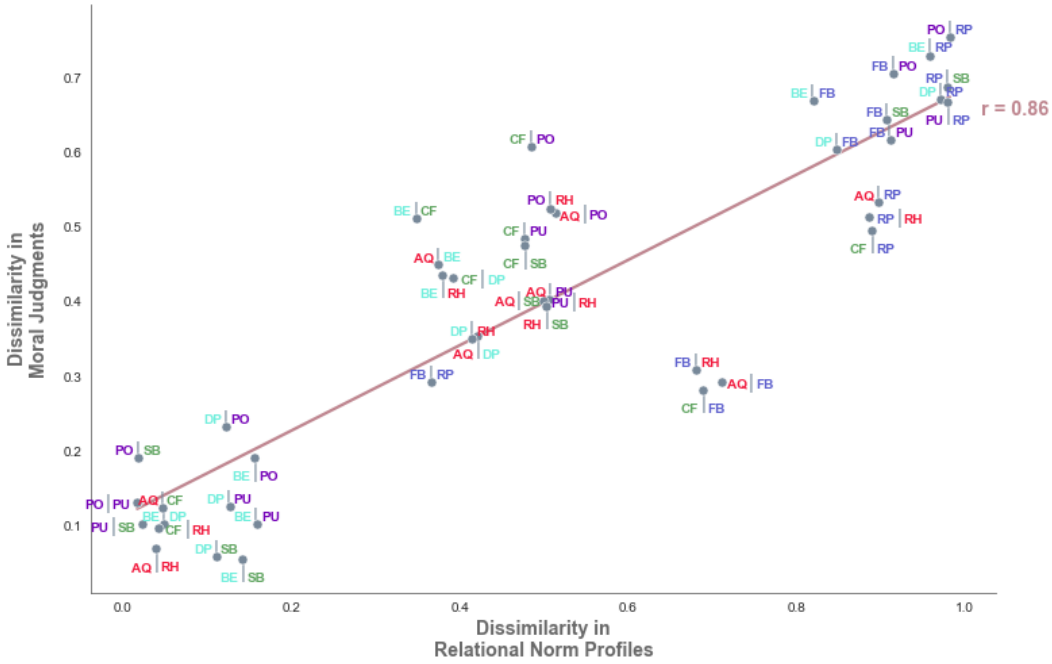
a. Care



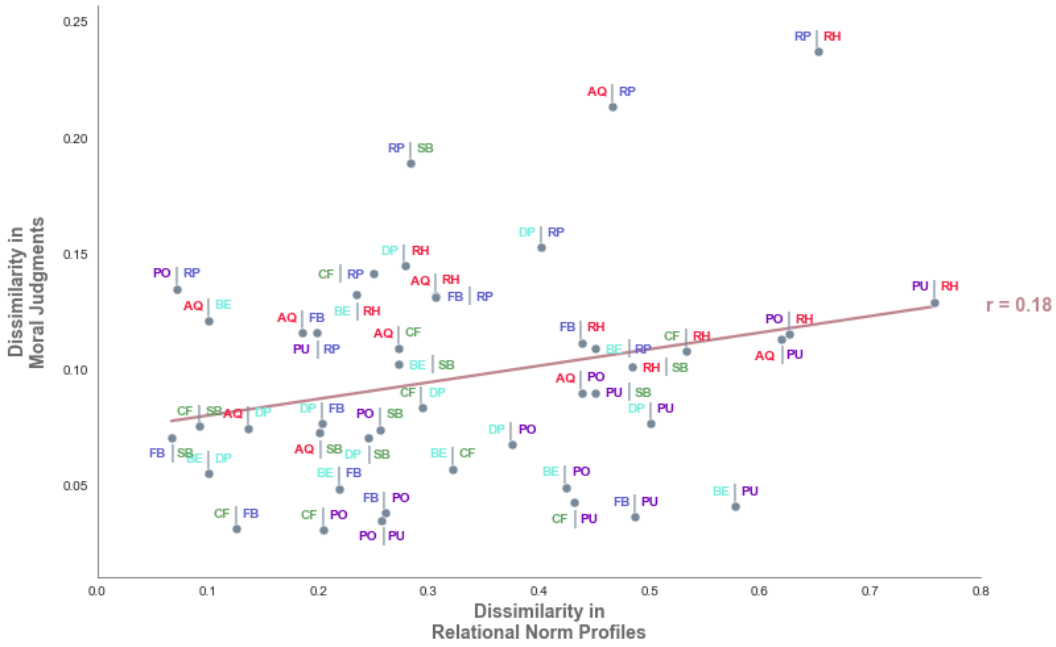
b. Hierarchy



c. Mating



d. Transaction



Figures 11a - 11d. Relational norm and moral judgment dissimilarity for function-strengthening actions. Scatterplots showing the correlation in K-S distance between each pair of relationship dyads in relational norm space (x-axis) and the K-S distance between those same dyads in moral judgment space (y-axis) for each relational function. Spearman's *r* is reported in each case. Note that the color of each relationship reflects the cluster in which it is located from Figures 5a-5b.

Predicting moral judgments: between-cluster comparison approach. As mentioned previously, in addition to the linear regression and K-S distance correlation approaches, we pre-registered a third analysis as follows. "For each function, we are interested in whether moral judgments will differ for actions previously judged to characteristically strengthen the function depending on whether the action occurs in relationships with relatively higher or lower normative expectations for that function (based on Study 1 ratings). For each function, we will divide relationships into 2+ clusters (e.g., 'low,' 'medium,' 'high') representing relative normative expectations for that function using [the same hierarchical clustering algorithm from Study 1] and conduct between-cluster statistical comparisons for moral judgments. We predict that the more a function is normatively expected within a cluster of relationships, the higher the moral judgment rating for strengthening that function will be on the blame-to-praise scale."

In adopting this approach, our aim was to test a simpler, exploratory hypothesis in a relatively straightforward way. Rather than comparing every relationship to every other relationship in both relational norm and moral judgment space -- as in the K-S distance correlation approach -- the idea here was to make a smaller number of comparisons, between *groups* of relationships, clustered together on the basis of their relative normative expectations for each cooperative function (from Study 1).

As noted in the Introduction, we expected that the mating function would show the clearest between-cluster pattern of results on this approach. This is because there are 3 highly distinct clusters of relationships spanning the full range of relational norm ratings for mating, as can be seen in Figure 3 from Study 1. In particular, there is a small cluster of relationships for which mating is strongly positively expected or

prescribed (long-term romantic partners and friends-with-benefits), a much larger cluster of relationships for which mating is strongly negatively expected or proscribed (e.g., kinship relationships, professional relationships), and a third cluster of relationships with a relatively large proportion of judgments around the middle of the scale, representing "It depends" (e.g., acquaintances, close friends). It follows from our theoretical expectations that strengthening the mating function in the first cluster of relationships should be seen as morally positive, eliciting judgments of praiseworthiness, whereas doing so in the second cluster of relationships should be seen as morally negative, eliciting judgments of blameworthiness, while judgments for the third cluster of relationships should fall somewhere in between. As seen in Figure 12a, this is precisely what we find.

In particular, a Mann-Whitney U test reveals a statistically significant difference in median moral judgments of mating behavior between, on the one hand, the relatively high-expectation (mating prescribed) cluster of relationships (M_{dn} = 39.00; that is, mating was judged to be moderately praiseworthy within these relationships) and, on the other hand, the relatively low-expectation (mating proscribed) cluster of relationships (M_{dn} = -66.33; that is, mating was judged to be highly blameworthy within these relationships), U = 30,270.5, $p < .001$ (note that all statistically significant results reported in this section survive a Bonferroni correction). Moreover, both judgments differ in turn from the median moral judgment regarding the intermediate cluster of relationships (M_{dn} = -11.33), that is, between the low-expectation cluster and the medium-expectation cluster (U = 50,017, $p < .001$) and between the medium-expectation cluster and the high-expectation cluster (U = 14,574, $p = .003$).

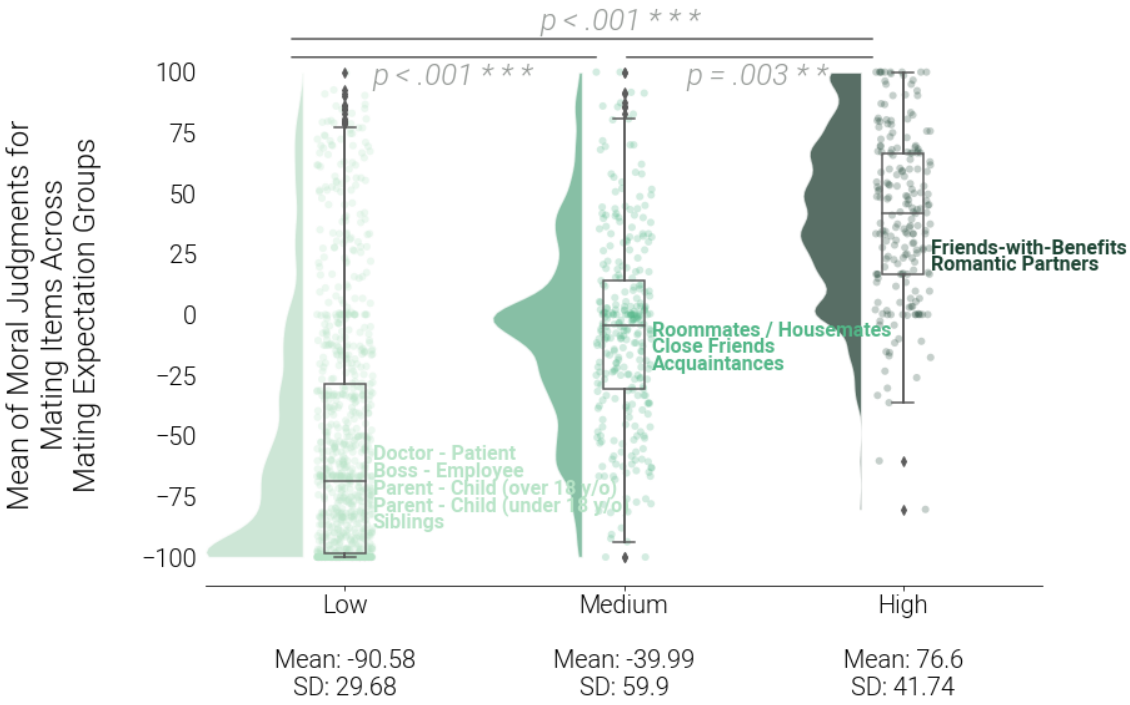


Figure 12a. Mating: moral judgments (y-axis) for strengthening mating across three clusters of relationships identified by hierarchical clustering analysis as having relatively low, medium, and high normative expectations for mating based on Study 1 ratings (x-axis). The y-axis runs from -100 ("very blameworthy") to +100 ("very praiseworthy"). The means displayed under the x-axis are based on the raw relational norm ratings (normative functional expectations) for the dyads within each cluster and have been added to aid interpretation; the cluster analysis, by contrast, was based on the K-S coefficient representing the distance between each pairing of relationships in relational norm space. The *p* values reflect the median statistical differences in *moral judgments* (y-axis) between clusters, according a Mann-Whitney U test (i.e., they do not represent differences in the mean normative functional expectations listed below the x-axis).

For hierarchy and transaction (see Figures 12b and 12c), the same basic pattern is observed, but with proportionally fewer judgments of outright blameworthiness for strengthening the functions in the 'wrong' relationship. This might be explained by the milder nature of the proscriptions (based on Study 1 relational norm ratings) regarding these functions, compared to the analogous proscriptions regarding mating, in the relationships for which each function is relatively non-normative.

In particular, for hierarchy, a Mann-Whitney U test reveals a statistically significant difference in median moral judgments regarding hierarchical (i.e., self-subordinating) behavior between, on the one hand, the relatively high-expectation

(hierarchy prescribed) cluster of relationships (M_{dn} = 45.33; that is, such behavior was judged to be praiseworthy within these relationships) and, on the other hand, the relatively low-expectation (hierarchy proscribed) cluster of relationships (M_{dn} = 40.5; that is, the behavior was judged to be less praiseworthy within these relationships), U = 82,410.5, $p < .001$. Please note that, for hierarchy, only two groups were identified by the cluster analysis, as opposed to three, as was the case for mating.

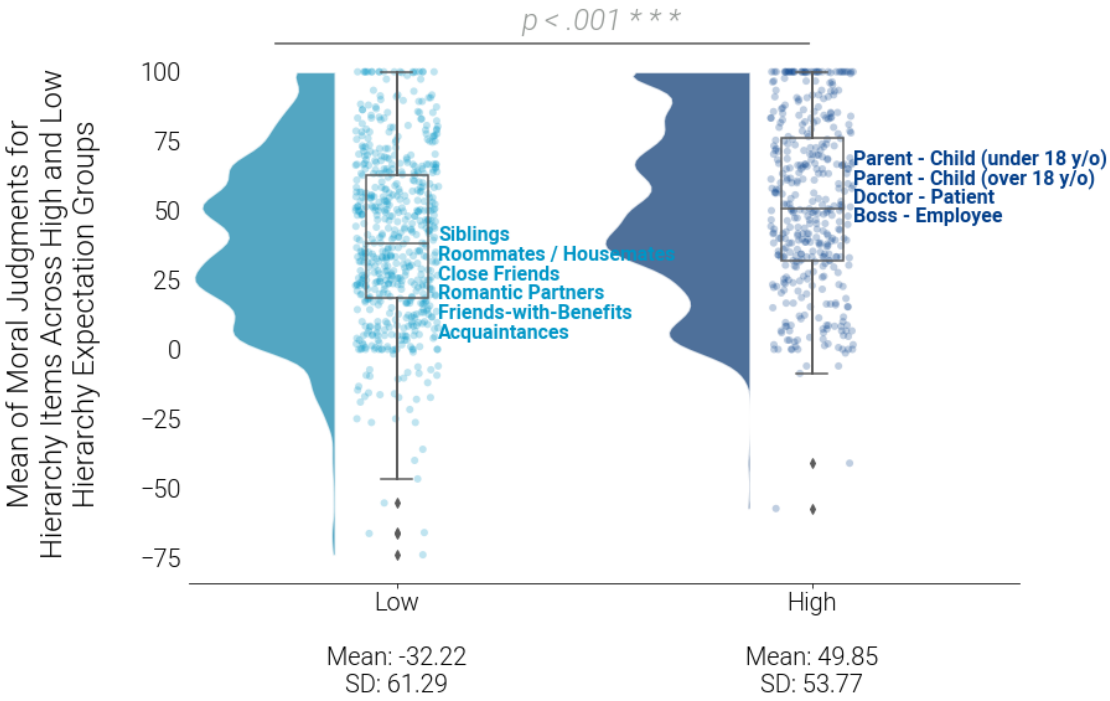


Figure 12b. Hierarchy: moral judgments (y-axis) for strengthening hierarchy across two clusters of relationships identified by hierarchical clustering analysis as having relatively low versus high normative expectations for hierarchy based on Study 1 ratings (x-axis). See Figure 12a for more information.

For transaction, like mating, a Mann-Whitney U test reveals a statistically significant difference in in median moral judgments regarding transactional behavior between, on the one hand, the relatively high-expectation (transaction prescribed) cluster of relationships (M_{dn} = 38.83; that is, transactional behavior was judged to be moderately praiseworthy within these relationships) and, on the other hand, the relatively low-expectation (transaction proscribed) cluster of relationships (M_{dn} =

9.00; that is, such behavior was judged to be much less praiseworthy within these relationships), U = 9,231.5, $p < .001$. Moreover, the median moral judgment regarding the high-expectation group differs in turn from the median moral judgment regarding transactional behavior in the intermediate cluster ($M_{dn} = 14.67$), that is, between the high-expectation cluster and the medium-expectation cluster (U = 16,285.5, $p < .001$), but, unlike mating, not between the medium-expectation cluster and the low-expectation cluster (U = 88,289, $p = .36$).

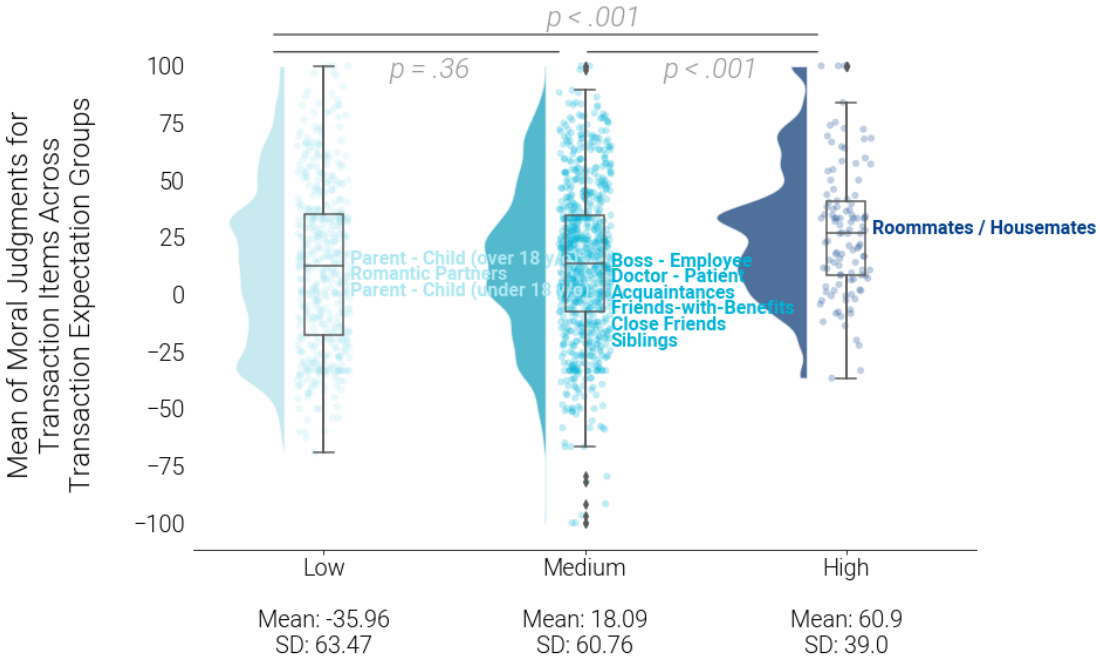


Figure 12c. Transaction: moral judgments (y-axis) for strengthening transaction across three clusters of relationships identified by hierarchical clustering analysis as having relatively low, medium, and high normative expectations for transaction based on Study 1 ratings (x-axis). See Figure 12a for more information.

Finally, as anticipated in the Introduction, the care function does stand apart. In Study 1, there were no relationships, apart from the customer-seller relationship (not included in the subset chosen for Studies 2 and 3), for which the mean relational norm rating for care was negative rather than positive. Instead, as we noted, care is typically welcomed in any relational context -- and if anything, it might be seen as even *more* praiseworthy to show care in those relationships for which it is least

normatively expected (as such behavior can plausibly be interpreted as going above and beyond the call of duty). As can be seen in Figure 12d, no statistically significant difference in median praiseworthiness ratings could be detected between clusters of relationships within which care is relatively less (M_{dn} = 68.00) versus more (M_{dn} = 77.67) normatively expected (p = .65), possibly due to a ceiling effect.

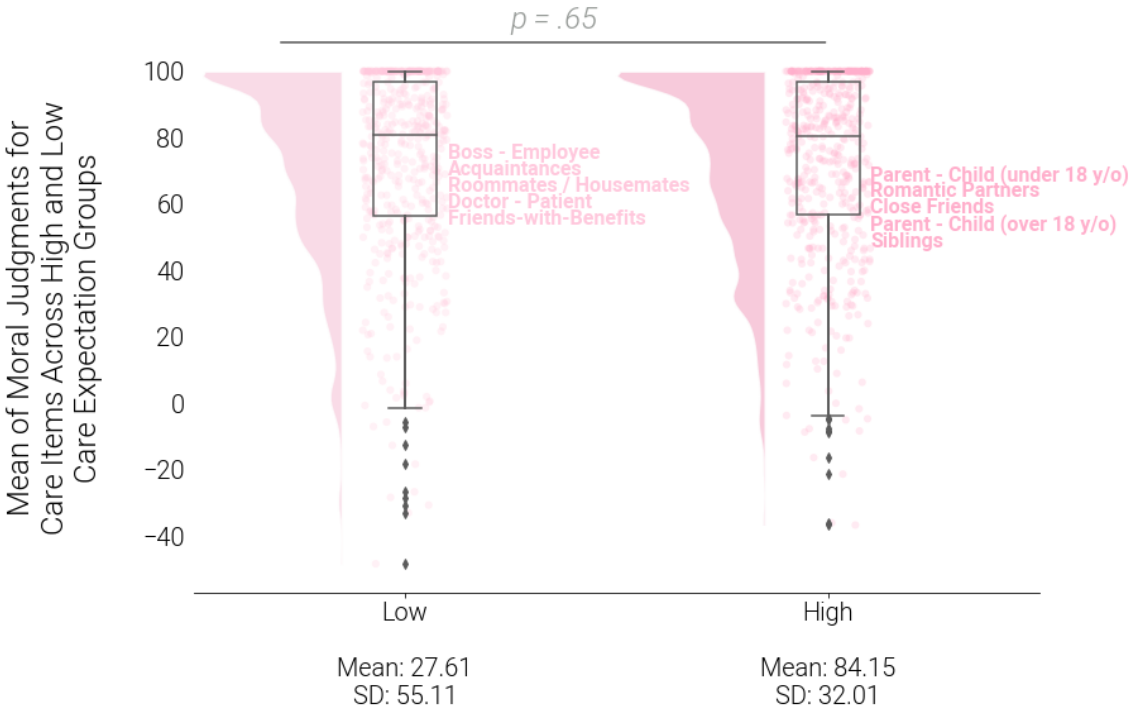


Figure 12d. Care: moral judgments (y-axis) for strengthening care across two clusters of relationships identified by hierarchical clustering analysis as having relatively low versus high normative expectations for care based on Study 1 ratings (x-axis). See Figure 12a for more information.

Please note: for each of the above figures, the "low," "medium," and "high" designations are *relative to each function*, and are derived from a hierarchical cluster analysis of the K-S distances between all included relations in terms of their Study 1 normative expectations for the function in question; whereas, the displayed means (and SDs) below the x-axis are based on the raw relational norm ratings for each set of relationships so identified, and have been added solely for the purpose of making the figures easier to interpret. Finally, to reiterate, the p values relate to statistical differences in median moral judgment ratings between clusters (the dependent

measure; see y-axis), not relative strength of relational norms (the independent measure; see x-axis).

Discussion.

In this study, we sought to determine whether judgments of blame and praise regarding characteristically function-strengthening actions could be predicted from previously measured relational norms. In Study 2, which focused on function-weakening actions, we found that we were able to predict such moral judgments for all four functions -- care, hierarchy, mating, and transaction -- both in a linear mixed regression model and by correlating K-S distances between relationships in relational norm and moral judgment space, respectively. In this study, however, which focused on function-strengthening actions, the predictive relationship between relational norms and moral judgments was less consistent.

For mating and hierarchy, relational norms did successfully predict moral judgments using the linear regression, K-S distance correlation, and the new between-cluster comparison approaches. However, for care, only the linear regression approach was successful, and for transaction, only the between-cluster comparison approach was successful.

Focusing first on transaction, we found that, consistent with our pre-registered hypothesis, the most positive moral judgments regarding transaction-strengthening behavior were observed within the relationship-cluster for which transaction was rated, in Study 1, as being the most strongly normatively expected (i.e., the roommates/housemates relationship). We also found that, for the parent-child and romantic partner relationships (low-expectation cluster, mean expectation of transaction = -35.96, *SD* = 63.47), about a third of the distribution of moral judgments

extends below 0 into “blameworthy” territory. Although the median moral judgment for this cluster was slightly above zero (with wide confidence intervals), the observed substantial proportion of blameworthy judgments is consistent with past research suggesting that exchange-based, transactional behavior in certain paradigmatically close -- and thus normatively communal -- relationships will in many cases be judged negatively overall (McGraw & Tetlock, 2005). With respect to the linear regression analysis, although, for transaction, the result was not statistically significant, the slope of the relationship between Study 1 relational norms and Study 3 moral judgments was in the expected direction (i.e., the more transaction is normatively expected within a relationship, the more praiseworthy it is to behave in a transactional manner).

Turning now to care, it is remarkable that, with respect to the linear regression model, only care had a negative slope ($p < .001$, 95% CI [-.14, -.04]). In other words, only in the case of care did a greater normative expectation for the function correspond to *less* positive moral judgment for strengthening the function in the relevant relational context. Although this result was not statistically significant in the between-cluster comparison model, that analysis did not allow us to control for the effect of other variables, such as action likelihood and target specificity. Whereas, when we do control for these factors in the pre-registered linear regression model, the negative relationship between normative expectations for care and degree of praiseworthiness for strengthening care emerges. Again, this finding suggests that it is *less* praiseworthy to behave in a caring manner in the context of relationships for which such behavior is strongly normatively expected than it is to do so in the context of relationships where care is less normatively expected (consistent with McManus et al., 2021).

General Discussion

We had three main goals with this research. The first was to test an updated relational norms model (RN 2.0), including a new transaction function (replacing reciprocity in RN 1.0). The second was to see whether we could predict, not only negative moral judgments regarding various actions in social-relational context, as in our previous model, but also positive moral judgments of praiseworthiness. The third was to determine whether our updated relational norms model could predict moral judgments not only of actions that characteristically *weaken* one or more cooperative functions (care, hierarchy, mating, and transaction) but also actions that characteristically *strengthen* one or more such functions.

With respect to the first two aims, we showed that moral judgments of both blame and praise regarding characteristic function-weakening actions can successfully be predicted from relational norms for all four of care, hierarchy, mating, and transaction. In particular, the more strongly a given relationship was normatively expected to serve each cooperative function, the more blameworthy it was judged to weaken the function; whereas, the more strongly a function was negatively expected (i.e., proscribed) within a relationship, the more positive the moral judgment for weakening the function.

With respect to the third aim, we found that, for hierarchy and mating, the more strongly prescribed the function, the more positive the moral judgment of function-strengthening behavior, regardless of the analytical approach used; whereas, the more strongly proscribed the function (e.g., mating within family relationships), the more negative the moral judgment. The same basic pattern was observed for the transaction function, but only one pre-registered analysis yielded a statistically significant result. Finally, for care, a striking departure from the other functions was

observed: in the pre-registered linear regression model, the stronger the normative expectation of care in a given relationship, the *less* praiseworthy it was judged to be to behave in a caring manner; whereas, the weaker the normative expectation of care (e.g., in the acquaintance relationship), the *more* praiseworthy it was judged to be to show care. This finding is consistent with other recent empirical work in relational moral psychology, for example, the study by McManus et al (2021), wherein an agent who helped a stranger (relatively low normative expectation of care) was judged as a morally better agent than one who helped a family member (relatively high normative expectation of care).

One immediate lesson from this pattern of results for care compared to the other cooperative functions is that an overly narrow focus on harming vs. helping behavior, as is typical for studies in this area, will likely result in an impoverished, and even potentially misleading, understanding of moral judgment in relational context. Although care is, on any view, a significant component of human morality, it is not the only component (Haidt & Joseph, 2007; Rai & Fiske, 2011b; Shweder et al., 1997). Rather, hierarchy, mating, and transaction are also important cooperative functions that help us solve common coordination problems. Moreover, as we have seen, actions that characteristically strengthen these functions do not seem to elicit the same patterns of moral judgment as do actions that strengthen care in relational context. Whereas caring behavior was rated as more praiseworthy the *less* normatively expected the care function, for the other functions, the opposite pattern was observed. In other words, hierarchical, mating, and transactional behaviors were rated as more praiseworthy the *more* normatively expected the respective function. Such a finding highlights the need to study multiple cooperative functions together --

i.e., within a single study or paradigm -- to unearth both similarities and differences between them while holding all else equal.

Limitations and future directions.

We believe this study marks a significant step forward in our understanding of how human social relationships shape moral judgment, both negative (as in judgments of blameworthiness) and positive (as in judgments of praiseworthiness). Yet there are also significant limitations. One of them concerns generalizability, as our samples were drawn exclusively from the U.S. population. Accordingly, an important future direction for this research will be to see how relational norms may differ across cultural contexts, and, if so, whether these differences can help us model and predict moral judgments (including potential cross-cultural moral disagreements). Another potential limitation is the small number of trained judges who rated candidate action items on their characteristicness (i.e., as function-weakeners or function-strengtheners). An advantage of the smaller number is that we could ensure adequate training for the judges, so as to increase the likelihood of getting high-quality ratings; however, it is unclear how representative these ratings are, even of the U.S. context. In future work, we plan to develop a larger pool of candidate items and cross-culturally validate them, while also recruiting and training a larger number of judges drawn from representative samples.

Acknowledgments. Thank you to Liane Young and members of the Young Lab at Boston College -- including Ryan McManus, Julia Marshall, and Gordon Kraft-Todd -- for very helpful feedback on this work. Thanks also to members of the Crockett and Bargh/Clark labs at Yale University. Finally, a sincere thank you to the 12 anonymous individuals who helped with stimulus rating, a very time-consuming task.

Chapter 4

True Love: A Normative Relational Concept

Abstract

In the previous chapters, we looked at moral norms that are embedded in a wide range of social relationships, such as siblings, neighbors, acquaintances, and romantic partners. In this chapter, I will zoom in on the last of these, the romantic partner relationship – given its special importance to so many of our lives – and explore how normativity is embedded within the very concepts we use to describe such relationship, focusing on the ordinary concept of true love.

When we say that what two people feel for each other is 'true love,' we seem to be doing more than simply clarifying that it is in fact love they feel, as opposed to something else. That is, an experience or relationship might be a genuine or actual instance of love without necessarily being an instance of true love. But what criteria do people use to determine whether something counts as true love?

This chapter explores three hypotheses. The first holds that the ordinary concept of true love picks out love that is highly prototypical. The second, that it picks out love that is especially good or valuable. The third, that people distinguish between psychological states that are 'real' or not, and that it picks out love that is real. Two experiments provide evidence against the first hypothesis and in favor of the second and third. Implications for real-life disagreements about love are also discussed.

Introduction

There's a difference, it seems, between *love* and *true love*. Just pick your favorite love story, from a book or a movie, or real life, where you find yourself most convinced of the special connection between the lovers. Where, however cynical or unromantic you may be, you might still be tempted to say such things as "They were made for each other," and mean it as more than a cliché. And now imagine that one of them dies. The other one grieves, for a good long while. Enough time passes, and the living partner starts a relationship with someone new.

Imagine that this new person is no mere rebound. They are deeply kind, attractive, intelligent, loyal. The surviving half of our original duo falls in love with them. And suppose they really are in love. In other words, what the two of them feel for each other, or what they have between them, counts as genuine (romantic) love on any plausible view. Even so, you might find yourself thinking, with a touch of sadness perhaps, that no matter how wonderful and worthy this new love-relationship is, the only time our protagonist experienced *true love* was with the one who died.

If you can get yourself to think that (you may have to use your imagination to fill in certain details), then you may be inclined to think that the concept *true love* is in some way distinct from the concept *love*. At least, that is how it seems to us: that you can have or experience the latter without the former. Indeed, people use the phrase 'true love' in ordinary discourse—in pop songs, poems, and private confessions—as though it expressed a concept all its own, and they seem to think this concept is getting at something important. Something that might justify a marriage, or cause an affair, or inspire a move between countries or a change of careers.

Much seems to hang on this concept, but what are its contours? What (if anything) does it refer to? There has been a mountain of scholarship, in philosophy

and other disciplines, on the nature of love, but there has been relatively little work on true love as a topic in its own right.

Of course, that is not to say that existing philosophical work never uses the phrase 'true love.' This phrase has occasionally appeared within existing work, but most of these uses are not invoking the concept that will be our primary concern here. Rather, the aim is often to distinguish actual cases of love from phenomena that may superficially appear to be love, but which are really something else: lust, say, or infatuation, or an unhealthy desire to possess the other person. For example, Velleman (1999) writes: "Students and teachers may of course feel desires for intimacy with one another, but such desires are unlikely to be an expression of true love in this context; usually, they express transference-love, in which the other is a target of fantasies" (p. 362). Similarly, Anglin (1991) argues that if an apparent case of love is the result of some deterministic process, "then it is not true love but mere love-behavior" (p. 20). In these examples, we suggest, the aim is not to explore a distinct concept of *true love* but is rather to understand the concept *love* and, specifically, to do so by distinguishing between actual love and the mere appearance of love.

We suspect there is more to true love than this. More, that is, than the mere marking of a boundary between genuine instances of love and its sundry pretenders. And if you bought into our opening example, you should agree. But if the 'true' in true love is not a mere synonym for 'actual' and suchlike—what is it?

There are various ways of tackling this question. To keep things focused, we will be looking at one particular kind of love—so-called romantic love—as illustrated by our opening example. This is not to say that the love between a parent and child, for instance, could never appropriately be described as 'true'. Perhaps it could, and pursuing this suggestion might ultimately shed light on the *scope* of the concept of

true love: that is, on the range of cases or kinds of love to which the concept applies. But even within the category of romantic love, it seems to us that some examples are liable to be described as 'true', while other examples, though still counting as legitimate (i.e., actual) cases of romantic love, are not liable to be described that way. We are interested in what distinguishes these two sorts of cases.

As an additional constraint, we will concern ourselves with one particular aspect of this puzzle, namely, with the *ordinary* concept of true love as it applies within this romantic domain. By this, we mean the concept as it exists in the minds of everyday speakers of English, as revealed by the criteria they use to determine which things count as true love and which do not. To make progress on this question, we will be exploring the patterns in people's ordinary judgments about true love.

Naturally, this will involve looking both at cases of agreement and at cases of disagreement in such judgments. In some cases, people overwhelmingly agree as to whether something counts as true love or not, and in those cases, an account of the ordinary concept should explain why people make the judgments they do. But of course, when it comes to questions of true love, we also often find considerable disagreement. Often -- as we shall see in our results below -- different people look at the very same phenomenon and make opposite judgments about whether it counts as true love. An account of the ordinary concept should also help us understand what it is that people are disagreeing about in these cases. This will be a core aspect of our inquiry.

If we do successfully uncover at least some of the criteria implicit in the ordinary concept, we immediately face a further question as to whether these criteria are the right ones or whether there might be reason to revise them or perhaps to abandon them, or even abandon the concept itself. These are important questions, and

we will turn to them in the final section of our paper. But before we can ask whether the ordinary criteria are right or wrong, we will need to have a better understanding of what those ordinary criteria actually are.

Three hypotheses

In our attempt to understand the ordinary concept of true (romantic) love, we will consider three main hypotheses. The first hypothesis says that true love, on the ordinary concept, is highly *prototypical* love; the second hypothesis says that it is especially *good, valuable,* or *praiseworthy* love, whether or not it is prototypical; the third hypothesis says that, independent of goodness or prototypicality, true love is love that is rooted in the *real*, in a sense we will be discussing further below. We begin by simply laying out these three hypotheses.

Hypothesis 1: Prototypicality. One hypothesis would be that true love is simply highly prototypical love. On this hypothesis, the criteria associated with the concept of love itself are best understood as a matter of degree. If a relationship or experience satisfies these criteria to a certain degree, people might be willing to say that it is an instance of love. But to count as *true love*, it would not be enough just to scrape over some minimal threshold; the relationship or experience would have to satisfy those criteria to a far greater degree.

According to prototype theory—by way of a brief review—members of a category are picked out by a number of features, each of which has a certain amount of weight (the greater the weight, the more important for category membership). Roughly speaking, the more features with the more weight an entity has, the more prototypical it is (Rosch & Mervis, 1975; Smith & Medin, 2013). So if true love is

prototypical love, it would be an instance of love that has most or all of the prototypical features of love that carry the most weight.

As an analogy, think of the concept of a *true jock*. Plausibly, the concept *jock* is a prototype concept. As such, the concept is associated with various features that count in favor of someone's being a member of the category (prioritizing athletics over other activities, holding certain objectifying attitudes towards women, not being particularly invested in high culture, and so on). One natural hypothesis would be that to be a true jock, one has to be a prototypical jock. On this hypothesis, if a person showed many of the features associated with the concept but not most or all, we might be willing on the whole to consider the person a jock, but we would not be willing to consider the person a true jock. Only a person who showed most or all of the features, and showed those features to a high degree, could be a true jock.

A question now arises as to whether a similar approach could be applied to the concept of true love. In support of the view that it can, research both in philosophy and in psychology has converged on the claim that the concept *love* is indeed a prototype concept (see below). There is now a good deal of evidence in favor of that claim. The key issue then is whether the concept *true love* is best understood in terms of this prototype.

Within philosophy, Chappell (2018) has defended an account of romantic love that distinguishes 'paradigm' cases from what she calls 'secondary' or 'marginal' cases. She provides strong arguments for the view that this distinction helps us make sense of certain core questions surrounding love. For example, it helps us tell whether someone is really experiencing romantic love in the fullest sense. Take a case in which someone feels strongly benevolent toward another but lacks intimacy or perhaps commitment. Chappell notes that "benevolence is one thing that we call

love," but goes on to argue that benevolence alone would not count as "full-blown love" (Chappell, 2018). Full-blown or paradigmatic love, she suggests, would require something more.

Research in psychology has provided evidence that supports this view. Such research suggests that the ordinary concept of love is indeed a prototype concept, and that it has a number of features apart from just benevolence. Among ordinary people, the most significant of these features appear to be *intimacy*, *passion*, and *commitment* (Aron & Westbay, 1996). Roughly speaking, *intimacy* involves feelings of closeness and connectedness, and a motive to promote the well-being of the other (i.e., a motive of benevolence). *Passion* encompasses romantic feelings, including physical attraction and sexual desire. And *commitment* refers to the promise or intention to stay together despite obstacles, along with the belief that the relationship will last (Sternberg, 1986).

What then does it mean for a person or couple to experience true love? In keeping with the jock analogy, as we noted, one hypothesis is that the person or couple experiences prototypical love. Perhaps people would be willing to categorize a relationship that exhibited just a few of the prototypical features of love as an instance of love, but only a relationship that had all of the features, and to high degree, as an instance of true love.[7]

[7] Note that this hypothesis is not committed to any specific view about which features are included in the prototype. For example, there are subtle but real differences between the account in Chappell (2018) versus Aron and Westbay (1996), and these accounts generate different predictions about which specific qualities of a relationship will most strongly influence people's judgments about its prototypicality. The hypothesis under discussion here does not rest on this, however. Rather, it says that the features of a relationship that influence people's prototypical love judgments—whatever those features turn out to be—will be the very same features that influence people's true love judgments (in the same way and to the same degree). So, although we happen to use the features of love unearthed by Aron and Westbay's classic empirical work to test this hypothesis, we might just as well have used the features proposed by Chappell, or even other features not included in either account (Earp & Savulescu, 2020; C. Jenkins, 2017). The key point is that, if prototypical love and true love are in fact the same concept, then, whatever the effect of a given set of features on judgments about the former, it should be roughly the same as the effect of equivalent features on judgments about the latter.

Let's try this idea out. Imagine a young couple. The partners are consumed by passionate, sexual feelings for each other, and they can't imagine the relationship ever ending. But they don't really know each other at a deeper level, so their feelings of intimacy and commitment are not well grounded. It might be right to say that there is at least some sense in which what they feel for each other is love—perhaps they are even 'in love' in a way that is often valorized in pop songs and movies[8]—but at the same time, without their having developed a stronger sense of mutual understanding and emotional closeness sufficient to ground a more durable commitment, it might be hard to characterize their relationship as an instance of *true love*.

Conversely, imagine a long-married couple that has considerable commitment toward their relationship, as evidenced by its sheer longevity, but who have emotionally drifted apart over the years and have a waning sense of romantic passion. Their relationship might well be an instance of love, but again, this would probably not be the first couple you would choose to illustrate the concept of true love.

By contrast, a couple that is emotionally intimate, profoundly committed, and smoldering with passion even after the so-called honeymoon phase—that is, a couple that strongly exhibits each of the most central, prototypical dimensions of the ordinary love concept—would seem to be a couple that experiences true love on almost any reasonable conception. Our first candidate hypothesis, then, is that true love is highly prototypical love.

Hypothesis 2: Goodness. The hypothesis that true love is prototypical love is a plausible first pass, or so we think. But upon reflection, it may not be the whole picture. Instead, it seems that we can imagine loving relationships that are not at all prototypical in the way we just described, but which, if you closely examine them and

[8] For a critical discussion of love being conceived this way, see (Cottingham, 2017).

come to appreciate what makes them valuable, good, or praiseworthy, would still seem to count as true love.

To illustrate this idea, we will tell you about a couple who escaped to the United States from Poland together after the invasion of the Nazis. They were set up by their respective families when they were younger, and went along with what was expected of them. They got married, moved in together, and developed a simple routine that became familiar. Their relationship didn't involve much deep conversation, and sexual contact was strictly biblical. But by the time the Nazis came, they had built a contented life together. No passion, not much in the way of (overt) emotional disclosure, but a committed partnership nevertheless.

Now imagine their harrowing escape; the miles they traveled together under harsh conditions; what they risked to keep each other alive; what they sacrificed in the way of personal freedom to make sure they found safety as a couple. At several points, we can suppose, each one had the opportunity to abandon the other for a more secure path forward. But they didn't hesitate to risk their lives to protect their relationship. Clearly something about their bond was profound.

Now, it seems clear that this is not a *prototypical* case of romantic love: the couple never poured their hearts out to each other, and sexual passion was never a feature of their relationship. But something about their quiet commitment, and the lengths they went to in order to keep each other safe from harm—and to preserve their way of life in a new country—might seem to warrant the claim that what they had between them was, nevertheless, true love. If our intuitions about this case are not idiosyncratic, there must be more to the concept of true love than mere prototypicality.

What might that something more be? One possibility is that it is something normative: something tied to the notion of goodness or praiseworthiness. In other words, when we say that what this couple has is true love, we are, perhaps among other things, expressing a favorable moral attitude toward their love or toward their relationship more broadly.

The notion that love simpliciter might be a normative concept has support from the existing literature. As Jenkins (2017) has noted, "the word 'love' packs a powerful rhetorical punch [and] its associated valence is typically positive rather than negative." To use the word 'love' in reference to an unhealthy or otherwise dysfunctional relationship, Jenkins argues, can be a "dangerously rhetorically effective way of concealing how bad" the relationship really is (pp. 94-95). Espousing a similar view, hooks (2000) argues that love requires honesty, trust, and respect, and is fundamentally inconsistent with certain negative attitudes or behaviors: "Abuse and neglect," hooks argues, "negate love" whereas care and affirmation, which are "the opposite of abuse and humiliation, are the foundation of love. No one can rightfully claim to be loving when behaving abusively" (p. 22).

Inspired by these ideas, one natural hypothesis would be that people reserve the phrase 'true love' for instances of love that excel along this normative dimension. In other words, perhaps people use this phrase only for instances of love that are especially admirable, or that most fully embody what is valuable, good, or praiseworthy about love. Beyond that, perhaps any clearly negative characteristic of a relationship rules out the applicability of the label "true love."

This hypothesis immediately generates predictions for our question about when people will agree versus disagree about whether something counts as true love. In certain cases, almost everyone will think that a certain instance of love manifests

something of deep value (perhaps our story about a couple escaping the Nazis would generate this reaction), and in those cases, the hypothesis predicts that almost everyone should agree that this instance counts as true love. By contrast, in other cases, people with opposing values will have correspondingly opposing views about whether a given instance of love manifests something of deep value. In those cases, the hypothesis predicts that different people should have very different judgments about whether the instance counts as true love. Those people who think that the case manifests something of deep value should say that it is true love, while those who think that it does not should disagree and say that it is not true love.

Importantly, however—and this something we will be testing below—the hypothesis predicts a substantial amount of agreement about whether something is true love *among those who agree about whether it is good or bad*. For example, among those people who think that a given instance of love is wrong or depraved, there should be strong agreement as to whether that instance of love counts as true love (i.e., agreement that it does not).

Hypothesis 3: Realness. Although there is certainly something tempting about the hypothesis that people use the phrase 'true love' only for relationships that they believe to be valuable, good, or praiseworthy, certain strands within existing research suggest a subtler view. As May (2013) has argued, there is a rich tradition in Western thought according to which love, and romantic love in particular, may be risky and all-consuming: dangerous to oneself or others and even threatening to the very fabric of society. Love can be a sort of madness. In fact, the idea that a bond must be 'healthy,' consistent with the well-being of the lovers, or something that is fit to be praised to count as love is in some respects a recent innovation. Could there be

relationships that are not good—or even highly dysfunctional in certain respects—where it would still be right to say that the couple experienced true love?

Consider Morgan and Robin. Until meeting one another, each of their prior romantic relationships had all been fairly uninspired. Suddenly here was a person who made them feel totally alive, filling them with an electric, almost addictive desire. They were that couple at the party who seem so in tune with one another that it makes you wonder about your own relationship. And yet, their love was also tumultuous. A day might begin happily and end in a bitter argument. Their fights occasionally spun out of control (once, Morgan had all the locks changed and Robin couldn't get back into the apartment for three days). But even in the darkest of times, they felt a passionate connection. Both were convinced that no one else could ever understand them—in all their unique peculiarity—quite so well; and they felt that if they weren't together, they would be missing out on what was most essential in life.

Suppose that, one day, exhausted from all the drama, they decide to break up for good. They both feel it is time to start building a stable future—to start looking for the kind of partner of whom their parents would approve. They don't feel an immediate connection to these new prospects, and they find themselves putting a lot more effort into enjoying one another's company. (Is it really necessary to spend multiple weekends together going in detail over potential mutual funds?) Although these relationships lack the intensity they once felt for each other, they are invested in making things work, and over the years, they come to really value their new lives. They can't help but marvel at how much happier they are now. And they aren't faking their feelings: they have in fact grown to love their new partners. Even so, we can imagine them thinking to themselves from time to time, perhaps lying awake at night

reflecting on old memories, that the other was their 'one true love.' Like the couple from the beginning of this paper.

If they would be reasonable in thinking that, how could this be explained? We can imagine different potential answers, but here is one to try: Although their relationship was in many respects unstable and unhealthy, what Robin and Morgan felt for each other was very *real*. Indeed, one can imagine them looking back at the time they spent together and thinking: "That was such a painful period, but even so, it was the only time in my life I felt fully in touch with something real." Perhaps this notion of what we will call 'realness' plays a role in people's ordinary concept of true love.

In saying this, we do not mean to be introducing a new technical term. Rather, the suggestion is that people ordinarily distinguish between psychological states, ways of relating, or even periods of their lives that are, in a particular sense, 'real' and those that are not. People might mark this distinction by using sentences like: "I was so angry about what happened, but at least I was feeling something *real*." Or: "I thought I was doing something meaningful with my life, but it was only when I quit that other job and started working full-time as an artist that I truly experienced anything *real*." Although this distinction can be applied to the case of love, or so we propose, the distinction itself does not seem specific to that emotion. Instead, it is a distinction that people can apply to a range of phenomena, including different psychological states (desire, happiness, sadness, hatred, and so forth).

Suppose we go with this hypothesis for the moment. The question that immediately arises is: How do people distinguish between those experiences, for example of love, that are real as opposed to not real—or perhaps less real? One approach to answering this question might be to invoke the notion of a 'true self.' A

body of empirical work suggests that people quite naturally think that some emotions, thoughts, or actions reflect an agent's true self, while others do not (Christy et al., 2019; De Freitas & Cikara, 2018; Strohminger et al., 2017). Very roughly, this research suggests that a person's true self is typically regarded as some fundamental part of who they are: not something due to mere socialization, or a desire to fit in, for example.

If people think that a given psychological state does not reflect the agent's true self, they will see that state as having a peculiar status. Take, for example, the experience of happiness, where this is judged not to reflect the agent's true self. Typically, people will say that there is a sense in which the agent is in fact happy—they don't deny that basic description—but they will also say that there is a deeper sense in which she isn't happy: the happiness is not rooted in her truest self.

Researchers have not reached a consensus about how best to make sense of this sort of judgment, and, beyond that, it is an open question whether judgments about the 'realness' of an experience should be understood in terms of the true self at all. We will not be attempting to address those issues here. Rather, we are raising the notion of a true self to give a sense of how one might try to explain what people *mean* when they judge that a psychological state is (or isn't) 'real'. But giving such an explanation is not the aim of this paper. Instead, our focus is on the more basic question of whether people's ordinary concept of true love is structured around such realness judgments.

Even in the absence of a detailed account of what realness is, however, the realness hypothesis makes certain testable predictions. Suppose people agree that what Robin and Morgan feel for each other is love, and our goal is to predict whether they will think it counts as true love. According to the realness hypothesis, their

judgments about this question should be predicted by their judgments about the realness of what Robin and Morgan feel. Moreover, judgments of realness should predict judgments of true love even controlling for prototypicality and goodness. To see this, suppose that people determine that Robin and Morgan's relationship is not a prototypical example of love and that, ultimately, it is not even good. It might seem, then, that they should also fail to regard the relationship, or perhaps what Robin and Morgan feel for each other within the context of the relationship, as an instance of true love. But the realness hypothesis makes a different prediction. It holds that there is a further sort of judgment people can make—a judgment about the realness of what Robin and Morgan feel—and to the extent that people judge this feeling to be real, they should judge that it is true love after all.

To bring out what is surprising and important in this hypothesis, it might be helpful to contrast the phrase 'true love' with other phrases that use the word 'true.' Suppose that John appears to be in some sense a jock, and we are wondering whether people will agree that he is a 'true jock.' Clearly, people's judgments about this would have nothing to do with whether they agreed with a statement like: 'John is real.' It is perfectly obvious that John himself is real, and the only question is whether he falls into a certain category. Thus, the best way to predict whether people think John is a true jock might be to see whether they agree with a statement like: 'John is an especially clear and paradigmatic example of a jock.'

On the realness hypothesis, the phrase 'true love' should be understood very differently. Suppose again that what Robin and Morgan feel for each other is in some sense love, and we want to predict whether people will judge that it is true love. The realness hypothesis predicts that such judgments will *not* turn on whether people think their feelings fit into some category (e.g., the category of love). Instead, it predicts

that people's judgments will depend on whether they think the feelings Robin and Morgan have for each other are *real*. In other words, people's judgments would not best be predicted by their agreement with a statement like: 'What Robin and Morgan feel for each other is an especially clear and paradigmatic example of love.' Rather, they should be predicted by agreement with a statement like: 'What Robin and Morgan feel for each other is real.'

The Present Studies

We have presented three hypotheses. The first is that true love, on the ordinary concept, is highly *prototypical* love. The second is that true love is love that is fundamentally *good*. The third is that true love is love that is *real*.

Although these three hypotheses differ from one another at a deeper theoretical level, they will often overlap in practice. For example, since the prototypical features of love are themselves typically considered good, our first and second hypotheses will lead to similar predictions in most cases. And our second and third hypotheses will lead to similar predictions in most cases as well: presumably, people will think that if a couple is experiencing love that is real, they are experiencing something good. They might even think that experiencing something real is good in itself.

To tease these hypotheses apart, then, it will be necessary to examine certain cases where prototypicality, goodness, and realness do not coincide, or where they independently vary, and assess the relative contribution of each dimension to intuitive judgments about the existence of true love in a given relationship. That is what we set out to do in a pair of empirical studies.

Study 1

Our first study looked at prototypicality and realness. We manipulated three features that were associated with prototypical love in previous studies (intimacy, passion, commitment) and also independently manipulated realness. Participants were then asked (a) whether the relationship was an example of prototypical love and (b) whether the relationship was an example of true love.

On the prototypicality hypothesis, according to which true love just is prototypical love, judgments about true love should show the same pattern as judgments about prototypical love. By contrast, on the realness hypothesis, judgments about true love might come apart from judgments about prototypical love, and we should instead find that such judgments are especially influenced by realness.

Method.

Open science. This study, including planned analyses and exclusion criteria, was pre-registered at http://aspredicted.org/blind.php?x=z68ka6. The open data and materials are available at https://osf.io/ezysq.

Participants. Eight hundred and four US participants were recruited on Mechanical Turk (MTurk) and received $0.35 for their time. Participants were excluded from the final sample prior to data analysis if they completed the survey in under 100 seconds (n = 74), provided an incorrect answer to a comprehension check (n = 269), or provided an incorrect answer to a test to ensure the participants were human rather than a 'bot' (n = 50). Our final sample included 481 participants (228 female, 248 male, 5 other; M_{age} = 35.94, SD = 11.06).

Procedure. Participants completed an online survey with a between-subjects design. In the first section, we familiarized participants with the notion of a "prototype" by presenting them with examples of more or less prototypical chairs (see the exact study materials online at the above link for specifics). In the next two sections, they read descriptions of hypothetical entities and judged the extent to which each entity is a prototypical example of a certain concept. They rated prototypicality on a sliding scale from 0-100 (0 = Not at all a prototypical x; 100 = Completely prototypical x). Participants also were asked to make an additional judgment about each entity unrelated to prototypically (also using a 100-point sliding scale). The purpose of these two sections was to ensure that participants were comfortable making prototypicality judgments before moving onto the main section of the survey. We also wanted them to expect a second, variable question that was unrelated to prototypicality so that the "true love" question would not stand out when they came to it. In the main section of the survey, participants read about a hypothetical relationship between Mario and Jasmine. Each participant was presented with one of sixteen conditions, which varied along four dimensions – intimacy, passion, commitment, and realness. See Table 1.

Table 1
Vignettes used in Study 1. Each participant was randomly assigned to receive one version of each of the four paragraphs, yielding sixteen different possibilities in total.

	High	**Low**
Intimacy	Mario and Jasmine have a warm, close, and comfortable relationship, where they trust each other and actively support each other's emotional needs. They communicate well and know they can count on each other when times get tough. And they often share deeply personal information, so they feel they really understand each other.	Mario and Jasmine don't always feel that warm and close in their relationship. They definitely care about each other's emotional needs, but they tend to wait for specific problems to come up before offering their support. They also struggle a bit with communication, and have some trust issues around this, mostly to do with feeling misunderstood. So sharing information that is too deep or personal can feel uncomfortable.

Passion	On top of that, they find each other very physically attractive. Even just seeing each other fills them with excitement – it feels almost magical. So they're always on each other's minds and they often fantasize about each other when they're apart. They have a hard time imagining life without each other, or anyone who could make them happier.	On top of that, they aren't particularly attracted to each other physically. They definitely enjoy seeing each other – it just doesn't have that sense of magic or excitement about it. So, while they have fond thoughts now and then when they cross each other's minds, fantasies are pretty rare. Sometimes, they find themselves imagining what a relationship would be like with someone they had more romantic feelings for.
Comm-itment	When they reflect on things, they realize they are committed to maintaining their connection, despite potential temptations, and even when they find each other hard to deal with. They feel confident in their love for each other and, somewhere deep down, believe it will last for the rest of their lives. At the end of the day, they feel a strong sense of responsibility for each other and plan to continue their relationship as long as they can.	When they reflect on things, they realize they are somewhat unsure about their actual commitment to the relationship. They understand that things might get rocky, or that others might come between them, and they don't want to set unrealistic expectations. Still, they love each other, and things feel pretty stable for now. But who knows about the rest of their lives? At the end of the day, they feel a certain amount of responsibility for each other, and they plan to continue their relationship as long as it works out.
Realness	One day, Mario was talking with his best friend Aaron. Aaron was telling him about an important event in his life from a couple of years back. "It's the only time in my life where things just felt really real, you know?" Aaron then asked Mario if his relationship with Jasmine made him feel that way. "You know what? Yes. Looking back on everything I've experienced in my life, I sometimes feel like my relationship with Jasmine is the only thing that's real."	One day, Mario was talking with his best friend Aaron. Aaron was telling him about an important event in his life from a couple of years back. "It's the only time in my life where things just felt really real, you know?" Aaron then asked Mario if his relationship with Jasmine made him feel that way. "You know what? That's a good question. I guess I need to think about it a little bit."

After reading the vignette, participants were asked to judge the extent to which Mario and Jasmine's relationship is an example of prototypical love, and the extent to which their relationship is an example of true love. Both questions were presented at the same time on the same page.

Prototypicality. *To what extent would you say that Mario and Jasmine's relationship is an example of prototypical love?*

True Love. *To what extent would you say that Mario and Jasmine's relationship is an example of true love?*

Participants rated prototypicality on a sliding scale from 0-100 (0 = Not at all prototypical love; 100 = Completely prototypical love). Similarly, they rated true love on a sliding scale from 0-100 (0 = Not at all true love; 100 = Completely true love).

Participants then completed a comprehension check in which they were asked whether Mario felt certain that his relationship with Jasmine was 'real.' They could either answer 'Yes' or 'No.' Because we are interested in the effect of realness on true love judgments and prototypicality judgments, it was essential that participants answered this question correctly for their given vignette. Those who answered incorrectly were excluded from the final sample, as noted above.

Finally, participants provided information about gender, age, and political orientation. They also completed a test to prove that they are human. Those who answered incorrectly were excluded from the final sample.

Results.

Although these data could be analyzed in a number of different ways, our concern here was with one specific question. The study looked at the influence of four different factors (intimacy, passion, commitment, realness) on judgments about two different questions (prototypical love, true love). For each of the different factors, we wanted to know whether it had the same impact on the two questions or whether it had different impacts.

We therefore used a mixed-model repeated measures ANOVA, with question type (prototypicality vs. true love) as a within-subjects factor and intimacy, passion, commitment, and realness as between-subjects factors. Our pre-registered prediction

was that the effect of realness would be greater on true love judgments than on prototypicality judgments.

There were significant main effects of question type, $F(1,464) = 6.55$, $p = .011$, $\eta p^2 = .014$, intimacy, $F(1,465) = 31.95$, $p < .001$, $\eta p^2 = .064$, passion, $F(1,465) = 54.31$, $p < .001$, $\eta p^2 = .105$, and realness, $F(1,465) = 91.63$, $p < .001$, $\eta p^2 = .165$. These were qualified by significant two-way interactions between intimacy and passion, $F(1,465) = 5.64$, $p = .018$, $\eta p^2 = .012$, and passion and realness, $F(1,465) = 10.94$, $p = .001$, $\eta p^2 = .023$. There were no other interactions nor main effects for the between-subjects comparisons.

Turning now to the key research question, we looked to see whether there were any interactions between question type and the other factors. As predicted, there was a significant interaction between question type and realness, $F(1,465) = 16.716$, $p < .001$, $\eta p^2 = .035$. There was also an interaction between question type and intimacy, $F(1,465) = 7.34$, $p = .007$, $\eta p^2 = .016$. To decompose these interactions, we conducted two separate 2 (realness: high, low) x 2 (intimacy: high, low) x 2 (passion: high, low) x 2 (commitment: high, low) ANOVAs on each question type (prototypicality, true love).

The effect sizes for each factor on judgments of prototypicality and true love are depicted in Figure 1. The panel on the left shows the degree to which each factor impacted people's judgments about prototypical love; the panel on the right shows the degree to which each factor impacted judgments about true love.

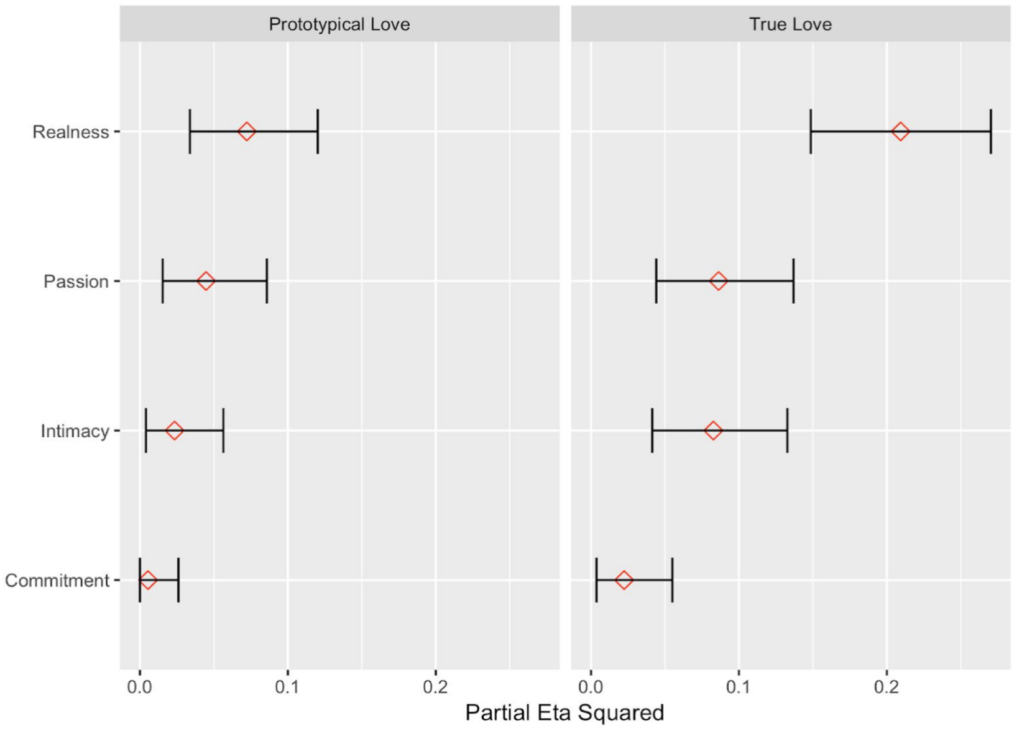


Figure 1. Effect sizes (ηp) of realness, passion, intimacy, and commitment on judgments of prototypicality and trueness in Study 1. Error bars show 95% confidence interval.

As the figure shows, the effect of intimacy on true love judgments, $F(1,465) = 43.30$, $p < .001$, $\eta p^2 = .085$, was greater than its effect on prototypicality judgments, $F(1,465) = 10.35$, $p < .001$, $\eta p^2 = .022$. And as predicted, the effect of realness on true love judgments, $F(1,465) = 118.08$, $p < .001$, $\eta p^2 = .203$, was much greater than its effect on prototypicality judgments, $F(1,465) = 32.46$, $p < .001$, $\eta p^2 = .065$.

Discussion.

In this first study, we found that the pattern of people's judgments about true love was quite different from the pattern of people's judgments about prototypical love. This finding provides strong evidence against the prototypicality hypothesis. Given the substantial difference between the pattern found for true love judgments

and the pattern found for prototypical love judgments, it is unlikely that the concept of true love is simply the concept of prototypical love.

Our data revealed two different respects in which the pattern of people's true love judgments departed from that of their prototypical love judgments. First, as predicted, realness had a far larger impact on true love judgments than on prototypical love judgments. Second, intimacy had a somewhat larger impact on true love judgments than on prototypical love judgments. It is possible that these are best understood as two independent effects, but it is also possible that the effect for intimacy could be understood as a byproduct of the effect on realness. That is, it might be that intimacy has a somewhat larger impact on true love judgments because intimacy is itself regarded, at least to some extent, as a cue to realness.

The fact that realness had such a large impact on true love judgments—far larger than the impact of any other factor—provides at least some prima facie support for the realness hypothesis. However, one might also think that this result is misleading. After all, as we alluded to above, realness could itself be regarded as something good, at least within the domain of love, so even if the goodness hypothesis were correct, one might still expect to find an impact of realness on true love judgments. We explore this issue more directly in the next study.

Study 2

In this second study, we turned to a different approach. We constructed a set of cases about which we expected to find a large amount of disagreement, with some participants saying that a given case was clearly an example of true love and other participants saying that the very same case was clearly not an example of true love.

We then asked whether each individual participant's true love judgment in these cases could be predicted by that participant's own judgments of goodness and of realness.

This method allows us to disentangle these two factors in a way that would not be possible with the method used in our previous study. If we simply tell participants in one condition that a couple is experiencing something real, the participants in that condition will presumably show a tendency on the whole to infer that the couple is experiencing something good, and vice versa. This fact limits our ability to distinguish the influence of these two factors. By contrast, in the present design, we can take advantage of the natural variance across participants in judgments of goodness and realness. In some cases, for example, we might find that some participants agree about whether a given case exhibits goodness, but disagree about whether it exhibits realness. We can then ask whether this natural variance in each type of judgment predicts attributions of true love.

The design of this second study sets up three potential predictions. One possibility is that, once one controls for goodness, the effect of realness on true love judgments is no longer significant. This would suggest that it is really the goodness of a relationship, rather than its realness, that is at the heart of such judgments. A second prediction is the inverse: that once one controls for realness, the effect of goodness disappears. This would suggest that realness is the driving factor. A third possibility is that each factor has an independent effect, even when controlling for the other. This would suggest that both factors actually play a role.

Method.

Open science. This study, including planned analyses and exclusion criteria, was pre-registered at http://aspredicted.org/blind.php?x=sr2ri7. The open data and materials are available at https://osf.io/ezysq.

Participants. Three hundred and fifty US participants were recruited on Mechanical Turk (Mturk) and received $0.35 for their time. Data from were excluded from the final sample prior to data analysis if they failed to complete the survey (n = 0), provided an incorrect answer to a comprehension check (n = 60), or provided an incorrect answer to a test to ensure the participants were human rather than a 'bot' (n = 11). Our final sample included 285 participants (134 female, 150 male, 1 other; M_{age} = 34.43, SD = 11.17).

Procedure. Participants were assigned randomly to one of three vignettes detailing a hypothetical relationship between Mario and Jasmine. The *abuse* vignette describes a passionate relationship interspersed with physical aggression. The *puppy love* vignette describes a simple but happy relationship between two elementary school children, unencumbered by the complexities of adult relationships. The *age difference* vignette describes a forbidden relationship between a professor and a student who seem to understand each other on a deeper level. The actual text of these vignettes was as follows:

> **Puppy Love.** When Jasmine was in 6th grade, she fell head over heels for a boy named Mario. Every day after school, they would take a walk in the park and let their imaginations run wild. Seeing each other was always the highlight of their day. Their bond was solidified during a school trip to France. They would sneak out in the dead of night and explore the streets of Paris together. Near the end of the trip, after a string of exhilarating escapades, they shared their first kiss. It felt so natural, so safe. Simultaneously innocent and totally electric.
>
> Nothing about their relationship was ever complicated. They never had to endure hardships together or make real sacrifices for each other. They never worried about whether they shared the same values or whether their life trajectories were in line. Such things never occurred to them. At that young age, the notions of sexual intimacy and long-term commitment weren't even on their radar. Just being together in the moment was enough. Everything was so simple and felt so fun and beautiful.

Now Jasmine is an adult, and in a committed relationship with a man named Jim. With Jim, things are not so simple. They care deeply about each other and feel warmly about each other on most days. They support each other through difficult times. But there is the usual mess of adult life to deal with: paying bills, getting along with in-laws, quarreling over little things after a long day at work. When she finds herself exhausted from all the tensions and complexities of her current relationship, Jasmine often thinks about her relationship with Mario from all those years back. She knows it seems silly, but sometimes, she feels as though her relationship with Mario was the only time she was ever really in love. It was pure in a way her adult relationships never were, or even could be.

Abuse. Jasmine has been in a romantic relationship with Mario for seven years. Mario is tough. It's part of why she was attracted to him in the first place. His brooding eyes, his physical strength. She knows that he would protect her from danger. When other men objectify her or make suggestive comments, Mario steps in without hesitation, and sends them scampering away at the mere sight of his imposing frame. He is loyal. A man of few words. But when he speaks, it is with intention. He also has deep practical knowledge, a way of being in tune with the environment. When Jasmine and Mario make love, it's like two parallel universes coming together and they lose themselves in the ecstasy of connection. Jasmine has never felt this alive with another man—a feeling of intensity and fullness that infuses her life with indescribable energy and meaning.

Mario is completely devoted to Jasmine. He has never had eyes for anyone else. He is usually kind and gentle, but sometimes, his emotions get the better of him. He punched a wall in their apartment once, breaking through the plaster (he quickly apologized and then repaired the wall himself). On another occasion, he knocked over a piece of furniture in frustration, causing a piece to crack. One time, Mario even hit Jasmine when he was really angry about something she had said, leaving a scar above one of her eyebrows. At first, she was in shock. She considered leaving him. But she decided to stay when he broke down and told her about his own abusive childhood and agreed to work on his anger.

In time, Jasmine came to think of Mario's aggressive episodes as somehow bound up with his protective nature. A kind of misdirection of the very strength and decisiveness that made her feel so safe when they weren't fighting. She even grew to like the little scar above her eyebrow—a reminder of Mario's ability to overpower her. This makes her feel vulnerable in a way that resonates with something deep inside her. His unpredictable aggression, interrupting long periods of quiet care and

companionship, makes her want to surrender herself to him, to give herself over to him completely. There is an ever-present, charged tension between them, part eroticism, part fear, part mutual obsession.

Age Gap. Mario is a 50-year-old professor at a prestigious university. He recently got to know a very bright 21-year-old undergraduate student from one of his classes named Jasmine. When they first met to discuss her senior thesis research over coffee, they immediately realized just how much chemistry they had, despite their very different ages and life experiences. Throughout his whole career, Mario has always felt distant from other people given his eccentric personality and unusual worldview. Most of his colleagues don't know what to make of him, but Jasmine seems to understand him on a deeper level.

Everything he says just clicks with her and she appreciates all of his strange idiosyncrasies. Furthermore, Mario is incredibly impressed by Jasmine's insight. (Her friends have always called her an 'old soul' and consider her wise beyond her years.) He often forgets that he is in the presence of an undergraduate student and views her as an equal. He has always fantasized about being with a much younger woman, and Jasmine has always had a thing for older men. Every time they met up, there was sexual tension in the air. One thing led to another, and now they're in a discreet romantic relationship.

Mario and Jasmine both know that they are violating university policy – especially given Mario's supervisory role over Jasmine – and they go to great lengths to conceal their relationship from other students, colleagues, and administrators. Ultimately, they feel that whatever might be met with disapproval about their relationship is overshadowed by the level of sync they feel together – intellectually, emotionally, spiritually, and physically.

After reading one of the above vignettes, participants were asked to judge the extent to which the relationship between the couple, named Mario and Jasmine in each vignette, was an example of true love.

True Love. *To what extent would you say that Mario and Jasmine's relationship is an example of true love?*

Participants made their ratings on a sliding scale from 0-100 (0 = Not at all true love; 100 = Completely true love). On the next page, they were asked to judge the extent to which Mario and Jasmine's relationship was characterized by realness and goodness.

> **Realness.** *When thinking about Jasmine and Mario's relationship, people might have different intuitions. Some people might think that their relationship is, in some respects, unconventional, but still that what they have between them is ultimately real. Others might disagree and say that, despite appearances, Jasmine and Mario aren't actually connecting on a real level. What do you think? Do you think that what Jasmine and Mario have between them is real?*

> **Goodness.** *When thinking about Jasmine and Mario's relationship, people might have different intuitions. Some people might think that there are certain flaws in how they relate to each other, but that, ultimately, their relationship is good. Others might disagree, and say that, although their relationship is positive in certain ways, ultimately, they have a bad relationship. What do you think? Do you think that what Jasmine and Mario have between them is good?*

Participants rated realness on a sliding scale from 0-100 (0 = Completely not real; 100 = Completely real) and goodness on a sliding scale from 0-100 (0 = Completely bad; 100 = Completely good). Participants then completed a comprehension check in which they were asked to judge whether a statement about the vignette was true or false. Those who answered incorrectly were excluded from the final sample. Finally, participants provided information about gender, age, and political orientation. They also completed a test to prove that they are human. Those who answered incorrectly were excluded from the final sample.

Results.

Data were analyzed using linear mixed effect models, with goodness and realness as fixed effects and vignette as a random effect (random intercepts only). All analyses were conducted in R using the lme4 and lmerTest packages.

There was a significant effect such that participants who gave higher goodness judgments also gave higher true love judgments, $B = 0.54$, $SE = 0.06$, $t = 8.91$, $p < 0.001$, $CI = [0.42, 0.66]$. However, even controlling for the effect of goodness, there was still a significant effect of realness on true love judgments: $B = 0.40$, $SE = 0.05$, $t = 7.32$, $p < 0.001$, $CI = [0.29, 0.50]$.

Figure 2 shows the results for all three variables. Looking at this figure, one can get a more qualitative sense of the patterns in people's judgments. For example, consider the puppy love vignette. In that vignette, almost all participants thought that the relationship was a very good one (i.e., the vast majority of points are toward the right-hand side on the x-axis). However, even among these participants, there was considerable disagreement about whether the couple had true love (as seen in the large amount of spread on the y-axis). Judgments of these cases were then predicted by realness (shown in the color of each point). That is, even among participants who agreed that the relationship was a good one, those who thought the couple were experiencing something real tended to say that they had true love, while those who thought that they were not experiencing something real tended to say that they did not have true love.

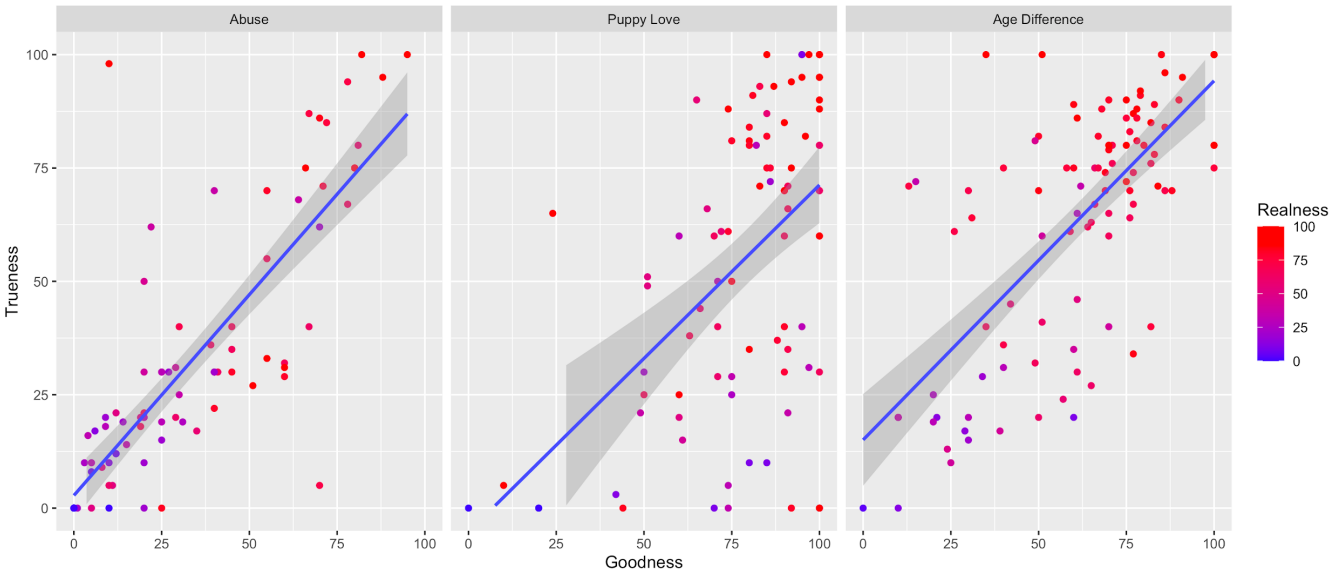


Figure 2. Scatterplot showing results from Study 2. X-axis shows goodness. Y-axis shows true love. Color shows realness.

Discussion.

In this second study, we looked at cases in which there was substantial disagreement between different participants as to whether something was an example of true love. We then asked whether participants' judgments in those cases were predicted by their goodness judgments and by their realness judgments. The results showed two different effects.

First, true love judgments were predicted by goodness judgments. This effect is very much in keeping with existing theoretical work on love (hooks, 2000; C. S. I. Jenkins, 2017) and provides evidence that existing theories are getting at something important about people's ordinary attributions.

Second, and notably, even controlling for goodness judgments, true love judgments were predicted by realness judgments. So we can tentatively conclude that, over and above the role of goodness in people's ordinary judgments of true love, there is also an important role for realness.

General Discussion

We began by noting that there is a conceptual difference between love and true love. Although the phrase 'true love' may sometimes be used to distinguish actual cases of love from merely apparent ones, we argued that *true love* is a concept in its own right, and a seemingly important one in many of our lives. How should this concept be understood? To answer this question, we tested three main hypotheses.

First, the hypothesis that true love is simply prototypical love. As we noted in the Introduction, previous work in both philosophy and psychology has argued that *love* is a prototype concept. The results of Study 1 strongly support this view: the more a relationship was characterized by paradigmatically loving features, the more the relationship was judged to be an instance of prototypical love. But equally strongly, the results of our first study contradict the hypothesis that *true love* and *prototypical love* are themselves the same concept: rather, these concepts are markedly distinct. Most notably, our manipulation of realness had very different effects on judgments about whether a relationship was an instance of prototypical versus true love. Since people's application of these concepts responded differently to the same manipulation, we have reason to reject the view that they are the same concept.

Second, the hypothesis that true love is love that is especially good or valuable. We found that perceived relationship goodness positively predicts judgments of true love, even controlling for perceived realness. This is exactly what should be expected given existing accounts of the normative significance of describing something as 'love' (hooks, 2000; C. S. I. Jenkins, 2017). Our results provide support for these accounts, and also for the claim that this same point applies to people's use of the phrase 'true love.' Further research should continue to explore

this effect. One key question will be whether the effect of goodness is best understood as reflecting something about the nature of people's very concept of true love or whether it is more a matter of people simply being reluctant to apply the words 'true love' to something they regard as bad.

Third, the hypothesis that true love is love that is real. The present findings provide strong support for this third hypothesis. In Study 1, the manipulation of realness had by far the largest effect on judgments of true love, going beyond such features as intimacy, passion, and commitment. In Study 2, realness judgments predicted true love judgments even when controlling for goodness judgments. Taken together, then, the results of these studies suggest a link between the ordinary concept of true love and judgments of realness.

Note that our results point to something distinctive about phrases like 'true love' that would not be seen with other sorts of phrases that include the word 'true.' For example, in Study 2, participants were not asked to judge the extent to which Mario and Jasmine have 'real love.' Instead, they were simply asked whether what Mario and Jasmine have between them is 'real.' In other words, participants who did not see their relationship as an instance of true love tended to think that what they had between them was just not real. By contrast, this sort of judgment would not make sense for other phrases that include the word 'true.' As we noted above, if people think that John is not a true jock, this would not be explained by their thinking that John himself is not real. Similarly, if people think that a certain sculpture is not a true work of art, it is likely not because they think the sculpture itself is not real, and so on.

It is an open question how we should understand people's judgments that certain emotions or experiences are not 'real.' We suggested earlier that one way to understand such judgments could be in terms of the notion of a 'true self' and we

sketched out a potential explanation along those lines. But we also noted that researchers disagree about how best to interpret 'true self' judgments, and we stated that we were not proposing to take a stand on whether people's ordinary judgments of realness should actually be understood in terms of this notion. We expect that the best approach to addressing such questions will be to expand the inquiry beyond the concept of true love and explore judgments of realness in other domains, or with respect to other kinds of emotions. That is, instead of just looking at judgments of realness insofar as they are relevant to the concept of true love, one might want to explore more generally why people see certain experiences as 'real' and others as 'not real' (or 'less real'). This is an important issue for further research.[9]

However, even in the absence of a fully worked-out account of realness, it seems that we can use the observed link between judgments of realness and judgments of true love to explain certain otherwise puzzling aspects of the ordinary concept of true love. Consider the different examples of true love we sketched out at the beginning of this paper: between the Polish couple and between Robin and Morgan. A remarkable fact about these relationships is that they had very different features, even seeming to be near-opposites. The Polish couple had little in the way of emotional closeness or intimacy, and virtually no romantic passion, yet were extraordinarily committed to the relationship. Robin and Morgan, by contrast, were extremely close emotionally and practically burning with romantic passion, yet ultimately, chose to end the relationship in order to find stability and calm with others. If we assume that the concept of true love is closely linked to judgments of

[9] In particular, it might be helpful to look at judgments of realness insofar as they are related to people's ordinary judgments of happiness. Existing studies show that people are reluctant to say that an agent is happy when that agent has a morally bad life (Phillips et al., 2017) and studies find that this tendency is mediated in part by judgments about whether agents actually are happy deep down in their true selves (Newman et al., 2015). This effect seems likely to be related in some important way to the ones we have been exploring in the present paper. For further discussion, see: (Phillips et al., 2011).

realness, we can begin to see why these apparently radically different relationships may both be seen as examples of true love. Though the two relationships differ when it comes to many of their salient features (intimacy, passion, commitment, and so on), there is another respect in which they are deeply similar. In both cases, the love that the people feel for each other seems to be real.

Conclusion.

The concept of true love is important. It matters to people's lives, and it is often cited as a justification for decisions or behaviors that might (otherwise) be seen as extreme or unwarranted. "Why did you leave your spouse of thirty years?" "Because I found true love with someone else." "Why did you quit your job and move to Europe?" "Because I found true love with someone who lives in Portugal." People will disagree about whether, or to what extent, such appeals can in fact justify certain acts or choices. And they will disagree about which relationships qualify as true love.

The present findings do not directly resolve these disagreements, but they do shed light on the nature of the disagreements themselves. Moreover, as we will discuss in the following chapter, these findings help us understand the criteria underlying the disagreements found in ordinary life, and they help us understand what we would be seeking to modify if we sought to modify those criteria. Putting this point in a slightly different way: the findings help us understand what we disagree about when we disagree about true love.

Acknowledgments. See the online, published version, available at https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199395729.001.0001/oxfordhb-9780199395729-e-38.

Chapter 5
Conclusion

Abstract

In this conclusion to the dissertation, I attempt to tie together the previous chapters, summarizing major findings from this line of work, discussing key limitations, and pointing to future directions. I also explore how empirical findings in relational moral psychology, such as those described in the present dissertation, can contribute to the analysis of normative questions (i.e., substantive moral questions of right and wrong) regarding social relationships and concepts used to characterize such relationships, such as the concept of 'true love.'

With respect to the last of these aims, I begin by noting that the dissertation has advanced a descriptive account of the sorts of moral judgments people make in different relationships, both positive and negative; and it has offered a functional explanation as to why they make those judgments. But are those the right judgments to make? How, if at all, can experimental moral psychologists help in answering such a value-laden question? One possibility is: by identifying factors that influence or underpin people's moral judgments about various cases, we may provide a basis for philosophers to determine whether these judgments should be accorded substantive normative weight, for example as part of a process of reflective equilibrium.

However, as I will discuss, making such a determination always requires granting certain prior, or more basic, moral and philosophical commitments, and/or assumptions that are held constant for the sake of the analysis or which are not in question in the given discourse. For example, it might be granted that the most prevalent moral judgments of a group of stakeholders should be accorded at least some prima facie normative weight in formulating a moral argument. Then, if a philosopher decides to reject such a common, robust moral judgment, she will have to provide an adequate error theory: i.e., an explanation of why people make the judgment despite its failure to yield (what the philosopher takes to be) the right normative conclusion. Here, I summarize and evaluate some recently proposed strategies for negotiating such decisions, highlighting potential pathways for reaching substantive normative conclusions from argumentative premises that include empirical claims about the moral-relational mind.

Introduction

Recent years have seen an explosion of research on the psychology of human moral judgments. Such judgments of right and wrong, good or bad, and praiseworthiness or blameworthiness, shape individuals and institutions regarding matters of great consequence, such as decisions to marry or divorce, decisions to hire or fire employees, criminal sentencing, foundational legal frameworks, political polarization, and global policy priorities. Most extant research on moral judgments has relied on paradigms that examine intergroup conflict versus cooperation or "prosocial" versus "antisocial" behavior. These paradigms involve, for example, decisions made in the context of economic games between anonymous strangers, or judgments about hypothetical agents, often taking place in unusual or life-threatening circumstances (as in the case of extensively studied moral dilemmas concerning the sacrifice of one individual to save many others).

Of course, many important moral judgments do concern ingroups and outgroups, and people do sometimes encounter strangers in need of help, albeit not usually in life-or-death situations. However, as emphasized in this dissertation, most of our everyday moral judgments concern actions (or failures to act) taking place between specific individuals who occupy a range of more or less well-defined social roles. Typically, these individuals know each other and interact repeatedly, and they stand in particular relationships with one another (such as boss-employee, spouses, student-teacher, parent-child, customer-seller, and so on). Moreover, moral judgments in different relational contexts expectably will differ from one another in systematic ways, such that one and the same action might be judged quite differently depending

on the relational context. Any adequate theory of moral psychology must be able to explain these differences precisely and systematically.

In this dissertation, I have presented work proposing that particular social relationships (like those between romantic partners, housemates, or siblings) are widely expected to serve one or more distinct *cooperative functions* – including care, transaction, hierarchy, and mating – to different degrees. Such multi-dimensional, relationship-specific patterns of prescribed cooperative functions ("relational norms") were hypothesized to influence downstream moral judgments: both moral wrongness (or blame) judgments in response to norm-violating actions, and judgments of praise for actions that fulfil or strengthen prescribed relational norms.

My collaborators and I found support for these hypotheses across a series of studies. In Chapter 2, I reported that relational norms for care, hierarchy, reciprocity, and mating allowed for highly precise out-of-sample predictions of moral wrongness judgments for norm violations in relational context. Moreover, our model explained more of the variance in relationally-situated moral judgments than models relying on other features of social relationships including genetic relatedness, social closeness (i.e., the depth of mutual understanding and acceptance between partners and the strength of their motivation to promote each other's well-being), or interdependence (i.e., the strength, breadth, and frequency of influence each partner has on the other's thoughts, feelings, and behaviors). Then, as reported in Chapter 3, we tested an improved model that replaced the reciprocity function with a more carefully defined 'transaction' function, allowing us to successfully predict out-of-sample judgments of both blame and praise for actions rated as characteristically weakening one or more cooperative functions.

When turning to moral judgments for actions rated as characteristically *strengthening* one or more cooperative functions, however, we found a striking inconsistency between functions that was not observed in the previous studies centered on function-weakening actions. In contrast to mating, hierarchy, and transaction, in the case of care, we found that the more strongly this function was normatively expected, the *less* positively it was judged to be to strengthen the function in the relevant relational context, when controlling for other relevant factors (e.g., the likelihood of the behavior in question). Put the other way around, the less that care was normatively expected within a relationship, the more positive the moral judgment for providing care in that relationship. How might this seeming paradox be explained? Perhaps such 'unexpected' provision of care within the relatively socially distant relationships (e.g., among acquaintances) was seen as superogatory behavior – that is, going above and beyond the call of duty – prompting more positive moral judgments. This is something we would like to test directly in future work.

Another issue that requires further analysis is the different patterning of judgments we observed for praise and blame in Chapter 3: although it was evident from our results that praise and blame are not simply mirror images of each other (consistent with Anderson et al., 2020), we were not able to make systematic statistical comparisons between praise and blame judgments across samples (i.e., controlling for relationship type while comparing moral judgments for weakening versus strengthening behavior). This leaves many open questions ripe for investigation. We might expect, for example, that providing care in a relationship for which care is strongly proscribed (e.g., the parent-child relationship) will not be seen as positive/praiseworthy to the same extent that failing to provide care in the same relationship would be seen as negative/blameworthy. This prediction is based on the

idea that merely adhering to what is normatively expected should not be as surprising as failing to adhere to, or actively violating, what is normatively expected, consistent with the "bad is stronger than good" dynamic of social cognition (Baumeister et al., 2001). How this predicted asymmetry in praise/blame judgments plays out across different relational contexts – for example, in the case of relationships for which a given cooperative function is more versus less strongly prescribed – will be an important research question going forward.

As we noted in Chapters 2 and 3, a major limitation of the current work is that all samples were drawn from the U.S. population, preventing assessment of cross-cultural generalizability of the findings. Moreover, most analyses focused on group-level responses (taking the mean or median rating, for instance), with less attention paid to individual differences in normative expectations or moral judgments. In the following section, I will consider some potential future directions for this work that would allow us to expand, and evaluate, our model along both cross-cultural and individual-difference dimensions. I will begin by considering individual differences.

Similarities and Differences Between Individuals and Cultures

Our model of multiple cooperative functions prescribed for various social relationships – and how these prescriptions shape relationally contextualized moral judgments – would be incomplete without accounting for *variation that exists between individuals* in both cooperative expectations and associated moral judgments. One possibility is that such variation will be systematically related to existing individual differences that characterize beneficial or harmful interactions within different types of relationships.

In Chapter 4, for example, we addressed the romantic partner relationship, finding substantial disagreement among participants as to the qualities such a relationship should have for it to count as an instance of 'true love' (a normative concept that we found tracks, in part, the extent to which a relationship is seen as morally good). Strikingly, some participants were willing to characterize even a physically abusive relationship in such terms, while others refused such a characterization. Given that the romantic partner relationship is, as our data show, strongly normatively expected to be governed by a norm of care, with very little variance in this judgment across participants (see Chapters 2 and 3), perhaps the observed variance in 'true love' ascriptions for the relationship described in our *abuse* vignette (Chapter 4) has to do with individual differences in beliefs about the extent to which abuse and care can coexist within a loving relationship (hooks, 2000; Taylor, 2017). Exploring such potential differences could shed light on how abuse is sometimes rationalized within romantic partnerships (Earp et al., 2017; Jenkins, 2017).

We will return to this question of disagreement about the moral status of romantic relationships later on, when discussing in more detail our 'true love' findings from Chapter 4. First, however, we will consider how individual differences in personality traits might relate – more generally – to differences in normative expectations for a wider range of social relationships. Consider individual differences in attachment style as an example. One plausible prediction is that the construct of avoidant attachment (Wei et al., 2007), which indexes abnormally low desire for dependency on others and avoidance of others' dependence on oneself, will predict variation relevant to the care function. For example, individuals who are high on avoidance might have lower functional expectations of care across a range of

relationships as compared to the group norm. This difference, in turn, might have downstream consequences for how these individuals perceive the moral status of behaviors that strengthen or weaken care in the relevant relationships. For example, they might regard a failure to show care within family relationships as less morally wrong than would those who are more securely attached.

We might also predict that social dominance orientation (SDO; Pratto et al., 1994), which indexes abnormally high desires for dominance over other individuals and groups, will predict variation relevant to the hierarchy function. Here, we might predict that individuals high in SDO will view hierarchy as more appropriate across a broad range of relationships compared to the group norm, with analogous downstream consequences for moral judgment. Finally, individual differences in sociosexuality (Simpson & Gangestad, 1992), which indexes sexual engagement with multiple partners and low commitment to them, might predict variation relevant to the mating function. For example, for individuals high in sociosexuality, we might expect that relationships such as close friends, colleagues, or acquaintances will be seen as appropriate candidates for serving the mating function to a greater degree than the group norm. This, in turn, might lead to less negative moral judgment of attempted mating behavior within such relationships (e.g., asking a colleague to go on a romantic date).

Beyond demonstrating an ability to predict variation on these measures, it will also be important to consider the implications of such variation for an individual's social and interpersonal functioning. Does having functional expectations for social relationships that more closely adhere to the group-level cooperative norm predict better social adjustment and well-being? Literature on the concept of cultural fit, or the relative match between a person’s individual characteristics or attitudes and their

cultural group (Searle & Ward, 1990), is suggestive. For example, cultural misfit of values (Stephens et al., 2012) and emotional responses (De Leersnyder et al., 2014) are predictive of poor adjustment and well-being. Here we might propose to approach the relational norms of individuals from a cultural fit perspective. That is, we could examine whether individuals whose personal relational norms fit better with those of their cultural group are better adjusted, both overall and in the context of particular social relationships.

Finally, what about differences between cultures? Societies across the globe vary in their patterns of relational norms: i.e., in how cooperative functions normatively are 'distributed' across, or embedded within, different social relationships (Miller et al., 1990; Miller & Bersoff, 2016). Indeed, there is a growing awareness of the limitations of psychological science which almost exclusively focuses on participants from Western, Educated, Industrialized, Rich Democracies (WEIRD; Henrich et al., 2010). Work in certain non-WEIRD contexts[10] reveals that people vary systematically across societies in the meaning structures and values that they use to guide social-relational behavior, including in the moral domain (e.g., Awad et al., 2018). As Fiske and Rai (2014) argue, "The most fundamental finding of anthropological research is the descriptive fact that morals are culturally relative … many actions that people judge to be right in any given culture are judged to be wrong in many others" (p. 7). This is no less true, of course, for relationally-situated moral judgments as it is of other potential kinds of moral judgments.

At the same time, at least some relational moral judgments are also *shared* across cultures: for example, the judgment that it is typically wrong for parents to fail to care for their child (assuming they are in a position to do so). Examining how

[10] For a critique of the "WEIRD vs. non-WEIRD" dichotomy, however, see Ghai (2021).

social relationships shape moral judgments across different societies thus represents an important opportunity to examine both cultural universals and cultural uniqueness in relational moral psychology. In future work, we propose to test the robustness of our model by sampling participants from a variety of cultures in which, we believe, all of the cooperative functions we include in our model must be served somehow or to some extent, but with variance in how each function is distributed between and/or embedded within relationships according to the culture.

Previous work has characterized several important cultural dimensions on which societies differ. Intriguingly, several of these dimensions show striking parallels with the cooperative functions that our findings suggest are important for explaining moral judgments in relational context. We plan to sample societies varying along the dimensions of relational mobility (Thomson et al., 2018), individualism/collectivism (Triandis & Gelfand, 1998), power distance (Hofstede, 1979; Rinne et al., 2012), and gender egalitarianism (McDaniel, 2008). We will also sample societies that vary along the dimension of tightness-looseness, which reflects the extent to which a society is characterized by strong norms that are enforced (Gelfand et al., 2006, 2011).

We anticipate that the relationships examined will be normatively expected to characteristically serve at least one cooperative function, demonstrating the universal nature of these dyadic functions. To illustrate, a relationship such as the one between neighbors might primarily be expected to serve the care function in one society and the transaction function in another society, but it will be expected to serve at least one cooperative function in all societies sampled. In addition, we predict that the *specific patterns* of cooperative functional expectations (i.e., relational norms) for each relationship will vary cross-culturally. For example, in highly relationally mobile

societies, in which individuals are easily able to select in and out of relationships based on their personal preferences, we would expect the care function to apply to a more diverse set of relationships. This is because, in such contexts, it is relatively risky to rely on too narrow a set of relationships for having one's needs met, as these relationships might change, end, or be replaced by others (Kito et al., 2017). Finally, we predict that, within each culture, moral judgments across relationships will track with the extent to which behaviors violate the relational norms of each relationship as prescribed by the culture in question (as we have found in our preliminary studies of U.S. participants). In this manner, our model holds promise for predicting variation in moral judgments both within and between cultures by identifying the culture-specific structure of relational norms.

To summarize, our broad predictions for future work are as follows: we expect that the "fit" between individual-level functional expectations and group-level cooperative norms will have consequences for social adjustment and well-being. And we expect that, across societies, we will be able to replicate the general finding that cooperative-functional similarity between relationships corresponds to similarity in moral judgments between relationships. Alongside this universal relationship, we also expect to observe variation across societies in cooperative relational norms and corresponding moral judgments, and that this cross-cultural variation, in turn, will be predicted by variation in broadscale cultural dimensions (e.g., relational mobility, individual/collectivism, power distance).

Of course, another way to describe such predicted 'variation' between cultures in prevailing relational norms and associated moral judgments is to speak of cross-cultural *disagreement* about substantive questions of right or wrong. Such substantive moral disagreements can occur within a single society as well, as we saw in the case

of between-participant differences in 'true love' ascriptions applied to an abusive relationship (Chapter 4). In the following sections, I will consider whether empirical work in relational moral psychology – such as that what has been presented in this dissertation – can contribute, not only to understanding the nature of such disagreements (i.e., what is at stake in them), but also, perhaps, to resolving them in certain cases by factoring into arguments for moral conclusions. Put differently, how, if at all, can experimental moral psychologists fruitfully contribute the project of answering substantive moral questions about (behavior within) relationships?

Normative Implications

The dissertation has, so far, advanced a descriptive account of certain characteristic moral judgments – both positive and negative – that people make in different social relationships; and it has offered a scientific, cooperative-functional explanation as to why it is they make those judgments. But are those the right judgments to make? The question is not straightforward. As Farber (1994) notes, "to go beyond description, to enter the arena of the normative, that is, to say what *ought* to be, involves an important shift that requires justification" (p. 156, emphasis added). But, he argues, scientific-explanatory accounts of people's moral judgments – the kind put forward in this dissertation – offer "no new basis, no new foundation, no new hope" of providing such "normative" justification (Farber, 1994, p. 156). Or as Korsgaard (1996) puts it: "When we seek a philosophical foundation for morality we are not looking merely for an *explanation* of moral practices. We are asking what *justifies* the claims that morality makes on us" (pp. 9-10, emphasis added). She argues that when we engage in moral philosophy, "[we] do not merely want to know why

those peculiar animals, human beings, think that they ought to do certain things. We want to know what, if anything, we really ought to do" (Korsgaard, 1996, p. 13).

Consider our finding from Chapters 2 and 3 that U.S. participants, on the whole, believe that certain social relationships, such as the parent-child relationship, *should* be governed by certain cooperative functions (e.g., care), more so than should other social relationships, such as the one between strangers. We might predict, then, that on pain of inconsistency, they would agree with the following statement as well: "Holding everything else equal, such as the magnitude of the need in question and the capacity of the individual to meet it, it would be *morally wrong* for a parent to care for stranger in need instead of caring for their very own child" (consistent with the findings of McManus et al., 2021). This is just another way of saying that, when forced to choose between helping a stranger or helping their own child in such a circumstance, a parent *ought* to help their child. What should we say about such a belief?

Two types of projects.

If we accept the concerns of Farber and Korsgaard, in confronting such a belief, there are two distinct projects we could pursue: an *explanatory* project (asking why participants would be inclined to believe such a thing), and a *justificatory* project (asking whether such a belief is justified from a moral perspective – or perhaps more controversially, whether such a belief is *true*). They are right that these are two different projects. The first one, presumably, would involve an appeal to evolutionary theory, possibly as inflected by biocultural psychology (Barrett et al., 2015). Something like: ancestral parents, from whom we have inherited certain relevant dispositions and/or

associated cultural practices,[11] who held such a belief and found it motivationally compelling would have had on average more offspring survive to sexual maturity. They would, therefore, have had a greater chance, compared to those ancestral parents who lacked such a belief or motivational pull, of passing on the relevant phenomena (genes, dispositions, motivations, cultural practices, and so on) (Bloom, 2011; Gellner et al., 2020). But what might the second sort of project involve?

One potential answer comes from an essay by Curry (2006) entitled, "Who's afraid of the naturalistic fallacy?" I discuss this fallacy in the following section.

Avoiding the naturalistic fallacy.

The naturalistic fallacy has been described and interpreted in many different ways over the centuries, but a common formulation holds that there is no logically valid way to reason directly from empirical facts (such as facts about what people believe a parent is morally required to do in the above situation) to moral or so-called 'normative' conclusions (such as what, if anything, a parent is in fact morally required to do in such a situation). Suppose we grant that, Curry (2016) says. We can still engage in means-ends reasoning to arrive at normative conclusions, so long as we assume the validity of at least some moral ends that are not likely to be in dispute.

To take a non-moral example, suppose we want to get to Grand Central Station within the hour, and the only way to do this is to take the Number 4 train. Well then – if we are going to satisfy our valued goal, namely, arriving at Grand Central Station on time – we *ought* to take the Number 4 train. That much should be uncontroversial.

[11] That is, practices we are disposed to develop and pass down.

Now let's try a moral example. Suppose that, instead of getting to Grand Central Station, what we want (i.e., our valued goal) is to adopt a set of social practices that will maximize our ability flourish together, given the kinds of creatures that we are. Further suppose that a major impediment to our flourishing is a suite of recurrent coordination problems that, if left unsolved, will reliably undermine our well-being (e.g., by preventing us from meeting our needs without the constant threat of violence or other harm due to competition over limited resources).

Finally, suppose that the only (or best) way to effectively solve these recurrent coordination problems – i.e., in a manner acceptable to all concerned[12] — is to imbue the various social-relational roles we occupy with certain patterns of prescribed[13] cooperative functions (i.e., relational norms); and, moreover, to distribute these norms among our various relationships in a way that accommodates to the particular constrains and affordances (economic, geographical, institutional, historical, political, etc.) our group faces. Well then, it follows that we *ought* to adopt – and indeed follow – the relevant set of relational norms.

Critiquing relational norms.

An advantage of this approach to grounding normativity (in the philosophers' sense), is that it gives one a potential foothold for *critiquing* certain relational norms as they exist in a given society. Suppose that a society has developed, through a combination of passive learning/cultural evolution and also social activism/political agitation, a set of relational norms that are distributed in a particular way at a given

[12] Or, perhaps, in a manner that *would* be acceptable to all concerned, if they were perfectly rational and well-informed and deciding about such matters from behind a veil of ignorance as to (among other things) the social roles they would end up occupying in a given society (Rawls, 1999).
[13] And as necessary, socially policed.

historical moment. If one *assumes* that the general flourishing of society members is a legitimate moral goal, and if this goal is stymied by persistent failures of cooperation in the face of recurring coordination problems, then it may be the case that the currently-prevailing relational norms for certain social roles (and/or how these norms are distributed among the various available social roles), are *functionally deficient* – that is, they do not successfully solve one or more coordination problems in a mutually satisfactory way.

Imagine, for example, that the teacher-student relationship in this society is prescriptively extremely hierarchical, with students normatively expected to follow the teacher's instructions in every situation regardless of the apparent cost. And now suppose, for the sake of argument, that such a prescription could in principle be functionally successful under very specific historical or group-survival conditions (i.e., conducive to the flourishing of all concerned, all things considered), but that those conditions no longer apply. Well then, it might be the case that students in such a situation *ought not* to follow every instruction their teacher gives them, but rather practice a kind of 'civil disobedience' in hope of changing the norm.

Again, such an analysis assumes that we are entitled to take for granted certain moral ends, and then use these to justify (or refute) more specific moral claims or behaviors as they apply to a given situation. But, it might be protested, why should we think we are in fact so entitled? Just because we human beings may *believe* that we ought to adopt social practices that will maximize our ability flourish together, this does not necessarily mean that the belief is justified in the sense demanded by Farber and Korsgaard. Why should we think that the promotion of human well-being, or the

avoidance of human suffering,[14] is *morally* significant – the sort of thing we can use to justify other, 'downstream' moral conclusions (e.g., about what we ought to do, or refrain from doing, in a given social-relational context). Put another way: just because we may *value* our own well-being (an empirical claim) does not mean that our well-being is in fact *valuable* in some ultimate moral sense (a normative claim). How might we respond to such a challenge?

Naturalizing ethics.

The challenge is, of course, an old one. As Hume (1777) famously argued:

> Ask a man *why he uses exercise*; he will answer, *because he desires to keep his health*. If you then enquire, *why he desires health*, he will readily reply, *because sickness is painful*. If you push your enquiries farther, and desire reason *why he hates pain*, it is impossible he can ever give any. This is an ultimate end, and is never referred to any other object. . . And beyond this it is an absurdity to ask for a reason. It is impossible there can be a progress *in infinitum*; and that one thing can always be a reason why another is desired. Something must be desirable on its own account, and because of its immediate accord or agreement with human sentiment and affection. (Hume, 1777, quoted in Curry, 2016)

Hume's approach to morality can be seen as a precursor to what is now called "ethical naturalism," the position that "any prescriptive ethics must be based on the needs, desires, and goods that people are naturally predisposed toward" (Fiske & Rai, 2014, pp. 289-290). According to Fiske and Rai (2014), for ethical naturalists, "empirical

[14] Much less the well-being or suffering of non-human animals – with whom, it should be noticed, we also stand in a kind of relationship and to whom we presumably have certain obligations. However, the ethics of human and non-human animal relations goes beyond the scope of this dissertation.

science plays a crucial role in any prescriptive ethics because it has the power to identify the basic human goods that people are naturally predisposed toward, as well as the conditions that support those goods" (p. 290). To support this point, they quote Flanagan et al. (2008): "the ends of creatures constrain what is good for them … morality cannot seek to instantiate behavior that no human beings have the propensity to seek [and] there are a limited number of goods that human beings seek given their nature and potentialities" (pp. 15-16).

Suppose we agree with all of that (as in fact I do). It still leaves open many, more specific questions, about how experimental moral psychologists who are interested in the normative structure of social relationships can summon the relevant data, i.e., to aid moral philosophers in the kind of project described (above) by Korsgaard. Recall, she said that those who take up moral philosophy "do not merely want to know why those peculiar animals, human beings, think that they ought to do certain things. We want to know what, if anything, we really ought to do" (Korsgaard, 1996, p. 13).

Let us, then, consider the practice of moral philosophy and how it might relate to psychological phenomena of the sort that are amenable to scientific study.

'Doing' moral philosophy.

Typically, when moral philosophers come up with their accounts of right and wrong, they rely on their own psychological intuitions and judgments – invariably formed within a particular cultural context with its attendant relational norms – about what their normative account entails in particular cases (Kagan, 2001). If it seems that their account yields a highly counterintuitive answer (for example, that it's morally

permissible for a mother to enact the mating function with her child), they will often go back and revise their account to get the intuitively 'right' answer. Other times, their commitment to a philosophical principle will force them to relinquish or override their intuitive judgment about a particular case; and back-and-forth they may go, trying to reach a reflective equilibrium (Cath, 2016). What, if anything, can experimental psychologists contribute to this process?

Here is one possibility: by giving an explanation of why certain relationally-situated moral judgments seem intuitive in a given context, psychologists may provide a basis for philosophers to determine whether a given judgment should in fact be given substantive normative weight – granting, of course, certain prior, or more basic, value commitments that are not in question in the given discourse (or which are held constant for the sake of argument). We will consider some examples of this below.

Alternatively, if a philosopher decides to reject a common, robust moral judgment shown to hold among a relevant group of stakeholders, she will have to give an adequate error theory: i.e., an explanation of why people have such a strong and consistent moral intuition despite its failure to yield (what the philosopher takes to be) the right normative conclusion (Singer, 2005).

In what follows, we will look at some recently proposed strategies for negotiating these kinds of decisions (adapted from Earp et al., 2020), highlighting potential pathways for reaching normative conclusions from argumentative premises that include empirical claims about the moral mind in relational context. First, however, we will return to our discussion of normative disagreements regarding one particular kind of relationship, namely the romantic partner relationship, based on our work on the ordinary concept of true love.

Normative Disagreements and Romantic Relationships

In Chapter 4, we noted that people often disagree about whether a given romantic relationship is an instance of true love, where this concept is at least partly normative: our findings demonstrated that the perceived *goodness* of a relationship is one important factor for grounding membership in the 'true love' category. The other important factor we identified was perceived *realness* – which we speculated might have to do with the sense that a given mental state is rooted in one's true self (Strohminger et al., 2017). What should we say about cases in which, even within a given society or culture, two people look at the very same romantic relationship – possessed of the very same facts about it – and yet reach opposite conclusions about whether it is morally good or exhibits realness and (hence) whether it is an instance of true love?

In subsequent sections, as I mentioned earlier, I will go beyond merely attempting to understand such disagreements to discussing strategies for potentially resolving them (i.e., on the way to reaching substantive moral conclusions). As a more modest aim, however, I will start here by discussing a role for empirical studies into ordinary people's use of relational concepts in allowing us to better pinpoint – and understand – the nature of such normative disagreements, using the example of romantic relationships.

Understanding relational moral disagreements.

People, we have said, often disagree about true love: what it is, whether it exists, who has it, and so on. For a concrete example, consider our *age difference*

vignette (from Chapter 4), which concerns a relationship between an older professor and his young undergraduate student. Many people responded that this was clearly a case of true love, while many others responded that it was clearly not a case of true love. Disagreements like this one seem to point to something fundamental about the concept of true love and the role it plays in the way people understand the normative dimensions of their lives and relationships.

The data from our empirical studies cannot directly tell us which of the opposing views in such cases is the correct one, but they can provide valuable insight into the nature of the disagreement itself. Imagine a person who accepts that there is something very wrong in the relationship described by the age difference vignette, but who nevertheless maintains that the characters in it are experiencing true love. Now imagine a critic who disagrees with this person, asserting that what the characters feel for each other in the vignette is not true love. In light of our reported findings, it seems that there are two distinct ways in which such a critic could argue for her view.

One approach would be to draw on the criteria associated with the ordinary concept of true love. In this first approach, the critic would accept the criteria revealed in the studies we reported, and she would then argue that the case in question doesn't actually fulfil those criteria. For example, focusing on the realness criterion (see above), she could say: "You may think that they are experiencing something real, but you are suffering from a delusion. No relationship between an older professor and a much younger student—especially one he directly supervises—can be rooted in the kind of realness that is necessary for true love."

Alternatively, the critic could argue against the criteria themselves. For example, she could argue that the ordinary criteria for applying the concept of true love are themselves flawed, and that we should instead adopt criteria according to

which nothing can count as true love without being (sufficiently) good. She might then say: “It may well be that their feelings for each other are real. And I recognize that realness is one of the main criteria we ordinarily use to decide whether something counts as true love. But their relationship is deeply wrong, and for that reason, we should reject any criterion according to which their feelings for one another count as true love.”

In short, there are at least two different ways in which people might disagree about true love. First, they might disagree about whether a particular relationship or experience fulfills the criteria associated with the ordinary concept. And second, they might disagree on a deeper level: they might disagree about the criteria themselves. Let us now take a closer look at each kind of disagreement in turn.

Disagreement about fulfilling criteria. The results of our empirical studies shed at least some light on the sorts of disagreements about true love that are rife in ordinary life. In Study 2 of Chapter 4, we found considerable disagreement between participants about whether the characters in each vignette were experiencing true love, but most of this disagreement simply mirrored the disagreement they showed on the questions about goodness and realness. Among participants who agreed about those other questions, there was relatively little disagreement about whether what the characters had between them was an instance of true love.

These results provide some support for a broader picture of the nature of ordinary disagreements regarding true love. On this picture, most of the disagreement is of the first of the two types described above. People share an understanding of the criteria that a relationship has to fulfill to count as true love, but they disagree about whether individual cases do or do not fulfill these criteria. When it comes to

judgments of goodness, for example, this disagreement would be straightforward. We can easily imagine a case in which two people agree that the criteria involve a role for goodness but just have radically different views about which things are good. On this view, people could have quite different views about which individual things count as 'true love' – but this would not simply be because they were using the phrase in completely different ways. Rather, it seems that people can share certain criteria for the use of this phrase, while being engaged in a substantive disagreement about which things fulfill those criteria.

We suggest this is something to look out for in cases of apparent moral disagreement, not only in relational contexts, but generally: Is the disagreement normatively substantive, or are the opposed parties simply employing moral concepts in different ways, thereby effectively talking past one another? Empirical studies can help to answer this question.

Disagreement about the criteria themselves. Now suppose that two people disagree, not about whether a given relationship meets some shared criterion for true love, but about whether a given criterion, such as realness, is the *right* criterion for picking out category members. There are at least two ways in which someone might take issue with the ordinary concept of true love by disagreeing about one or more of its criteria. Specifically, there could be a *naturalistic* disagreement about the criteria, and there could be a *normative* disagreement about the criteria.

A *naturalistic* disagreement would be premised on the belief that there really is such a thing as true love in the word, and that the ordinary concept of true love, in placing so much emphasis on realness, say, does not succeed in uniquely picking it out. A scientific reductionist, for example, might identify true love with some

biological process related to reproduction, or a particular brain state, and argue that it is *this* feature which ought to be central to the concept on grounds of descriptive accuracy. A proponent of this view, then, might then wish to engage in what has been called *naturalist conceptual engineering* (Veit & Browning, 2020). That is, the proponent might try to promote what they take to be a more *accurate* or finely discriminating conception of true love and encourage its wider adoption among ordinary people.

A *normative* disagreement would be premised on a different kind of belief. This would be a moral or sociopolitical belief that the ordinary concept of true love is not *desirable* in its current form, given certain normative ends. As Haslanger (2012) argues, the operative concept of X may be different from what she calls the 'manifest' concept (the concept people explicitly take themselves to be applying when they pick out X); and this in turn may be different from what she calls the 'target' concept—the concept people *should* apply when picking out X, all things considered (Haslanger, 2012).

To see what a normative disagreement about the concept of true love might look like, let us imagine someone speaking to a troubled friend, perhaps one of the characters in our *abuse* vignette (see Chapter 4). "If your partner abuses you," we'll imagine this person saying, "no matter how much you may feel affection for each other ... what you have between you is not *true love*." Now suppose this was a direct response to the other person saying: "I know the abuse is wrong, but what we have is *true love* and that is more important than anything else." We would have two different uses, then, of the same concept that are mutually incompatible.

Suppose that both of these (hypothetically) operative uses were circulating in the language community. Depending on our aims and values, we might think that it

would be normatively *better*—all things considered—if the use that excludes abuse became more intuitive and widely employed, while the use that is compatible with abuse became counterintuitive among most ordinary language users. Supposing that was our goal, we might wish to undertake what Haslanger calls an 'ameliorative' project, or what has recently been termed *moral conceptual engineering* (Veit & Browning, 2020). That is, we might try to promote the first use of true love and encourage its greater uptake among ordinary people.

Notice the phrase "Depending on our aims and values" in the previous paragraph. What this phrase indicates is that one can't reach a normative conclusion about how some relational concept – such as 'true love' – *should* be used in a given discourse unless one takes for granted certain moral norms or values as an argumentative starting point. In other words, recalling the above-stated concerns about the so-called 'naturalistic fallacy' (Frankena, 1939), there is no way to reach an 'ought' directly from an 'is' without smuggling in some normative premise along the way. However, some normative premises are more reasonable or widely shared than others, and these may serve as useful candidates for anchoring a moral argument regarding relational concepts or norms. Again, empirical studies may be useful for identifying which relational norms are in fact widely shared among a group of stakeholders (either within or between cultures) for purposes of grounding a more productive moral dialogue. In the following section, we will look at this process in greater detail.

Resolving Disagreements

Within moral philosophy, a common strategy for reaching normative conclusions – for example, about the moral duties implied by a given social relationship – is to take a coherence-seeking approach in which the practitioner tries to achieve a kind of "reflective equilibrium." That is, the moral philosopher attempts to "harmonize all the elements contributing to moral judgment, including intuitions about cases, moral principles, moral theories, and background theories of moral agency and social organization" (Arras, 2016, n.p.). Within this method, moral theories and principles function to "organize, explain, criticize, and extend our intuitive responses to cases," while at the same time, "those very responses can, in turn, help us to amend and sharpen our principles and theories when they prove inadequate to the complexities of emerging cases" (Arras, 2016, n.p.).

To better appreciate this general strategy, however, it is necessary to ask who the implied "we" is in the reference to "our" intuitive responses to cases. For the sake of clarity, the responses of interest here are moral judgments regarding particular relationally-situated behaviors—for example, judgments that it is worse for someone to fail to feed a hungry individual when that person is their child than when the person is a non-paying customer at a restaurant (Earp et al., 2021). In developing substantive normative theories, these sorts of judgments have, traditionally, been those of the moral philosopher considering various cases. In this way, the judgments of particular individuals have long provided a key source of data for "armchair" approaches to moral philosophy, including the coherence-seeking kind described above. But what if the judgments of moral philosophers differ from those of lay people within a given society (or from those of other moral philosophers in different societies)?

One possibility is that a philosopher's speculative reflection, especially on abstract or idealized cases, might fail adequately to capture the concrete normative

and empirical issues at stake in a given social-relational context (Mills, 2005; Tobin & Jaggar, 2013). This, in turn, might call into question the real-world relevance of such reflection for reaching normative conclusions. If the goal is to develop a normative position regarding actual social relationships as they are widely conducted within a society, might the judgments of ordinary individuals constitute equally, if not more, relevant data? And supposing that we have identified relevant empirical data pertaining to various stakeholder judgments: (how) can we draw normative inferences on the basis of that data?

Within the field of bioethics, which can be considered an applied branch of moral philosophy, several strategies have been developed toward this end. According to a recent systematic review (Davies et al., 2015), most methodologies employed in the field can be classed as either "dialogical" or "consultative." Dialogical approaches involve actual dialogues between researchers and stakeholders to reach a shared understanding and a joint resolution to a particular moral problem. Consultative approaches involve collecting empirical data relating to stakeholder views, attitudes, and experiences, and then using these as a basis for drawing normative conclusions. In terms of the majority of consultative approaches, the end goal is either the achievement of coherence between stakeholder data and moral theory ("narrow reflective equilibrium") or between stakeholder data and broader considerations, such as background theories, moral principles, "expert" intuitions, morally-relevant facts, and considered judgments ("wide reflective equilibrium").

According to Davies and colleagues (2015), the key difference between dialogical and consultative methods is the role of participants: whereas participants in dialogical approaches work together with researchers to analyze stakeholder data and develop normative conclusions regarding discrete problems on the basis of consensus,

participants in consultative approaches do not take part in the analysis or the process of forming normative conclusions (Davies et al., 2015). Furthermore, the aims of consultative approaches vary, “ranging from theory development to the generation of concrete answers to discrete problems" (Davies et al., 2015, p. 7). In addition, when consultative approaches aim to deliver normative conclusions regarding a specific problem, they tend to employ a coherence-based methodology like the ones described above (Davies et al., 2015).

Experimental philosophy (x-phi) is another field in which practitioners have attempted to glean (meta)philosophical insights, including but not limited to normative inferences, from empirical data. How they do this depends on how the practitioners interpret the purpose or function of x-phi in general. At least two main purposes have been identified, corresponding to two separate research programs, each of which can be understood in relation to the tradition of conceptual analysis in analytic philosophy (Alexander et al., 2010; Fisher, 2015; Knobe, 2016; Machery, 2017; Mukerji, 2019; Sosa, 2007).

The first program aims to make a positive contribution to conceptual analysis, though not necessarily through the provision of necessary and sufficient conditions for concept application. The second program engages negatively by providing evidence against the intuitive assumptions of more traditional approaches to conceptual analysis. However, as Knobe (2016) argues, regardless of which program they may claim to be pursuing, what experimental philosophers typically do in their studies is a kind of cognitive science: they investigate effects on psychological structures thought to underpin judgments held by participants (Knobe, 2016, p. 42). It is this characterization of x-phi as cognitive science that is especially useful for

understanding what the following strategies have been trying to achieve in terms of generating normative conclusions from premises that include empirical data.

Inferential Strategies

Within the strategies for drawing normative inferences discussed below, the running theme is that if we can get a deeper understanding of the criteria underlying stakeholder moral judgments regarding social-relational cases (e.g., Person A does X to Person B), this understanding can help us to address substantive normative questions about the actual moral status of the behavior or behaviors in question.

But we must start by dealing with a couple of red herrings. First, it should be uncontroversial that the most ethically justified conclusion – for example, about how one party to a given social relationship should behave toward the other – is not always, or simply, the most popular one based on common opinion within a given society (cue references to Nazi Germany). However, I will argue that, when certain conditions are met, researchers can legitimately appeal to the presence of prevalent or highly consistent stakeholder judgments revealed by an empirical study as one (ultimately defeasible) reason that counts in favor of a particular normative claim.

At the same time, it should be obvious that simply deferring to "ethical experts" such as trained moral philosophers – especially when their judgments or associated normative conclusions defy those of other, "ordinary" stakeholders – can also be problematic. For example, such knee-jerk deferral to a circumscribed group of people can enshrine prejudices and lead to dogmatism and parochialism (Machery, 2017). As Savulescu and colleagues observe (Savulescu et al., 2019), some laws and policies regarding, for example, the doctor-patient relationship do in fact run counter

to public preferences: bans on voluntary assisted dying in the U.S., U.K., and Australia, for instance, have been put in place despite large majority preferences for permitting assisted dying. Other laws and policies, however, may be grounded in "mere" public sentiment without necessarily appealing to more principled normative considerations.

Typically, neither a direct appeal to an *argumentum ad populum* nor a simple appeal to a single set of "expert" judgments will provide reliable guidance toward normative conclusions. To get out of this bind, Savulescu and colleagues have offered some preliminary proposals. First, they suggest that we need to identify the moral judgments and intuitions of those who have been careful in their reasoning and have a "clear understanding of the issues" (Savulescu et al., 2019, p. 1241). But, we might ask, who are these careful and reliable reasoners and how do we find them? They might be those "ethical experts" we have been alluding to: professional moral philosophers, bioethicists, legal professionals, and the like. However, members of these groups make up a tiny fraction of the population, and between them they may have idiosyncratic perspectives, conflicting judgments, moral disagreements, and incompatible moral inferences. Furthermore, debates regarding a specific moral-philosophical issue may have reached a stalemate with good reasons for adopting several positions and/or no adequate way for the "experts" to agree upon which position should be implemented in practice. Consequently, Savulescu et al. suggest that we need "refined expert intuitions" (along with guidance from formal ethical theories) *as well as* "widespread public responses" (Savulescu et al., 2019, p. 1242).

A potential problem with such proposals is that they do not spell out what to do when widespread public responses, such as common moral judgments, and those of putative experts diverge or, indeed, what to do when they seem to agree. More

generally, how do ethical theories, expert judgments, and the judgments of ordinary people relate to one another, and how can this information be integrated to draw normative conclusions about appropriate moral behavior in relationships? I will briefly discuss four potential strategies (summarized in Earp et al., 2021), as follows: *parsimony*, *debunking*, *triangulation*, and *pluralism*.

The Parsimony Approach: Appealing to Robust Effects or Consistent Judgments

The first strategy is based on a principle of parsimony. This view assumes that ordinary people's judgments about certain cases carry significant (albeit defeasible) normative weight, such that experts who wish to make claims about what ought to be done should begin by carefully studying those judgments. The strategy is parsimonious, then, in that it relies on the simplest possible model for deriving normative content from the moral judgments of ordinary people: it holds that those judgments should be given at least some positive normative weight. In short:

> **Parsimony**. If relevant stakeholders consistently make a judgment *p* which encodes moral claim M, then M has prima facie normative weight.

One of the aims of studies employing this strategy is to gather data relating to stakeholder judgments, often with the assumption that no matter what these judgments are, they are normatively significant. Once the data have been gathered, a proponent of the parsimony strategy might then identity the most consistent (e.g., common or robust) moral judgments revealed by a study and give these *prima facie*

normative weight when deciding on a solution to an associated normative question. Note that the normative weight accorded to such judgments need not be especially strong. The parsimony strategy requires only that these judgments, to the extent that they are consistent or prevalent, be viewed as providing some normative weight in a moral-philosophical argument. They will never be enough on their own to deliver an all-things-considered normative conclusion.

In this way, studies that employ the parsimony strategy are consistent with consultative approaches in bioethics, mentioned earlier, insofar as the latter rely on the robust judgments of some group of stakeholders as a basis for arbitrating between competing normative claims. However, consultative approaches in bioethics tend to be concerned with identifying the most prevalent judgments (or attitudes, preferences, etc.) of the group, primarily through observational or cross-sectional methods. By contrast, experimental studies that adopt a parsimony strategy might look beyond the mere prevalence of a judgment and instead emphasize the robustness of an experimental finding (e.g., across methods, materials, or operationalizations of a causal stimulus) regarding a given effect on participant responses (see Rueda et al., 2020). For an example applied to the parent-child relationship, where the moral question at stake concerns the nature of a parent's duty to care for a child by donating tissue to meet the child's health-based needs, see Case Study 1.

Case study 1. Parsimony. What grounds moral duties in the parent-child relationship with respect to the care function, in the case of meeting certain medical needs of the child?

Consider a child who needs a tissue donation to survive. Suppose that their biological parent could donate the needed tissue. Insofar as it seems intuitive that the parent has a moral responsibility to donate the tissue, what drives this judgment? Is it the biological relation between the donor and recipient (McMahan, 2003) or the fact that the donor is uniquely suited to provide tissue that will work for the recipient (Beverley, 2016)? Beverley and Beebe, in a study involving a series of contrastive vignettes, found that "unique ability rather than biological relatedness was the primary predictor of people's judgments of moral responsibility" (Beverley & Beebe, 2018, p. 92). To distill the normative relevance of this finding, the authors adopt a meta-philosophical stance: folk judgments need not "rigidly constrain philosophical theorizing" but counterintuitive normative views (e.g., that moral responsibility stems from biological relatedness) carry an explanatory burden (Beverley & Beebe, 2018). As such, the "parsimony" model would advocate that the "unique ability" judgment be assigned *prima facie* normative weight.

There are two interrelated concerns with employing this approach that critics might think to raise. First, it is highly likely that traditional moral philosophers would hesitate to accept such a method on the basis that it seems to derive normative conclusions from empirical premises without necessarily appealing to more principled normative considerations. This relates to a second potential concern, which is that the "parsimony" approach seems to reduce moral reasoning about contested social-relational issues to a popularity contest.

We think this concern is unwarranted. As already observed, the mere fact that consistent judgments are revealed by a study does not entail that the associated moral issues are conclusively settled. Rather, the identification of consistent judgments is just one factor that counts in favor of the relevant moral claim, and the normative weight accorded need not be strong. Indeed, reasons for granting more normative weight to a particular set of judgments are, ultimately, defeasible if, for example, it

can be convincingly shown that these judgments are unreliable (see below). In short, this approach puts the burden of proof on those who would claim that we should *not* respond to the particular ethical issue in question by according at least some normative weight to the most consistent judgments of relevant stakeholders.

The Debunking Approach

In contrast to the "parsimony" approach, which assigns *prima facie* (though necessarily not strong) normative weight to stakeholder judgments, one might wish to argue that a certain judgment should *not* be accorded normative weight when considering a solution to a social-relational moral problem. And one might do this by testing whether the judgment is the output of a psychological process that, for example, has been substantially influenced by prejudice, bias, or morally irrelevant differences in the ways in which a moral dilemma is presented (e.g., it is subject to misleading "framing effects") (Demaree-Cotton, 2016; Petrinovich & O'Neill, 1996). To investigate whether a judgment should *not* be accorded normative weight in a moral argument, one might wish to pursue a "debunking" strategy derived from the following general argumentation scheme (Mukerji, 2019, pp. 31-56).

Debunking.

(P1) Judgment *p* is the output of a psychological process that possesses the empirical property of being substantially influenced by factor F. (Empirical premise)

(P2) If a judgment is the output of a psychological process that possesses the empirical property of being substantially influenced by factor F, then it is *pro tanto* unreliable. (Bridging normative premise)

(C) Judgment *p* is *pro tanto* unreliable.

Such an approach can be employed to assess whether ordinary people revise their judgments under various "treatment" conditions. Take, for example, exposure to a particular philosophical argument: say, the famous argument of Singer (1972) that we are no less obligated to help distant strangers in far off lands than we are to help members of our own community, assuming we are equally capable of providing the requisite help in either case. Most people are likely to find this claim counter-intuitive, consistent with our findings (from Chapters 2 and 3) that participants consistently evinced a stronger normative expectation of care for socially close relationships (e.g., family members) compared to socially distant ones (e.g., strangers). But if people were to update their judgment about this case having reflected on a philosophical argument (see, for a related approach, Schwitzgebel et al., 2020), then at least two points might follow: (1) they had not previously considered the philosophical argument in question; and (2) once they did, they abandoned their original judgment. This might suggest that their original judgment was not particularly robust – or was insufficiently informed by careful reflection – and this could be a reason to downgrade its normative significance.

The susceptibility of a moral judgment to so-called framing effects also has been proposed as a factor that should weaken our confidence in the judgment from a normative perspective. As Andow (2016, p. 908) observes, the substantive influence of morally irrelevant factors – such as a mere change in logically-equivalent framing of a case – on judgments is important because "it is capable of radically altering the moral position that one ends up endorsing." And the claim here is, if a person's judgment about some case is the output of a psychological process that has been

substantively influenced by such a morally irrelevant factor, then we have a *prima facie* reason to doubt the judgment. But more than this, we have a reason to believe that a process of reflection *based on* this judgment "will only lead one deeper into error" (Wedgwood, 2007, p. 244). At least, the *pro tanto* unreliability of a judgment is one factor that counts against accepting it as a premise in a normative argument.

But we must remain cautious. Even if stakeholders hold a judgment that has been shown to be *pro tanto* unreliable in a specific instance, alternative explanations should be explored for why the target population holds that judgment. After all, we can never be sure that a particular judgment can be debunked *in general*: debunking proceeds by looking at isolated, specific ways in which the psychological processes outputting certain judgments can be deficient. Thus, experimentalists can play a valuable role in cases where there are plausible alternative explanations for what appears to be a normatively unjustified judgment. In particular, they can conduct experiments to test the alternative explanation(s), often by carefully manipulating relevant cognitive factors. For example, they might see whether a *pro tanto* unreliable judgment persists when participants are asked to consider a hypothetical society in which a given social-relational norms differ in some relevant way from their own. Depending on the results, such an experiment could provide evidence for or against the alternative explanation for the original suspect judgment.

In the next two examples, we will see how the debunking approach might be pursued in practice, using two separate case studies focused on the doctor-patient relationship: (1) an apparently *failed* debunking attempt (providing evidence that ordinary people's judgments, in a specific instance, are not largely biased by a particular factor that would have made the judgments unreliable; however, see the footnote at the end of this paragraph for some qualifications) (case study 2); and (2)

an apparently *successful* debunking attempt (providing evidence that ordinary people's judgments, in a specific instance, *are* largely biased by a particular irrelevant factor, and so should *not* be trusted at least in this instance) (case study 3).[15]

Case study 2. Debunking: failed. What grounds moral duties in the doctor-patient relationship regarding killing vs. letting die?

Laypeople distinguish killing and letting die by evaluating the morality of the physician's intervention (Cushman et al., 2008). For example, doctors who observe a terminal patient's wishes are seen as allowing them to die, whereas doctors who disregard the patient's wishes are seen as killing them (Rodríguez-Arias et al., 2020). The judgments of ordinary people may afford little normative insight here, in part because they lack the requisite understanding of the medical and clinical issues in play. This objection makes a straightforward empirical prediction: if laypeople acquired the relevant medical knowledge, they would abandon their untrained judgments in favor of the canonical distinction between killing and letting die as commissive versus omissive life-ending acts, respectively. However, Rodríguez-Arias and colleagues found no evidence of this: doctors, medical students, and laypeople revealed strikingly similar judgments about end-of-life cases (Rodríguez-Arias et al., 2020). The determining factor appears to be whether the patient wished to live or die, and not how the patient's death was brought about (i.e., via action or omission). Thus, the ordinary judgment could not be debunked on grounds of ignorance of clinically relevant details.

[15] There is an asymmetry between these two cases. If there really is conclusive evidence that a popular judgment is grounded in some factor that undermines its normative force (successful debunking attempt), then the judgment should be set aside or discounted. But if one fails to debunk a given judgment, this does not automatically entail that it should be trusted. Rather, it might still be normatively suspect on other grounds that have not yet been tested. So, one should try to test, and rule out, the most *plausible* debunking explanations, and if one reliably fails in this, it becomes reasonable to treat the judgment as carrying *prima facie* normative weight (a process akin to Popperian falsificationism, notwithstanding its various shortcomings; see, e.g., Earp (2020); Lakatos (1970).

Case study 3. Debunking: successful. Within the doctor-patient relationship, (when) is it morally wrong to treat the patient differently on account of their sex or gender?

For another example concerning the doctor-patient relationship, it is relevant to determine whether people have a gender bias in assessing children's pain. We conducted an experiment in which we manipulated the perceived gender of a young child getting a finger-stick to draw blood (Earp et al., 2019, based on Cohen et al., 2014). To keep the experiment as controlled as possible, participants viewed a single video stimulus of a child whose sex could not be visually determined (i.e., the same video in both conditions). In one condition, participants were told the child's name was "Samuel," and in the other, "Samantha." Participants then watched the video and rated how much pain the child experienced. We found that participants rated the child named "Samuel" as experiencing more pain than the child named "Samantha." Thus, perceived gender alone appeared to bias observer interpretations of felt pain (for alternative explanations, see Earp & Boerner, 2019). Such evidence plausibly undermines the trustworthiness of a physician's moral judgments that, say, boys and girls *should* receive different pain treatment given a comparable injury.

The Triangulation Approach

Suppose that the normative judgments of moral philosophers differ from those of lay people with respect to some social-relational dilemma (e.g., Peter Singer's judgment about our relative duties to family members versus strangers, compared with the typical judgment of ordinary people regarding this case). Or suppose that moral philosophers from different cultures disagree with one another. What should be done about such divergences? In such cases, experimental findings could be employed as part of a coherence-seeking strategy of "narrow" reflective equilibrium discussed previously. Here, the coherence being sought is between competing expert judgments and/or between expert judgments and those of lay stakeholders. We refer to this approach as a type of triangulation:[16]

[16] Scientists and certain philosophers of science employ the term "triangulation" to refer to the use of multiple and independent sources of evidence to generate causal inferences from data to phenomena (as

> **Triangulation.** Divergence among the judgments of various groups of experts and/or between expert and lay judgments requires the following: adjusting, pruning, or supplementing the normative conclusions derived from either expert or lay judgments in order to accommodate the normative implications of the opposing views.

Experimentalists can perform three important roles in pursuing a triangulation strategy: first, using empirical means, they can identify the judgments of various experts and lay stakeholders in response to a specific normative problem, ensuring that the judgments respond to relevant features of ecologically valid contexts. Second, using the aforementioned argumentation strategies, they can experimentally investigate the cognitive mechanisms underpinning these judgments, ensuring that various expert and lay judgments are not *pro tanto* unreliable (and/or setting aside or discounting those judgments that are convincingly shown to be *pro tanto* unreliable). Finally, they can help to execute trade-offs among the respective *pro tanto* reliable judgments, revising normative conclusions as coherence and mutual support seem to require.

According to the standards of reflective equilibrium, the normative conclusions arrived at through this process, together with the revisions to the competing judgments, will be justified if and only if there is reason to believe that they will maximize the coherence of the overall set of relevant considerations. However, in order to avoid the standard objection that the equilibrium arrived at "may be no more than a reshuffling of moral prejudices" (Brandt, 1979, p. 21), the triangulation approach might better be characterized as a coherence-seeking

---

opposed to phenomena-to-theory deductive inferences) (Kuorikoski & Marchionni, 2016). Analogously, triangulation in psychologically empirically informed moral philosophy is one of the means by which *normative* inferences might be generated on the basis of multiple, independent, and *pro tanto* reliable pieces of empirical evidence.

methodology based on a "moderate foundationalism" (BonJour, 1985, pp. 26-30). The problem that Richard Brandt identifies is that the coherence constraint on its own may not succeed in correcting for all the errors or biases in the respective judgments (R. B. Brandt, 1979). As already observed in the section on debunking, it will not succeed if the antecedent judgments are so unreliable that further reflection on these judgments will only lead moral philosophers deeper into error. As a result, proponents must also explain how it is antecedently or independently rational for us to regard some or all of these competing judgments as (*pro tanto*) reliable (Wedgwood, 2007).

According to Scanlon, in carrying out the process of reflective equilibrium, we should ask whether there is more reason to revise a normative conclusion in the light of conflicting judgments, or to give up the judgments that conflict with it (Scanlon, 2014). Ultimately, as Ralph Wedgwood suggests, what is being proposed is a thoroughgoing form of fallibilism (Wedgwood, 2007, pp. 64-65). On this approach, we can never have any guarantee that we will not be rationally required to revisit and reconsider, and perhaps revise, the normative conclusions derived from the process of triangulation if and when further reliable empirical evidence is identified. However, in practical moral decision-making, we will often have to "bite the bullet" and commit to a specific normative claim based on available empirical data and the degree of coherence we have been able to achieve. As Scanlon suggests, such a commitment must be based on the best reasons for counting in favor of a specific claim (Scanlon, 2014).

When proceeding with a triangulation strategy, it will not necessarily be as straightforward as seeking a simple compromise or accommodation. Rather, it will often be necessary to ask, in the case of judgments about normative relational concepts – such as 'true love' – what we want the concept to do, that is, its (desired)

function (as opposed to meaning) (Lewis, 2020a). As Haslanger asks, "what is the point of having these concepts? What cognitive or practical task do they (or should they) enable us to accomplish? Are they effective tools to accomplish our (legitimate) purposes; if not, what concepts would serve these purposes better?" (Haslanger, 2000, p. 33).[17]

Once again, we can envisage a vital role for experimentalists in answering these questions. As Nado argues, "the experimental philosopher's focus on underlying psychological mechanisms seems to be a promising route (though of course not the only possible route) for discovering the purposes our concepts serve, and the means by which these purposes are achieved" (Nado, 2019, pp. 16-17). On this approach, we should empirically investigate our judgments about a normative concept because we want to know whether the normative concept in question is already fulfilling its intended functions to a reasonably good degree.

Take, for example, the concept of *parent.* Originally, this concept might have served the function of picking out those individuals who stand in a biological parent-child relationship with an offspring. However, nowadays, we recognize that individuals who adopt a child to whom they are biologically unrelated should also be counted as members of the 'parent' conceptual category, perhaps in part because they serve the *cooperative functions* that are normatively embedded within the 'parent-child' relationship in most societies. Thus, it seems that a social-functional understanding of 'parent' helps us to identify those individuals who play a particular role in bringing up vulnerable members of society, whether or not they are biologically related to one another.

[17] For other broadly functionalist, experimentally-based approaches to navigating conceptual disagreement, see: (Fisher, 2015; Lindauer, 2020; Machery, 2017; Nado, 2019; Thomasson, 2012).

The Pluralism Approach

This brings us to a final strategy. Suppose that the best reasons count in favor of preserving two diverging expert judgments (e.g., across cultural contexts) or the competing judgments of experts and lay stakeholders (e.g., within a given culture). In other words, there are equally good reasons for adopting two or more judgments as the basis for (competing) normative conclusions with no better reasons for adjusting, pruning, or supplementing the different positions. In some cases, such a scenario might justify a pluralistic response:

> **Pluralism.** In cases where expert and lay stakeholders hold conflicting, yet *pro tanto* reliable, judgments or where multiple and independent communities each reveal persistent disagreement between two or more conflicting, yet *pro tanto* reliable, judgments, these judgments may all have comparable normative weight.

The pluralism approach is similar to, and consistent with, the Shared Decision Making approach that has recently become an important part of clinical practice and health policy – another example centered on the doctor-patient relationship. To be successful, Shared Decision Making relies on two sources of expertise: (1) the health professional as an expert on the effectiveness, probable benefits, and potential harms of different treatment options; and (2) the patient as an expert on themselves, their social and personal circumstances, attitudes to illness and risk, tolerances for pain and discomfort, long-term outlooks, broader values, and preferences (Lewis, 2020b). As Lewis notes, Shared Decision Making is most appropriately applied under conditions of *uncertainty*, which arise because a treatment decision is preference-sensitive, that is, because medical evidence and clinical expertise suggest that there is more than one medically reasonable option, and the choice of which option is best for a given patient

depends on their values and preferences (Lewis, 2020b). In short, according to this approach, so long as the patient can fulfil certain conditions of autonomy, then she should choose the particular intervention that best satisfies her attitudes and preferences (Lewis, 2020b; Notini et al., 2020).

There is a more general lesson here: in many cases, there may be no single 'best' answer to a normative question about, for example, what course of action is morally preferable to take within a given social relationship. Often, there will be a number of options that different stakeholders favor, or to which they assign comparable weights, and we should be open to the possibility that more than one option is normatively justifiable.

For example, consider the boss-employee relationship, which, in our studies, was consistently rated as being normatively expected to serve a hierarchical cooperative function. This means that, ideally, the individual in the boss role exercises good leadership (e.g., being appropriately decisive as called for by the situation) while the individual in the employee role exercises good followership (e.g., ultimately deferring to the boss's instructions, even if the employee might have preferred to perform a task a different way). But even granting this general framework, there might be conflicting visions – between individuals or cultures – as to what good leadership or followership actually looks like or entails in practical terms.

For example, in one society (or company-specific subculture), good followership might be thought to require a highly deferential attitude toward one's superiors (e.g., bowing upon seeing them and adopting a formal manner of address); whereas, in another society or company subculture, good followership might be characterized by a more relaxed mode of relating (e.g., greeting one's boss with a wave or a handshake and calling them by their first name). When making a moral

judgment about, for example, whether an employee's behavior is disrespectful, it will often be necessary to relativize the judgment to the society or company-culture in question. Trying to determine whether there is one normatively preferable standard for how to exhibit followership, by contrast, would likely not be justified in such a case.

Final Thoughts

After giving an overview of the dissertation and discussing some key findings, limitations, and possible future directions for empirical work on relational moral psychology, I turned to an examination of potential normative implications of this work. Noting that people often disagree, both within and between cultures, about substantive moral issues in relational context, I explored a number of strategies for achieving two mains aims: (1) better understanding the underlying nature of such disagreement, including by experimentally probing people's ordinary language use of normative relational concepts (such as 'true love'), and (2) potentially resolving such disagreement, using parsimony, debunking, and triangulation strategies – or adopting a pluralistic attitude which escapes the need to resolve certain moral disagreements by finding a single normative solution. This discussion has by no means been exhaustive, but I hope to have given a sense of some of the exciting work on relational morality – both empirical/scientific and normative/philosophical – that lies ahead in the coming years.

References

Alexander, J., Mallon, R., & Weinberg, J. M. (2010). Accentuate the negative. *Review of Philosophy and Psychology*, *1*(2), 297–314. https://doi.org/10.1007/s13164-009-0015-2

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in Cognitive Sciences*, *24*(9), 694–703. https://doi.org/10.1016/j.tics.2020.06.008

Andow, J. (2016). Reliable but not home free? What framing effects mean for moral intuitions. *Philosophical Psychology*, *29*(6), 904–911. https://doi.org/10.1080/09515089.2016.1168794

Anglin, W. S. (1991). *Free Will and the Christian Faith*. Oxford University Press.

Archer, A. (2016). Are acts of supererogation always praiseworthy? *Theoria*, *82*(3), 238–255. https://doi.org/10.1111/theo.12085

Argyle, M., Henderson, M., Bond, M., Iizuka, Y., & Contarello, A. (1986). Cross-cultural variations in relationship rules. *International Journal of Psychology*, *21*(1–4), 287–315. https://doi.org/10.1080/00207598608247591

Aron, A., & Westbay, L. (1996). Dimensions of the prototype of love. *Journal of Personality and Social Psychology*, *70*(3), 535–551.

Arras, J. (2016). Theory and bioethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entriesheory-bioethics/

Ashford, E., & Mulgan, T. (2018). Contractualism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/contractualism/

Atari, M., Graham, J., & Dehghani, M. (2020). Foundations of morality in Iran. *Evolution and Human Behavior*, *41*(5), 367–384. https://doi.org/10.1016/j.evolhumbehav.2020.07.014

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Barrett, L., Pollet, T., & Stulp, G. (2015). Evolved biocultural beings (who invented computers). *Frontiers in Psychology*, *6*, 1047. https://doi.org/10.3389/fpsyg.2015.01047

Bartkowski, J. P. (1997). Debating patriarchy: Discursive disputes over spousal authority among evangelical family commentators. *Journal for the Scientific Study of Religion*, *36*(3), 393–410. https://doi.org/10.2307/1387857

Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology*, *72*(6), 1335–1348. https://doi.org/10.1037/0022-3514.72.6.1335

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, *8*(9), 536–554. https://doi.org/10.1111/spc3.12131

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*(01), 59–78. https://doi.org/10.1017/S0140525X11002202

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037//1089-2680.5.4.323

Bayer, J. B., Lewis, N. A., & Stahl, J. L. (2020). Who comes to mind? Dynamic construction of social networks. *Current Directions in Psychological Science*, online ahead of print. https://doi.org/10.1177/0963721420915866

Bentham, J. (1789). *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press.

Berg, M. K., Kitayama, S., & Kross, E. (2021). How relationships bias moral reasoning: Neural and self-report evidence. *Journal of Experimental Social Psychology*, *95*(104156), 1–8. https://doi.org/10.1016/j.jesp.2021.104156

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs*, *37*(4), 293–329. https://doi.org/10.1111/j.1088-4963.2009.01164.x

Berscheid, E., Snyder, M., & Omoto, A. (1989). The Relationship Closeness Inventory: Assessing the closeness of interpersonal relationships. *Journal of Personality and Social Psychology*, *57*(5), 792–807.

Beverley, J. (2016). The ties that undermine. *Bioethics*, *30*(5), 304–311. https://doi.org/10.1111/bioe.12213

Beverley, J., & Beebe, J. (2018). Judgments of moral responsibility in tissue donation cases. *Bioethics*, *32*(2), 83–93. https://doi.org/10.1111/bioe.12412

Bisson, M. A., & Levine, T. R. (2009). Negotiating a friends with benefits relationship. *Archives of Sexual Behavior*, *38*(1), 66–73. https://doi.org/10.1007/s10508-007-9211-2

Bloom, P. (2011). Family, community, trolley problems, and the crisis in moral psychology. *The Yale Review*, *99*(2), 26–43. https://doi.org/10.1111/j.1467-9736.2011.00701.x

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Harvard University Press.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science*, *29*(7), 1084–1093. https://doi.org/10.1177/0956797617752640

Bowlby, J. (1982). *Loss: Sadness and Depression*. Basic Books.

Bowles, S. (2008). Policies designed for self-interested citizens may undermine "the moral sentiments": Evidence from economic experiments. *Science*, *320*(5883), 1605–1609. https://doi.org/10.1126/science.1152110

Brandt, R. B. (1979). *A Theory of the Good and the Right*. Clarendon Press.

Brinberg, M., Ram, N., Conroy, D. E., Pincus, A. L., & Gerstorf, D. (2021). Dyadic analysis and the reciprocal one-with-many model: Extending the study of interpersonal processes with intensive longitudinal data. *Psychological Methods*, online ahead of print. https://doi.org/10.1037/met0000380

Bugental, D. B. (2000). Acquisition of the algorithms of social life: A domain-based approach. *Psychological Bulletin*, *126*(2), 187–219. https://doi.org/10.1037/0033-2909.126.2.187

Burnstein, E., Crandall, C., & Kitayama, S. (1994). Some neo-Darwinian decision rules for altruism: Weighing cues for inclusive fitness as a function of the

biological importance of the decision. *Journal of Personality and Social Psychology*, *67*(5), 773–789. https://doi.org/10.1037/0022-3514.67.5.773

Cath, Y. (2016). Reflective equilibrium. In H. Cappelen, T. S. Gendler, & J. Hawthorn (Eds.), *The Oxford Handbook of Philosophical Methodology* (pp. 213–230). Oxford University Press.

Chalik, L., & Dunham, Y. (2020). Beliefs about moral obligation structure children's social category-based expectations. *Child Development*, *91*(1), e108–e119. https://doi.org/10.1111/cdev.13165

Chappell, S. G. (2018). Love and knowledge. In C. Grau & A. Smuts (Eds.), *The Oxford Handbook of Philosophy of Love* (online, pp. 1–19). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199395729.013.31

Christy, A. G., Schlegel, R. J., & Cimpian, A. (2019). Why do people believe in a "true self"? The role of essentialist reasoning about personal identity and the self. *Journal of Personality and Social Psychology*, *117*(2), 386–416. https://doi.org/10.1037/pspp0000254

Chuang, Y.-C. (1998). The cognitive structure of role norms in Taiwan. *Asian Journal of Social Psychology*, *1*, 239–251.

Clark, M. S., & Boothby, E. (2013). A strange(r) analysis of morality: A consideration of relational context and the broader literature is needed. *Behavioral and Brain Sciences*, *36*(1), 85–86. https://doi.org/10.1017/S0140525X12000751

Clark, M. S., Boothby, E., Clark-Polner, E., & Reis, H. (2015). Understanding prosocial behavior requires understanding relational context. In D. A. Schroeder & W. G. Graziano (Eds.), *The Oxford Handbook of Prosocial*

Behavior. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399813.013.37

Clark, M. S., Earp, B. D., & Crockett, M. J. (2020). Who are "we" and why are we cooperating? Insights from social psychology. *Behavioral and Brain Sciences*, *43*(e66), 21–23. https://doi.org/10.1017/S0140525X19002528

Clark, M. S., Lemay, E. P., & Reis, H. T. (2018). Other people as situations: Relational context shapes psychological phenomena. In J. F. Rauthmann, R. Sherman, & D. C. Funder (Eds.), *The Oxford Handbook of Psychological Situations* (p. online ahead of print). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190263348.013.5

Clark, M. S., & Mills, J. (1993). The difference between communal and exchange relationships: What it is and is not. *Personality and Social Psychology Bulletin*, *19*(6), 684–691. https://doi.org/10.1177/0146167293196003

Clark, M. S., & Mills, J. R. (1979). Interpersonal attraction in communal and exchange relationships. *Journal of Personality and Social Psychology*, *37*(1), 12–24.

Clark, M. S., & Mills, J. R. (2012). A theory of communal (and exchange) relationships. In P. A. M. V. Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of Theories of Social Psychology: Volume Two* (pp. 232–250). SAGE.

Clark, M. S., Mills, J. R., & Corcoran, D. M. (1989). Keeping track of needs and inputs of friends and strangers. *Personality & Social Psychology Bulletin*, *15*(4), 533–542.

Clark, M. S., & Taraban, C. (1991). Reactions to and willingness to express emotion in communal and exchange relationships. *Journal of Experimental Social Psychology*, *27*(4), 324–336. https://doi.org/10.1016/0022-1031(91)90029-6

Clark, M. S., & Waddell, B. (1985). Perceptions of exploitation in communal and exchange relationships. *Journal of Social and Personal Relationships*, *2*(4), 403–418. https://doi.org/10.1177/0265407585024002

Cohen, L. L., Cobb, J., & Martin, S. R. (2014). Gender biases in adult ratings of pediatric pain. *Children's Health Care*, *43*(2), 87–95. https://doi.org/10.1080/02739615.2014.849918

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition*, *179*, 241–265. https://doi.org/10.1016/j.cognition.2018.04.018

Cottingham, J. (2017). Love and religion. In C. Grau & A. Smuts (Eds.), *The Oxford Handbook of Philosophy of Love* (online, pp. 1–18). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199395729.013.33

Crimston, C. R., Bain, P. G., Hornsey, M. J., & Bastian, B. (2016). Moral expansiveness: Examining variability in the extension of the moral world. *Journal of Personality and Social Psychology*, *111*(4), 636–653. https://doi.org/10.1037/pspp0000086

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366. https://doi.org/10.1016/j.tics.2013.06.005

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings*

of the National Academy of Sciences, *111*(48), 17320–17325. https://doi.org/10.1073/pnas.1408988111

Curry, O. (2006). Who's afraid of the naturalistic fallacy? *Evolutionary Psychology*, *4*(1), 234–247. https://doi.org/10.1177/147470490600400120

Curry, O. S., Jones Chesters, M., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality*, *78*, 106–124. https://doi.org/10.1016/j.jrp.2018.10.008

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47–69. https://doi.org/10.1086/701478

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292. https://doi.org/10.1177/1088868313495594

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition*, *108*(1), 281–289. https://doi.org/10.1016/j.cognition.2008.02.005

Darwall, S. (2009). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Harvard University Press.

Darwall, S. (Ed.). (2010). Precis: The second-person standpoint. *Philosophy and Phenomenological Research*, *81*(1), 216–228. JSTOR.

Darwall, S. (2018). "Second-personal morality" and morality. *Philosophical Psychology*, *31*(5), 804–816. https://doi.org/10.1080/09515089.2018.1486603

Davies, R., Ives, J., & Dunn, M. (2015). A systematic review of empirical bioethics methodologies. *BMC Medical Ethics*, *16*(15), 1–13. https://doi.org/10.1186/s12910-015-0010-3

De Freitas, J., & Cikara, M. (2018). Deep down my enemy is good: Thinking about the true self reduces intergroup bias. *Journal of Experimental Social Psychology*, *74*(1), 307–316. https://doi.org/10.1016/j.jesp.2017.10.006

De Leersnyder, J., Mesquita, B., Kim, H., Eom, K., & Choi, H. (2014). Emotional fit with culture: A predictor of individual differences in relational well-being. *Emotion*, *14*(2), 241–245. https://doi.org/10.1037/a0035296

Demaree-Cotton, J. (2016). Do framing effects make moral intuitions unreliable? *Philosophical Psychology*, *29*(1), 1–22. https://doi.org/10.1080/09515089.2014.989967

Demaree-Cotton, J., & Kahane, G. (2018). The neuroscience of moral judgment. In K. Jones, M. Timmons, & A. Zimmerman (Eds.), *The Routledge Handbook of Moral Epistemology* (p. in press). Routledge.

Deng, Y., Wang, C. S., Aime, F., Wang, L., Sivanathan, N., & Kim, Y. C. (Karina). (2021). Culture and patterns of reciprocity: The role of exchange type, regulatory focus, and emotions. *Personality and Social Psychology Bulletin*, *47*(1), 20–41. https://doi.org/10.1177/0146167220913694

Dill, B., & Darwall, S. (2014). Moral psychology as accountability. In J. D'Arms & D. Jacobson (Eds.), *Moral Psychology and Human Agency* (pp. 40–83). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198717812.003.0003

Dwyer, S., Huebner, B., & Hauser, M. D. (2010). The linguistic analogy: Motivations, results, and speculations. *Topics in Cognitive Science*, *2*(3), 486–510. https://doi.org/10.1111/j.1756-8765.2009.01064.x

Earp, B. D. (2020). Falsification: How does it relate to reproducibility? In J.-F. Morin, C. Olsson, & E. Ö. Atikcan (Eds.), *Research Methdos in the Social Sciences: An A-Z of Key Concepts* (pp. 119–123). Oxford University Press.

Earp, B. D., & Boerner, K. E. (2019, April 4). *Does gender bias influence how people assess children's pain?* OUPblog. https://blog.oup.com/2019/04/gender-bias-children-pain/

Earp, B. D., Douglas, T., & Savulescu, J. (2017). Moral neuroenhancement. In S. Johnson & Rommelfanger (Eds.), *Routledge Handbook of Neuroethics* (pp. 166–184). Routledge.

Earp, B. D., Foddy, B., Wudarczyk, O. A., & Savulescu, J. (2017). Love addiction: Reply to Jenkins and Levy. *Philosophy, Psychiatry, & Psychology*, *24*(1), 101–103. https://doi.org/10.1353/ppp.2017.0014

Earp, B. D., Lewis, J., Dranseika, V., & Hannikainen, I. R. (2020). Experimental philosophical bioethics and normative inference. *Theoretical Medicine and Bioethics*, forthcoming.

Earp, B. D., McLoughlin, K., Monrad, J., Clark, M. S., & Crockett, M. (2021). How social relationships shape moral wrongness judgments. *Nature Communications*, *12*(5776), 1–13. https://doi.org/10.1038/s41467-021-26067-4

Earp, B. D., Monrad, J. T., LaFrance, M., Bargh, J. A., Cohen, L. L., & Richeson, J. A. (2019). Gender bias in pediatric pain assessment. *Journal of Pediatric Psychology*, *44*(4), 403–414. https://doi.org/10.1093/jpepsy/jsy104

Earp, B. D., & Savulescu, J. (2020). Love's dimensions. In *Love Drugs: The Chemical Future of Relationships* (pp. 16–35). Stanford University Press.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*(4), 583–610. https://doi.org/10.1007/s10683-011-9283-7

Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200–216. https://doi.org/10.1016/j.jesp.2018.07.004

Fabbri, R., & De León, F. G. (2017). A statistical distance derived from the Kolmogorov-Smirnov test: Specification, reference measures (benchmarks) and example uses. *ArXiv*, 1711.00761. https://arxiv.org/abs/1711.00761

Farber, P. L. (1994). *The Temptations of Evolutionary Ethics*. University of California Press.

Fei, X. (1992). *From the Soil: The Foundations of Chinese Society* (G. G. Hamilton & W. Zheng, Trans.). University of California Press.

Fisher, J. C. (2015). Pragmatic experimental philosophy. *Philosophical Psychology*, *28*(3), 412–433. https://doi.org/10.1080/09515089.2013.870546

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, *99*(4), 689–723.

Fiske, A. P., & Rai, T. S. (2014). *Virtuous Violence: Hurting and Killing to Create, Sustain, End, and Honor Social Relationships*. Cambridge University Press.

Fiske, A. P., & Tetlock, P. E. (1997). Taboo trade-offs: Reactions to transactions that transgress the spheres of justice. *Political Psychology*, *18*(2), 255–297. https://doi.org/10.1111/0162-895X.00058

Flanagan, O., Sarkissian, H., & Wong, D. (2008). Naturalizing ethics. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vol. 1, pp. 1–26). MIT Press. https://philarchive.org/archive/FLAN-4v1

Foot, P. (1972). Morality as a system of hypothetical imperatives. *The Philosophical Review*, *81*(3), 305–316. JSTOR. https://doi.org/10.2307/2184328

Foster, K. R., Wenseleers, T., & Ratnieks, F. L. W. (2006). Kin selection is the key to altruism. *Trends in Ecology & Evolution*, *21*(2), 57–60. https://doi.org/10.1016/j.tree.2005.11.020

Frankena, W. K. (1939). The naturalistic fallacy. *Mind*, *48*(192), 464–477.

Fried, B. H. (2012). What does matter? The case for killing the trolley problem (or letting it die). *The Philosophical Quarterly*, *62*(248), 505–529. https://doi.org/10.1111/j.1467-9213.2012.00061.x

Gelfand, M., Nishii, L., & Raver, J. (2006). On the nature and importance of cultural tightness-looseness. *Journal of Applied Psychology*, *91*(6), 1225–1244. https://doi.org/10.1037/0021-9010.91.6.1225

Gelfand, M., Raver, J. L., Nishii, L., Leslie, L., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Ferrer, M., Fischlmayr, I. C., … Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*(6033), 1100–1104. https://doi.org/10.1126/science.1197754

Gellner, D. N., Curry, O. S., Cook, J., Alfano, M., & Venkatesan, S. (2020). Morality is fundamentally an evolved solution to problems of social co-operation. *Journal of the Royal Anthropological Institute*, *26*(2), 415–427. https://doi.org/10.1111/1467-9655.13255

Ghai, S. (2021). It's time to reimagine sample diversity and retire the WEIRD dichotomy. *Nature Human Behaviour*, *5*(8), 971–972. https://doi.org/10.1038/s41562-021-01175-9

Gilead, M., David, Y. B., & Ecker, Y. (2018). Not our fault: Judgments of apathy versus harm toward socially proximal versus distant others. *Social Psychological and Personality Science*, *9*(5), 568–575. https://doi.org/10.1177/1948550617714583

Gillon, R. (1986). Doctors and patients. *British Medical Journal*, *292*(6518), 466–469.

Gold, N., Pulford, B. D., & Colman, A. M. (2014). The outlandish, the realistic, and the real: Contextual manipulation and agent role effects in trolley problems. *Frontiers in Psychology*, *5*(35), 1–10. https://doi.org/10.3389/fpsyg.2014.00035

Goldstein-Greenwood, J., Conway, P., Summerville, A., & Johnson, B. N. (2020). (How) do you regret killing one to save five? Affective and cognitive regret differ after utilitarian and deontological decisions. *Personality and Social Psychology Bulletin*, *46*(9), 1303–1317. https://doi.org/10.1177/0146167219897662

Gopnik, A. (2009). *The Philosophical Baby*. Macmillan.

Gray, K., & Keeney, J. E. (2015). Impure or just weird? Scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science*, *6*(8), 859–868. https://doi.org/10.1177/1948550615592241

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong & C. B. Miller (Eds.), *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (pp. 35–79). MIT Press.

Greene, J. D. (2015). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *The Law & Ethics of Human Rights*, *9*(2), 141–172. https://doi.org/10.1515/lehr-2015-0011

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. https://doi.org/10.1016/j.neuron.2004.09.027

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. https://doi.org/10.1126/science.1062872

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLOS ONE*, *14*(3), e0213544. https://doi.org/10.1371/journal.pone.0213544

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998–1002. https://doi.org/10.1126/science.1137651

Haidt, J., & Baron, J. (1996). Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology*, *26*(2), 201–218. https://doi.org/10.1002/(SICI)1099-0992(199603)26:2<201::AID-EJSP745>3.0.CO;2-J

Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology*, *31*(1), 191–221. https://doi.org/10.1111/j.1559-1816.2001.tb02489.x

Haidt, J., & Joseph, C. (2007). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence, & S. Stitch (Eds.), *The Innate Mind* (Vol. 3, pp. 367–391). Oxford University Press.

Hamilton, V. L., & Sanders, J. (1981). The effect of roles and deeds on responsibility judgments: The normative structure of wrongdoing. *Social Psychology Quarterly*, *44*(3), 237–254. https://doi.org/10.2307/3033836

Harman, G. (2008). Using a linguistic analogy to study morality. In W. Sinnott-Armstrong & C. B. Miller (Eds.), *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (pp. 107–143). MIT Press.

Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be? *Noûs*, *34*(1), 31–55.

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique*. OUP USA.

Hauser, M. D., Young, L. L., & Cushman, F. (2008). Reviving Rawls's linguistic analogy: Operative principles and the causal structure of moral action. In W. Sinnott-Armstrong & C. B. Miller (Eds.), *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (pp. 107–143). MIT Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hester, N., & Gray, K. (2020). The moral psychology of raceless genderless strangers. *Perspectives on Psychological Science*, *15*(2), 16–230. https://doi.org/10.1177/1745691619885840

Hirsch, J. L., & Clark, M. S. (2019). Multiple paths to belonging that we should study together. *Perspectives on Psychological Science*, *14*(2), 238–255. https://doi.org/10.1177/1745691618803629

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560

Hofstede, G. (1979). Hierarchical power distance in forty countries. In *Organizations Alike and Unlike*. Routledge.

hooks, bell. (2000). *All About Love: New Visions*. Harper.

Hume, D. (1777). *An Enquiry Concerning the Principles of Morals* (G. "Azra" Hernández, Ed.). Project Gutenberg.

Isern-Mas, C., & Gomila, A. (2020). Naturalizing Darwall's second person standpoint. *Integrative Psychological and Behavioral Science*, online ahead of print. https://doi.org/10.1007/s12124-020-09547-y

James-Hawkins, L., Qutteina, Y., & Yount, K. M. (2017). The patriarchal bargain in a context of rapid changes to normative gender roles: Young Arab women's role conflict in Qatar. *Sex Roles*, *77*(3), 155–168. https://doi.org/10.1007/s11199-016-0708-9

Jenkins, C. (2017). *What Love Is: And What it Could Be*. Basic Books.

Jenkins, C. S. I. (2017). “Addicted”? To “love”? *Philosophy, Psychiatry, & Psychology*, *24*(1), 93–96. https://doi.org/10.1353/ppp.2017.0012

Jonason, P. K., Hatfield, E., & Boler, V. M. (2015). Who engages in serious and casual sex relationships? An individual differences perspective. *Personality and Individual Differences*, *75*, 205–209. https://doi.org/10.1016/j.paid.2014.11.042

Jones, E. F., Parker, B. L., Joyner, M. H., & Ulku-Steiner, B. (1999). The influences of behavior valence and actor race on black and white children’s moral and liking judgments. *The Journal of Psychology*, *133*(2), 194–204. https://doi.org/10.1080/00223989909599733

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: A meta-analytic test of their relative validity. *Journal of Applied Psychology*, *89*(5), 755–768. https://doi.org/10.1037/0021-9010.89.5.755

Kagan, S. (2001). Thinking about cases. *Social Philosophy and Policy*, *18*(2), 44–63.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*(5), 551–560. https://doi.org/10.1080/17470919.2015.1023400

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018a). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164. https://doi.org/10.1037/rev0000093

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018b). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164. https://doi.org/10.1037/rev0000093

Kahane, G., Everett, J. A. C., 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. https://doi.org/10.1016/j.cognition.2014.10.005

Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgement. *Mind & Language*, *25*(5), 561–582. https://doi.org/10.1111/j.1468-0017.2010.01401.x

Kant, I., & Paton, H. J. (1785). *The Moral Law Groundwork of the Metaphysics of Morals*. Routledge. http://www.tandfebooks.com/isbn/9780203981948

Kaspar, K., Newen, A., Dratsch, T., de Bruin, L., Al-Issa, A., & Bente, G. (2016). Whom to blame and whom to praise: Two cross-cultural studies on the appraisal of positive and negative side effects of company activities. *International Journal of Cross Cultural Management*, *16*(3), 341–365. https://doi.org/10.1177/1470595816670427

Kelley, H. H. (1979). *Personal Relationships: Their Structures and Processes*. Lawrence Erlbaum Associates Publishers.

Kelley, H. H., Holmes, J. G., Kerr, N. L., Reis, H. T., Rusbult, C. E., & Van Lange, P. A. M. (2003). *An Atlas of Interpersonal Situations*. Cambridge University Press. https://doi.org/10.1017/CBO9780511499845

Kito, M., Yuki, M., & Thomson, R. (2017). Relational mobility and close relationships: A socioecological approach to explain cross-cultural differences. *Personal Relationships*, *24*(1), 114–130. https://doi.org/10.1111/pere.12174

Kneer, M., & Hannikainen, I. R. (2020). *Triage dilemmas: A window into (ecologically valid) moral cognition*. PsyArXiv. https://doi.org/10.31234/osf.io/v87sb

Knobe, J. (2016). Experimental philosophy is cognitive science. In J. Sytsma & W. Buckwalter (Eds.), *A Companion to Experimental Philosophy* (pp. 37–52). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118661666.ch3

Ko, A., Pick, C. M., Kwon, J. Y., Barlev, M., Krems, J. A., Varnum, M. E. W., Neel, R., Peysha, M., Boonyasiriwat, W., Brandstätter, E., Crispim, A. C., Cruz, J. E., David, D., David, O. A., de Felipe, R. P., Fetvadjiev, V. H., Fischer, R., Galdi, S., Galindo, O., … Kenrick, D. T. (2020). Family matters: Rethinking the psychology of human social motivation. *Perspectives on Psychological Science*, *15*(1), 173–201. https://doi.org/10.1177/1745691619872986

Koleva, S., Selterman, D., Iyer, R., Ditto, P., & Graham, J. (2014). The moral compass of insecurity: Anxious and avoidant attachment predict moral judgment. *Social Psychological and Personality Science*, *5*(2), 185–194. https://doi.org/10.1177/1948550613490965

Korsgaard, C. M. (1993). The reasons we can share: An attack on the distinction between agent-relative and agent-neutral values. *Social Philosophy and Policy*, *10*(1), 24–51. https://doi.org/10.1017/S0265052500004003

Korsgaard, C. M. (1996). *The Sources of Normativity*. Cambridge University Press.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4), 1–28. https://doi.org/10.3389/neuro.06.004.2008

Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. *Philosophy of Science*, *83*(2), 227–247. https://doi.org/10.1086/684960

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & Musgrave (Eds.), *Criticism and the Growth of Knowledge* (pp. 91–196). Cambridge University Press.

Law, K. F., Campbell, D., & Gaesser, B. (2021). Biased benevolence: The perceived morality of effective altruism across social distance. *Personality and Social Psychology Bulletin*, online ahead of print. https://doi.org/10.1177/01461672211002773

Lee, J., & Holyoak, K. J. (2018). "But he's my brother." How family obligation impacts moral judgments. In C. Kalish, M. Rau, J. Zhu, & T. T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1–7). Cognitive Science Society.

Lee, J., & Holyoak, K. J. (2020). "But he's my brother": The impact of family obligation on moral judgments and decisions. *Memory & Cognition*, *48*(1), 158–170. https://doi.org/10.3758/s13421-019-00969-7

Lewin, K. (1951). *Field Theory in Social Science*. Harper.

Lewis, J. (2020a). From x-phi to bioxphi: Lessons in conceptual analysis 2.0. *AJOB Empirical Bioethics*, *11*(1), 34–36. https://doi.org/10.1080/23294515.2019.1705430

Lewis, J. (2020b). Getting obligations right: Autonomy and shared decision making. *Journal of Applied Philosophy*, *37*(1), 118–140. https://doi.org/10.1111/japp.12383

Lindauer, M. (2020). Experimental philosophy and the fruitfulness of normative concepts. *Philosophical Studies*, *177*(8), 2129–2152. https://doi.org/10.1007/s11098-019-01302-3

Lindström, B., Jangard, S., Selbing, I., & Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General*, *147*(2), 228–242. https://doi.org/10.1037/xge0000365

Machery, E. (2017). *Philosophy Within Its Proper Bounds*. Oxford University Press.

Magee, J. C., & Galinsky, A. D. (2008). Social hierarchy: The self-reinforcing nature of power and status. *Academy of Management Annals*, *2*(1), 351–398. https://doi.org/10.5465/19416520802211628

Mammen, M., Köymen, B., & Tomasello, M. (2021). Young children's moral judgments depend on the social relationship between agents. *Cognitive Development*, *57*, 100973. https://doi.org/10.1016/j.cogdev.2020.100973

Mammen, M., Köymen, B., & Tomasello, M. (in press). Young children's moral judgments depend on the social relationship between agents. *Cognitive Development*.

Mark, K. P., Garcia, J. R., & Fisher, H. E. (2015). Perceived emotional and sexual satisfaction across sexual relationship contexts: Gender and sexual orientation differences and similarities. *The Canadian Journal of Human Sexuality*, *24*(2), 120–130. https://doi.org/10.3138/cjhs.242-A8

Marshall, J., Mermin-Bunnell, K., & Bloom, P. (2020). Developing judgments about peers' obligation to intervene. *Cognition*, *201*, 104215. https://doi.org/10.1016/j.cognition.2020.104215

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. https://doi.org/10.1111/cdev.13390

Mason, M. (2014). Reactivity and refuge. In D. Shoemaker & N. Tognazzini (Eds.), *Oxford Studies in Agency and Responsibility, Volume 2* (pp. 143–162). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198722120.003.0008

May, S. (2013). *Love: A History*. Yale University Press.

McDaniel, A. E. (2008). Measuring gender egalitarianism: The attitudinal difference between men and women. *International Journal of Sociology*, *38*(1), 58–80. https://doi.org/10.2753/IJS0020-7659380103

McGraw, A. P., & Tetlock, P. E. (2005). Taboo trade-offs, relational framing, and the acceptability of exchanges. *Journal of Consumer Psychology*, *15*(1), 2–15. https://doi.org/10.1207/s15327663jcp1501_2

McMahan, J. (2003). *The Ethics of Killing: Problems at the Margins of Life*. Oxford University Press.

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

McManus, R. M., Mason, J. E., & Young, L. (2021). Re-examining the role of family relationships in structuring perceived helping obligations, and their impact on moral evaluation. *Journal of Experimental Social Psychology*, *96*, 104182. https://doi.org/10.1016/j.jesp.2021.104182

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. https://doi.org/10.1016/j.tics.2006.12.007

Mill, J. S. (1863). *Utilitarianism*. Parker, Son and Bourn.

Miller, J. G., & Bersoff, D. M. (2016). Cultural influences on the moral status of reciprocity and the discounting of endogenous motivation. *Personality and Social Psychology Bulletin*, *20*(5), 592–602. https://doi.org/10.1177/0146167294205015

Miller, J. G., Bersoff, D. M., & Harwood, R. L. (1990). Perceptions of social responsibilities in India and in the United States: Moral imperatives or personal decisions? *Journal of Personality and Social Psychology*, *58*(1), 33–47. https://doi.org/10.1037/0022-3514.58.1.33

Mills, C. W. (2005). "Ideal theory" as ideology. *Hypatia*, *20*(3), 165–184.

Mills, J., Clark, M. S., Ford, T. E., & Johnson, M. (2004). Measurement of communal strength. *Personal Relationships*, *11*(2), 213–230. https://doi.org/10.1111/j.1475-6811.2004.00079.x

Mukerji, M. (2019). *Expermental Philosophy: A Critical Study*. Rowman & Littlefield.

Nado, J. (2019). Conceptual engineering via experimental philosophy. *Inquiry*, online ahead of print. https://doi.org/10.1080/0020174X.2019.1667870

Nakagawa, S., & Parker, T. H. (2015). Replicating research in ecology and evolution: Feasibility, incentives, and the cost-benefit conundrum. *BMC Biology*, *13*(88), 1–6. https://doi.org/10.1186/s12915-015-0196-3

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, *39*(1), 96–125. https://doi.org/10.1111/cogs.12134

Noddings, N. (2013). *Caring: A Relational Approach to Ethics and Moral Education*. University of California Press.

Notini, L., Earp, B. D., Gillam, L., McDougall, R. ., Savulescu, J., Telfer, M., & Pang, K. C. (2020). Forever young? The ethics of ongoing puberty suppression for non-binary adults. *Journal of Medical Ethics*, *46*(1), 743–752. https://doi.org/10.1136/medethics-2019-106012

Pataki, S. P., Fathelbab, S., Clark, M. S., & Malinowski, C. H. (2013). Communal strength norms in the United States and Egypt. *Interpersona*, *7*(1), 77–87. http://dx.doi.org/10.23668/psycharchives.2162

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Calò, M., Silani, G., Cikara, M., & Cushman, F. (2019). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *PsyArXiv*.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, *17*(3), 145–171. https://doi.org/10.1016/0162-3095(96)00041-6

Phillips, J., Misenheimer, L., & Knobe, J. (2011). The ordinary concept of happiness (and others like it). *Emotion Review*, *3*(3), 320–322. https://doi.org/10.1177/1754073911402385

Phillips, J., Mott, C., Freitas, J. D., Gruber, J., & Knobe, J. (2017). True happiness: The role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General*, *146*(2), 165–181.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741–763. https://doi.org/10.1037/0022-3514.67.4.741

Rai, T. S., & Fiske, A. P. (2011a). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57–75. https://doi.org/10.1037/a0021867

Rai, T. S., & Fiske, A. P. (2011b). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57–75. https://doi.org/10.1037/a0021867

Randall, T. E. (2019). Justifying partiality in care ethics. *Res Publica*, online ahead of print. https://doi.org/10.1007/s11158-019-09416-5

Rawls, J. (1999). *A Theory of Justice* (2nd edition). Belknap Press; r.

Rinne, T., Steel, G. D., & Fairweather, J. (2012). Hofstede and Shane revisited: The role of power distance and individualism in national-level innovation success. *Cross-Cultural Research*, *46*(2), 91–108. https://doi.org/10.1177/1069397111423898

Rodríguez-Arias, D., López, B. R., Monasterio-Astobiza, A., & Hannikainen, I. R. (2020). How do people use 'killing', 'letting die' and related bioethical concepts? Contrasting descriptive and normative hypotheses. *Bioethics*, *34*(5), 509–518. https://doi.org/10.1111/bioe.12707

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rowe, S. J., Vonasch, A. J., & Turp, M.-J. (2020). Unjustifiably irresponsible: The effects of social roles on attributions of intent. *Social Psychological and Personality Science*, online ahead of print. https://doi.org/10.1177/1948550620971086

Rueda, J., Hannikainen, I. R., Hortal-Carmona, J., & Rodriguez-Arias, D. (2020). Examining public trust in categorical versus comprehensive triage criteria. *The American Journal of Bioethics*, *20*(7), 106–109. https://doi.org/10.1080/15265161.2020.1779867

Savulescu, J., Kahane, G., & Gyngell, C. (2019). From public preferences to ethical policy. *Nature Human Behaviour*, *published online*(August 26), 1–3. https://doi.org/10.1038/s41562-019-0711-6

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Harvard University Press.

Scanlon, T. M. (2014). *Being Realistic about Reasons*. Oxford University Press.

Schein, C. (2020). The importance of context in moral judgments. *Perspectives on Psychological Science*, *15*(2), 207–215. https://doi.org/10.1177/1745691620904083

Schwitzgebel, E., Cokelet, B., & Singer, P. (2020). Do ethics classes influence student behavior? Case study: Teaching the ethics of eating meat. *Cognition*, *203*, 104397.

Searle, W., & Ward, C. (1990). The prediction of psychological and sociocultural adjustment during cross-cultural transitions. *International Journal of Intercultural Relations*, *14*(4), 449–464. https://doi.org/10.1016/0147-1767(90)90030-Z

Selterman, D., Moors, A. C., & Koleva, S. (2018). Moral judgment toward relationship betrayals and those who commit them. *Personal Relationships*, *25*(1), 65–86. https://doi.org/10.1111/pere.12228

Shweder, R. A. (1992). Ghostbusters in anthropology. In R. G. D'Andrade & C. Strauss (Eds.), *Human Motives and Cultural Models* (pp. 45–57). Cambridge University Press.

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The "Big Three" of morality (autonomy, community, divinity) and the "Big Three" explanations of suffering. In A. M. Brandt & P. Rozin (Eds.), *Morality and Health* (pp. 119–169). Psychology Press.

Simpson, A., & Laham, S. M. (2015). Individual differences in relational construal are associated with variability in moral judgment. *Personality and Individual Differences*, *74*, 49–54. https://doi.org/10.1016/j.paid.2014.09.044

Simpson, A., Laham, S. M., & Fiske, A. P. (2016). Wrongness in different relationships: Relational context effects on moral judgment. *The Journal of Social Psychology*, *156*(6), 594–609. https://doi.org/10.1080/00224545.2016.1140118

Simpson, J. A., & Gangestad, S. W. (1992). Sociosexuality and romantic partner choice. *Journal of Personality*, *60*(1), 31–51. https://doi.org/10.1111/j.1467-6494.1992.tb00264.x

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, *1*(3), 229–243.

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, *9*(3), 331–352. https://doi.org/10.1007/s10892-005-3508-y

Siraj, A. (2010). "Because I'm the man! I'm the head": British married Muslims and the patriarchal family structure. *Contemporary Islam*, *4*(2), 195–214.

Smith, E. E., & Medin, D. L. (2013). Categories and Concepts. In *Categories and Concepts*. Harvard University Press.

Sommers, T. (2007). The objective attitude. *The Philosophical Quarterly (1950-)*, *57*(228), 321–341. JSTOR.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, *132*(1), 99–107. https://doi.org/10.1007/s11098-006-9050-3

Stephens, N. M., Townsend, S. S. M., Markus, H. R., & Phillips, L. T. (2012). A cultural mismatch: Independent cultural norms produce greater increases in cortisol and more negative emotions among first-generation college students. *Journal of Experimental Social Psychology*, *48*(6), 1389–1393. https://doi.org/10.1016/j.jesp.2012.07.008

Sternberg, R. J. (1986). A triangular theory of love. *Psychological Review*, *93*(2), 119–135. https://doi.org/10.1037/0033-295X.93.2.119

Sterri, A. B., & Moen, O. M. (2020). The ethics of emergencies. *Philosophical Studies*, online ahead of print. https://doi.org/10.1007/s11098-020-01566-0

Strawson, P. F. (1962). *Freedom and Resentment and Other Essays*. Routledge.

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, *12*(4), 551–560. https://doi.org/10.1177/1745691616689495

Suhler, C. L., & Churchland, P. (2011). Can innate, modular "foundations" explain morality? Challenges for Haidt's Moral Foundations Theory. *Journal of Cognitive Neuroscience*, *23*(9), 2103–2116. https://doi.org/10.1162/jocn.2011.21637

Sunar, D., Cesur, S., Piyale, Z. E., Tepe, B., Biten, A. F., Hill, C. T., & Koç, Y. (2021). People respond with different moral emotions to violations in different relational models: A cross-cultural comparison. *Emotion*, *21*(4), 693–706. https://doi.org/10.1037/emo0000736

Sznycer, D., & Lukaszewski, A. W. (2019). The emotion–valuation constellation: Multiple emotions are governed by a common grammar of social valuation. *Evolution and Human Behavior*, *40*(4), 395–404. https://doi.org/10.1016/j.evolhumbehav.2019.05.002

Tan, D., & Snell, R. S. (2002). The third eye: Exploring guanxi and relational morality in the workplace. *Journal of Business Ethics*, *41*(4), 361–384. https://doi.org/10.1023/A:1021217027814

Taylor, S. E. (2017). *Loving the ones who are violent to us: An existential phenomenological study* [Psy.D., The Chicago School of Professional Psychology]. https://search.proquest.com/docview/1901897769/abstract/1E6D534012824BC7PQ/1

Tepe, B., & Aydınlı-Karakulak, A. (2019). Beyond harmfulness and impurity: Moral wrongness as a violation of relational motivations. *Journal of Personality and Social Psychology*, *117*(2), 310–337.

Thomasson, A. L. (2012). Experimental philosophy and the methods of ontology. *The Monist*, *95*(2), 175–199. https://doi.org/10.5840/monist201295211

Thomson, R., Yuki, M., Talhelm, T., Schug, J., Kito, M., Ayanian, A. H., Becker, J. C., Becker, M., Chiu, C., Choi, H.-S., Ferreira, C. M., Fülöp, M., Gul, P., Houghton-Illera, A. M., Joasoo, M., Jong, J., Kavanagh, C. M., Khutkyy, D., Manzi, C., … Visserman, M. L. (2018). Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat. *Proceedings of the National Academy of Sciences*, *115*(29), 7521–7526. https://doi.org/10.1073/pnas.1713191115

Tobin, T. W., & Jaggar, A. M. (2013). Naturalizing moral justification: Rethinking the method of moral epistemology. *Metaphilosophy*, *44*(4), 409–439. https://doi.org/10.1111/meta.12050

Tomasello, M. (2020). The moral psychology of obligation. *Behavioral and Brain Sciences*, *43*(e56), 1–58. https://doi.org/10.1017/ S0140525X19001742

Triandis, H. C., & Gelfand, M. J. (1998). Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology*, *74*(1), 118–128. https://doi.org/10.1037/0022-3514.74.1.118

Turiel, E. (2008). Thought about actions in social domains: Morality, social conventions, and social interactions. *Cognitive Development*, *23*(1), 136–154. https://doi.org/10.1016/j.cogdev.2007.04.001

Veit, W., & Browning, H. (2020). Two kinds of conceptual engineering. *PhilSciArchive*, 1–20. http://philsci-archive.pitt.edu/17452/

Velleman, J. D. (1999). Love as a moral emotion. *Ethics*, *109*(2), 338–374.

Volk, A. A., & Atkinson, J. A. (2013). Infant and child death in the human environment of evolutionary adaptation. *Evolution and Human Behavior*, *34*(3), 182–192. https://doi.org/10.1016/j.evolhumbehav.2012.11.007

Voorhees, E. M. (1985). The cluster hypothesis revisited. *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 188–196. https://doi.org/10.1145/253495.253524

Waytz, A., Dungan, J., & Young, L. (2013). The whistleblower's dilemma and the fairness–loyalty tradeoff. *Journal of Experimental Social Psychology*, *49*(6), 1027–1033. https://doi.org/10.1016/j.jesp.2013.07.002

Wedgwood, R. (2007). *The Nature of Normativity*. Clarendon Press.

Wei, M., Russell, D. W., Mallinckrodt, B., & Vogel, D. L. (2007). The Experiences in Close Relationship Scale (ECR)-Short Form: Reliability, validity, and factor structure. *Journal of Personality Assessment*, *88*(2), 187–204.

Weidman, A. C., Sowden, W. J., Berg, M. K., & Kross, E. (2020). Punish or protect? How close relationships shape responses to moral violations. *Personality and Social Psychology Bulletin*, *46*(5), 693–708. https://doi.org/10.1177/0146167219873485

Williams, C. L., Giuffre, P. A., & Dellinger, K. (1999). Sexuality in the workplace: Organizational control, sexual harassment, and the pursuit of pleasure. *Annual Review of Sociology*, *25*(1), 73–93. https://doi.org/10.1146/annurev.soc.25.1.73

Yudkin, D. A., Gantman, A. P., Hofmann, W., & Quoidbach, J. (2021). Binding moral values gain importance in the presence of close others. *Nature Communications*, *12*(1), 2718. https://doi.org/10.1038/s41467-021-22566-6

Appendix 1

Supplementary information for:

How Social Relationships Shape Moral Wrongness Judgments

Stage 1

Method.

Pre-registration #26400 on aspredicted.org. Full materials, raw data, and code available at https://osf.io/zxjt6/?view_only=66c1211300974dd68e97b88269fec4a3.

1.1. Participants. Using an online polling software (https://www.nbrii.com/our-process/sample-size-calculator/), we calculated that 385 participants would be required to have a representative U.S. sample with a 5% margin of error and 95% confidence level. Based on exclusion rates of previous studies conducted in our lab, we over-recruited by about 15% and recruited 450 U.S. participants representative for age, race, and gender via the Prolific Academic platform (Prolific). Ultimately 493 participants took some portion of the survey; not all of them finished. Participants were paid $0.60 to complete a training session in which they learned about the relational functions of care, coalition, hierarchy, reciprocity, and mating, plus a $2.00 bonus for completing the rest of the survey (rating each relationship on the extent to which it should ideally serve or not serve each function). To ensure high quality data, we included several attention, comprehension, and bot checks. Seventy (70) participants were excluded based on pre-registered exclusion criteria (see Supplementary Table 1). This left us with a final sample of 423 participants (217 female, 201 male, 4 other/nonbinary) ranging in age from 18 to 79 (M_{age} = 44.25, SD_{age} = 15.67); see Supplementary Table 2 for complete demographic information.

1.1.1. Supplementary Table 1
Summary of pre-registered exclusion criteria for Stage 1

Exclusion criteria met	Type of check	Excluded *N*
Did not reach main portion of survey (i.e., did not pass training)	Comprehension check	10

Did not type in the word 'FRIDAY'	Bot check	6
Did not move slider to (at least) 1 of 2 specified positions	Attention check	54

Note: some participants met more than one criterion.

1.1.2. Supplementary Table 2
Demographics of Sample 1 participants

Age	N (%)	Race	N (%)	Gender	N (%)
18 - 27	82 (19.39%)	White	296 (69.98%)	Female	217 (51.30%)
28 - 37	78 (18.44%)	Black/ African-American	60 (14.18%)	Male	201 (47.52%)
38 - 47	70 (16.55%)	Asian	29 (6.86%)	Other/ Non-binary	4 (0.95%)
48 - 57	75 (17.73%)	Hispanic/Latinx	21 (4.96%)		
58+	117 (27.66%)	Other	10 (2.36%)		
Missing	1 (0.24%)	American Indian/ Alaska Native	4 (0.95%)		
		Hawaiian/ Pacific Islander	2 (0.47%)		

1.2. Procedure. Participants completed a brief online survey through the Qualtrics interface. Before getting to the main part of the survey, participants were shown the full descriptions of each of the five main cooperative functions, as shown in Supplementary Table 3.

1.2.1. Supplementary Table 3:
Descriptions of relationship functions

Function	Description
Care	**Full version:** The main purpose of this kind of relationship is to make sure that a person's basic well-being is secure, without any strings attached to the giving or receiving of support (like expecting compensation or favors in return, or feeling a debt). In other words, it is to make sure that people have someone in their corner on whom they can truly count for care and support, in good times and bad.

	Note that there are two roles assumed by this relationship: the caregiving role (the person that can be truly counted on), and the care-seeking role (the person who may need unconditional support). **Brief version:** the function of giving or receiving unconditional support
Hierarchy	**Full version:** The main purpose of this kind of relationship is to help coordinate behavior between people with different status (often they have unequal power or responsibility). In many situations, it is most effective for one person to be the 'leader' or have final say about what happens. So, hierarchical relationships involve assigning people to different roles based on their status or power in a given situation, to help coordinate behavior and accomplish goals. There are two main roles in such relationships: the leader role and the follower role. The person in the leader role has 'final say' over what happens, while the person in the follower role ultimately must go along with what the leader decides. **Brief version:** the function of coordinating behavior between people of different status, power, or responsibility
Mating	**Full version:** The main purpose of this kind of relationship is to find and maintain a sexual partner. For our ancestors, the ultimate point of mating was to produce healthy offspring, so that we could pass on our genes and continue as a species. Of course, today we have birth control, and people often have sexual relationships without consciously planning to have children. But the underlying "logic" of the mating relationship -- in terms of the feelings and motivations it tends to inspire -- remains the same: to attract and secure a mate and stay with that person long enough to at least potentially have children together. **Brief version:** the function of establishing and maintaining a sexual partnership
Reciprocity	**Full version:** The main purpose of this kind of relationship is to coordinate behavior between people who can mutually benefit, where they each have equal say in a given situation or activity. So, this can be a kind of ‘tit-for-tat’ arrangement between people, where each one says, “I'll scratch your back if you scratch mine.” It can also be a way to keep things fair between people exchanging favors, goods, or services. It can even be a way to coordinate activities for mutual enjoyment, such as playing a game together. In any case, it involves making sure that the scales between people don't get too far out of balance. **Brief version:** the function of coordinating behavior between people with equal say in a situation and keeping things fair
Coalition	**Full version:** The main purpose of this kind of relationship is to form and maintain a group identity, so group members can work toward a common goal. People in the same group look out for each other and try to promote their own group's interests over the interests of competing groups: “us versus them.” This involves having shared expectations for what's normal or appropriate behavior and potentially making sacrifices for the good of the group, especially when it's in danger. **Brief version:** the function of forming and maintaining a group identity for a common goal: us versus them

To ensure that participants were paying attention and were thinking of the functions in the way we wanted them to, each description was followed by a multiple-choice question about the definitions of the functions. Participants were not allowed to advance to the main part of the survey if they failed to answer this check correctly.

Once we had introduced the participants to the five cooperative functions, we gave them instructions for the main task of the survey. For each of 20 relationships, we asked participants how much the relationship ideally should serve each of the five cooperative functions. We specified that, by 'ideally,' we meant that "if this kind of relationship was the best possible relationship of its kind it could be," how much should it serve each of those five functions?

Participants were then presented 20 blocks of questions, one for each relationship, in random order. For each relationship, we included a specific description of what we meant by that relationship – see Supplementary Table 4 for the descriptions. Then, for each combination of relationship and cooperative function, participants rated how much the relationship ideally should serve the given function, with the function now presented in its brief form as a reminder (see Supplementary Table 3). For instance, if the relationship and cooperative function pair was siblings/care, participants would be asked: "To what extent should the relationship between siblings ideally serve the function of **giving or receiving unconditional support**? (Care function.)" Responses were recorded on a sliding scale ranging from 'Definitely SHOULD NOT' (-100) through 'Neutral' (0) to 'Definitely SHOULD' (+100). Each of the recruited participants responded to questions about all five functions for all 20 relationships, yielding 100 data points per participant. Finally, we collected a battery of demographic measures: gender, age, race, ethnicity, income, level of education, English fluency, political leanings on social and economic issues, and religiosity.

1.2.2. Supplementary Table 4
Descriptions of relationships

Relationship	Description
Siblings	This refers to brothers and sisters. It includes adoptive as well as biological brothers and sisters.
Long-term romantic partners	This refers to romantic partners that have a commitment to each other, meaning they intend to remain together for the long term. It includes married

	partners like spouses, but also long-term romantic partners who aren't married but still have a commitment.
Close friends	This refers to people who seek one another out to spend time together, and who refer to each other as best friends, close friends, or good friends. It includes two people of any gender combination who are committed to remaining friends for the long term, and whose interest in each other is NOT primarily romantic.
Work colleagues or classmates	This refers to people who interact with each other on a regular basis at work/school or in work/school related activities. It includes only people who are on the same career or schooling level as each other.
Boss and employee	This refers to any workplace relationship in which one person directly supervises the other and that same person has some decision-making control over the other's activities and outcomes. It includes only people who interact with each other in person in a workplace.
Teacher and student	This refers to relationships in which someone with more experience guides someone with less experience, such as in school, sports or other extracurricular activity. It includes teacher-student or coach-player relationships, for example.
Doctor and patient	This refers to relationships in which one person provides expert medical or therapeutic services to the other person. It includes medical doctor-patient relationships or therapist-client relationships, for example.
Extended family members	This refers to relationships with relatives who are not in the immediate family (so, not parents or siblings). It includes relationships with a cousin, aunt, uncle, grandfather or grandmother.
Teammates	This refers to a relationship between two people who are in a clearly defined group together working towards a common goal. It includes teammates on a sports team or members of a theater troupe, for example.
Customer and seller	This refers to any relationship in which someone sells something and another person buys that thing, and they interact with each other directly. It includes a local baker and the baker's customer, or a house cleaner and the person who owns the house, for example.
Neighbors	This refers to people who live in the same section of an apartment building, or within the same block of houses on a street. It includes only people who know each other and at least occasionally interact.
Friends with benefits	This refers to people who know each other and interact sexually on a somewhat regular basis, but who refer to each other as friends rather than romantic partners. It includes friends who interact sexually without committing to a monogamous romantic relationship.
Acquaintances	This refers to people who know each other, and interact now and then, but don't consider one another to be friends. It includes people one might know from work, school, or the neighborhood.
Members of a political party	This refers to people who are active, registered members of the same political party. They likely hold similar ideological positions and tend to vote for the same candidates in elections. They may not necessarily know each other for any other reason than their common political activities and party affiliation.
Roommates/housemates	This refers to anyone who lives together who are not family members or in a romantic relationship. This includes college room- or suite mates, apartment mates, or housemates.

Father and child (under 18)	This refers to fathers and their non-adult (meaning under age 18) children specifically. It includes both biological and adoptive fathers/children.
Mother and child (under 18)	This refers to mothers and their non-adult (meaning under age 18) children specifically. It includes both biological and adoptive parents/children.
Father and child (over 18)	This refers to fathers and their adult (meaning over age 18) children specifically. It includes both biological and adoptive fathers/children.
Mother and child (over 18)	This refers to mothers and their adult (meaning over age 18) children specifically. It includes both biological and adoptive mothers/children.
Strangers	This refers to people who encounter each other in any setting for the first time. It includes people who don't know each other from another context and don't anticipate interacting again in the future.

1.3. Data preparation and analysis details. Raw data files (.csv) were prepared and analyzed using Python, within a Jupyter Notebook environment. Primary packages used: numpy, scipy, statsmodels, matplotlib, seaborn, pandas. For data files and all coding scripts, see the OSF link above.

1.4. Supplementary results.

1.4.1. Supplementary Table 5
Most to least functionally polarized relationships

	Functional Expectations				
	Care	Hier.	Mate.	Recip.	Across all functions
	M	*M*	*M*	*M*	*SD*
Siblings	84.06	-0.83	-98.47	74.96	74.86
Mother/child (under 18)	95.53	65.39	-98.22	9.05	74.10
Father/child (under 18)	93.99	67.39	-98.37	12.93	73.99
Father/child (over 18)	86.58	35.97	-97.59	53.62	70.32
Mother/child (over 18)	88.05	28.02	-97.69	49.75	69.79
Boss and employee	7.86	84.75	-92.17	29.14	65.45
Extended family members	64.65	15.22	-96.40	57.26	65.00
Teacher and student	42.97	72.77	-95.61	24.49	64.30
Teammates	50.53	31.86	-73.00	75.43	63.72
Doctor and patient	53.75	41.63	-95.31	30.40	61.21
Close friends	79.39	-20.32	-50.31	79.96	59.18

Members of a political party	7.60	45.24	-66.04	63.82	55.39
Customer and seller	-18.37	22.04	-81.77	60.39	53.16
Colleagues or classmates	17.38	18.92	-60.00	77.29	50.93
Roommates or housemates	24.90	-4.48	-52.39	87.30	50.71
Neighbors	13.08	-16.25	-57.99	67.57	46.35
Romantic partners	92.43	-14.07	95.12	84.95	45.80
Friends with benefits	28.13	-30.36	58.43	59.87	39.87
Acquaintances	-2.74	-3.54	-46.82	51.38	34.85
Strangers	-26.62	-11.00	-55.87	34.95	33.46

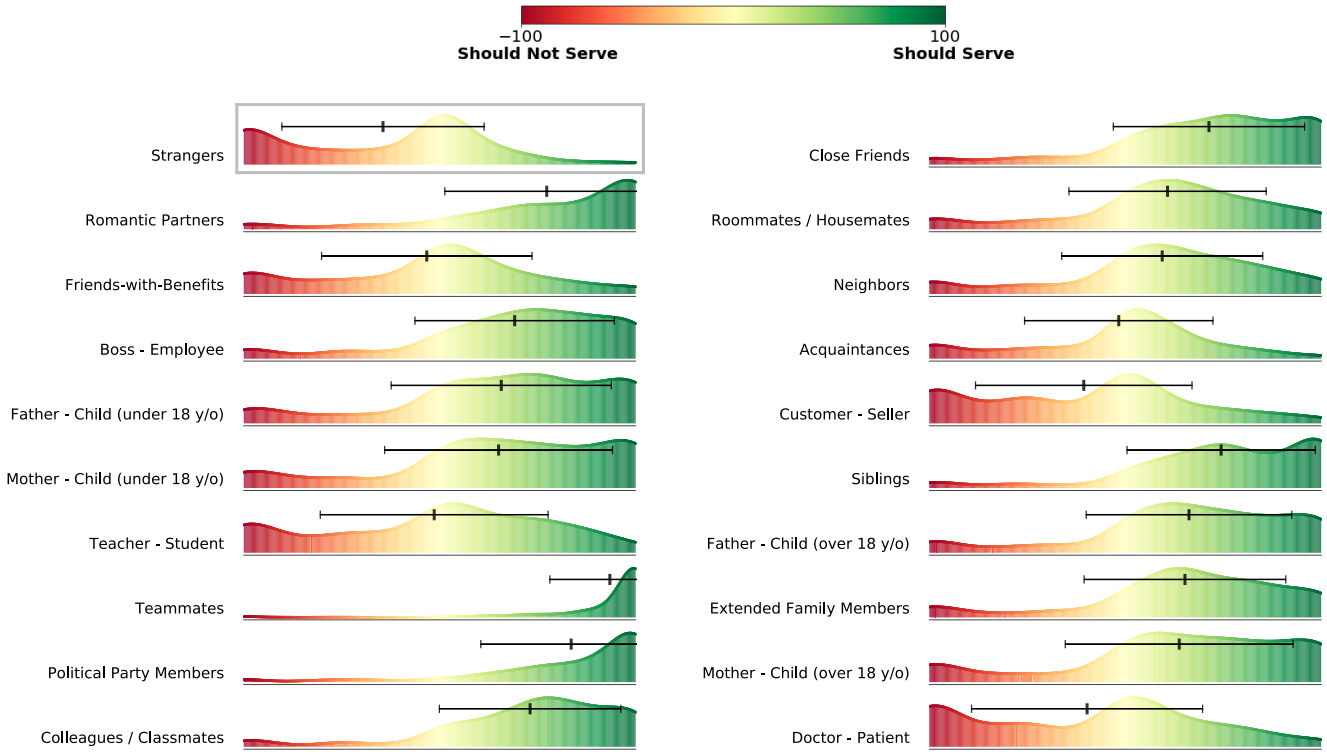
1.4.2. Supplementary Table 6

Most to least functionally specific relationships

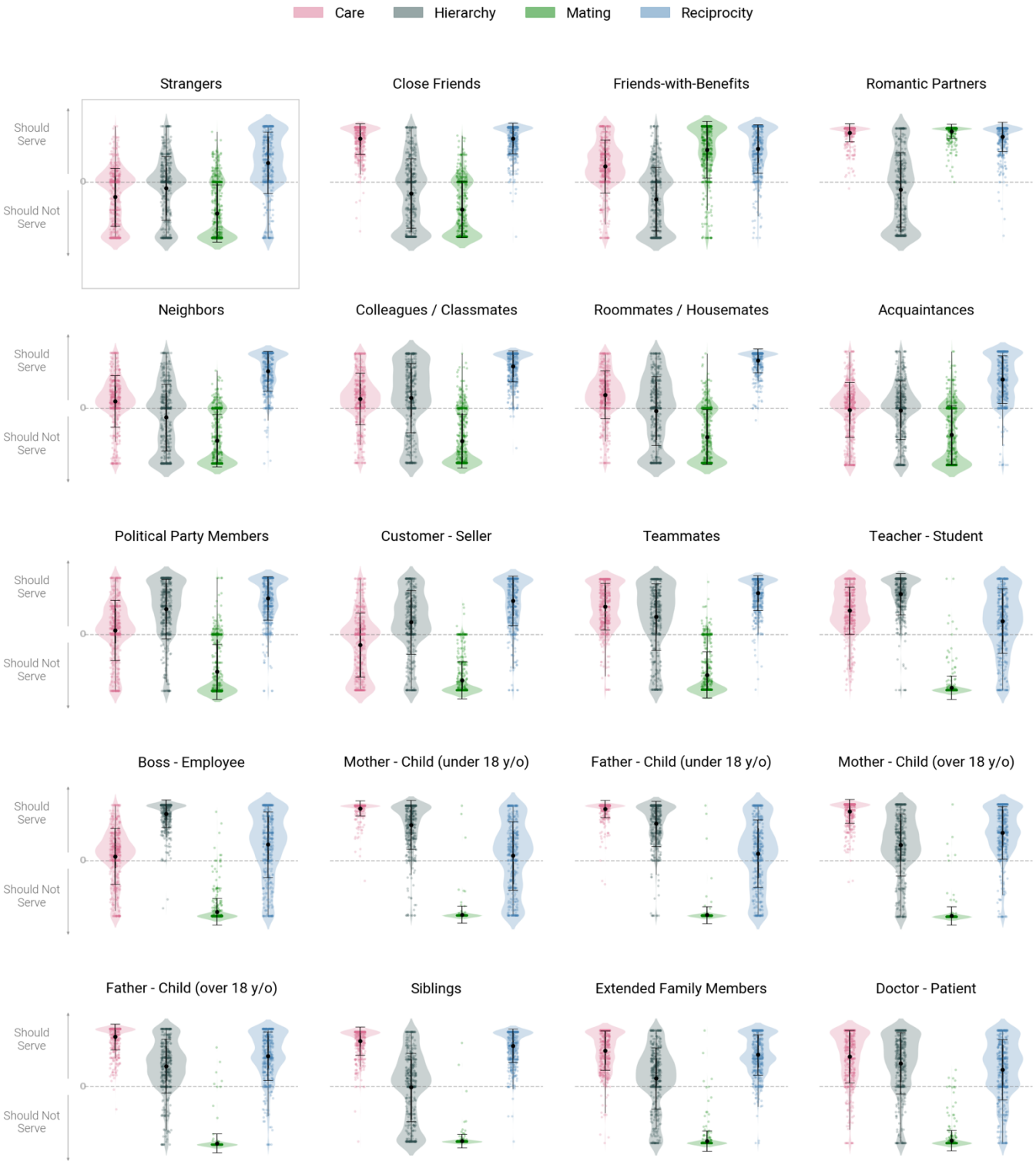
	Functional Expectations							
	Care	Hier.	Mate.	Recip.	Highest *M* Function	Highest *M* Value	Other Sum	Max Other Difference
	M	*M*	*M*	*M*		*M*		
Boss and employee	7.86	84.75	-92.17	29.14	Hierarchy	84.75	-55.17	139.92
Customer and seller	-18.37	22.04	-81.77	60.39	Reciprocity	60.39	-78.09	138.49
Neighbors	13.08	-16.25	-57.99	67.57	Reciprocity	67.57	-61.15	128.72
Strangers	-26.62	-11.00	-55.87	34.95	Reciprocity	34.95	-93.49	128.44
Mother/child (under 18)	95.53	65.39	-98.22	9.05	Care	95.53	-23.78	119.31
Roommates or housemates	24.90	-4.48	-52.39	87.30	Reciprocity	87.30	-31.96	119.25
Father/child (under 18)	93.99	67.39	-98.37	12.93	Care	93.99	-18.05	112.04
Siblings	84.06	-0.83	-98.47	74.96	Care	84.06	-24.34	108.40
Mother/child (over 18)	88.05	28.02	-97.69	49.75	Care	88.05	-19.91	107.96
Acquaintances	-2.74	-3.54	-46.82	51.38	Reciprocity	51.38	-53.09	104.47
Colleagues or classmates	17.38	18.92	-60.00	77.29	Reciprocity	77.29	-23.70	100.99
Teacher and student	42.97	72.77	-95.61	24.49	Hierarchy	72.77	-28.14	100.91
Father/child (over 18)	86.58	35.97	-97.59	53.62	Care	86.58	-8.00	94.59

Extended family members	64.65	15.22	-96.40	57.26	Care	64.65	-23.92	88.57
Doctor and patient	53.75	41.63	-95.31	30.40	Care	53.75	-23.29	77.04
Members of a political party	7.60	45.24	-66.04	63.82	Reciprocity	63.82	-13.20	77.02
Close friends	79.39	-20.32	-50.31	79.96	Reciprocity	79.96	8.77	71.19
Teammates	50.53	31.86	-73.00	75.43	Reciprocity	75.43	9.38	66.05
Friends with benefits	28.13	-30.36	58.43	59.87	Reciprocity	59.87	56.20	3.68
Romantic partners	92.43	-14.07	95.12	84.95	Mating	95.12	163.31	-68.18

1.4.3. *Supplementary Figure 1*. Coalition ratings. Kernel density plot of functional expectations for coalition only for 20 common relationships. Dot represents the mean; cap represents +/- 1 standard deviation. The height of the curve represents density: the likely proportions of scores (relative to each function) that fall within the given range along the x-axis. Source data are provided as a Source Data file.



1.4.4. *Supplementary Figure 2.* Relational norm profiles: violin plots for all 20 relationships. Error bars represent the mean (dot) and +/- 1 SD (caps). Source data are provided as a Source Data file.



1.4.5. Supplementary Tables 7a-7e.
Complete demographic analyses for Stage 1/Sample 1: regression tables

These regression tables show the results of the mixed effects linear regression models where participant and relationship are random factors, described in the main text. Note that participants who reported gender as 'other' were excluded from the analysis as this was not a big enough sub-group to make statistical comparisons possible. Table a = overall model; b = gender and other demographic effects on functional expectations for mating; c = same for care; d = same for reciprocity; e = same for hierarchy. Source data are provided as a Source Data file.

a. Overall

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	13.74	1.00	[11.8, 15.69]	< .001	
Gender (Female)	1.00	.84	[-.66, 2.65]	.24	
Income (High)	-1.17	.82	[-2.79, .45]	.16	
Economic Ideology (Liberal)	-2.10	1.12	[-4.3, .10]	.06	
Social Ideology (Liberal)	2.49	1.14	[.27, 4.72]	.03	
Religiosity (Very Religious)	.61	.88	[-1.13, 2.34]	.49	
					Conditional R^2 = .53

Note. *LL* and *UL* indicate the lower and upper limits of a 95% confidence interval, respectively. The same applies to all other relevant tables. Here, the coefficients and confidence intervals are based on the raw data, whereas the corresponding analyses reported in the main text used functional expectations scaled to each participant.

b. Mating

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	-57.42	1.66	[-60.67, -54.17]	< .001	
Gender (Female)	-7.15	1.40	[-9.9, -4.40]	< .001	
Income	-3.61	1.37	[-6.30, -.92]	.01	

(High)				
Economic Ideology (Liberal)	-5.65	1.87	[-9.32, -1.99]	.003
Social Ideology (Liberal)	7.23	1.89	[3.52, 10.93]	< .001
Religiosity (Very Religious)	-2.99	1.47	[-5.88, -.12]	.04
				Conditional R^2 = .72

c. Care

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	39.84	1.5	[36.89, 42.78]	< .001	
Gender (Female)	6.36	1.27	[3.87, 8.86]	< .001	
Income (High)	-.57	1.24	[-3.01, 1.87]	.65	
Economic Ideology (Liberal)	-2.10	1.70	[-5.43, 1.22]	.22	
Social Ideology (Liberal)	.64	1.72	[-2.72, 4.00]	.71	
Religiosity (Very Religious)	5.05	1.33	[2.43, 7.66]	< .001	
					Conditional R^2 = .62

d. Reciprocity

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	51.10	1.33	[48.49, 53.71]	< .001	
Gender (Female)	4.80	1.13	[2.59, 7.01]	< .001	
Income (High)	-1.97	1.10	[-4.13, .19]	.07	

Economic Ideology (Liberal)	1.01	1.50	[-1.94, 3.95]	.50
Social Ideology (Liberal)	.62	1.52	[-2.36, 3.60]	.68
Religiosity (Very Religious)	1.64	1.18	[-.68, 3.96]	.17

Conditional R^2 = .37

e. Hierarchy

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	21.46	1.71	[18.10, 24.81]	< .001	
Gender (Female)	-.04	1.45	[-2.88, 2.80]	.98	
Income (High)	1.46	1.42	[-1.32, 4.24]	.30	
Economic Ideology (Liberal)	-1.65	1.93	[-5.44, 2.14]	.39	
Social Ideology (Liberal)	1.48	1.95	[-2.35, 5.31]	.45	
Religiosity (Very Religious)	-1.27	1.52	[-4.24, 1.71]	.41	

Conditional R^2 = .46

Stage 2

Method.

Pre-registration #31592 on aspredicted.org. Full materials, raw data, and code available at https://osf.io/zxjt6/?view_only=66c1211300974dd68e97b88269fec4a3.

2.1. Materials creation.

Selection of relationships. In Stage 1, we collected data for 20 different relationships. Our analysis of the Kolmogorov-Smirnov (K-S) distance between relationships (see the main text) revealed that several of these relationships were highly similar in terms of their prescribed cooperative functions (i.e., relational norm profiles). To avoid redundancy in Stage 2, we decided to use a subset of 10 relationships for the following study that were among the most functionally dissimilar to each other. Our procedure for selecting these relationships is described next.

As noted in the main manuscript, we first excluded the data pertaining to the coalition function. We then calculated the mean K-S distance score for every pairing of relationships across the four remaining functions. We set up 10 slots for "face-offs" between pairs of relationships with low K-S distances. For each face-off, we dropped the relationship with the lowest mean K-S distance from all other relationships (i.e., the more redundant of the two relationships considered in the context of the entire set). To enable gender comparisons, however, we first instituted a rule such that, if a father/mother relationship faced off, each was retained. Then, we moved sequentially from the lowest K-S score pairings to the highest K-S score pairings until all 10 slots were filled, dropping relationships along the way according to the first rule (if a relationship faced off with another relationship that had already been eliminated, it was retained by default). The final set of relationships identified by this method were: long-term romantic partners, friends with benefits, boss and employee, colleagues or classmates, mother/father and under-18 child, siblings, close friends, roommates or housemates, teammates, and strangers.

Selection of action statements. In this part of the study, we sought to test the hypothesis that the relational norm profile of a given relationship (based on the prescribed cooperative functions identified in Stage 1) would predict moral judgments for violations of associated functions in the context of that relationship. To this end, we created a set of 86 actions that we thought would plausibly weaken or impair one or more cooperative functions. To determine the extent to which each action would characteristically weaken (or strengthen) each of the four dyadic cooperative functions, we had 15 trained judges rate each of the 86 actions in our set. These judges were recruited among lab members and colleagues and were given extensive training either in person or over Skype to ensure that their ratings reflected only the functional implications of each action (i.e., according to the cooperative logic of the functions) rather than being about moral judgments of any kind.

The trained judges completed an online survey, which included the same descriptions of cooperative functions that we used in Stage 1. After reading these descriptions and completing multiple comprehension checks, the judges were shown the 86 action statements, all of which were of the form "Person A does X to Person B" (see original materials at the OSF link above for the full list of actions). For each action, the judges were asked about the extent to which the action would weaken or strengthen each of the five functions. Their responses were recorded on a sliding scale, ranging from "Would characteristically **weaken**" (-100) through "It depends / Would neither weaken nor strengthen" (0) to "Would characteristically **strengthen**" (100). Thus, we obtained a mean rating between -100 and 100 for each action-function pair.

Based on the judges' ratings, we used an algorithm (described in the main text)[18] to identify 12 actions (three for each function) that were most characteristic as individual function-weakeners, while maintaining roughly equal "characteristicness" of items across functions.[19] This process resulted in a final set of 12 function-weakening action statements, depicted in Supplementary Table 8.

2.1.1. Supplementary Table 8
Action statements used in Stage 2

Function	Action
Care	- Person A sees Person B crying and walks away from them - Person A keeps checking their cellphone while Person B tells a sad personal story - Person A watches passively while Person B carries several heavy boxes up the stairs, even though they could easily help
Hierarchy	- Person A refuses to follow a reasonable order from Person B - Person A repeatedly interrupts Person B while they are speaking - Person A decides to skip a meeting scheduled with Person B without a good excuse
Mating	- Person A refuses to have sex with Person B - Person A repeatedly turns down Person B's offer to go on a romantic date - Person A invests time and energy in a romantic relationship with someone other than Person B

[18] MATLAB code is available at https://osf.io/j435q/.
[19] Three items from the 86-action set were excluded before running the algorithm due to having misleading or ambiguous wording. After running the code, we also dropped the item "Person A mocks Person B for a poor performance, when Person B tried their best," as we realized that the word "performance" could have different meanings across relationships (e.g., job performance versus an artistic performance).

Reciprocity	- Person A decides not to pay Person B back, hoping Person B won't remember - Person A decides not to return Person B's nice favor - Person A charges Person B $50 for an item worth $25.

2.2. Participants. In Stage 1, we powered for 95% confidence in a 5% margin of error for a nationally representative sample across age, race, and gender. This required that we have 385 observations per distribution (with each participant rating all 20 original relationships on all 5 original functions). To ensure that our Sample 2 distributions would be comparable to those from Sample 1, we powered for the same confidence in the same margin of error. In the first sample, every participant gave one rating per function for all of 20 relationships; in this study, Sample 2 participants would give three ratings for just one relationship out of a smaller set of 13 relationships (13 because questions regarding the non-symmetrical relationships, e.g., boss-employee, were asked in both directions). To achieve parity, then, we multiplied the previous target sample of 385 (per distribution) by 13 (accounting for the switch to a between-subjects design) and divided by three (accounting for three ratings per function in the current sample, compared to just one in Sample 1), yielding a required sample of 1,551. As in Stage 1, we over-recruited by about 10%, aiming for 1,706 participants for Stage 2. Ultimately, 1,822 participants took at least part of the survey (not all finished).

Once again, the participants were recruited online and were paid were paid $1.00 to complete the survey. Five hundred and two (502) participants were excluded based on pre-registered exclusion criteria (see Supplementary Table 9 for criteria), leaving us with a final sample of 1,320 participants (553 female, 758 male, 6 other/non-binary) ranging in age from 18 to 73 (M_{age} = 35.32, SD_{age} = 10.56). Se Supplementary Table 10 for complete demographic information.

2.2.1. Supplementary Table 9
Summary of exclusion criteria for Stage 2

Exclusion criteria met	Type of check	Excluded *N*
Failed question about survey instructions OR answered multiple-choice question incorrectly (correct answer includes the word "embarrassment")	Comprehension/attention check	246

Did not move slider to (at least) 1 of 2 specified position	Attention check	171
Failed bot-checker test	Bot check	0
Failed text-entry test	Bot check	53
Being younger than 18	Demographic check	2
Not fluent English speaker	Demographic check	12
Finished survey in < 4 min.	Quality check	73

Note: some participants met more than one criterion.

2.2.2. Supplementary Table 10
Demographics of Sample 2 participants

Age	**N (%)**	**Race**	**N (%)**	**Gender**	**N (%)**
18 - 27	329 (24.96%)	White	929 (70.49%)	Female	553 (41.96%)
28-37	561 (42.56%)	Black/ African-American	175 (13.28%)	Male	758 (57.51%)
38-47	230 (17.45%)	Asian	92 (6.98%)	Other/ Non-binary	6 (0.46%)
48-57	135 (10.24%)	Hispanic/ Latinx	88 (6.68%)		
58+	63 (4.78%)	Other	19 (1.44%)		
Missing	0 (0.00%)	American Indian/ Alaska Native	12 (0.91%)		
		Hawaiian/ Pacific Islander	1 (0.08%)		

2.3. Procedure. Each participant was assigned to one of 13 relationship-pairs (13 because questions about the non-symmetrical relationships, e.g., boss-employee, were asked separately in both directions), and were shown a brief description of their assigned relationship (see Supplementary Table 4). We informed participants that they would be asked to consider various actions in the context of their assigned relationship and to answer how morally wrong each of those actions would be. To orient them to the rating scale, we clarified that none of the actions they would see would be extreme (e.g., murder), but rather would all be actions that might plausibly occur within the course of day-to-day life. We then 'anchored' their expectations by showing them a list of actions comparable to the ones included in the task.

Following instructions and attention checks, participants were shown the 12 (three for each of the four functions) hypothetical actions selected by our algorithmic approach described above. For instance, for the sibling relationship, participants were asked, e.g., "Imagine that someone keeps checking their cellphone while their sibling tells a sad personal story. How morally wrong would that be, if at all?" Responses were recorded on a sliding scale from "Not at all morally wrong" (0) to "Very morally wrong" (100). Finally, we collected exploratory data about how likely it is that each action would occur in real life, plus the same demographic measures that we collected from Sample 1 participants.

2.4. Data preparation and analysis details. Raw data files (.csv) were prepared and analyzed using Python, within a Jupyter Notebook environment. Primary packages used: numpy, scipy, statsmodels, matplotlib, seaborn, pandas. For data files and all coding scripts, see the OSF link above.

2.5. Supplementary results.

2.5.1. Supplementary Tables 11a-11e.
Complete demographic analyses for Stage 2/Sample 2: regression tables.
These regression tables show the results of the mixed effects linear regression models where participant and relationship are random factors, described in the main text. Note that participants who reported gender as 'other' were excluded from the analysis as this was not a big enough sub-group to make statistical comparisons possible. For each set of regression results below, an Anderson-Darling test indicated that the outcome variable was non-normally distributed (all *p*s < .001). However, Q-Q plots of standardized residuals against fitted values indicated that this did not impact model

performance.[20] Table a = overall model; b = gender and other demographic effects on moral wrongness judgments for mating; c = same for care; d = same for reciprocity; e = same for hierarchy.

a. Overall

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	39.85	1.35	[37.21, 42.50]	< .001	
Gender (Female)	-.12	1.22	[-2.51, 2.28]	.92	
Income (High)	1.49	1.20	[-.87, 3.85]	.22	
Economic Ideology (Liberal)	-.48	1.54	[-3.50, 2.53]	.75	
Social Ideology (Liberal)	1.06	1.55	[-1.98, 4.10]	.50	
Religiosity (Very Religious)	8.93	1.22	[6.54, 11.32]	< .001	
					Conditional R^2 = .56

b. Mating

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	-.07	1.83	[-3.66, 3.52]	.97	
Gender (Female)	-2.94	1.66	[-6.19, .31]	.08	

[20] The Q-Q plot (quantile-quantile plot) is a visualization that assesses whether data plausibly came from a theoretical distribution, such as a normal distribution. Q-Qs can be used to inspect whether the assumption of a normally distributed outcome measure required by certain statistical tests is violated. The graph plots two sets of quantiles or percentiles – thresholds below which certain points of our data fall – against one another: the observed quantiles and those of the theoretical (normal) distribution. If both sets of quantiles came from the same distribution, the points should form a line that is roughly straight. In the case of our linear mixed effects model predicting moral wrongness judgments from functional expectations, the Q-Q plot forms a nearly perfect straight line (see the code listed at https://osf.io/zxjt6/). This suggests that the observed non-normality of our dependent measure did not violate the normality assumption of the model. Gelman and Hill (2007) note that the normality or otherwise of residuals doesn't affect the parameter estimates in multilevel models. They therefore advise against normality tests of regression residuals (p. 46). Gelman, A., Hill, J., 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Income (High)	5.09	1.63	[1.89, 8.29]	.002
Economic Ideology (Liberal)	2.00	2.09	[-2.09, 6.10]	.34
Social Ideology (Liberal)	3.17	2.10	[-.95, 7.30]	.13
Religiosity (Very Religious)	16.53	1.66	[13.28, 19.78]	< .001

Conditional R^2 = .44

c. Care

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	57.20	2.73	[51.85, 62.55]	< .001	
Gender (Female)	3.51	2.49	[-1.37, 8.39]	.16	
Income (High)	.24	2.43	[-4.53, 5.01]	.92	
Economic Ideology (Liberal)	1.65	3.49	[-5.20, 8.50]	.64	
Social Ideology (Liberal)	-.19	3.52	[-7.09, 6.71]	.96	
Religiosity (Very Religious)	3.52	2.45	[-1.29, 8.32]	.15	

Conditional R^2 = .42

d. Reciprocity

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	64.27	1.63	[61.08, 67.46]	< .001	
Gender (Female)	.26	1.47	[-2.63, 3.15]	.86	
Income (High)	-.95	1.45	[-3.79, 1.89]	.51	

Economic Ideology (Liberal)	-2.04	1.86	[-5.68, 1.60]	.27
Social Ideology (Liberal)	.20	1.87	[-3.46, 3.86]	.91
Religiosity (Very Religious)	4.01	1.47	[1.13, 6.90]	.006
				Conditional R^2 = .06

e. Hierarchy

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	53.76	3.07	[47.74, 59.78]	< .001	
Gender (Female)	1.51	2.77	[-3.91, 6.93]	.59	
Income (High)	.75	2.74	[-4.61, 6.11]	.78	
Economic Ideology (Liberal)	-3.78	3.23	[-10.12, 2.55]	.24	
Social Ideology (Liberal)	-1.34	3.24	[-7.69, 5.01]	.68	
Religiosity (Very Religious)	8.44	2.81	[2.95, 13.94]	.003	
					Conditional R^2 = .58

2.5.2. Supplementary Table 11f.

Full regression table for the main analysis, controlling for demographic information

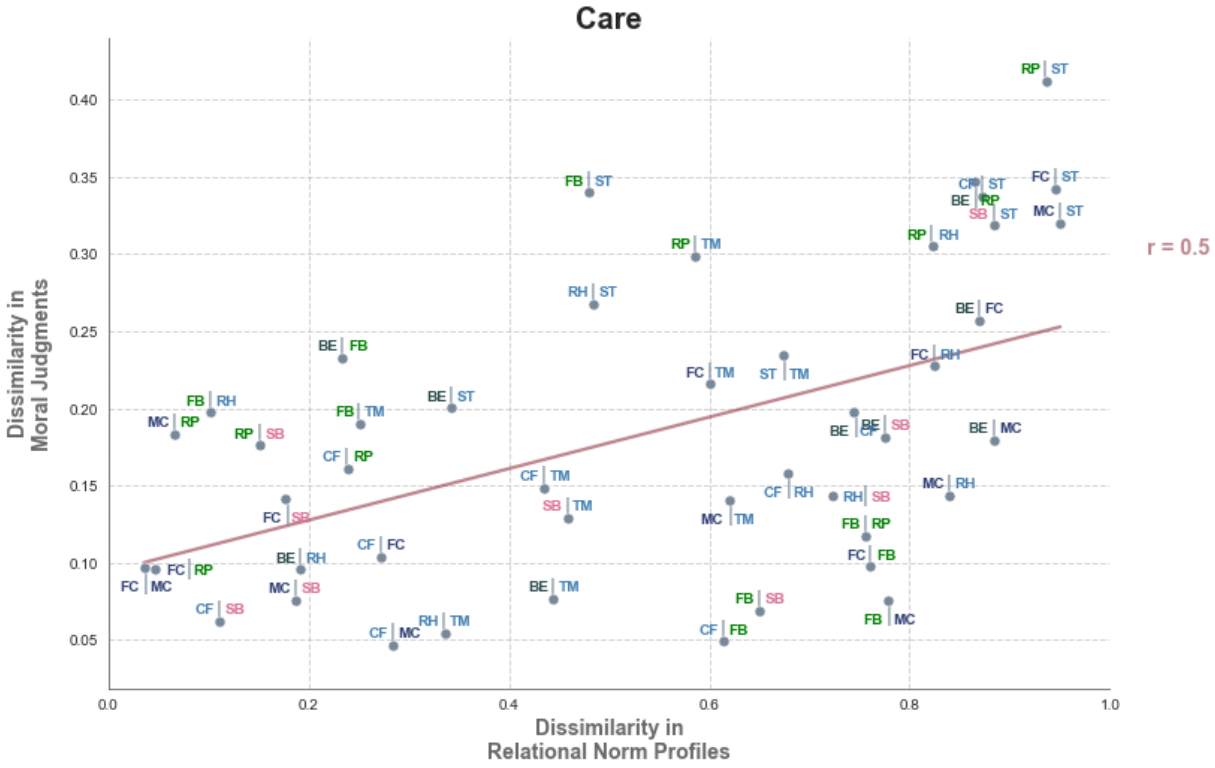
Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	76.95	1.31	[72.94, 79.42]	< .001	
Relational Norms	16.26	.32	[15.64, 16.88]	< .001	
Action Likelihood	-.20	.01	[-.21, -.18]	< .001	
Target Specificity	.37	.01	[.34, .40]	<.001	
Gender	-.94	.81	[-2.53, .67]	.25	

(Female)				
Income (High)	.50	.82	[-1.44, 1.98]	.54
Economic Ideology (Liberal)	2.67	1.05	[.61, 4.71]	.01
Social Ideology (Liberal)	-1.86	1.05	[-3.92, .21]	.08
Religiosity (Very Religious)	9.16	.82	[7.55, 10.77]	< .001

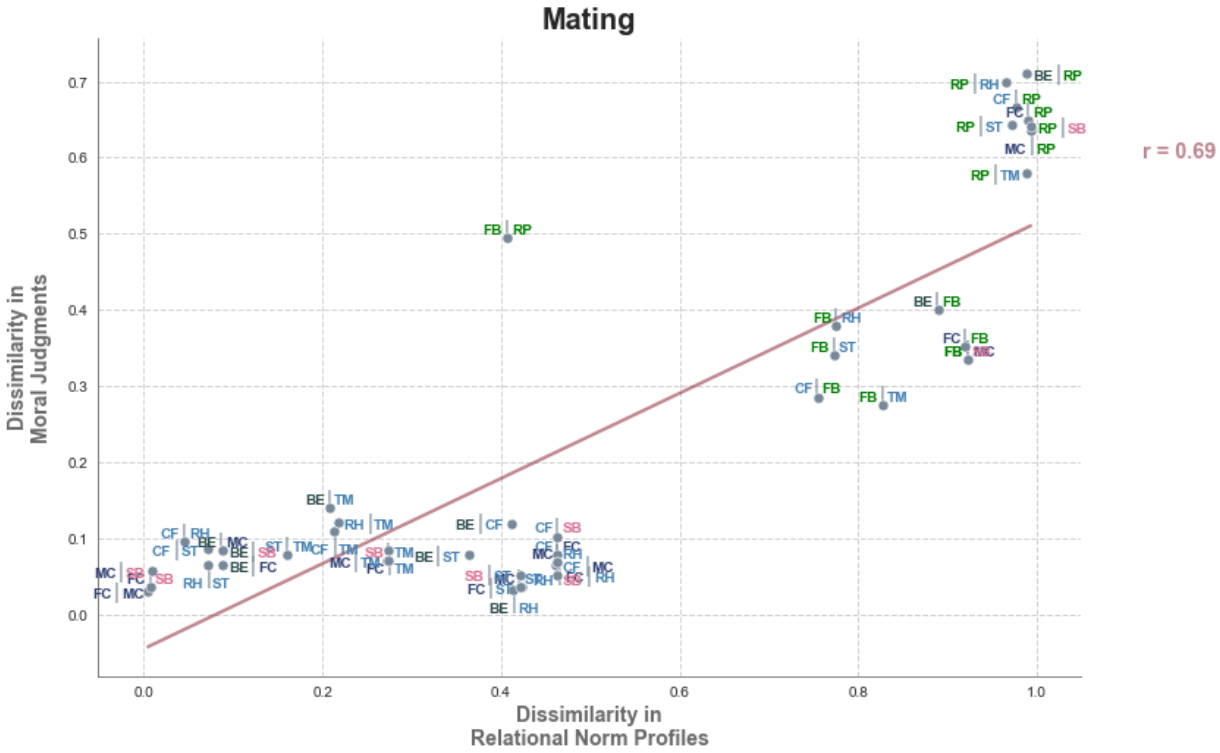
Conditional R^2 = .63

2.5.3. *Supplementary Figures 3a-3d*. Scatterplots of function-specific correlations between KS distance scores in relational norm and moral judgment space respectively: a = care, b = mating, c = hierarchy, d = reciprocity. Spearman's r is the reported value. For the legend describing the relationship labels and colors, please see the main text. *See next page.*

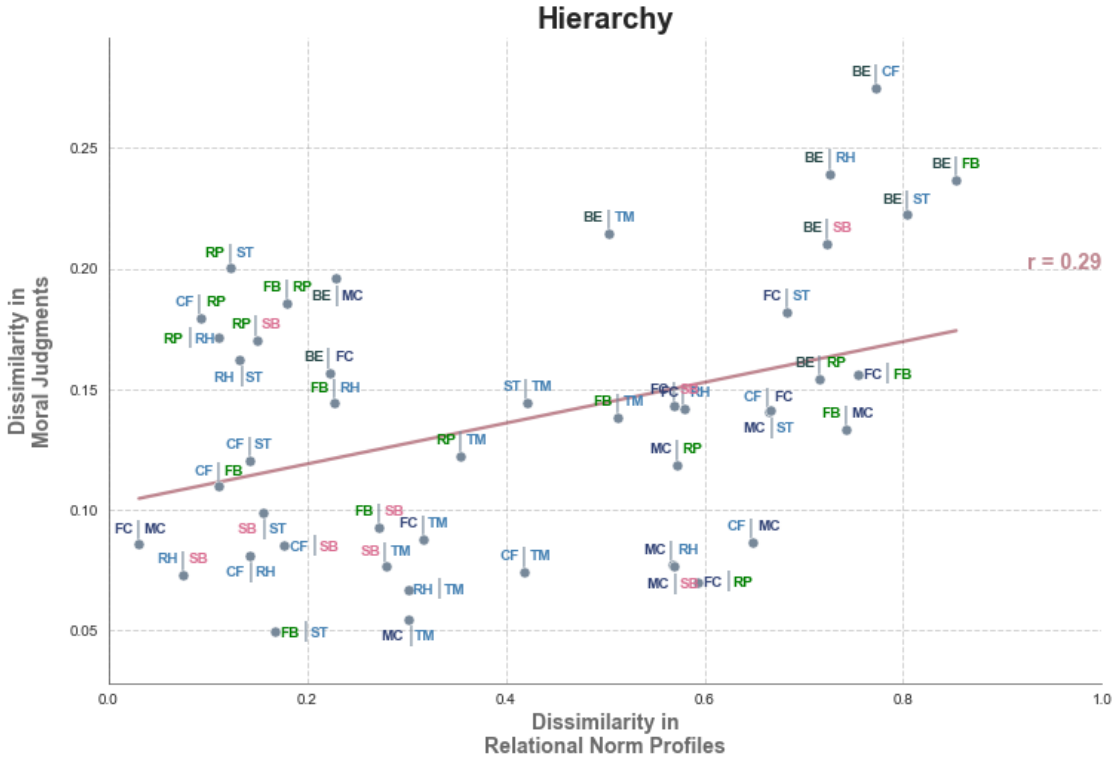
a.



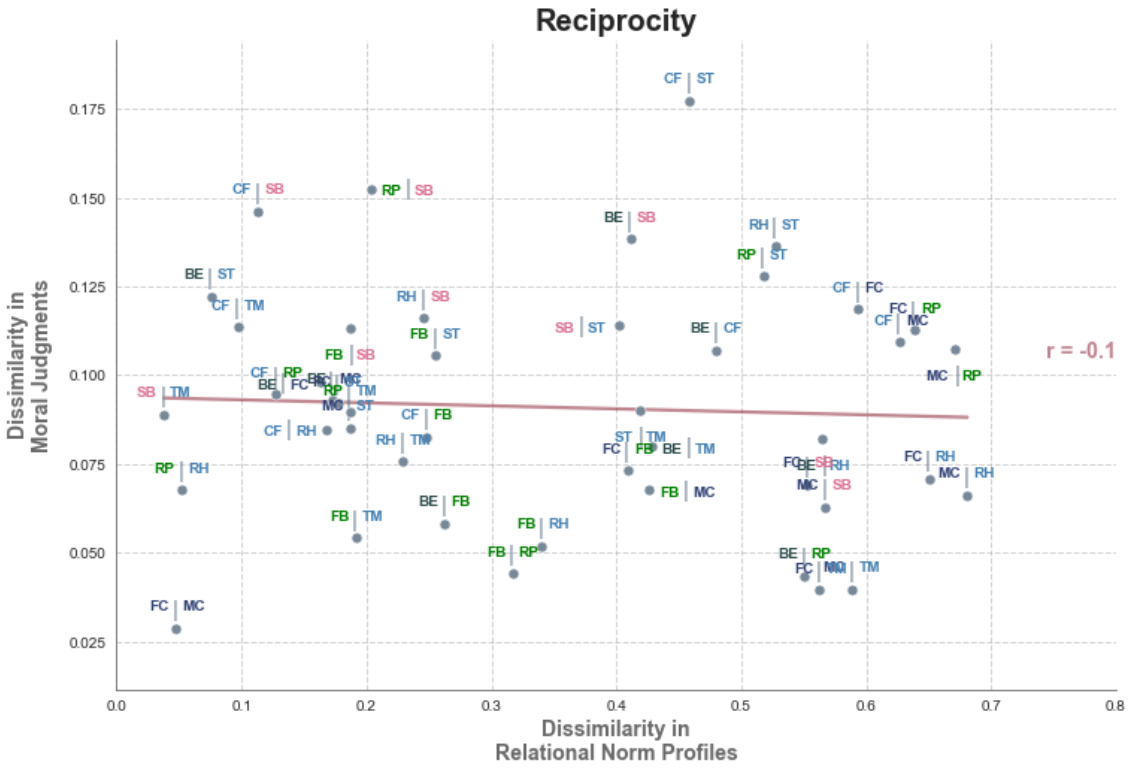
b.


Mating
Dissimilarity in Moral Judgments
Dissimilarity in Relational Norm Profiles
r = 0.69

c.


Hierarchy
Dissimilarity in Moral Judgments
Dissimilarity in Relational Norm Profiles
r = 0.29

d.



Stage 3

Method.

3.1. Participants. For Stage 3, we aimed to have as many observations per distribution as we had in Stage 1. For Stage 1, which was powered for a nationally representative sample across age, race, and gender, we recruited for 450 observations with a final sample of 423 (Sample 1). In the current study, each participant would have to make three ratings for each measured construct rather than one (as in Stage 1), so we determined to recruit a third of the number of target participants as we had in our first sample to achieve similar statistical power. Accordingly, we recruited 150 participants, with 167 participants ultimately taking at least part of the survey (not all finished). Each participant who completed the survey[21] was paid $1.00. Sixty-four participants (64) were excluded on the basis of the pre-registered exclusion criteria (see Supplementary Table 12), leaving us with a final sample of N = 85 (38 female, 46 male, 1 other/non-binary) ranging in age from 21 to 69 (M_{age} = 34.82, SD_{age} = 10.66); see Supplementary Table 13 for complete demographic information.

[21] 18 "participants" in the raw data file had no location information recorded and the data posted to Qualtrics nearly a week after the completion of the survey. We assumed these were bots and deleted those 18 lines of data prior to applying the exclusion criterion. Hence the n = 149 "completed" participants reported in the manuscript: 167-18 = 149.

3.1.1. Supplementary Table 12
Summary of exclusion criteria for Stage 3

Exclusion criteria met	Type of check	Excluded *N*
Did not move sliders to (at least) 1 of 2 specified positions	Attention check	64

3.1.2. Supplementary Table 13:
Demographics of Sample 3 participants

Age	N (%)	Race	N (%)	Gender	N (%)
18 - 27	23 (27.06%)	White	59 (69.41%)	Female	38 (44.71%)
28 - 37	37 (43.53%)	Black/ African-American	12 (14.12%)	Male	46 (54.12%)
38 - 47	14 (16.47%)	Asian	5 (5.88%)	Other/ Non-binary	1 (1.18%)
48 - 57	6 (7.06%)	Hispanic/ Latinx	4 (4.71%)		
58+	5 (5.88%)	Multiracial	3 (3.53%)		
		Native American	2 (2.35%).		

3.2. Procedure. We informed participants that they would be shown a set of relationships and asked some questions about each. Then, participants were shown, in random order, each of the 10 relationship pairs that were used in Stage 2 along with a brief description. Because this study was conducted during the COVID-19 pandemic, we were concerned that participants' judgments about relationships might be affected by the unprecedented circumstances. To address this concern, we included a note asking participants to think of the relationships as they would be under more 'normal' circumstances.

For each of the 10 relationship pairs, participants were asked to rate it along three dimensions of social closeness and three dimensions of interdependence: see Supplementary Table 14 for the exact descriptions of each. As an example, for the sibling relationship and one of the social closeness dimensions, participants were asked, e.g, "In an ideal, well-functioning relationship between siblings, to what extent would the relationship be characterized by **deeply understanding each other**?" Responses were recorded on a sliding scale from 0 to 100 (see endpoint labels in Supplementary Table 14 below). Finally, we included similar individual demographic measures as in the previous samples.

3.2.1. Supplementary Table 14
Dimensions of social closeness and interdependence

Construct	Dimension
Social Closeness	- In an ideal, well-functioning relationship between [relationship pair], to what extent would the relationship be characterized by **deeply understanding each other**? (0 = Not at all; 100 = A great deal) - In an ideal, well-functioning relationship between [relationship pair], to what extent would the relationship be characterized by **accepting and validating each other's natures**? (0 = Not at all; 100 = A great deal) - In an ideal, well-functioning relationship between [relationship pair], to what extent would the relationship be characterized by **striving to care for and promote each other's overall well-being**? (0 = Not at all; 100 = A great deal)
Inter-dependence	- In an ideal, well-functioning relationship between [relationship pair], how **frequently** would they affect each other's thoughts, feelings, and behaviors? (0 = Not at all frequently; 100 = Very frequently) - In an ideal, well-functioning relationship between [relationship pair], in **how many different ways** would they affect each other's thoughts, feelings, and behaviors across different situations? (0 = Very few ways; 100 = A great variety of ways) - In an ideal, well-functioning relationship between [relationship pair], how **strongly** would they affect each other's thoughts, feelings, and behaviors? (0 = Not at all strongly; 100 = Very strongly)

3.3. Data preparation and analysis details. Raw data files (.csv) were prepared and analyzed using Python, within a Jupyter Notebook environment. Primary packages used: numpy, scipy, statsmodels, matplotlib, seaborn, pandas. For data files and all coding scripts, see the OSF link above.

3.4. Supplementary results.

3.4.1. Supplementary Table 15
Full-exclusion regression model for Stage 3

This table shows the full results for the mixed effects linear regression model described in the main text. Note: because there are three samples, the data are not at the participant level, which precludes controlling for demographic information. Source data are provided as a Source Data file.

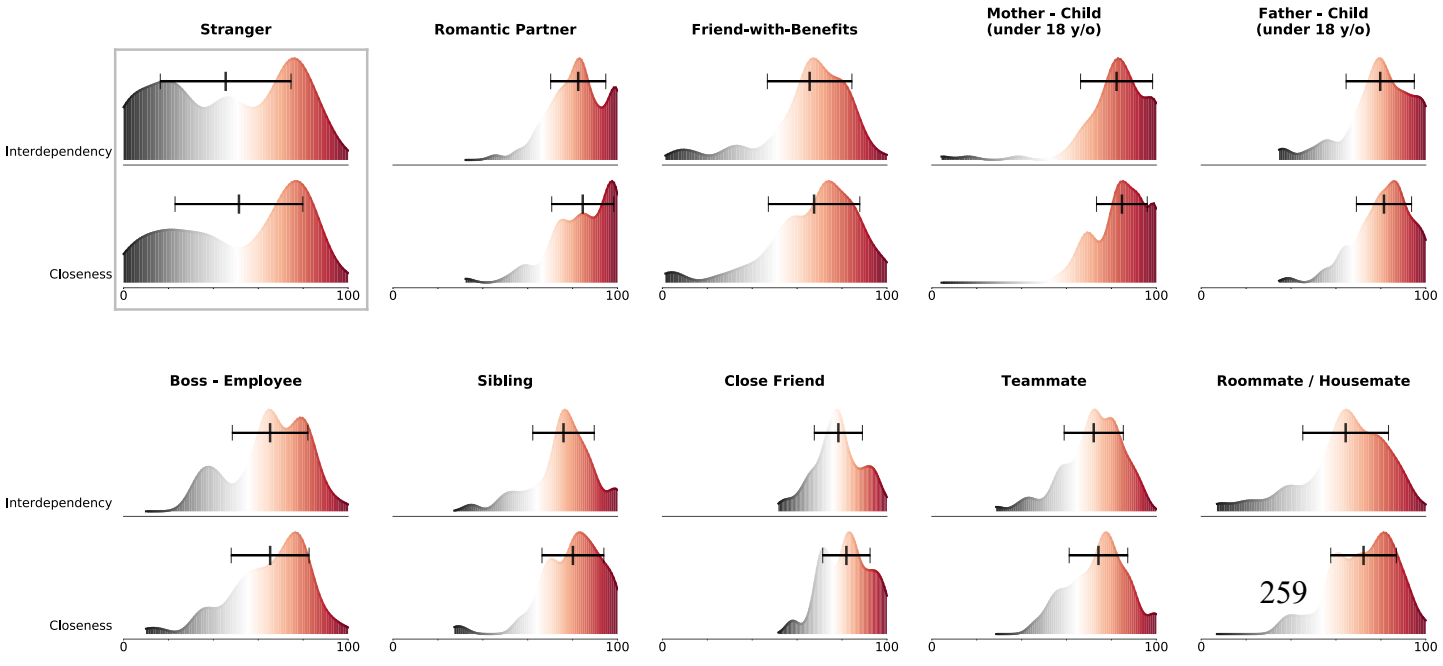
Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	103.12	13.15	[44.62, 50.89]	< .001	
Relational Norms	.12	.01	[-5.23, .07]	< .001	

Social Closeness	.24	.35	[2.98, 8.87]	.48	
Interdependency	-.20	.33	[-2.32, 3.01]	.55	
Genetic Relatedness	1.13	3.52	[-1.95, 4.72]	.75	
Action Likelihood	-.51	.09	[-6.70, .03]	< .001	
Target Specificity	.52	.18	[4.34, 9.75]	< .01	Conditional R^2 = .89

3.4.2. Supplementary Table 16
No exclusions regression model for Stage 3.

Predictor	β	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	104.38	15.54	[74.49, 134.26]	< .001	
Relational Norms	9.36	.96	[7.51, 11.21]	< .001	
Social Closeness	.44	.73	[-.96, 1.84]	.55	
Interdependency	-.36	.70	[-1.71, .1]	.61	
Genetic Relatedness	.89	3.41	[-5.67, 7.46]	.79	
Action Likelihood	-.53	.09	[-.70, -.36]	< .001	
Target Specificity	.54	.18	[.19, .88]	< .01	Conditional R^2 = .90

3.4.3. *Supplementary Figure 4: Study 3 results.* Kernel density plot of expectations for social closeness and interdependency for 10 common relationships. Dot represents the mean, with caps representing +/- 1 standard deviation. The height of the curve represents density: the likely proportions of scores (relative to each function) that fall within the given range along the x-axis.



Appendix 2

Supplementary information for:

Praise and Blame in Relational Context

Study 1

1.1. Method details.

1.1.1. Participants. See main text.

Supplementary Table 1
Summary of pre-registered exclusion criteria for Study 1

Exclusion criteria met	**Type of check**	**Excluded *N***
Did not reach main portion of survey (i.e., did not pass training)	Comprehension check	11
Did not move slider to (at least) 1 of 2 specified position	Attention check	41
Did not type in the word 'FRIDAY'	Bot check	21
Finished survey in < 15 min.	Quality check	32

Note: some participants met more than one criterion.

1.1.2. Procedure details. Participants completed a brief online survey through the Qualtrics interface. Before getting to the main part of the survey, participants were shown the full descriptions of each of the four main relationship functions, as shown in Table 1 in the main text. To ensure that participants were paying attention and were thinking of the functions in the way we wanted them to, each description was followed by a multiple-choice question about the definitions of the functions. Participants were not allowed to advance to the main part of the survey if they failed to answer this check correctly.

Once we had introduced the participants to the four relationship functions, we gave them instructions for the main task of the survey. For each of 20 relationships, we asked participants how much a good relationship of that kind would serve each of the four relationship functions. We specified that, by 'good relationship,' we meant: "it is a relationship that works well for both people, given their respective roles in the relationship." Participants were then presented 20 blocks of questions, one for each relationship, in random order. For each relationship, we included a specific description of what we meant by that relationship—see Supplementary Table 2 for the descriptions. Then, for each combination of relationship and relationship-function, participants rated how much a good relationship of each kind would serve the given function, with the function now presented in its brief form as a reminder (see Table 1 in the main text). For instance, if the relationship and relationship-function pair was siblings/care, participants would be asked: "To what extent would a good relationship between adult siblings typically serve the function of **giving or receiving support based on need, without creating a debt**? (Care function.)" Responses were recorded on a sliding scale ranging from 'Definitely WOULD NOT' (-100) through 'Neutral' (0) to 'Definitely WOULD' (+100). Each of the recruited participants responded to questions about all four functions for all 20 relationships, yielding 80 data points per participant. Finally, we collected a battery of demographic measures: gender, age, race, ethnicity, relationship status, income, level of education, political leanings on social and economic issues, and religiosity. We also collected individual differences measures for use in a separate study, namely the Experiences in Close Relationships questionnaire[22] and the Oxford Utilitarianism Scale.[23]

Supplementary Table 2
Descriptions of relationships

Relationship	Description
Siblings	This refers to brothers and sisters over the age of 18. It includes both biological and adoptive brothers and sisters.

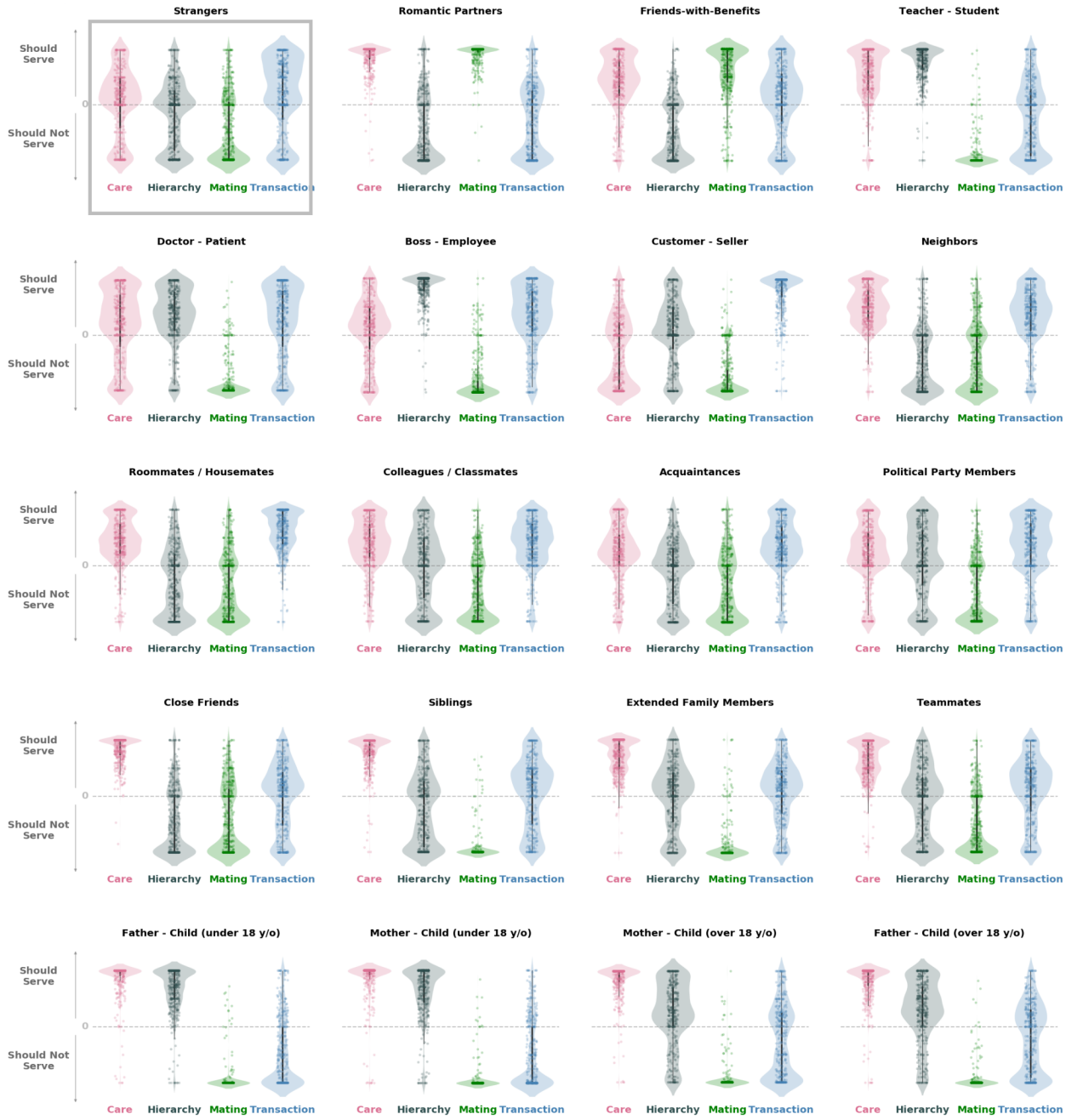
[22] Fraley, R. C., Heffernan, M. E., Vicary, A. M., & Brumbaugh, C. C. (2011). The Experiences in Close Relationships-Relationship Structures questionnaire: A method for assessing attachment orientations across relationships. *Psychological Assessment, 23*, 615-625.
[23] Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131-164.

Long-term romantic partners	This refers to romantic partners that have a commitment to each other, meaning they intend to remain together for the long term. It includes married partners like spouses, but also long-term romantic partners who aren't married but still have a commitment.
Close friends	This refers to people who seek one another out to spend time together, and who refer to each other as best friends, close friends, or good friends. It includes two people of any gender combination who are committed to remaining friends for the long term, and whose interest in each other is NOT primarily romantic.
Work colleagues or classmates	This refers to any workplace relationship between people who are on the same career or schooling level as each other. It includes only people who interact with each other in person in a workplace or at a school.
Boss and employee	This refers to any workplace relationship in which one person directly supervises the other and that same person has some decision-making control over the other's activities and outcomes. It includes only people who interact with each other in person in a workplace.
Teacher and student	This refers to relationships in which someone with more experience guides someone with less experience, such as in school, sports or other extracurricular activity. It includes teacher-student or coach-player relationships, for example.
Doctor and patient	This refers to relationships in which one person provides expert medical or therapeutic services to the other person. It includes medical doctor-patient or therapist-client relationships, for example.
Extended family members	This refers to relationships with relatives who are not in the immediate family (so, not parents or siblings). It includes relationships with a cousin, aunt, uncle, grandfather or grandmother.
Teammates	This refers to a relationship between two people who are in a clearly defined group together working towards a common goal. It includes teammates on a sports team or members of a theater troupe, for example.
Customer and seller	This refers to any relationship in which someone sells something and another person buys that thing, and they interact with each other directly. It includes a local baker and the baker's customer, or a house cleaner and the person who owns the house, for example.

Neighbors	This refers to people who live in the same section of an apartment building, or within the same block of houses on a street. It includes only people who know each other and at least occasionally interact.
Friends-with-benefits	This refers to people who know each other and interact sexually on a somewhat regular basis, but who refer to each other as friends rather than romantic partners. It includes friends who interact sexually without committing to a monogamous romantic relationship.
Acquaintances	This refers to people who know each other, and interact now and then, but don't consider one another to be friends. It includes people one might know from work, school, or the neighborhood.
Members of a political party	This refers to people who are active, registered members of the same political party. They likely hold similar ideological positions and tend to vote for the same candidates in elections. They may not necessarily know each other for any other reason than their common political activities and party affiliation.
Roommates/housemates	This refers to relationships between people who live together and are not family members or in a romantic relationship. This includes college room- or suite mates, apartment mates, or housemates, for example.
Father and child (under 18)	This refers to fathers and their non-adult (meaning under age 18) children. It includes both biological and adoptive fathers/children.
Mother and child (under 18)	This refers to mothers and their non-adult (meaning under age 18) children. It includes both biological and adoptive parents/children.
Father and child (over 18)	This refers to fathers and their adult (meaning over age 18) children. It includes both biological and adoptive fathers/children.
Mother and child (over 18)	This refers to mothers and their adult (meaning over age 18) children. It includes both biological and adoptive mothers/children.
Strangers	This refers to people who encounter each other in any setting for the first time. It includes people who don't know each other from another context and don't anticipate interacting again in the future.

1.1.3. Data preparation and analysis details. Raw data files (.csv) were prepared and analyzed using Python, within a Jupyter Notebook environment. Primary packages used: numpy, scipy, statsmodels, matplotlib, seaborn, pandas. For data files and all coding scripts, see https://osf.io/zxjt6/.

1.2. Supplementary results



Supplementary Figure 1. Relational norm profiles for all 20 relationships. Pink represents care, black represents hierarchy, green represents mating, blue represents reciprocity. The raw data are shown in individual dots; error bars represent the mean (dot) and +/- 1 SD (caps). Note: Mother/Father and under-18 child and Mother/Father over-18 child have each been combined into a single plot, because these relationships were not significantly different from one another in terms of expected cooperative functions: Mann-Whitney U tests, all ps > .07.

1.2.1. Demographic analyses. In order to determine whether there were demographic differences in functional expectations for any of the functions, we built a

mixed linear effects regression model with age, gender, education, self-ascribed socioeconomic status, romantic relationship status, parental status, religiosity, and social and economic political ideology entered as fixed effects, and participant and relationship type entered as random effects. See Tables 3a-3d for the results.

Supplementary Tables 3a–3d.
Complete demographic analyses for Study 1: regression tables

Table a = gender and other demographic effects on normative functional expectations for mating; b = same for care; c = same for transaction; d = same for hierarchy. *LL* and *UL* indicate the lower and upper limits of a 95% confidence interval, respectively.

a. Mating

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	-58.02	8.34	[-74.35, 41.69]	< .001	
Gender (Male)	8.5	2.59	[3.41, 13.59]	.001	
Gender (Other)	6.71	8.07	[-9.15, 22.58]	.41	
Age	-.46	.09	[-.64, -.28]	< .001	
Relationship Status (Yes)	-1.08	2.8	[-6.59, 4.43]	.70	
Parental Status (Yes)	3.27	3.12	[-2.87, 9.40]	.30	
Education	3.38	.95	[1.52, 5.25]	< .001	
Socioeconomic Status	.19	.74	[-1.26, 1.64]	.79	
Social Ideology	-.06	08	[-.22, .10]	.44	

Economic Ideology	.10	.08	[-.06, .25]	.22
Religiosity	.05	.04	[-.02, .12]	.14

Conditional R^2 = .89

b. Care

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	48.05	7.28	[33.78, 62.31]	< .001	
Gender (Male)	-.61	2.26	[-5.06, 3.84]	.79	
Gender (Other)	-4.84	7.05	[-18.69, 9.02]	.49	
Age	.04	.08	[-0.12, 0.19]	.65	
Relationship Status (Yes)	.26	2.45	[-4.55, 5.07]	.92	
Parental Status (Yes)	-.26	2.73	[-5.61, 5.10]	.93	
Education	-.79	.83	[-2.42, 0.84]	.34	
Socioeconomic Status	.10	.64	[-1.17, 1.36]	.88	
Social Ideology	.004	.07	[-.14, .14]	.96	

Economic Ideology	.03	.07	[-.10, .17]	.61
Religiosity	.03	.03	[-.03, .09]	.38

Conditional R^2 = .89

c. Transaction

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	12.32	9.83	[-6.94, 31.58]	.21	
Gender (Male)	11.23	3.06	[5.22, 17.23]	< .001	
Gender (Other)	9.04	9.52	[-9.67, 27.74]	0.34	
Age	-.52	.11	[-.73, -.30]	< .001	
Relationship Status (Yes)	4.18	3.31	[-2.31, 10.68]	.21	
Parental Status (Yes)	.45	3.68	[-6.78, 7.68]	.90	
Education	1.26	1.12	[-.94, 3.46]	.26	
Socioeconomic Status	-.99	.87	[-2.70, .72]	.26	

Social Ideology	.11	.10	[-.08, .30]	.25
Economic Ideology	.01	.09	[-.17, .19]	.94
Religiosity	.11	.04	[.03, .20]	.009

Conditional R^2 = .89

d. Hierarchy

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	2.5	9.15	[-15.43, 20.42]	.78	
Gender (Male)	9.69	2.85	[4.1, 15.28]	< .001	
Gender (Other)	4.98	8.86	[-12.44, 22.39]	.57	
Age	-.31	.10	[-.51, -.11]	.003	
Relationship Status (Yes)	8.38	3.08	[2.34, 14.43]	.007	
Parental Status (Yes)	-.74	3.42	[-7.47, 5.99]	.83	
Education	2.03	1.04	[-.02, 4.08]	.05	
Socioeconomic Status	-.42	.81	[-2.01, 1.17]	.60	
Social Ideology	.11	.10	[-.08, .30]	.25	
Economic Ideology	.05	.09	[-.12, .21]	.60	

Religiosity	.15	.04	[.07, .22]	< .001

Conditional R^2 = .88

1.2.2. Selection of relationships for use in Studies 2 and 3. In Study 1, we collected data for 20 different relationships. Our analysis of the Kolmogorov-Smirnov (K-S) distance between relationships (see main text) revealed that several of these relationships were highly similar in terms of expected cooperative functions. To avoid redundancy in Studies 2 and 3, and to allow for fewer between-subjects conditions, we decided to use a subset of 10 relationships for those studies that were among the most functionally dissimilar to each other. Our procedure for selecting these relationships is described next.

First, we entered the K-S distance into a hierarchical clustering algorithm as described in the main text. We then excluded the customer/seller relationship because it is defined by monetary exchange, specifically, whereas the transaction function on our conception encompasses a wider range of interactions other than purely commercial ones. Second, because we are focusing on dyadic relationships for this study, we excluded relationships primarily defined in terms of group membership, including political party members, extended family members, teammates, and colleagues/classmates relationships. As noted in the main text, there were no significant differences between the mother-child and father-child relationships, so we collapsed these into parent-child relationships; however, in order to track potential functional changes in the parent-child relationship over time, we decided to include both the parent and under-18 child and parent and over-18 child dyads in our final analysis.

After applying these exclusions, there were 13 relationships remaining. To ensure we had a representative selection of relationships across all clusters, we selected the two most distinctive remaining relationships from each cluster. Based on the previously instituted exclusions, most clusters only had two remaining relationships. To choose relationships from the highest-level cluster that still included more than two, we selected those from within the lowest-level clusters that had the highest K-S distance from all other relationships (i.e., least redundant). After these

exclusions, we were left with 10 relationships. The final set of relationships identified by this method were: long-term romantic partners, friends-with-benefits, parent and over-18 child, parent and under-18 child, doctor-patient, boss-employee, acquaintances, roommates/housemates, close friends, and siblings.

Study 2

2.1. Method details.

2.1.1. Selection of action statements. After first excluding several items that we determined were inadequate on theoretical grounds, our algorithm ran for 10,000 iterations. On each iteration, 3 items from the remaining pool of candidate function-weakening items were randomly selected for each function, such that the set of 3 would be among the most characteristic items for weakening the function based on the judges' ratings. The algorithm then computed the mean rating for each function across the set of 3 candidate items, and stored the mean and standard deviation (SD) of the means across all 4 functions.

We implemented a constraint for the preceding step, such that the mean of the absolute means for each set of three items had to be less than or equal to 85 and greater than or equal to 70. We set this floor and ceiling based on exploratory analyses regarding potential function-strengthening items for use in Study 3, as we wanted to ensure that the final set of function-weakening and function-strengthening items would be comparably characteristic across studies. If that constraint was satisfied, the SD of the means across all functions was compared to the lowest such value found across the previous iterations of the algorithm; if the new SD was found to be lower than the previous lowest value, then the randomly-selected items were stored as a new reference point for subsequent iterations of the algorithm.

At the 10,000th iteration, the algorithm returned the last stored set of items. This set of 12 items represented those with the lowest SD of means between the sets of 3 across all 4 functions, ensuring that each set was as similar to all others in terms of mean characteristicness as possible.

2.1.2. Participants. In Study 1, we powered for 95% confidence in a 5% margin of error for a U.S. nationally representative sample across age, race, and gender. This required that we have 385 observations per distribution (with each

participant rating all 20 original relationships on all 4 original functions). To ensure that our Study 2 distributions would be comparable to those from Study 1, we powered for the same confidence in the same margin of error. However, unlike in the previous study, (1) the present study was not nationally representative; we instead used a convenience sample through MTurk; and (2) instead of having each participant give ratings for all 20 relationships as in the previous study, each participant in the present study gave ratings for just one relationship out of a smaller set of 14 relationships in a between-subjects design (14 because questions regarding the non-symmetrical relationships, e.g., boss-employee, were asked in both directions). This means that, theoretically, we would need 14x as many participants for each distribution of ratings to be comparable to its counterpart from the previous study. However, in the present study, each participant gave three ratings for each function (i.e., praiseworthiness/blameworthiness ratings for weakening the function), whereas in the previous study, each participant gave just one rating for each function (i.e., extent to which a good instance of the relationship would or would not serve the function). So instead of multiplying by 14 to get a comparably powered distribution, we multiplied by 14 and then divided by 3. Hence: 385 X 14 / 3 = 1,796. We did not over-recruit this time to account for potential exclusions because we used a quality control measure in the CloudResearch platform ("Exclude low-quality participants"). We simply rounded up to recruit 1800 participants. Ultimately, 1,824 participants completed the survey, each of whom was paid $1.30. One hundred sixty-four (164) participants were then excluded prior to data analysis based on pre-registered exclusion criteria (see Supplementary Table 4 for details), leaving us with a final sample of 1,660 participants.

Supplementary Table 4
Summary of exclusion criteria for Study 2

Exclusion criteria met	**Type of check**	**Excluded *N***
Failed question about survey instructions OR answered multiple-choice question incorrectly (correct answer includes the word “embarrassment”)	Comprehension/attention check	33
Did not move slider to (at least) 1 of 2 specified position	Attention check	129

Failed bot-checker test or text-entry test	Bot check	71
Being younger than 18	Demographic check	36
Not fluent English speaker	Demographic check	8
Finished survey in < 4 min.	Quality check	34

Note: some participants met more than one criterion.

2.1.3. Procedure details. Each participant was randomly assigned to one of 10 relationships, albeit with 14 conditions in total. As noted above, this is because the asymmetrical dyads were presented separately in both directions (e.g., a boss acting toward their employee in one condition, and an employee acting toward their boss in a separate condition). For purposes of analysis, however, only the subordinate-to-dominant condition was used for hierarchy ratings, and only the dominant-to-subordinate condition was used for care ratings. After condition assignment, each participant was shown a brief description of their relationship (see Supplementary Table 2). We informed participants that they would be asked to consider various actions in the context of their assigned relationship and to answer how blameworthy or praiseworthy each of those actions would be. To orient them to the rating scale, we clarified that none of the actions they would see would be extreme (e.g., murder), but rather would all be actions that might plausibly occur within the course of day-to-day life. We then 'anchored' their expectations by showing them a list of actions comparable to the ones included in the task. Following further instructions and attention checks, participants were shown all 12 actions (three for each of the four functions) in random order, as described in the main text.

2.1.4. Data preparation and analysis details. Raw data files (.csv) were prepared and analyzed using Python, within a Jupyter Notebook environment. Primary packages used: numpy, scipy, statsmodels, matplotlib, seaborn, pandas. For data files and all coding scripts, see the OSF link above.

2.2. Supplementary results.

Supplementary Tables 5a–5d.
Primary regression model including demographic predictors

a. Care

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	-55.41	6.38	[-67.89, -42.93]	< .001	
Relational Norms	-.21	.03	[-.27, -.14]	< .001	
Action Likelihood	.39	.03	[.33, .45]	< .001	
Target Specificity	.28	.06	[.17, .39]	< .001	
Gender (Female)	-8.56	2.20	[-12.88, -4.25]	< .001	
Gender (Other)	18.35	23.44	[-27.57, 64.27]	.43	
Income	.58	.48	[-.35, 1.52]	.22	
Education	4.37	1.36	[1.72, 7.03]	.001	
Social Ideology	.07	.06	[-.05, .19]	.24	
Economic Ideology	.02	.06	[-.09, .14]	.67	
Religiosity	.09	.03	[.03, .15]	.005	
					Conditional R^2 = .54

a. Hierarchy

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	-52.63	6.89	[-66.12, -39.14]	< .001	
Relational Norms	-.10	.02	[-.13, -.06]	< .001	

Action Likelihood	.38	.03	[.32, .45]	< .001
Target Specificity	.26	.09	[.08, .43]	.005
Gender (Female)	-2.64	2.00	[-6.57, 1.28]	.19
Gender (Other)	.44	29.86	[-58.04, 58.92]	.99
Income	-.24	.43	[-1.09, .61]	.58
Education	6.11	1.20	[3.77, 8.46]	< .001
Social Ideology	-.01	.06	[-.12, .09]	.80
Economic Ideology	.12	.05	[.02, .23]	.02
Religiosity	.01	.03	[-.05, .06]	.77

Conditional R^2 = .39

b. Mating

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	-91.14	5.46	[-101.84, -80.44]	< .001	
Relational Norms	-.33	.02	[-.36, -.30]	< .001	
Action Likelihood	.40	.03	[.34, .47]	< .001	
Target Specificity	-.70	.03	[-.76, -.65]	< .001	

Gender (Female)	-.54	1.91	[-4.29, 3.20]	.78
Gender (Other)	-8.56	24.07	[-55.70, 38.59]	.72
Income	.88	.41	[.08, 1.69]	.03
Education	.36	1.15	[-1.90, 2.62]	.76
Social Ideology	-.03	.05	[-.13, .08]	.60
Economic Ideology	.05	.05	[-.05, .15]	.33
Religiosity	.05	.03	[-.0004, .10]	.05

Conditional $R^2 = .51$

c. Transaction

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	-199.01	4.77	[-208.35, -189.68]	< .001	
Relational Norms	-.08	.02	[-.11, -.05]	< .001	
Action Likelihood	.38	.03	[.33, .44]	< .001	
Target Specificity	-2.23	.05	[-2.32, -2.14]	< .001	
Gender (Female)	-2.08	1.56	[-5.13, .97]	.18	
Gender (Other)	2.29	19.65	[-36.20, 40.77]	.91	

Income	.38	.34	[-.28, 1.04]	.26
Education	3.12	.94	[1.29, 4.96]	< .001
Social Ideology	-.02	.04	[-.11, .06]	.64
Economic Ideology	.04	.04	[-.04, .12]	.32
Religiosity	.07	.02	[.03, .12]	< .001

Conditional R^2 = .43

Study 3

3.1. Method details.

3.1.1. Participants. See main text.

Supplementary Table 6
Summary of exclusion criteria for Study 3

Exclusion criteria met	**Type of check**	**Excluded *N***
Failed question about survey instructions OR answered multiple-choice question incorrectly (correct answer includes the word "embarrassment")	Comprehension check	97
Did not move slider to (at least) 1 of 2 specified position	Attention check	370
Failed bot-checker test or text-entry test	Bot check	232
Being younger than 18	Demographic check	90
Not fluent English speaker	Demographic check	10
Finished survey in < 4 min.	Quality check	61

Note: some participants met more than one criterion.

3.1.2. Procedure. The procedure was identical to that employed in Study 2.

3.1.3. Data preparation and analysis details. Raw data files (.csv) were prepared and analyzed using Python, within a Jupyter Notebook environment. Primary packages used: numpy, scipy, statsmodels, matplotlib, seaborn, pandas. For data files and all coding scripts, see the OSF link above.

3.2. Supplementary results

Supplementary Table 7a–7d.
Primary regression models including demographic predictors for Study 3

a. Care

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	1,135.02	159.1	[823.55, 1,446.49]	< .001	
Relational Norms	-.08	.03	[-.13, -.03]	.001	
Action Likelihood	.37	.03	[.31, .42]	< .001	
Target Specificity	-16.53	2.42	[-21.28, -11.78]	< .001	
Gender (Female)	6.95	1.64	[3.72, 10.17]	< .001	
Gender (Other)	-2.70	13.07	[-28.30, 22.91]	.84	
Income	.53	.35	[-.16, 1.21]	.13	
Education	-2.56	.99	[-4.51, -.62]	.01	

Social Ideology	.04	.05	[-.06, .13]	.47
Economic Ideology	.01	.05	[-.08, .10]	.81
Religiosity	-.08	.02	[-.13, -.04]	< .001

Conditional R^2 = .45

b. Hierarchy

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	11.72	7.10	[-2.18, 25.63]	.10	
Relational Norms	.05	.02	[.01, .08]	.005	
Action Likelihood	.66	.03	[.59, .72]	< .001	
Target Specificity	-.24	.07	[-.39, -.10]	< .001	
Gender (Female)	2.74	1.84	[-.86, 6.33]	.14	
Gender (Other)	-3.42	14.51	[-31.85, 25.00]	.81	
Income	.20	.38	[-.55, .95]	.61	
Education	-1.81	1.11	[-3.99, .36]	.10	

Social Ideology	-.05	.06	[-.16, .06]	.37
Economic Ideology	.09	.05	[-.02, .19]	.10
Religiosity	.09	.03	[.04, .14]	< .001

Conditional R^2 = .40

c. Mating

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	-80.22	6.89	[-93.71, -66.72]	< .001	
Relational Norms	.23	.02	[.20, .26]	< .001	
Action Likelihood	.92	.03	[.87, .97]	< .001	
Target Specificity	.24	.07	[.10, .37]	< .001	
Gender (Female)	-12.40	1.91	[-16.15, -8.65]	< .001	
Gender (Other)	-1.22	14.67	[-29.96, 27.51]	.93	
Income	-.09	.40	[-.88, .70]	.82	
Education	3.05	1.16	[.77, 5.33]	.009	

Social Ideology	.08	.06	[-.04, .19]	.20
Economic Ideology	.01	.06	[-.10, .12]	.87
Religiosity	.15	.03	[.10, .20]	< .001

Conditional R^2 = .72

d. Transaction

Predictor	b	Std. Error	95% CI [LL, UL]	*p*	Fit
(Intercept)	-129.97	4.19	[-138.17, -121.78]	< .001	
Relational Norms	.03	.02	[-.01, .06]	.11	
Action Likelihood	.84	.02	[.79, .89]	< .001	
Target Specificity	1.60	.05	[1.50, 1.70]	< .001	
Gender (Female)	-6.65	1.54	[-9.67, -3.64]	< .001	
Gender (Other)	-.19	11.80	[-23.30, 22.92]	.99	
Income	-0.20	.33	[-.84, .44]	.54	
Education	2.73	.93	[.90, 4.56]	.003	

Social Ideology	.07	.05	[-.02, .16]	.13
Economic Ideology	-.004	.05	[-.09, .08]	.92
Religiosity	.08	.02	[.04, .12]	< .001

Conditional R^2 = .52