Cognition 250 (2024) 105790

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full length article

When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment

Sydney Levine[a,b,c,*], Max Kleiman-Weiner[d], Nick Chater[e], Fiery Cushman[b], Joshua B. Tenenbaum[c]

[a] Allen Institute for Artificial Intelligence, United States of America
[b] Department of Psychology, Harvard University, United States of America
[c] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, United States of America
[d] Foster School of Business, University of Washington, United States of America
[e] Warwick Business School, University of Warwick, United Kingdom

ARTICLE INFO

Dataset link: https://github.com/sydneylevine/blue-house

Keywords:
Moral judgment
Contractualism
Virtual bargaining

ABSTRACT

Rules help guide our behavior—particularly in complex social contexts. But rules sometimes give us the "wrong" answer. How do we know when it is okay to break the rules? In this paper, we argue that we sometimes use *contractualist* (agreement-based) mechanisms to determine when a rule can be broken. Our model draws on a theory of social interactions – "virtual bargaining" – that assumes that actors engage in a simulated bargaining process when navigating the social world. We present experimental data which suggests that rule-breaking decisions are sometimes driven by virtual bargaining and show that these data cannot be explained by more traditional rule-based or outcome-based approaches.

1. Introduction

For many years, there were only two rules at Boston's Commonwealth High School: "be kind" and "no roller skating in the hallways". Most organizations have more rules than that, though. Just a few miles away, at the University of Massachusetts Boston, students are presented with a 51-page book of regulations upon arriving on campus. Formal codes of conduct are common in many social and organizational settings. But no matter how simple or complex the rules are, how few or numerous, there are many cases where the appropriate thing to do is to *break* a rule.

Imagine, for instance, there is a company-wide rule that employees have to turn in weekly progress reports to their supervisor by 5pm on Friday. Imagine that you spent all day Friday helping your supervisor do damage-control when something went drastically wrong with a project. Might it be OK to turn in the report late? Or say you have been updating your supervisor about your progress multiple times a day and there is nothing additional to report. Or that your spouse goes into labor on Friday afternoon and you have to rush to the hospital. On the flip side, imagine that your supervisor has expressed concern that you are not getting enough done on a weekly basis. Or they are away this week and the report is due to your boss's boss instead. Or imagine that you are working on a high-stakes project that needs urgent feedback. For these latter cases, it seems unlikely that turning in the report late would be okay.

How do these intuitions about when it is okay to break the rules arise? You may have found yourself thinking about what *agreement* would be reached if you and your supervisor could discuss the issue at hand. The results of this imagined discussion depend on who is involved in the negotiation and who would be benefited or burdened by adhering to or breaking the rule. If, in your imagined discussion, you and your supervisor clearly would have agreed that breaking the rule is good for the relevant parties in this case, then probably you will conclude this is the type of situation in which breaking the rule is okay. If, in your imagined discussion, you are pretty sure that your supervisor would not agree to your request to break the rule, then probably it is not okay to break the rule. If it is not clear what the outcome of the hypothetical discussion might be, then you are left in a gray area, unsure whether it is okay to break the rule or not.

In this article, we explore this type of *contractualist* (or *agreement-based*) decision-making mechanism and propose that is sometimes used to determine when it is permissible to break a rule. That is, people sometimes imagine what would happen if everyone they believe would be impacted by that rule (e.g. not just you and your supervisor, but friends, neighbors, other managers, clients, colleagues or other stakeholders) could get together to bargain over the terms of the proposed

[*] Corresponding author at: Allen Institute for Artificial Intelligence, United States of America. sydneyl@allenai.org (S. Levine).

[1] For other recent approaches to understanding rule-breaking, see Bregant, Wellbery, and Shaw (2019), Bridgers, Schulz, and Ullman (2021).

https://doi.org/10.1016/j.cognition.2024.105790
Received 20 June 2022; Received in revised form 1 April 2024; Accepted 2 April 2024
Available online 21 June 2024

breach. The result of this imagined negotiation – or “virtual bargain” – tells a decision-maker whether it is permissible or not to break the rules.[1] This technique may be used when the rules are explicit—as they often are in organizational settings. But it is also sometimes used for rules that are generally followed in a particular group or society, though not codified in any particular way (such as “don’t lie” and “keep your promises”).

We posit both a *cognitive process* and a *cultural norm*, which together constitute a mechanism of human moral judgment and decision-making. That is, we propose both that people have a cognitive process that lets them think about hypothetical agreements, and that it is culturally acceptable (within the culture context of our participants) to use this process for moral judgment and decision-making.

1.1. Rules are central to moral cognition

Nearly every contemporary theory of moral psychology posits some role for moral rules (Baumard, 2016; Crockett, 2013; Cushman, 2013; Greene, 2014; Harsanyi, 1978; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Mikhail, 2011; Nichols & Mallon, 2006). Rules are clearly central to moral cognition. And for good reason. Rules help individuals act consistently, help groups coordinate around common standards, enable easier detection of free-riders, and allow for cognitively efficient decision-making (Hare, 1981). Here, we focus especially on the last of these benefits: Moral judgments can often be complex, so that relying on pre-established rules can be an efficient way to come to an answer that is pretty good most of the time.

Of course, this strategy is not limited to *moral* judgment and decision-making; we see rules used effectively across a wide range of decision-making contexts. Making an ideal decision can require gathering a lot of information and take a lot of time. Sometimes that is indeed what we should to do—when the stakes are high, for instance. But for many routine matters, it makes sense to use heuristics – or rules – instead. Relying on rules that are tuned to the decision-making environment – particularly when time, information and cognitive processing power are limited – can get good answers most of the time (Anderson, 1990; Chater & Oaksford, 1999; Gigerenzer & Gaissmaier, 2011; Simon, 1955; Sunstein & Ullmann-Margalit, 1999). Thus, rules can be thought of as a *resource-rational* strategy in the context of optimal decision-making (Chater & Oaksford, 1999; Gershman, Horvitz, & Tenenbaum, 2015; Levine, Chater, Tenenbaum, & Cushman, 2023; Lewis, Howes, & Singh, 2014; Simon, 1955).

This perspective is crucial to our analysis because it suggests a rational way to *break* rules. If rules are, in part, a cognitively efficient strategy for approximating ideal behavior, then one might rationally break a rule in situations where this approximation clearly generates an unacceptably high departure from the ideal.

As this perspective makes obvious, a key feature of a rule (moral or non-moral) is that it has the appropriate level of abstraction. That is, the rule cannot be written with such fine-grained detail that it will not generalize to novel situations (it is a waste of time to make a rule that will never be used again) but it cannot be written in such general terms that it fails to guide action (it will not save any time or effort). The “ideal” level of abstraction is one that trades off ensuring that all relevant “violations” are forbidden (capturing as many true positives as possible) while not unnecessarily restricting actions that are in fact permitted (avoiding false positives).[2] This trade off must be made while also taking into account the cost of codifying, storing, and deploying rules.

This becomes particularly apparent in organizational settings. For example, the highly-specific rule “Buy the more expensive printer paper with faster shipping if the client demands print-outs by Tuesday” might tell an employee exactly what to do in a specific circumstance, but does not help when a similar predicament comes up with different office supplies or a different deadline. On the other hand, while the general rule “maximize profits” might always be in force, it provides little direct guidance on what to do about the printer paper right now. Most rules have been optimized both directly by institutions and indirectly by cultural evolution to sit somewhere in the middle of these extremes, applying to many circumstances and directly guiding action.

Naturally, rules of this kind will sometimes give the wrong answer compared to an ideal outcome. There will be some “edge cases” that the rule was not designed to handle. These edge cases might arise from new or unforeseen circumstances that make the rule outdated and ineffective. A truly resource-rational strategy, therefore, would establish rules to be followed in most circumstances, but also enable flexibility. We should be able to recognize when the rules should be overridden. While there may be numerous mechanisms allowing us to figure out when it is permissible to break the rules, here we explore a *contractualist* approach to permissible rule-breaking.

1.2. Theoretical inspiration: Moral philosophy and contractualism

Theories of moral philosophy – while intended to be normative – can also offer a window into ordinary thinking. Indeed, the field of moral philosophy has inspired numerous successful theories of socio-moral decision-making (Greene, 2014; Mikhail, 2007; Nichols, 2004). Specifically, moral psychology has made great strides recently by considering how our minds might use processes that mirror the ideas at the center of two major classes of moral theories: consequentialism and deontology. Crudely characterized, consequentialism posits that moral permissibility is determined solely by the consequences of an act. Deontology posits that moral permissibility is determined by considering whether specific actions are permitted, often by reference to the rules, prohibitions, and duties that guide those actions. Our most promising theories of moral judgment and decision-making describe how our moral minds function using either psychological rules (Baron & Ritov, 2004; Cushman, Young, & Hauser, 2006; Mikhail, 2007; Nichols, 2004), calculations of the utilities of outcomes (FeldmanHall et al., 2016; Hsu, Anen, & Quartz, 2008; Lockwood, Klein-Flügge, Abdurahman, & Crockett, 2020), or both (Crockett, 2013; Cushman, 2013; Greene, 2014; Kleiman-Weiner et al., 2015).

Curiously, a third major family of philosophical moral theories – contractualism – has been much less explored in contemporary moral psychology (for notable exceptions, see André, Fitouchi, Debove, and Baumard (2022), Baumard (2016) for a game-theoretical account and Everett, Pizarro, and Crockett (2016) for experimental work).[3] Contractualism bases moral permissibility on the *agreement* of the affected parties. This might be literal agreement, but it more often involves idealized or hypothetical agreement (“what rational agents would agree to” etc.), as in the example of the imagined discussion with the supervisor above. In this paper, across several studies, we present evidence for “intuitive contractualism”—i.e., a dimension of our moral psychology that mirrors key features of contractualist moral philosophy.

There are various forms of contractualism which differ on their details (e.g., Gauthier, 1986; Habermas, 1990; Parfit, 2011; Rawls, 1971; Rehg, 1994; Scanlon, 1998)—though what is central and common to all these views is that the moral acceptability of an action is determined by

[2] Here, “violations” and “permitted” actions are used loosely to include any actions that are precluded or allowed by the rule, and does not necessarily imply that those actions are morally laden.

[3] Note that in studies of distributive justice – which look at judgments of resource distributions – recourse to contractualism is more common. See, for instance, Konow (2003), Michelbach, Scott, Matland, and Bornstein (2003), Mitchell, Tetlock, Mellers, and Ordonez (1993) and Baumard, André, and Sperber (2013) for review. We thank an anonymous reviewer for highlighting this point.

taking into account the interests of all the stakeholders and determining (using a wide range of techniques) what those parties would agree to.[4] Habermas (1996) proposes, for instance, that what we should do is determined by the outcome of a fully rational dialog that would take place between the affected parties. A fair conclusion would be reached by way of the "unforced force of the better argument" (Habermas, 1996). Scanlon (1998) on the other hand, imagines that each party would consider the policy being proposed and compare their personal burden under that policy against the benefit accrued to the actor. If the policy cannot be "reasonably rejected" by anyone, then it is morally permitted. Rawls (1971) imagines agreement emerging by putting the hypothetical bargainers behind a "veil of ignorance", a context in which the citizens of a society decide on the rules that will govern them without knowing what position in society they will occupy. Each of these scholars proposes a different method to implement contractualist reasoning, but their goals are aligned: to determine whether an action is morally permissible by considering who would agree to what and why.

From a psychological perspective, one attractive feature of contractualism is that it accounts for situational flexibility. While adherence to rules often guides our socio-moral behavior, we also seem quite able to treat rules as arrangements that can be renegotiated by common agreement. Such flexibility is useful because it accommodates unusual or novel cases, which may fall beyond the scope of rule-based approaches. Such unusual cases will arise in a dynamic and changing world. A contractualist approach is therefore well-suited to explain how and when we decide to break the rules.

We suspect that there are many contractualist methods of moral decision-making used by the mind (Levine et al., 2023). In this paper, we describe one such process: virtual bargaining (Misyak & Chater, 2014).

1.3. Virtual bargaining

Our model of contractualist rule-breaking is an extension of the "virtual bargaining" (VB) approach (Misyak & Chater, 2014; Misyak, Melkonyan, Zeitoun, & Chater, 2014), an account of social interaction that explains how agents select mutually advantageous actions when their actions are interdependent. The VB approach models social interactions as economic games with multiple Nash equilibria.[5] In such games, players often settle on one equilibrium over the other, which Nash's model alone cannot explain Misyak and Chater (2014). Moreover, the account correctly predicts that players often coordinate on states that are not Nash equilibria at all, but are instead mutually advantageous and enforceable (non-exploitable) options that are better than the equilibria available to them. The VB approach suggests that players decide by imagining what would happen if they could openly negotiate by communicating with one another. Importantly, players achieve this outcome without *actually* communicating in any way; rather, the communication is inferred or imagined, hence *virtual* bargaining.

The account captures notable trademarks of actual bargaining. For instance, consider a two-action (A and B) two-player coordination game with two Nash equilibria: If both players play A, then player 1 gets 10 and player 2 gets 9. If both players play B, then player 1 gets 1 and player 2 gets 11. If players fail to coordinate (i.e.: one plays A, and one plays B), then both get 0. The pure-strategy Nash equilibria are both players playing A or both playing B. Under VB both players can infer that player 1 would not agree in a negotiation to play B (which massively disadvantages player 1), so both players play A (Misyak & Chater, 2014). VB applications to other economic games show that participants are not simply choosing the outcomes with the greatest summed payoffs (see Misyak & Chater, 2014), but instead choose actions that would be mutually agreed upon.

Like virtual bargaining, our model assumes that agents simulate what would happen in an actual negotiation if the affected parties could discuss the potential benefits and harms at stake. Notice how the *normative* question of whether it is acceptable to break a rule is explained in terms of a *descriptive* question: would the affected parties agree that the rule should be broken, were they able to communicate? Put another way, we propose grounding certain moral judgments in the bargains that rational agents would strike, on a standard view of rational agency supported elsewhere in the literature (Chater & Oaksford, 1999; Gershman et al., 2015; Levine et al., 2023; Lewis et al., 2014; Misyak & Chater, 2014; Misyak et al., 2014; Simon, 1955).

In the General Discussion we discuss how possible limitations on our cognitive abilities to simulate negotiations may impact the outcomes of virtual bargaining and cause errors and sub-optimality when deciding to break the rules.

1.4. Experimental approach

In these experiments, we employ a series of carefully controlled (albeit fantastical) scenarios. These scenarios were designed with two primary goals in mind. First, they allow us to control and manipulate particular parameters of interest to investigate their impact on acceptability judgments, allowing precise predictions of moral acceptability. Second, they demonstrate the ability of agreement-based thinking to generate decisions in novel rule-violation scenarios where participants have little prior experience. The cases elicit judgments that participants have not "pre-compiled"—participants cannot easily pattern match their answers onto previous moral situations they have encountered or judged.

Overall, our data provide empirical evidence for contractualist moral judgment through the mechanism of virtual bargaining. Participants judgments show patterns characteristic of contractualist theories as opposed to purely rule-based or utility-based theories.

Throughout this paper, we ask participants to make "moral acceptability" judgments. However, we do not mean to draw a sharp distinction between moral decisions and other kinds of social normative decisions.[6] In many social contexts, decisions often have intersecting and conflicting normative elements, though not all of them recognizably "moral" on traditional definitions (Levine, Rottman, et al., 2020). For instance, in an organizational context, an employee may need to balance loyalty to colleagues, following correct procedure, keeping a promise, avoiding deception, not missing good opportunities for the organization, doing what everyone else would do, doing what they are told by the manager, reporting wrong-doing, and so on. Our theory aims to capture decisions made in any of these social normative contexts that involve rules and rule-breaking.

2. Study 1

Participants read a vignette in which a mysterious stranger appears and asks the protagonist, Hank, if he would be willing to alter his

[4] Some scholars draw a distinction between *contractualism* and *contractarianism*. When defined narrowly, contractualism is typically associated with T.M. Scanlon and his (Kant-inspired) view that an act's moral permissibility is based on whether the policy guiding the act could be reasonably rejected by anyone affected. In contrast, contractarianism finds its roots in Hobbes' writings; these views take contracts to be the agreements of self-interested actors. In this paper, we use the term "contractualist" in the broad sense, covering both these views, and referring to the general class of theories that derives moral permissibility from agreement.

[5] A Nash equilibrium is a state in which neither player can improve their payoff in the game by unilaterally switching their answer. Put another way, each agent makes a best response to the action of the other player.

[6] Indeed, there may be no coherent "intuitive" or universally accepted definition of what counts as moral at all (Levine, Rottman, et al., 2020; Stich, 2018).

neighbor's property in some way in exchange for a certain sum of money. For example, the stranger might request that Hank paint his neighbor's house blue in exchange for $1 million. Or, he might request that Hank spill bleach on the neighbor's lawn in exchange for $1000.

On the face of things, one simple rule governs this case: only the owner of a property has the right to use, modify, or destroy it. (This rule is sometimes referred to in legal theory as "the right to exclusion". See Stonehouse and Friedman (2021) for further references and discussion.) Hank should therefore refuse the stranger's offer and not touch his neighbor's property. But, almost immediately, other options may come to mind. After all, this rule protecting property rights has been established because in most situations it is best for us to be in control of our property and be secure in the knowledge that things will remain as we left them. But the case of the mysterious stranger seems to be an exception; the property-rights rule did not anticipate mysterious strangers and their offers of large cash rewards.

In fact, it seems that *breaking* the rule may be what the neighbor would prefer in some of these cases, especially if he would get to keep a sizeable portion of the money. To ignore this possibility, and instead follow the general rule, seems misguided. Indeed, there is already some evidence that violations of the simple rule about property rights often seem intuitively acceptable. Direct improvements to someone else's property, for instance, are often judged permissible (Stonehouse & Friedman, 2021). We hypothesize that participants sometimes use an agreement-based process to determine permissibility in these cases. In our case specifically, taking the stranger's offer and breaking the rule should be permissible when the neighbor and Hank would be able to reach an agreement if they could sit down and talk about it.

2.1. Predictions of an agreement-based model and alternate models

In this section we describe the predictions made by four moral permissibility model classes. These four models are:

1. Our **agreement-based** model
2. Two **rule-based** models
    (a) One **strict**
    (b) And one more **lenient**
3. Two **utility-maximization** models (or "outcome-based" models)
    (a) One **simple**
    (b) One more **sophisticated**
4. A **preference-based** model

2.1.1. Model predictions of permissibility data

Our **agreement-based model** suggests that the permissibility of breaking a rule hinges on whether or not the two parties would agree to it if, hypothetically, they bargained. How do participants infer whether an agreement would be reached? Our model assumes that the victim (the neighbor in this case) would accept some amount of money to have the harm (property violation) voluntarily done to him. If the actor is confident that he can compensate the victim this amount with the money on offer from the stranger, then he assumes an agreement could be reached. Violating the rule would therefore be permissible with the presumption that the actor would compensate the victim with a side-payment (discussed further below). If no agreement would be reached, then violating the rule is not permissible. Put another way, any given property damage is more likely to be permissible given a larger offer from the stranger and a smaller compensation demand (see Fig. 1C).

In contrast, a **strict rule-based model** of moral judgment posits that it is never permissible to break basic moral rules for some material benefit (Tetlock, Kristel, Elson, Green, & Lerner, 2000). This model predicts that there would be no significant effect of offer amount or compensation requests on acceptability judgments (see Fig. 1A). A slightly more **lenient rule-based approach** might predict that moral permissibility is related to the extent of the damage done (i.e. doing more damage is a larger violation of the property-rights rule) but not to the size of stranger's offer (see Fig. 1B).

However, a **utility-maximization model** of moral judgment makes similar predictions to our agreement-based model about the acceptability of Hank's actions. An overall utility-maximizer will think it is acceptable to break the property-rights rule as long as overall utility is increased; such a judge would consider it acceptable to accept the stranger's offer as long as the amount of money on offer exceeds the costs imposed by the property damage. That is, as long as, when all is said and done, utility has been increased, the action was morally acceptable—no matter who ends up better or worse off than their starting position. (This idea is commonplace in policy-making, e.g. Boardman, Greenberg, Vining, and Weimer (2017), though it is also represented in leading psychological theories, e.g. Greene, 2014.) Like the agreement-based model, the utility-maximization model predicts that as the difference between offer and compensation requirements increases, moral permissibility will increase as well (Fig. 1C).

A **preference-based** approach to moral judgment also makes similar predictions. This model assumes that an action is judged acceptable as long as the preferences of the actor are satisfied. In our cases, the actor stands to benefit from the stranger's offer, so many of the cases will be rendered acceptable. However, this model also assumes that the actor has social preferences: i.e., preferences over the utilities of other people (Cosmides & Tooby, 2013; Delton & Robertson, 2016). Here, Hank may care about his neighbor, so some amount of damage to the neighbor's property will reduce Hank's own welfare. Therefore, this model also anticipates that moral acceptability will be a function of offer amount and compensation owed (which is a proxy for the extent of the utility loss to the neighbor). (Again, see Fig. 1C.)

While there are many versions of preference-based models, it is useful to consider the specific predictions of one by way of illustration. Sell's "recalibrational theory of anger" is one such model that puts *welfare trade-off ratios* (WTRs) at its core (Sell et al., 2017). This theory proposes that the evolutionary function of anger is to bargain for better treatment. Anger is a response to the angry individual's assessment that the offender has put too little weight on the angry person's welfare and the expression of anger is an attempt to increase the weight the offender places on the angry person's welfare (that is, increase their WTR). This view suggests two clear predictions about our case.[7] First, "holding the benefit the offender received constant, anger will become more intense as the cost imposed on the angry person increases" (Sell et al., 2017, pg 113). In our case, this suggests that as the offer from the stranger remains constant, permissibility should decrease with increasing property damage. Second, "holding the cost imposed constant, anger will become less intense as the benefit the target received increases" (Sell et al., 2017, pg 114). In our case, when property violation is held constant, permissibility should increase as the offer from the stranger increases. (Note that is this is precisely what is depicted schematically in Fig. 1C.)

2.1.2. Model predictions of side-payment data

In order to distinguish between the agreement-based, outcome-based, and preference-based models, we turn to another feature of these cases: the amount of money Hank should give his neighbor, or the "side-payment" owed.

On an **agreement-based account** of moral judgment, it might be acceptable for Hank to accept the offer only if he compensates the neighbor, for instance by giving him some amount of the money he is offered (a "side-payment"). How much? Depending on how the agent reasons, there are several points of agreement that could be envisioned.

[7] Assuming – importantly – that increasing levels of anger are directly correlated with judgments of increasing impermissibility/decreasing levels of permissibility.

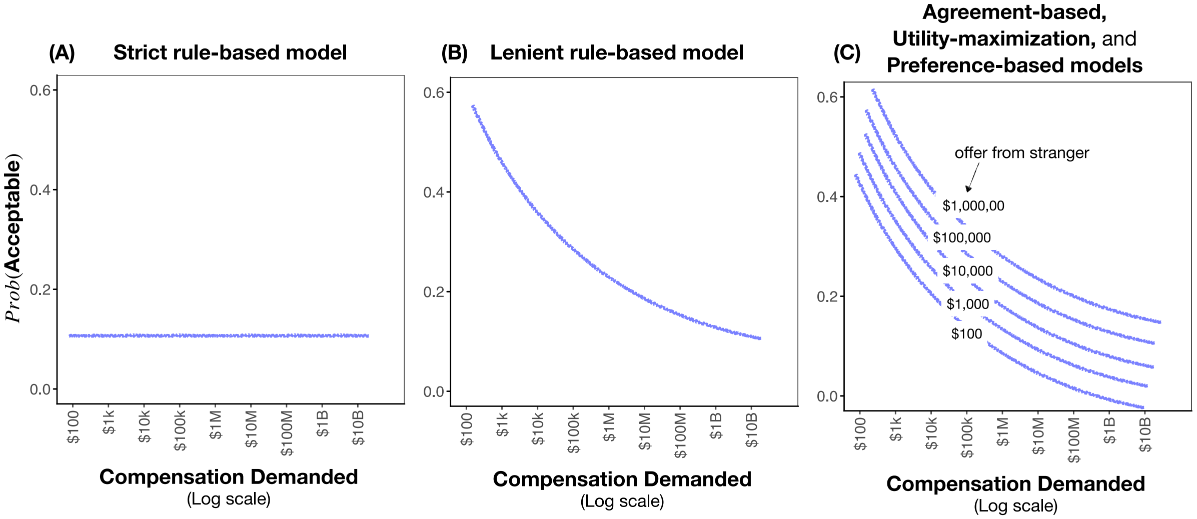

Fig. 1. Model predictions for Study 1. Graphs illustrate moral acceptability as a function of the compensation demanded. Compensation demanded is a function of the amount of harm done for each property damage.

Perhaps Hank and the neighbor would agree that, as long as Hank pays to reverse the damage and restore his neighbor's property to its original condition, Hank can keep the rest of the money. After all, no harm was done. Alternatively, it may be the case that the neighbor has more bargaining power than that. Without the neighbor's agreement, after all, the deal should not go through. So the neighbor may be able to demand an even split of the money, a pattern detected in ultimatum and other bargaining games (Güth, Schmittberger, & Schwarze, 1982). The expected pattern of side-payment judgments for an agreement-based model is therefore a mixture with several modes, as depicted in Fig. 2C. Specifically, the model predicts a pattern of data with three modes: one that represents an even split of the money, one that represents compensating the neighbor for the damage done (which would be a different amount for each property damage, though in general less than a quarter of the offer), and one that represents giving the full offer (largely in cases where the damage done exceeds the offer amount).[8] See below for further analysis of the predicted "compensation mode".

On the other hand, a **strict-utility maximization** model does not make clear predictions about how the money should be distributed. As long as utility is being increased (as it would for any case where the offer amount exceeds the cost of the damage), Hank is free to keep the money for himself or give any amount he wishes to the neighbor (Fig. 2A). A somewhat **more sophisticated utility maximization model** would take into account diminishing marginal returns of the money and how the money would effect the overall wealth levels of Hank and the neighbor. Assuming approximately equal levels of wealth across Hank and his neighbor, then splitting the money in half would maximize utility (Fig. 2A). Note that side-payments on this model should be independent of the property damage incurred, because most of the property damages incurred do not change the approximate wealth level of either individual. That is, if we assume that neighbors tend to have similar wealth levels – but that the exact wealth of each is unknown – then one neighbor having his mailbox painted blue might have him out $50 to paint it back, but it is still reasonable to assume that the neighbors have approximately equal levels of wealth. The amount it would cost to reverse any of the damages is generally within the margin of error when considering overall wealth levels—so maximizing utility would require splitting the money evenly.[9] It is also worth noting a highly counterintuitive prediction of this type of account: that Hank should spontaneously split money with the neighbor even if the stranger merely gave Hank an unexpected windfall donation, with no damage to, or even mention of, the neighbor's property. Indeed, if people were willing to give money to their neighbors in the present scenario on purely utility-maximizing grounds, then it would be difficult to understand why people do not feel similarly obligated to redistribute significant portions of their money to other people in daily life.

Finally, a **preference-based approach** assumes Hank has a preference over his neighbor's utility. Hank's own utility goes up a certain amount (based on Hank's *welfare trade-off ratio* ("WTR", Cosmides and Tooby (2013), Delton and Robertson (2016)) when his neighbor's utility goes up and Hank's utility goes down when his neighbor's utility goes down. However, Hank values his own utility more than his neighbor's (to an extent specified by the WTR). So, if Hank's neighbor has some of his property damaged, Hank should give some of his windfall profit to the neighbor purely in service of maximizing Hank's own utility. That amount should be a function of the damage done to the neighbor's property, times Hank's welfare trade-off ratio (assuming diminishing marginal returns on Hank's windfall profit). The expected pattern of side-payment judgments for this model is depicted in Fig. 2B. The model predicts distributions with two modes: one that represents compensating the neighbor for the damage done, and one that represents giving the full offer. These correspond to two of the modes expected by the agreement-based model, but the even-split mode predicted by agreement is not predicted by the preference-based model.[10]

[8] This analysis assumes that the maximum side-payment is equal to the offer amount. We assume, for instance, that if Hank accepts the offer to paint his neighbor's house blue in exchange for $100, Hank can only pay the neighbor a side-payment of $100 and not the full amount required to paint the house back to the original color. The fact that the compensation necessary to reverse the painting is less than the maximum possible side-payment is partially what accounts for the judgment that this offer is unacceptable. Moreover, the fact that there is no "compensation mode" between 0.5 and 1 is not fundamental to this class of problems, but a feature of the specific damages and offer amounts chosen.

[9] The main exception to this claim is destroying the neighbor's house. On this model, Hank should compensate the neighbor for the destruction of the house and then split the remainder of the offer with them.

[10] Note that the predictions of the agreement-based model are coincident with the possibility that participants are using a combination of the preference-based model and the sophisticated utility-maximization model. We take up this possibility in Study 4.

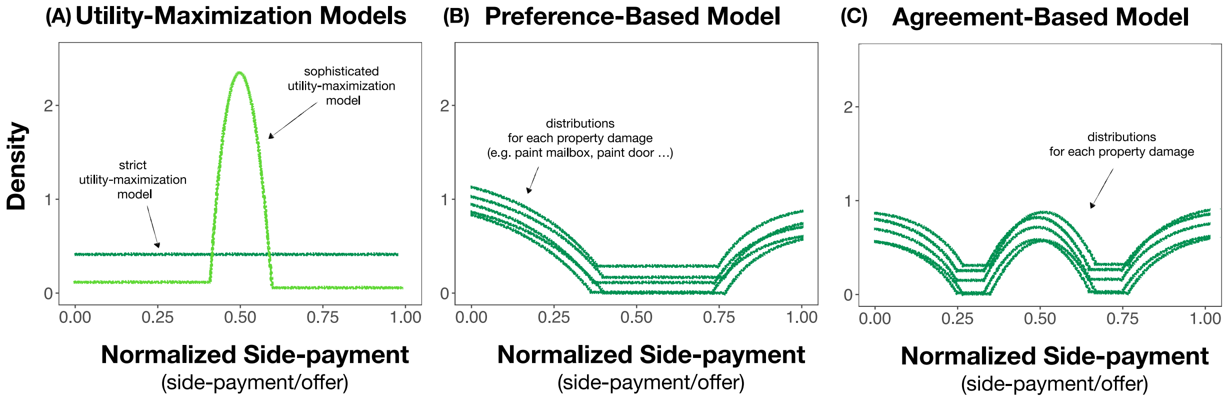

Fig. 2. Model predictions for Study 1. Graphs illustrate predicted side-payments distributions. The mode at 0.5 corresponds to equally splitting the windfall.

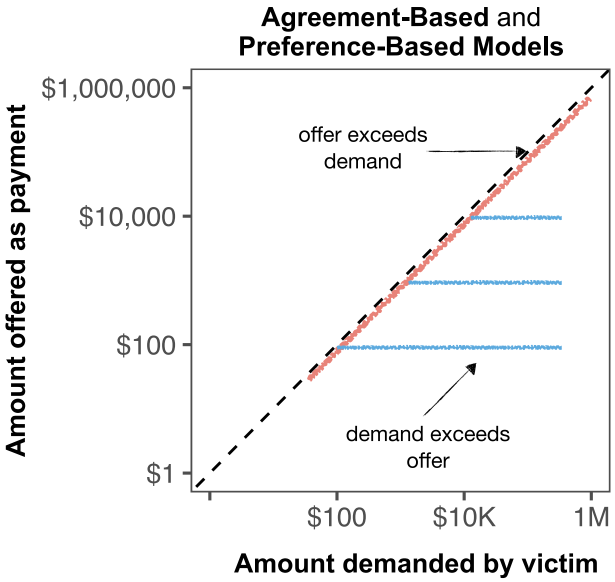

Fig. 3. Model predictions for Study 1. Graph illustrates predicted side-payments for agreement-based and preference-based models. Each model predicts that (at least some) side-payments should be strongly correlated with the amount of money that the neighbor ("victim") would demand as compensation for the damage done to his property (red line). This relationship should only hold when the offer by the stranger (that is, the maximum possible side-payment) exceeds the demand. When the demand exceeds the offer, the side-payment should equal the demand (blue lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Both the **agreement-based** and **preference-based** models predict that some side-payments should be proportional to the damage done to the neighbor. In Fig. 2, this shows up as a mode towards the left of the graphs; but this way of depicting the data conceals the systematic patterns predicted by these models. Specifically, the model predicts side-payments that are strongly correlated with the amount of money that would be required to reverse the damage. Moreover, this pattern should only hold for cases where the maximum possible side-payment (that is, the entire amount of money on offer from the stranger) is greater than the amount required to reverse the damage. If the offer is less than the cost of the damage, then the whole offer should be given to the neighbor. This predicted pattern of data is depicted in Fig. 3.

2.2. Materials and methods

A version of the following vignette was shown to each subject: "A mysterious but very wealthy stranger arrives at Hank's doorstep with a strange request. He asks Hank whether he would be willing to do something particular to Hank's neighbor's property. If Hank agrees, he will receive a sum of [$1 million] which Hank can use however he likes. Otherwise, the stranger will leave and everything remains as before. The man shows the [$1 million] dollars in cash to Hank right then and there. Hank can either take the [$1 million] and carry out the stranger's request or he can turn the stranger away".

The amount of money on offer from the stranger varied between participants: The offers were $100; $1000 $10,000; $100,000; $1,000,000, or a threat against the life of Hank's son ("gun" henceforth).

Each subject (n=360, MTURK participants, 60 participants in each offer condition) was then prompted to imagine that the stranger requested the following property damages: painting the neighbor's mailbox blue, painting the outside of the neighbor's front door blue, painting the inside of the neighbor's front door blue, painting the neighbor's house blue, cutting down a tree in the neighbor's yard, breaking all the windows in the neighbor's house, spilling several gallons of bleach on the neighbor's lawn, smearing dog poop on the neighbor's front steps, painting over a mural created by neighbor's daughter, or entirely demolishing the neighbor's house. After each item, participants were asked "Would it be morally acceptable for Hank to carry out the stranger's request? (Yes/No)" Each participant judged 10 cases: each level of property damage crossed with a single offer amount.

Following this, participants were asked to assume that Hank accepted the stranger's offer and receives the money. Participants were then asked to indicate how much money Hank owes the neighbor given what he did. For instance, "Imagine that Hank painted his neighbor's mailbox blue. How much money, if any, should Hank give his neighbor?" (See Appendix for further details.)

A separate group of participants (n=100, MTURK participants) made judgments about how much compensation they would have to be offered to have the property damages voluntarily done to them ("compensation demanded"). Participants in this group read the following prompt, in which the stranger offers money in compensation to carry out the damage to *Hank's own* property:

"A mysterious but very wealthy stranger arrives at Hank's doorstep with a strange request. The stranger says that he will pay Hank a sum of money if Hank agrees to have something particular done to his property. After that, the stranger will go away and nothing else will happen to Hank or his property. What is the minimum amount of money you think the stranger would have to offer for Hank to agree to let the stranger do the following things to his property?" Participants then saw the full list of property damages.

2.3. Results

Data for this and all experiments can be found at Levine, Kleiman-Weiner, Chater, Cushman, and Tenenbaum (2024).

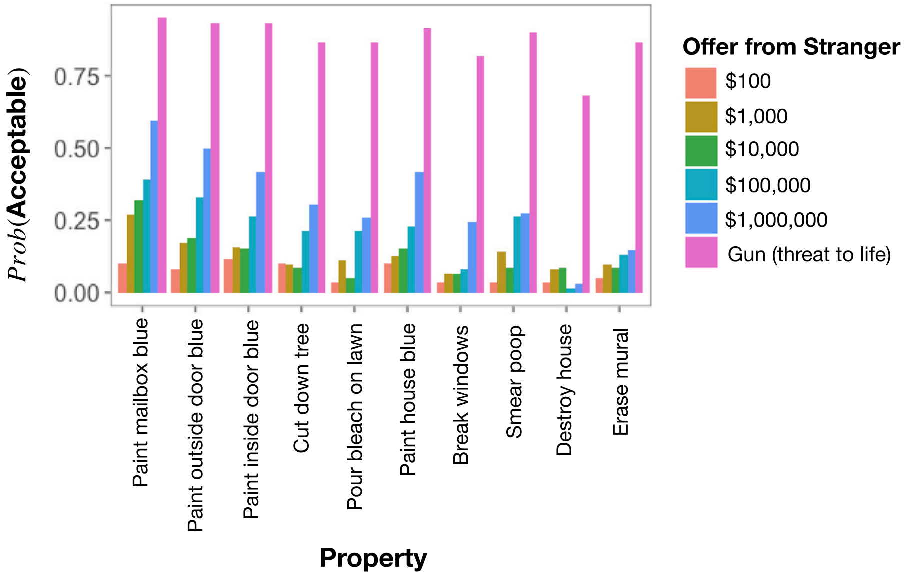

Fig. 4. Study 1 results: percent of participants who judged that agreeing to each offer would be morally acceptable. When the stranger's offer involves money, Hank receives the money if he commits the corresponding property damage. For the "gun" case, doing the property damage allows Hank to save his son's life. The data shows that as offer goes up, acceptability tends to increase. Likewise, as the extent of the property damage increases, acceptability decreases.

2.3.1. *Acceptability judgment analysis*

Fig. 4 shows subject's acceptability judgments for each offer/ property pair. There are two main trends to be accounted for: (1) As offer goes up, the probability of the offer being judged morally acceptable goes up. That is, taking the $1M offer is generally more acceptable than taking the $1K offer—regardless of the property damage at stake. (2) As harm goes up (as determined by participants' compensation demand judgments, see Fig. 5), the chance of the offer being morally acceptable goes down. For example, painting the mailbox blue is more acceptable on average than destroying the house.

Fig. 5 shows the distributions of compensation demands for each property. One feature of this data is that there is a wide range of compensation demands for each potential property damage (note the log scale). We assume that our participants represent a version of this distribution—they are aware that there is a range of compensation demands that someone might make for a given property. For example, one person might not be so attached to their mailbox (and may even enjoy getting a fresh coat of blue paint on it) and would therefore demand little if any compensation for the harm. On the other hand, someone else might have a mailbox that their grandfather handcrafted from pine wood gathered from his childhood backyard; such a person would ask for much more money to have his mailbox painted blue. Given that no specific information is given about the neighbor and their feelings about their property, we assume that subjects rely on their representation of the distribution of possible attitudes. That is, a range can be estimated using reasonable estimation about most people's feelings about their mailboxes; in general, people care a little about their mailboxes and its color, but not a tremendous amount. This estimation is reflected in the distributions of Fig. 5. In Study 3, we take up the question of how and whether participants take into account information about the specific person who would be party to the bargain, were it to actually take place.

We model the value for compensation that participants are using in their moral acceptability judgments as the 90th percentile of the compensation distributions. Put another way, we assume that participants choose a value for compensation that they are highly confident will be greater than the amount someone would demand in compensation for that particular harm when making moral acceptability judgments, though our findings are robust to choices over different percentiles.

Our *agreement-based model* predicts that permissibility judgments depend both on the stranger's offer and the compensation owed to the victim (refer back to Fig. 1(C)). In contrast, a *strict rule-based approach* assumes that there should be no significant effect of offer or compensation owed on permissibility judgments (Fig. 1(A)). A slightly *more lenient rule-based approach* predicts that moral permissibility is related to the extent of the damage done but not to the stranger's offer (Fig. 1(B)).

To distinguish between these models, our central analysis is as follows. We analyzed the acceptability data with a linear regression, predicting the log odds of mean subject response to each offer/property pair. We included in the regression the log of offer amount,[11] and the log of the 90th percentile of the distribution of compensation demands for each property violation (Fig. 5 blue dots). Results are shown in Fig. 6. Importantly, just as the agreement-based model predicts, there is a significant effect of both compensation demand ($\beta = -0.212, p < .0001$), and offer ($\beta = 0.136, p < .0001$) on participants' moral acceptability judgments. This stands in contrast to the predictions of the strict rule-based approach (which prohibits breaking rules under any circumstance and therefore denies that there should be an effect of either compensation demand or offer) and a slightly more sophisticated rule-based approach (which treats larger moral violations as more morally problematic and therefore predicts an effect of compensation demand but not offer).

Individual differences. Despite the central finding reported above that the rule-based approaches do not explain the main trends in participants' acceptability judgments, it is nonetheless likely that *some* rule-based thinking is at play in our participants' decision-making processes. This is hinted at by the fact that permissibility judgments for all conditions (offers) except "gun" rarely surpass 50%—and are often considerably lower. Indeed, a substantial proportion of subjects (59.7%) are well-characterized by the strict rule-based view—judging every case they were presented with as impermissible, regardless of the context (i.e., the pattern depicted in Fig. 1(A)). Importantly, once this pattern of responses is accounted for, the remaining data (n=123 participants) are still robustly explained by the class of models depicted

[11] In the quantitative analysis that follows, the "gun" condition was dropped, so that offer amounts could be entered into the regression as a continuous variable. However, the findings reported here are robust if the "gun" condition is assigned a range of values.

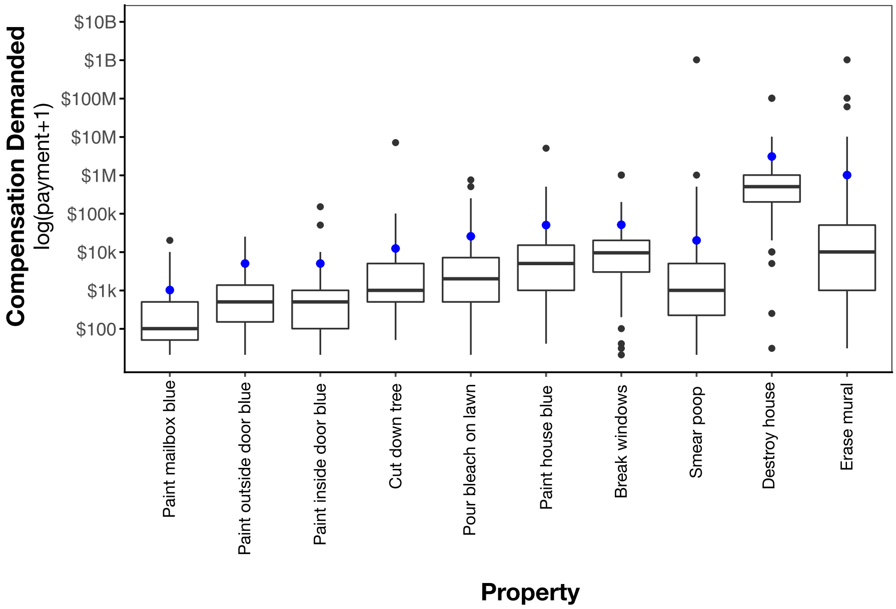

Fig. 5. Study 1 results: distributions of subject responses for the compensation they would demand to agree to have some property damage done to them. Boxes visualize five summary statistics: median, 25th percentile (lower hinge), 75th percentile (upper hinge), largest value no further than 1.5 * IQR from the upper or lower hinge (whiskers). Blue dots indicate the 90th percentile of the distribution, which is the value used for compensation in the model. Y-axis depicts the log of the side-payment + $1 (in order to allow the graph to depict the data when sidepayments are $0). IQR = inter-quartile range.

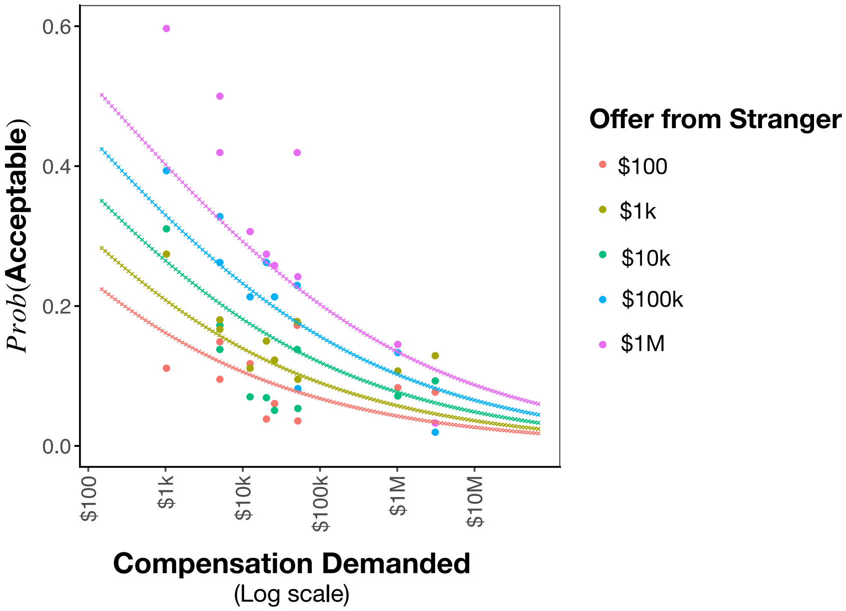

Fig. 6. Study 1 results: moral acceptability judgments. Solid lines are quantitatively precise predictions of the agreement-based model, as generated by the linear regression. Compensation demanded is the 90th percentile of the compensation distributions shown in Fig. 5. Compare to model predictions in Fig. 1.

in Fig. 1(C), namely, the agreement-based, utility-maximization, and preference-based models. We analyzed this subset of the data using the same model we described above—a linear regression, predicting the log odds of mean subject response to each offer/property pair. Predictors included the log of offer amount and the log of the 90th percentile of the distribution of compensation demands for each property violation. Results are shown in Fig. 11 in Appendix. There was a significant effect of compensation demand ($\beta = -0.359, p < .001$), and offer ($\beta = 0.0727, p < .05$) on participants' moral acceptability judgments. In the General Discussion, we further discuss how purely rule-based mechanisms might interact with agreement-based mechanisms in the moral mind.

2.3.2. Side-payment analysis

Side-payment analysis allows us to differentiate between our agreement-based model, a utility-maximization model, and a preference-based model. On an agreement-based account of moral judgment, if Hank accepts the offer, he should be willing to (a) give enough money to the neighbor to compensate for the damage or (b) to split the money 50/50. In contrast, a strict-utility maximization model does not make clear predictions about how the money should be distributed. A utility-based model that takes diminishing marginal returns into account predicts that side-payments should reflect offer and not compensation. Finally, a preference-based model assumes side-payments should reflect compensation and not offer.

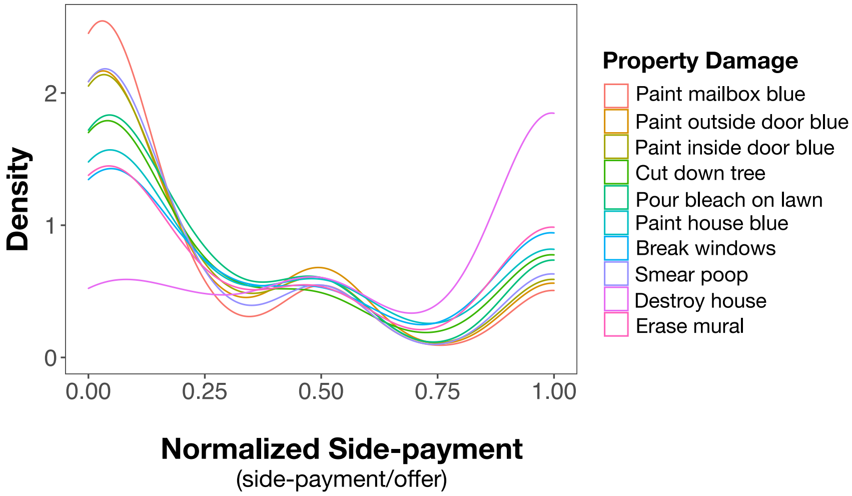

Fig. 7. Study 3 results: subject judgments of the amount of money (side-payment) that Hank owes the neighbor if he accepts the stranger's offer to do each of the property damages. Side-payments are normalized (side payment is shown as a fraction of offer amount). Side-payment decisions include from all participants, regardless of the moral acceptability judgment they rendered about the case. Compare to model predictions in Fig. 2.

Inspection of Fig. 7 provides initial support for the agreement-based model. A substantial proportion of participants choose a 50/50 split (side-payment/offer = 0.5, see Fig. 7) across all property damages and all offers. We also find a meaningful mode where the side-payment given is the entire offer amount (side-payment/offer = 1), see Fig. 7. This is driven by the conditions where the stranger offer is lower than what would be required to fully compensate the neighbor (e.g. $100 is on offer for demolishing the neighbor's house).

The agreement-based model predicts that some participants will be inclined to offer the amount of money that would be required to compensate the neighbor for the damage done. Therefore, this model predicts that there should be a mode in the side-payment distribution that corresponds to that amount. However, this strategy is only relevant when the amount of money on offer exceeds the amount of money that would be required to compensate the victim for the damage. Therefore, we expect to find this "compensation mode" just in the cases where offer exceeds the compensation required.

To test this hypothesis, we calculated the geometric mean of the side-payment data, excluding any offers that were 50%–55% of the offer (which are accounted for by the "even split" prediction of the agreement-based hypothesis). This can be thought of as the "secondary mode" of the distribution.[12]

Fig. 8 shows the relationship between the secondary mode of the side-payment distribution and the mean of the compensation demands (that is, participants' estimations of how much someone would want to be compensated to have some damage voluntarily done to them; see Fig. 5). For high offers ($1 million, $100,000, $10,000) there is a clear relationship between the secondary mode and compensation demands, indicating that this secondary mode likely reflects the desire to compensate the victim. This pattern is expected for the high offer conditions because high offer amounts exceed the (geometric) mean values for compensation demands for most of the property violations. Likewise, for low offers ($100, $1000) there is virtually no relationship between the secondary mean and compensation demands, again, as expected given that nearly all compensation demands are greater than the offer amount.

More formally, we predicted that when offer exceeds compensation demand, there should be a strong relationship (steep slope) between compensation and secondary mode. In contrast, when compensation demand exceeds offer, there should be a weaker relationship. We constructed a "Compare" variable which indicates whether or not offer exceeds compensation. We predicted a significant interaction between Compare and Compensation Demand as predictors of the (geometric mean of the) secondary side-payment mode. The secondary side-payment mode (for each offer/property pair) was predicted by a linear model that included Compensation Demand (geometric mean for each property damage), a Compare variable (indicating whether or not offer exceeds demand), and their interaction as fixed effects. Each of the fixed effects, as well as their interaction, was significant ($p < .0001$).

2.3.3. Summing up

The agreement-based model makes specific, unique predictions about how participants would respond to the acceptability and side-payment questions and these predictions are born out in the data. None of the other candidate models can account for all the patterns found in the data. However, we cannot yet rule out the possibility that instead of using an agreement-based model, participants employ a combination of two of the alternate models: the sophisticated utility-maximization model and the preference-based approach. Although neither of those models predicts the full pattern of acceptability judgments we find by itself, they can be combined to do so (see Fig. 1). Therefore, we cannot rule out the possibility that these methods of moral judgments are being used by participants in combination. We take up this issue in Study 2.

3. Study 2

Study 2 was designed to rule out the combined utility-maximization and preference-based models as an alternate account for the data in Study 1. If participants are using a *utility-maximization approach* to make moral judgments in the cases we used in Study 1, then they should endorse the possibility of Hank giving the money he acquires to charity instead of offering it as a side-payment to his neighbor or keeping it for himself. This is on the assumption that a unit of money improves aggregate welfare more when spent by a charity than when spent by a typical homeowner. Thus, in Study 2, we offered participants the option of Hank accepting the stranger's offer and then donating the money to charity.

It is also possible that some of the side-payment data from Study 1 can be explained by the *preference-based model*. It is possible that a

[12] Note that in this side-payment analysis, we truncated all distributions to include only payments that were less than or equal to the offer amount. We also excluded participants in the "gun" condition, owing to the difficulty of normalizing side-payments against offer in this case. Finally, we included all participants in this analysis, not just those who thought that accepting the stranger's offer was permissible.

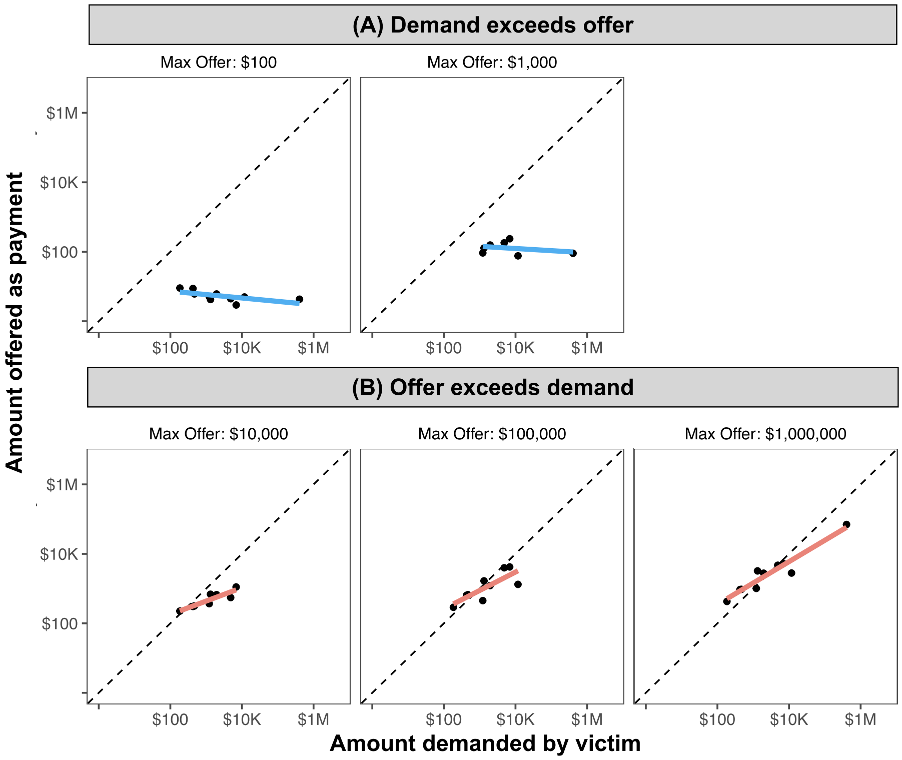

Fig. 8. Relationship between compensation demands and the "secondary mode" of the side-payment distribution for each property/offer pair. The secondary mode refers to the geometric mean of the side-payment data, excluding those payments that were equal to the offer value and those that were near even splits of the offer. The value used for compensation demands is the geometric mean of the compensation distribution for each property (see Fig. 7). If participants are using a compensation strategy to determine if the neighbor would agree to Hank's action, then we would expect a strong correlation between this secondary mode and compensation demands. This strategy only applies when offer exceeds compensation demand (Panel B, red lines) and is irrelevant when demand exceeds offer (Panel A, blue lines). As expected, there are strong and significant correlations between compensation demand and the secondary mode for high offers ($1,000,000, $100,000, and $10,000) and no correlation for low offers ($100 and $1000). Graphs are shown for the cases where 3 or more points within a given offer fell into that category (A or B). Compare to model predictions in Fig. 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tight correlation between compensation judgments and side-payments (as demonstrated in Fig. 8, panel B) can be explained if, coincidentally, Hank's preference for the neighbor's welfare, combined with his and the neighbor's diminishing marginal returns for wealth, happen to combine to require precise compensation of the damage to the neighbor's property. But, according to a preference-based model of this kind, there is nothing special about Hank's neighbor's welfare in virtue of *his house being painted blue*. Hank should be just as likely to give the money to anybody else he values equivalently, based on his social preferences. In contrast, an agreement-based model posits that, when Hank is determining whether he can paint his neighbor's house blue, it is *his neighbor's agreement* that he must consider—and, thus, there is a special reason for him to make a side-payment to his neighbor specifically.[13]

In order to distinguish between these alternatives, participants read a modified version of the vignette used in Study 1. In this story, Hank has the option of painting someone's house blue in exchange for a million dollars. This time the owner of the house, Steve, is not Hank's neighbor, but someone Hank has never met. Meanwhile, Hank also hears about another stranger, Carol, whose house was accidentally repainted without her knowledge. The preference-based model would predict that if Hank accepts the offer, then he should give equal amounts of money to both Steve and Carol (or, in any event, there is no special reason to think he would give more to one than the other). After all, both are strangers and both have had their houses repainted without their consent. Put another way, if the reason Hank gives money to Steve is that he is simply maximizing his own utility which involves considering other people's utilities (times a welfare trade-off ratio), then this argument should apply equally to Carol; so Hank should give the same amount of money to both. In contrast, an *agreement-based model* predicts that Hank should owe more money to Steve than to Carol. On this view, it is morally acceptable to paint Steve's house only if Steve would agree to it in a bargain. Steve would agree to it in a bargain if there was something in it for him (minimally, compensation for the house painting, but potentially an even split of the money, as discussed above). Hank has no such relationship with Carol and therefore does not necessarily owe her any money.

3.1. Materials and methods

Participants read a vignette similar to the one in Study 1, with a few key differences (see Appendix for details). First, participants learn that Hank reads a story in the newspaper about someone whose house is accidentally painted pink. Following this, and completely coincidentally, a stranger arrives at Hank's doorstep and tells Hank that he will give him $1 million if Hank agrees to pick someone's name randomly from the phone book and paint his house blue. Hank picks Steve's name out of the phone book. This is a person Hank has never met and cannot get in touch with before he has to make a decision. (House colors are counterbalanced across conditions.)

Participants are asked to choose what the most morally acceptable thing is for Hank to do. Participants are given the following options: accept the offer and give some of the money to Steve, accept the offer and give some of the money to Carol, accept the offer and give some of the money to Steve and some to Carol, accept the offer and donate the

[13] Note that this prediction does not necessarily hold for the recalibrational theory of anger (Sell et al., 2017), a specific preference-based model discussed in the Introduction. We discuss this further in the General Discussion.

money to charity, refuse the offer and turn the stranger away, accept the offer and keep all the money. If a subject chooses any of the options that indicate that Hank should give some of the money to Steve or Carol, they are asked how much. Participants then answered a series of comprehension checks and were excluded for incorrect answers.

209 participants (MTURK participants) completed the study. 46 participants were excluded for failing control questions, leaving 163 participants included in the study. 47% of participants reported being female, 52% male, and 1% non-binary. Mean age was 38.9 years, SD = 12.5, min = 19, max = 75.

3.2. Results

3.2.1. Moral judgments

Moral judgment data is reported in Fig. 9A. 42.3% of participants judged that Hank should accept the offer and give some money to just Steve, 0% judged that Hank should accept the offer and give some money to just Carol, 41.7% judged that Hank should refuse the offer entirely, 1.8% judged that Hank should accept the offer and donate the money, and 3.7% said that Hank should accept the offer and keep the money for himself. Note that in a conceptual replication of Study 1, we find strong evidence for both agreement-based thinking (in those who choose to accept the offer and give some money to Steve) and rule-based thinking (those who refuse the offer outright).

The first theory this study was designed to test was the *utility-maximization* view. If participants are impartially maximizing utility (with or without diminishing marginal returns), then they should opt to accept the offer and donate the money to charity. In fact, only a small minority of participants indicate that they think this is the most morally acceptable thing for Hank to do.

The second theory this study was designed to test was the *preference-based* approach, which predicted no difference in whether Hank owed Steve or Carol money. To answer this question, we analyzed the data as follows. If a subject judged that Hank should give money to Steve or Carol only, we coded a 1 for Steve or Carol, respectively. If a subject said that Hank should give money to both, we coded a 1 for both Steve and Carol. We then summed the scores for Steve and Carol. A one-sided binomial test determined that Steve's score (86) was significantly higher than Carol's (17) ($p < .0001$).

3.2.2. Side-payments

If participants judged that Hank should accept the offer and give money to Steve or Carol (or both), participants were then asked how much money they would give. A preference-based approach predicts that the amount of money given to each should be approximately the same. An agreement-based approach predicts that Hank should give more money to Steve than to Carol. A one-sided t-test confirms that the mean amount of money participants judged that Hank should give to Steve is greater than that for Carol ($t = 4.53, df = 66.6, p < .0001$, see Fig. 9B.)

4. Study 3a

Studies 1 and 2 provide initial evidence that participants are using virtual bargaining to make moral judgments. However, both studies leave open the question of whether subjects imagine the would-be rule-breaker to be bargaining with a generic, typical, or "reasonable" person (and therefore importing a range of assumptions about what such a person would agree to) or with the specific person who would be part of the negotiation were it to actually take place.

Contractualist theories of moral philosophy propose a variety of ways that attributes of individual bargainers may or may not be involved in reaching an agreement. On some accounts, the preferences and desires of the individual bargainers (as well as their assets, outside options, and so forth) are critical to determining what would be agreed to (e.g. Gauthier, 1986). Others propose that the bargainers should be ignorant of their individual identities during the bargaining process—a position famously advocated for by Rawls (1971) who described the bargainers as existing behind a *veil of ignorance*. Cross-cutting this distinction is the question of what sorts of *other-regarding* preferences ought to be allowed into the bargaining processes. Some bargainers might have *negative* other-regarding attitudes (such as jealousy, spite, and vengeance), which could warp the outcome (or prevent an agreement entirely), while others might have *positive* other-regarding attitudes such as care or the desire to please. Theorists differ as to whether these personal tendencies should be allowed to impact the ultimate agreement. (For a review of these ideas, see Cudd & Eftekhari, 2021.) Finally, nearly all views make some general assumptions about the bargainers' willingness to participate in the bargaining process and/or the norms of engagement that set the stage for the agreement-making process to proceed. For instance, some theories imagine bargainers that are rationally self-interested (caring only about benefits and burdens to themselves, e.g. Gauthier, 1986), while others posit that the bargainers each have a fundamental commitment to justify their actions to the others (Scanlon, 1998).

Studies 3a and 3b begin to investigate the importance of the individual attributes of bargainers in a contractualist moral psychology. Given our proposal that *virtual bargaining* is driving agreement-based moral decision-making, we chose to investigate whether manipulating the willingness of the parties to *engage in a bargaining process at all* impacted our participants' moral judgments. If virtual bargaining is operating with strong assumptions about bargainers coming to the hypothetical contracting situation as fully rational agents (or with other idealized characteristics that abstract away from personal tendencies), then this manipulation should have no impact on participants' judgments. If, on the other hand, virtual bargaining is sensitive to the personal characteristics of bargainers (and, in particular, the willingness of parties to engage in bargaining) then the agreement-based solution is less likely to be endorsed (and, instead, a reliance on previously established rules is likely).

4.1. Materials and methods

This study was preregistered (https://aspredicted.org/WGL_NY7) and approved by the MIT Institutional Review Board. Participants read a vignette similar to the one in Study 1. In this case, a mysterious stranger offers the protagonist, Hank, $1 million if he paints his neighbor's house blue. Critically, the neighbor was either described as reasonable or unreasonable. In the Reasonable Neighbor Condition, subjects were told: "Hank knows that his neighbor is a very reasonable person and would probably agree to have his house painted, especially if he thought it would leave him better off". In the Unreasonable Neighbor Condition, subjects were told: "Hank knows that his neighbor is a very unreasonable person and incredibly protective of his house and would probably refuse to have his house painted, even if he thought it would leave him better off". (See Appendix for details.)

Participants were asked what the most morally acceptable thing would be for Hank to do. Answer choices were as follows: Accept the offer and keep all the money; Accept the offer and donate the $1 million to charity; Accept the offer and give some of the money to his neighbor; Refuse the offer and turn the stranger away.

Sample size was determined by a power analysis of a pilot study. In anticipation of conducting a Chi-Square test to evaluate our central hypothesis (i.e., that there would be a difference in the number of participants choosing the contractualist "split" option), and having observed an effect size around 0.38 and aiming for a power of 0.95 and significance level of .05, we estimated that about 90 subjects were necessary.

150 subjects (MTURK participants) completed the study. 5 participants were excluded for failing control questions (see Appendix for exclusion criteria). Demographics of the participant sample were as

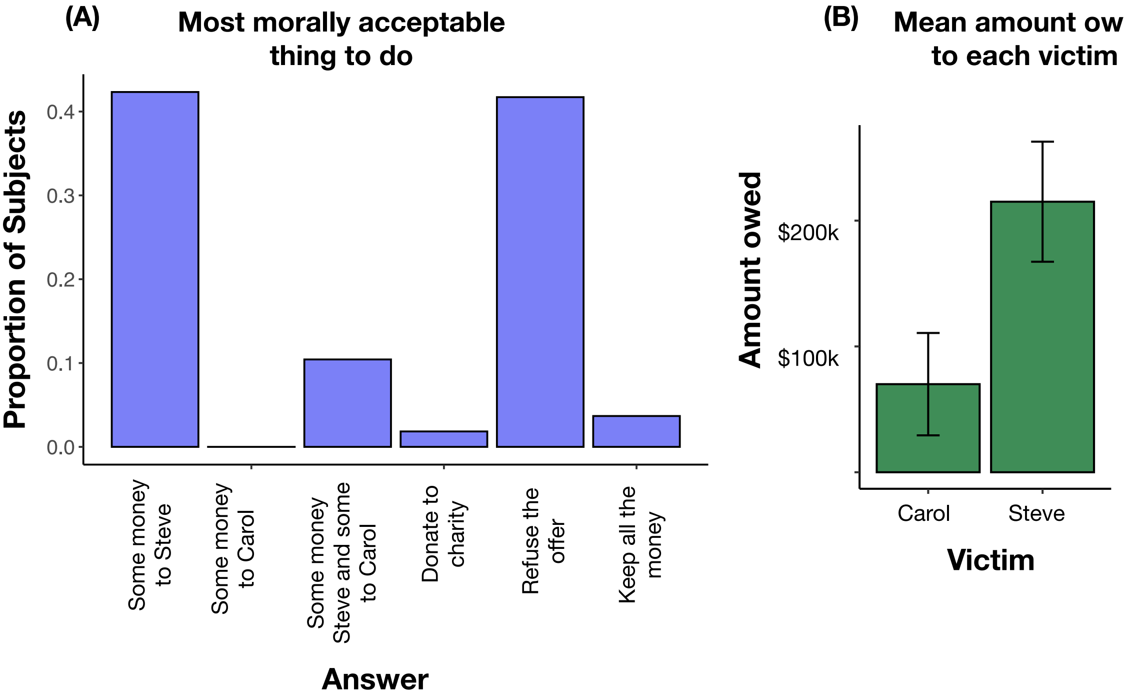

Fig. 9. Study 2 results. (A): Proportion of participants choosing each answer to the question "What is the most morally acceptable thing for Hank to do?" Steve is the owner of the house that Hank is considering painting in exchange for a million dollars. Carol accidentally had her house re-painted due to an unrelated mix-up. (B): If participants judged that Hank should accept the offer and give some money to Steve, Carol, or both, they were then asked how much Hank should give. The graph depicts the mean amount of money these participants judged was owed to Steve and Carol.

follows. Mean age: 41 years old. Gender: 55.9% of participants identified as female/woman, 41.4% identified as male/man, 2.7% identified as non-binary or trans. Race: 2.8% American Indian or Alaska Native, 4.8% Asian, 9.0% Black or African American, .7% Native Hawaiian or Other Pacific Islander, 7.6% Hispanic, Latino or Spanish Origin, 1.4% Middle Eastern or North African, 76.6% White, 4.1% Other or Prefer not to say. 98% of the sample reported that they considered English a primary language for them. Average political leaning was 3.2 on a scale ranging from 1 (extremely conservative) to 5 (extremely liberal).

4.2. Results

Subjects answers were coded in terms of whether they chose the "contractualist" answer (accept the offer and give some of the money to the neighbor; "split") or chose any other answer ("non-split"). χ^2 analysis comparing responses to the two conditions indicated that subjects chose the split option more frequently when the neighbor was described as reasonable than when described as unreasonable, but the effect was only marginally significant ($\chi^2 = 3.7, df = 1, p = 0.055$). χ^2 analysis was also used to compare the conditions when the answer choices were not collapsed, and again showed marginally significant differences between the conditions ($\chi^2 = 8.0, df = 3, p = 0.046$).

Given these inconclusive results, this study was replicated with a larger sample in Study 3b.

5. Study 3b

5.1. Materials and methods

This study was preregistered (https://aspredicted.org/PGW_J3Dd) and approved by the MIT Institutional Review Board. Materials and methods were identical to Study 3a, except for sample size.

Sample size was determined by a power analysis, given the effect size observed in Study 3a. In anticipation of conducting a Chi-Square test to test our central hypothesis, and having observed an effect size around 0.16 and aiming for a power of 0.98 and significance level of .05, we estimated that about 629 subjects were necessary. Study 3a had 97% inclusion rate, so the number of subjects recruited was not inflated beyond this.

627 subjects (MTURK participants) completed the study. 45 participants were excluded for failing control questions (see Appendix for exclusion criteria), leaving 582 subjects included in the analysis. Demographics of the participant sample were as follows. Mean age: 41 years old. Gender: 60.5% of participants identified as female/woman, 38.5% identified as male/man/"XY", .9% identified as non-binary, agender, genderqueer or trans. Race: 2.4% American Indian or Alaska Native, 7.4% Asian, 9.1% Black or African American, .3% Native Hawaiian or Other Pacific Islander, 10.1% Hispanic, Latino or Spanish Origin, .3% Middle Eastern or North African, 80.1% White, 1.7% Other or Prefer not to say. 99% of the sample reported that they considered English a primary language for them. Average political leaning was 3.3 on a scale ranging from 1 (extremely conservative) to 5 (extremely liberal).

5.2. Results

Subjects answers were coded in terms of whether they chose the "contractualist" answer (accept the offer and give some of the money to the neighbor; "split") or chose any other answer ("non-split"). χ^2 analysis comparing responses to the two conditions indicated that subjects chose the split option more frequently when the neighbor was described as reasonable than when described as unreasonable ($\chi^2 = 60.7, df = 1, p = 6.5 * 10^{-15}$). χ^2 analysis was also used to compare the conditions when the answer choices were not collapsed, and again showed significant differences between the conditions ($\chi^2 = 66.2, df = 3, p = 2.8 * 10^{-14}$).

We then combined the datasets from Study 3a and Study 3b (n=727 participants total) and repeated the same analyses (see Fig. 10). When answer choices were collapsed into split and non-split categories, χ^2 analysis comparing the two conditions indicated that subjects chose the split option more frequently when the neighbor was described as

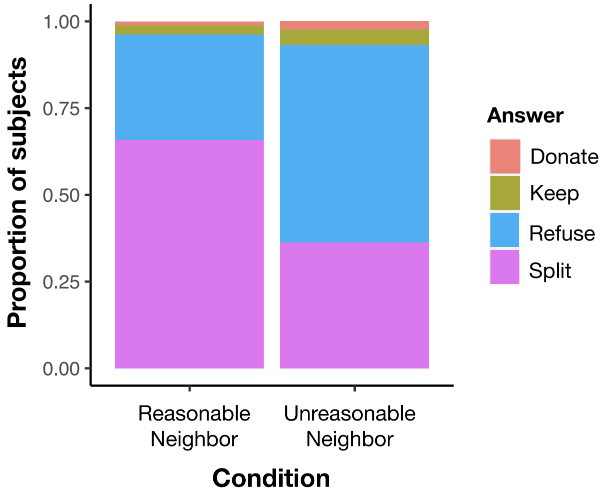

Fig. 10. Results of Studies 3a and 3b (combined data). When the neighbor is described as a reasonable person who would probably agree to a deal in a negotiation, participants are more likely to choose the contractualist option, namely, that the protagonist should paint the neighbor's house blue and give some of the money on offer to the neighbor. When the neighbor is described as unreasonable, participants are more likely to judge that a non-split option (such as refusing the offer entirely) is preferred. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

reasonable than when described as unreasonable ($\chi^2 = 62.505, df = 1, p = 2.7 * 10^{-15}$). When the answer choices were not collapsed, χ^2 analysis again showed significant differences between the conditions ($\chi^2 = 63.9, df = 3, p = 8.5 * 10^{-14}$).

5.3. Discussion

The results of Studies 3a and 3b suggest that participants are taking into account the specific person who would be part of the negotiation (and their reasonable or unreasonable nature) when using a virtual bargaining strategy to make moral judgments.

In addition to "reasonableness", there are many attributes that might impact the outcome of a bargaining process, as reviewed in the introduction to this study. In addition, it is also possible that people will take into account their personal relationship with the other affected parties. Personal relationships create both special obligations (for instance, one might be expected and obligated to care much more for close family members), and also special allowances (for instance, family members might willingly bear inconvenience in a way that strangers would not). Consistent with these possibilities, existing work has illustrated some of the nuanced ways that close personal relationships versus impersonal relationships create distinct expectations and obligations (Everett, Faber, Savulescu, & Crockett, 2018; Marshall, Wynn, & Bloom, 2020; McManus, Kleiman-Weiner, & Young, 2020). These obligations might change the ultimate agreement that arises from a virtual bargaining process. For instance, if one promises a friend to give them a ride home creates an obligation to that friend. If the one reneges on the promise, that will likely influence the relevant social compensation (an apology, a new promise, calling them a taxi, etc.). Exploring the intersection between relationships and contractualist moral decision-making remains an important topic for future research.

The results of these studies also leave open the question of how participants are imagining the parties of the negotiation when no specific information is given about them. One possibility is that most participants imagine the bargainers as reasonable by default, and draw inferences about their preferences and tendencies from there. Substantial recent work – both empirical and theoretical – explores the question of what the standard of reasonableness entails, with one compelling account suggesting that this notion is partially statistical and partially prescriptive (Tobia, 2018). (See also Kneer (2022) and works cited there.) Consistent with this suggestion is the possibility that at least some participants assume by default that the negotiators are *un*reasonable—a possibility we take up in the General Discussion. Future work should investigate the nature and scope of the reasonable person presumption in contractualist moral reasoning and the ways in which individual differences of the bargainers may or may not be able to override the reasonable person presumption.

6. General discussion

When should the rules be followed and when should they be broken? We suggest one possible answer to this question through the mechanism of "virtual bargaining", a kind of contractualist thinking. Rules should be broken when the affected parties would agree to it in a negotiation and they should be followed when an agreement on another course of action cannot be reached. In three studies we find that participants have intuitions about rule-breaking that are well-explained by appeal to contractualist thinking.

Although contractualist thinking offers the best account of the specific cases considered here, we do not argue that all decision-making concerning rule-breaking is contractualist. To the contrary, utility-maximization and purely rule-based patterns of thought are clearly evident in other cases (Cushman, 2013; Greene, 2014; Levine et al., 2023). Indeed our data could be interpreted to mean that not everyone uses contractualist reasoning, even in cases specifically designed to elicit it. The results from Study 1 make this clear: over 40% of the participants in all cases say that Hank should refuse the stranger's offer no matter how much money is on offer and how slight the damage would be to the neighbor's property. The one exception to this rule are the cases where the stranger is threatening to kill Hank's son. This pattern of results could be interpreted to mean that these participants are not willing to flexibly break the rules, instead employing a strict rule-based method for moral judgment in these cases.

Another possibility, however, is that these participants are in fact using contractualist mechanisms to make their judgments, but they place more weight on adherence to the rule than our model accounts for. Given the many benefits of wide-spread social rule-following – detecting cheaters, enabling planning, facilitating coordination – some people may place substantial weight on rule-adherence in general, even when the specific parties involved in an individual rule breach would plausibly agree to break the rule. For instance, in low-trust societies, creating an environment that discourages cheating (ie: one that demands adherence to the letter of the law) might feel more critical to an individual decision-maker than allowing agreement-based rule breaches (André et al., 2022). Put another way, a flexible, contractualist, virtual-bargaining-based mechanism may be at play in the data, though undetectable for some subjects who have a strong bias against rule breaking. Perhaps we would see the characteristic pattern of virtual bargaining emerge in more high-stakes cases where the down-sides of strict rule-following are more pronounced. Indeed, we see a hint of this in the data already: the vast majority of subjects deem it acceptable to break the rule about property rights when Hank's son's life is on the line. (See the "Triple Theory" section below for a further discussion of the interaction between rule-based and agreement-based processes.)

It is also interesting to consider what strict rule-following participants assume about the neighbor's likely reaction. It is possible, for example, that such participants assume that the neighbor would be outraged at a violation of their property, and indeed that they would not be mollified with side-payments. In the language of Study 3, these subjects might be simulating a negotiation, though also assuming that the neighbor is unreasonable. If so, the rule-based and agreement-based accounts would both potentially explain adherence to the rules. The accounts would give different predictions, though, were the participant informed that, as a matter of fact, the neighbor would be happy to accept the damage with suitable compensation (but they are not, of course, able to give their consent as they are out of town). According to a purely rule-based account, participants should still refrain from accepting the stranger's offer nonetheless, sticking to the rule that it is not permissible to damage another person's property without prior consent. But if the rule-followers believe that the neighbor would be likely to be outraged, then they should be reassured to hear that the neighbor would not be—and hence would be more likely to accept the stranger's offer. It is plausible that both styles of moral reasoning leading to adherence to the rules are in play for different participants, and even within a single participant. This issue arises quite generally of course: we may stick to the rules because "rule are rules" or because we may think that others would not agree to our making an exception. Where people are following this latter contractualist logic, they will be more likely to make an exception to a rule given evidence that other affected parties would not object.

Does contractualist-thinking sometimes backfire, causing someone to make a judgment that a rule-breach is permitted when, in fact, it should not be? Understanding that people are doing virtual bargaining allows us to better comprehend their occasional failures, and to diagnose what went wrong. One possibility, of course, is that people may manipulate the virtual bargain in a way that suits their interests (Baucus, Norton, Baucus, & Human, 2008; Gandz & Bird, 1996), e.g., "I bet my neighbor *wants* bleach spilled on his lawn! He never liked that lawn anyway! This will be the reason he needs finally to dig it up". However, another possibility is that people may be acting in good faith – that they earnestly believe it would be best for them to violate a rule – but they incorrectly infer what course of action would be agreed upon. The cases we employed purposely used simple, stripped-down scenarios in which agreement can be easily modeled using a straight-forward computation over a single inferred value (compensation demands). However, decisions that people are actually faced with may include multiple value inferences, projections of uncertain causal sequences of events, and require weighting the utilities of multiple stake-holders appropriately. Uncertainty over these values could lead to conflicts over rule-breaking.

A range of questions concerning contractualist thinking remain unanswered. In multi-agent scenarios, when some agents have directly conflicting interests, how are the interests of the parties taken into account and weighed against each other? How do we determine who is an affected party at all and with whom to (virtually) bargain? All of these provide a rich field for further investigation.

6.1. How flexible is too flexible?

One of the benefits of contractualist decision-making is its flexibility: it takes into account the nuances of a specific situation and the preferences and goals of the affected parties to determine if a rule should be overridden. But if this mechanism were employed constantly, the usefulness of rules – efficiency, predictability, and communicability – would be undermined. Ideally, contractualist rule-breaking should only be deployed when a rule gets the "wrong" answer. But checking whether the rule gets the wrong answer requires calculating both the rule- and the contractualist-based judgments of the case and comparing them. Again, the efficiency of having rules at all would be undermined if this were necessary. When, then, should contractualist reasoning be deployed? This kind of problem is known in the cognitive science and AI literatures as the "strategy selection problem" and has no straightforward solution (but for a promising approach see Lieder & Griffiths, 2017). It is possible that contractualist reasoning is deployed when faced with a particularly unusual situation, where the application of the typical rules and norms are questionable. One empirical implication of this suggestion is that if participants were encouraged to use virtual bargaining when they would not have otherwise, we would expect more instances of permissible rule-breaking judgments. More work remains to be done to test this hypothesis.

6.2. Scaling up, scaling out

The rule-breaking cases we considered involved two agents whose interests were at stake. Imagining what two people would agree to in a negotiation is a tractable problem. But many cases of rule-breaking impact more than just a few people. In an organizational setting, for instance, rule-breaking might impact a huge range of employees, clients, and shareholders. In an academic setting a rule breach may impact students, staff, faculty, and research participants. Is it tenable to imagine a negotiation among everyone affected by a rule breach that impacts so many people?

We suspect that virtual bargaining, as described here, is constrained to cases that involve a small number of affected parties. However, we suspect that virtual bargaining can be "scaled up" for cases that impact larger numbers of people—though it seems likely that short-cuts and heuristics will have to be used to simplify the process (Levine et al., 2023). For instance, when considering whether to break a rule, we might *universalize* our action and imagine what would happen if everyone felt at liberty to break the rule in the situation at hand (Levine, Kleiman-Weiner, Schulz, Tenenbaum, & Cushman, 2020). If things would be fine, then that suggests that those affected would agree to the action because the agent is not taking any special privilege for themselves that they could not grant to others. On the other hand, if it would lead to things going badly, that suggests that the action would not be agreed to. The use of universalization for making moral judgments has been studied in cases where no rule exists, and there is growing evidence that this mechanism is used to determine when it is permissible to break rules (Awad et al., In press; Kwon, Zhi-Xuan, Tenenbaum, & Levine, 2023).

Moreover, the cases considered in this paper were intentionally designed to be highly unusual for reasons enumerated in the introduction. However, a side-effect of this experimental choice was that the decisions made by Hank (and the judgements rendered thereof by our participants) are not necessarily being thought of as establishing a precedent for how these neighbors will interact with each other in the future. The bargains our participants imagine concern only the particular facts of these particular cases; we do not think that the conclusions rendered for these cases are made to be drawn on in future interactions. However, we suspect that there is a version of virtual bargaining that can be "scaled out" to involve negotiating agreements concerning how two people will treat each other when they interact repeatedly. After all, unusual cases come up all the time; rather than negotiating from scratch when faced with a novel case, we may be able to establish standards for interaction that help us navigate complex cases more easily. One tool that would be useful in this regard would be to establish how much two people value each other's utility. Therefore, people might virtually bargain over welfare trade-off ratios (Delton & Robertson, 2016; Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008) in order to make future decision-making more tractable. Indeed, recent research has shown that "special obligations" (e.g. to family members) sometimes are treated as many times more important than obligations to strangers—which has down-stream implications for the moral permissibility of helping one family member (for instance), when the same time or effort could have been used to help many strangers (Everett et al., 2018; McManus et al., 2020). It is possible that virtual bargaining is used to establish these sorts of welfare trade-off ratios (Levine et al., 2023).

The recalibrational theory of anger is an example of how such a view could get implemented through emotion (Sell et al., 2017). Notice that this view is in line with many of the predictions of the (one-off) virtual bargaining process described in this paper (refer back to Figs. 1 and 2).[14] Further empirical work should explore the different predictions made by these two accounts, particularly when manipulating the strength of the relationship between the two neighbors and likelihood that this situation (or any interaction at all) is likely to occur again.

[14] The recalibrational theory makes similar predictions about permissibility judgments as the virtual bargaining theory, as we explain in the Introduction

6.3. A psychological "triple theory": The relationship between agreement-based, rule-based, and outcome-based judgment

It has frequently been suggested that the point of morality (that is, morality's *ultimate* function) is to ensure mutual benefit in a society of many interacting agents with many interacting interests. If so, we might expect that the *proximate mechanisms* of moral judgement involve agreement—a tool that is well-designed to help people get along. Surprisingly, however, very little work has been done proposing that moral judgment is driven by agreement-based processes. In this paper, we have proposed one agreement-based process that drives moral judgment: virtual bargaining.

How do agreement-based processes interact with and relate to other mechanisms of moral judgment, such as rule-based and outcome-based processes? Throughout this paper we have proposed that agreement-based processes are sometimes used to override the use of previously-established rules. Future work should investigate the possibility that agreement-based processes also create and help update those rules and also support purely consequence-based moral reasoning (Levine et al., 2023).

7. Conclusion

As we start to think more seriously about how to program AI systems that act dynamically in the human world they will inevitably need to make decisions in social contexts that are guided by norms. AI engineers have attempted to solve this problem through the use of inflexible rules that constrain the decisions of the AI agent, yet this strategy is quickly becoming obsolete with no clear direction for future progress (for a recent failed attempt see Wurman et al. (2022); for a review of previous attempts and suggestions for future directions see Abel, MacGlashan, & Littman, 2016; Awad et al., In press). Using a computationally-grounded contractualist approach may enable AI systems to mimic the flexibility of the human moral mind (Chater, Misyak, Watson, Griffiths, & Mouzakitis, 2018).

CRediT authorship contribution statement

Sydney Levine: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Max Kleiman-Weiner:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Nick Chater:** Writing – review & editing, Funding acquisition, Conceptualization. **Fiery Cushman:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Joshua B. Tenenbaum:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Data availability

Data available at https://github.com/sydneylevine/blue-house.

Acknowledgments

The authors thank Max Bazerman for his feedback on an earlier draft of this paper and Adam Bear for his assistance with statistical modeling. NC gratefully acknowledges support from the European Union/UKRI under Horizon Europe Programme Grant Agreement 101120763 TANGO (UKRI grant 10082598 - 1389 Horizon Europe Guarantee Extension). Views of opinions expressed however those of the author(s) only and do not necessarily reflect those of the European Union of European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. We also thank John Templeton Foundation, United States of America (Grant 61061), Templeton World Charity Foundation, United States of America, and The Center for Brains, Minds and Machines, United States of America (NSF STC Award CCF-1231216) for financial support. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC-BY) licence to any Author Accepted Manuscript version arising from this submission.

Appendix. Methodological details

Study 1: Stimuli

Group 1: Moral acceptability judgments

Participants read the following vignette: "A mysterious but very wealthy stranger arrives at Hank's doorstep with a strange request. He asks Hank whether he would be willing to do something particular to Hank's neighbor's property. If Hank agrees, he will receive a sum of [$1 million] which Hank can use however he likes. Otherwise, the stranger will leave and everything remains as before. The man shows the [$1 million] dollars in cash to Hank right then and there. Hank can either take the [$1 million] and carry out the stranger's request or he can turn the stranger away.

Participants were randomly assigned to a condition that varied based on offer amount. Amounts included: $100; $1000 $10,000; $100,000; $1,000,000, or a threat against the life of Hank's son ("gun" henceforth).

Participants then saw the following series of ten questions:

"Imagine that the stranger asks Hank to... [Paint his neighbor's mailbox blue.] Would it be morally acceptable for Hank to carry out the stranger's request?".

Property damages were presented in random order and included: painting the neighbor's mailbox blue, painting the outside of the neighbor's front door blue, painting the inside of the neighbor's front door blue, painting the neighbor's house blue, cutting down a tree in the neighbor's yard, breaking all the windows in the neighbor's house, spilling several gallons of bleach on the neighbor's lawn, smearing dog poop on the neighbor's front steps, painting over a mural created by neighbor's daughter, or entirely demolishing the neighbor's house.

After each item, participants were asked "Would it be morally acceptable for Hank to carry out the stranger's request? (Yes/No)" Each participant judged 10 cases: each level of property damage crossed with a single offer amount.

Participants were then asked: "When you were trying to decide whether it was morally acceptable for Hank to carry out the stranger's request, did you think about the fact that Hank could pay his neighbor later? (Yes/No)".

Participants the read the following prompt: "Imagine that Hank accepts the stranger's offer. Hank carries out the stranger's request and the stranger gives him the [$1 million] as promised. On the following page, you will be asked how much money, if any, Hank now owes his neighbor, given what he did".

Participants then saw the following series of ten questions:

"Imagine that Hank... [Painted his neighbor's mailbox blue.] How much money, if any, should Hank give his neighbor? (Please enter a dollar amount in the box below.)".

(see Fig. 1C). It is somewhat less clear what the recalibrational theory predicts for side payment judgments. One possibility is that it would match the predictions of the other preference-based models (see Fig. 2B), though the concern of the potential offender in giving a side-payment would be to mitigate/erase the possibility of anger on the part of the would-be victim rather than calculating how much money he should give the neighbor in order to maximize his own utility. Under this assumption, the side-payment results from Study 1 (see Fig. 7) could be explained by concerns over ameliorating anger and separate concerns over an even split. Moreover, the recalibrational theory makes similar predictions to the virtual bargaining theory in Study 2 (and different predictions than the preference-based model explored in the main text). Again, assuming that side-payments can be used as a way of mitigating anger, the recalibrational theory would predict that Hank should give more money to Steve (whose house he chose to paint) rather than Carol (whose house coincidentally got painted because of an unrelated mix-up).

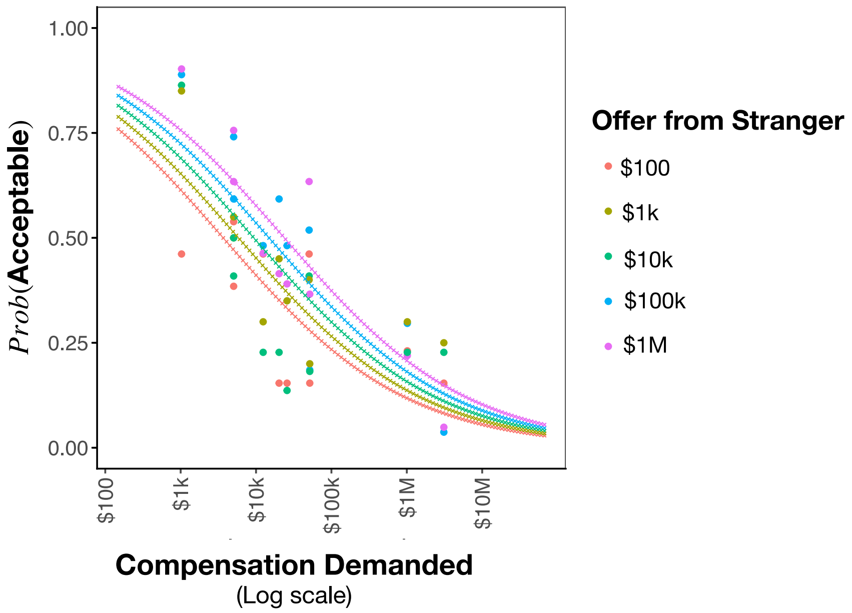

Fig. 11. Study 1: Moral acceptability judgments excluding participants who were characterized by the strict rule-based model (i.e., those who judged that every case was impermissible, regardless of the context). Solid lines are predictions of the agreement-based model, as generated by the regression described in the main text, but with parameters refit to this subset of participants (n=123). Compensation demanded is the 90th percentile of the compensation distributions shown in Fig. 5. Compare to model predictions in Fig. 1 and the model predictions and data associated with the full data set in Fig. 4.

Finally, participants answered the following question: "When thinking about how much money Hank should give his neighbor, which of the following did you take into account? (Check all that apply.)".

- What the stranger was offering Hank
- The amount of damage Hank's action would cause the neighbor
- Compensating the neighbor for inconvenience
- Compensating the neighbor for doing something to him without his permission
- None of these

Was there another factor, not listed above, that you took into account when you thought about how much money Hank should give his neighbor?" [Free Response].

Group 2: Compensation judgments

Participants in this group read the following prompt:

"A mysterious but very wealthy stranger arrives at Hank's doorstep with a strange request. The stranger says that he will pay Hank a sum of money if Hank agrees to have something particular done to his property. After that, the stranger will go away and nothing else will happen to Hank or his property. What is the minimum amount of money you think the stranger would have to offer for Hank to agree to let the stranger do the following things to his property? Please enter a dollar amount in each of the boxes below".

Participants then saw the full list of property damages, for instance, "Paint Hank's mailbox blue".

Study 1: Individual differences analysis of moral judgments: Supplemental results

See Fig. 11.

Study 2: Stimuli

Participants read the following prompt. The colors in the first and second paragraphs were counterbalanced between participants. (Half the participants read that Carol's house had been painted pink and half read that her house had been painted blue.).

"One morning, Hank reads a story in the newspaper about a funny mistake. A woman named Anne wanted to have her house painted pink while she was on vacation. Anne wrote down her address for the painting company, but accidentally put down the address of her neighbor, Carol, instead. Carol was also away on vacation and when she returned, she discovered that her house had been painted pink! "Completely coincidentally, later that day, a mysterious but very wealthy stranger arrives at Hank's doorstep with a strange request. He asks Hank whether he would be willing to pick an address at random from the phone book and paint that house blue. If Hank agrees, he will receive a sum of $1 million which Hank can use however he likes. Otherwise, the stranger will leave him alone and everything remains as before. The man shows the $1 million in cash to Hank right then and there. Hank takes out the phone book and picks an address at random—it happens to belong to a man named Steve who he does not know and does not have time to contact. Hank can either take the $1 million and paint Steve's house blue or he can turn the stranger away.

Which of the following is the most morally acceptable thing for Hank to do?"

Choices were randomly presented to participants in the order they are listed below or the reverse order.

- Accept the offer and give some of the money to Steve
- Accept the offer and give some of the money to Steve
- Accept the offer and give some of the money to Steve and some to Carol
- Accept the offer and donate the money to charity
- Refuse the offer and turn the stranger away
- Accept the offer and keep all the money

If "give some of the money to Steve", "give some of the money to Carol", or "give some of the money to Steve and some to Carol" was chosen, participants were asked how much of the money should be given to the relevant parties. (I.e. If a subject said that the money should be given to Carol, they were asked how much should be given to Carol only. If a participants said that money should be given to both Steve and Carol, they were asked how much should be given to each person.).

The following questions were used:

- How much money should Hank give to Steve? (Please enter a dollar amount in the box below. No dollar sign is necessary.)
- How much money should Hank give to Carol? (Please enter a dollar amount in the box below. No dollar sign is necessary.)

All participants were then asked the following question:

- Please explain your reasoning. [Free Response.]

On the next page of the survey, participants were shown the story again and were then asked a series of comprehension questions.

- What happens between Hank and Carol in the story? [Free Response.]
- What happens between Hank and Steve in the story? [Free Response.]
- How was the stranger involved in Carol's house being painted pink? [Free Response.]

Participants were excluded from analysis who responded incorrectly to any of the control questions.

Participants then answered a series of demographic questions concerning their gender, age, country of residence, U.S. state of residence, languages spoken in their home, whether or not English is a primary language for them, race/ethnicity, religion, level of religiosity, politics, income and level of education.

Study 3a and 3b: Stimuli

Participants read the following prompt.

A mysterious but very wealthy stranger arrives at Hank's doorstep with a strange request. He asks Hank whether he would be willing to paint his next-door neighbor's house blue. If Hank agrees, he will receive a sum of $1 million which Hank can use however he likes. Otherwise, the stranger will leave and everything remains as before. Rather inconveniently, Hank's neighbor is away on vacation, and cannot be communicated with for the next week—but the mysterious stranger requires an answer today. The man shows the $1 million dollars in cash to Hank right then and there. Hank can either take the $1 million and paint the house, without permission from his neighbor, or he can turn the stranger away.

[Reasonable Neighbor Condition:] Hank knows that his neighbor is a very reasonable person and would probably agree to have his house painted, especially if he thought it would leave him better off.

[Unreasonable Neighbor Condition:] Hank knows that his neighbor is a very unreasonable person and incredibly protective of his house and would probably refuse to have his house painted, even if he thought it would leave him better off.

[Dependent Variable:] Which of the following is the most morally acceptable thing for Hank to do?

- Accept the offer and keep all the money
- Accept the offer and donate the $1 million to charity
- Accept the offer and give some of the money to his neighbor
- Refuse the offer and turn the stranger away

[If the third answer choice is selected, the following question was asked:] How much money should Hank give to his neighbor? (Please enter a dollar amount in the box below. No dollar sign is necessary.).

[Comprehension check 1:] What does the stranger offer Hank if Hank agrees to paint his neighbor's house blue? *Subjects passed this comprehension check if they answered that the stranger offered money or gave the amount of money (1 million dollars).*

[Comprehension check 2:] What is Hank's opinion of his neighbor? *Subjects passed this comprehension check in the reasonable condition if they said neighbor was reasonable, rational, agreeable, obliging, easy-going, or the like. They were excluded if just said the opinion of the neighbor was unknown, opinion was simply good, positive, favorable or the like without also mentioning reasonable. They were also excluded if simply said neighbor was away or out of town. In the unreasonable condition, subjects were included if said that the neighbor was unreasonable, protective of his house, would refuse if asked if his house could be painted, not agreeable, stubborn and/or possessive. Subjects were not included if Hank's opinion was just described as poor, strained, unfavorable, not good, greedy, or unknown.*

[Demographic Questions.]

- What is your age?
- Which state are you participating from?
- Which country are you participating from?
- What languages are spoken in your household?
- Would you say English is the/a primary language for you?
- Which categories describe you? Select all that apply.
  - American Indian or Alaska Native
  - Asian
  - Black or African American
  - Native Hawaiian or Other Pacific Islander
  - Hispanic, Latino or Spanish Origin
  - Middle Eastern or North African
  - White
  - Other:
  - Prefer not to say
- What is your present religion, if any?
  - Protestant
  - Roman Catholic
  - Mormon
  - Greek or Russian Orthodox
  - Jewish
  - Muslim
  - Buddhist
  - Hindu
  - Atheist
  - Agnostic
  - Nothing in particular
  - Prefer not to say
  - Other:
- To what extent do you consider yourself to be religious?
  - Not religious
  - Slightly religious
  - Moderately religious
  - Very religious
  - Do not know
  - Prefer not to say
- How would you describe your political views? (5 point scale)
- Please estimate your total household income per year.
- What is the highest level of school you have completed or the highest degree you have received?
  - Less than high school degree
  - High school graduate (high school diploma or equivalent including GED)
  - Some college but no degree
  - Associate degree in college (2-year)
  - Bachelor's degree in college (4-year)
  - Master's degree
  - Doctoral degree
  - Professional degree (JD, MD)

References

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In *Workshops at the thirtieth AAAI conference on artificial intelligence*.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

André, J.-B., Fitouchi, L., Debove, S., & Baumard, N. (2022). An evolutionary contractualist theory of morality. http://dx.doi.org/10.31234/osf.io/2hxgu, osf.io/preprints/psyarxiv/2hxgu.

Awad, E., Levine, S., Loreggia, A., Mattei, N., Rahwan, I., Rossi, F., et al. (In press). When is it acceptable to break the rules? Knowledge representation of moral judgement based on empirical data, *Journal of Autonomous Agents and Multi-Agent Systems*.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74–85.

Baucus, M. S., Norton, W. I., Baucus, D. A., & Human, S. E. (2008). Fostering creativity and innovation without encouraging unethical behavior. *Journal of Business Ethics*, *81*(1), 97–115.

Baumard, N. (2016). *The origins of fairness: How evolution explains our moral nature*. Oxford University Press.

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

Boardman, A. E., Greenberg, D. H., Vining, A. R., & Weimer, D. L. (2017). *Cost-benefit analysis: Concepts and practice*. Cambridge University Press.

Bregant, J., Wellbery, I., & Shaw, A. (2019). Crime but not punishment? Children are more lenient toward rule-breaking when the "spirit of the law" is unbroken. *Journal of Experimental Child Psychology*, *178*, 266–282.

Bridgers, S., Schulz, L., & Ullman, T. D. (2021). Loopholes, a window into value alignment and the learning of meaning. In *Proceedings of the annual meeting of the cognitive science society*: *Vol. 43*.

Chater, N., Misyak, J., Watson, D., Griffiths, N., & Mouzakitis, A. (2018). Negotiating the traffic: Can cognitive science help make autonomous vehicles a reality? *Trends in Cognitive Sciences*, *22*(2), 93–95.

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

Cosmides, L., & Tooby, J. (2013). Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology*, *64*, 201–229.

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

Cudd, A., & Eftekhari, S. (2021). Contractarianism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2021 ed.). Metaphysics Research Lab, Stanford University.

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

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

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

Everett, J. A., 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.

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

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

Gandz, J., & Bird, F. G. (1996). The ethics of empowerment. *Journal of Business Ethics*, *15*(4), 383–392.

Gauthier, D. (1986). *Morals by agreement*. Oxford University Press.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

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

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388.

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

Habermas, J. (1996). *Between facts and norms, trans. William Rehg* (pp. 274–328). Oxford: Polity.

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

Harsanyi, J. C. (1978). Bayesian decision theory and utilitarian ethics. *The American Economic Review*, *68*(2), 223–228.

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

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Proceedings of the 37th annual conference of the cognitive science society*.

Kneer, M. (2022). Reasonableness on the clapham omnibus: Exploring the outcome-sensitive folk concept of reasonable. In *Judicial decision-making: integrating empirical and theoretical perspectives* (pp. 25–48). Springer.

Konow, J. (2003). Which is the fairest one of all? A positive analysis of justice theories. *Journal of Economic Literature*, *41*(4), 1188–1239.

Kwon, J., Zhi-Xuan, T., Tenenbaum, J., & Levine, S. (2023). When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments. http://dx.doi.org/10.31234/osf.io/n8bjr, osf.io/preprints/psyarxiv/n8bjr.

Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. A. (2023). Resource-rational contractualism: a triple theory of moral cognition. http://dx.doi.org/10.31234/osf.io/p48t7, osf.io/preprints/psyarxiv/p48t7.

Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F., & Tenenbaum, J. (2024). Dataset for the paper "When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment". https://github.com/sydneylevine/blue-house.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, *117*(42), 26158–26169.

Levine, S., Rottman, J., Davis, T., O'Neill, E., Stich, S., & Machery, E. (2020). Religious affiliation and conceptions of the moral domain. *Social Cognition*, *39*(1), 139–165.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*(2), 279–311.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762.

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

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

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

Michelbach, P. A., Scott, J. T., Matland, R. E., & Bornstein, B. H. (2003). Doing rawls justice: An experimental study of income distribution norms. *American Journal of Political Science*, *47*(3), 523–539.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152.

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

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

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

Mitchell, G., Tetlock, P. E., Mellers, B. A., & Ordonez, L. D. (1993). Judgments of social justice: Compromises between equality and efficiency. *Journal of Personality and Social Psychology*, *65*(4), 629.

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. Oxford University Press.

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

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

Rawls, J. (1971). *A theory of justice*. Belknap Press.

Rehg, W. (1994). *Insight and solidarity: The discourse ethics of Jürgen Habermas*: *Vol. 1*, Univ of California Press.

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

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

Simon, H. (1955). A behavioral model of bounded rationality. *Quarterly Journal of Economics*, *69*(1), 99–118.

Stich, S. (2018). The quest for the boundaries of morality. In *The Routledge handbook of moral epistemology*. New York: Taylor and Francis Group.

Stonehouse, E. E., & Friedman, O. (2021). Unsolicited but acceptable: Non-owners can access property if the owner benefits. *Journal of Experimental Psychology: General*, *150*(1), 135–144.

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

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*(5), 853.

Tobia, K. P. (2018). How people judge what is reasonable. *Alabama Law Review*, *70*, 293.

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

Wurman, P. R., Barrett, S., Kawamoto, K., MacGlashan, J., Subramanian, K., Walsh, T. J., et al. (2022). Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, *602*(7896), 223–228.