PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

A moral trade-off system produces intuitive judgments that are rational and coherent and strike a balance between conflicting moral values

Ricardo Andrés Guzmán[a,1,2], María Teresa Barbato[a,1], Daniel Sznycer[b,c], and Leda Cosmides[d,2]

Edited by Susan Fiske, Princeton University, Princeton, NJ; received August 16, 2022; accepted August 30, 2022

How does the mind make moral judgments when the only way to satisfy one moral value is to neglect another? Moral dilemmas posed a recurrent adaptive problem for ancestral hominins, whose cooperative social life created multiple responsibilities to others. For many dilemmas, striking a balance between two conflicting values (a compromise judgment) would have promoted fitness better than neglecting one value to fully satisfy the other (an extreme judgment). We propose that natural selection favored the evolution of a cognitive system designed for making trade-offs between conflicting moral values. Its nonconscious computations respond to dilemmas by constructing "rightness functions": temporary representations specific to the situation at hand. A rightness function represents, in compact form, an ordering of all the solutions that the mind can conceive of (whether feasible or not) in terms of moral rightness. An optimizing algorithm selects, among the feasible solutions, one with the highest level of rightness. The moral trade-off system hypothesis makes various novel predictions: People make compromise judgments, judgments respond to incentives, judgments respect the axioms of rational choice, and judgments respond coherently to morally relevant variables (such as willingness, fairness, and reciprocity). We successfully tested these predictions using a new trolley-like dilemma. This dilemma has two original features: It admits both extreme and compromise judgments, and it allows incentives—in this case, the human cost of saving lives—to be varied systematically. No other existing model predicts the experimental results, which contradict an influential dual-process model.

moral psychology | evolutionary psychology | moral dilemmas | judgment and decision-making | moral value pluralism

Significance

Intuitions about right and wrong clash in moral dilemmas. We report evidence that dilemmas activate a moral trade-off system: a cognitive system that is well designed for making trade-offs between conflicting moral values. When asked which option for resolving a dilemma is morally right, many people made compromise judgments, which strike a balance between conflicting moral values by partially satisfying both. Furthermore, their moral judgments satisfied a demanding standard of rational choice: the Generalized Axiom of Revealed Preferences. Deliberative reasoning cannot explain these results, nor can a tug-of-war between emotion and reason. The results are the signature of a cognitive system that weighs competing moral considerations and chooses the solution that maximizes rightness.

Moral dilemmas are an inescapable aspect of the human condition because no single principle regulates judgments in all social interactions (1–4). Natural selection wrote different rulebooks for siblings, parents and offspring, cooperative partners, and coalitional allies, to mention a few. Different cognitive systems evolved for navigating each of these relationships, including ones specialized for helping kin (5), trading goods and favors (1, 6–8), pooling risk in foraging (9–11), and cooperating in groups (3, 12, 13). Each cognitive system is equipped with different concepts and inferential mechanisms, which generate moral intuitions tailored to its domain.

In many situations, moral intuitions collide. A situation in which your friend and sibling both need help may be represented by two distinct systems, each generating different intuitions about the right thing to do. Loyalty to your allies might harm an old friend. The day may be too short to fulfill your duties at both work and home.

When moral intuitions collide, solutions that strike a balance between conflicting moral values are usually possible. Has natural selection produced a cognitive system for making moral trade-offs like these?

Past studies cannot answer this question because they use moral dilemmas that force extreme judgments: ones that fully satisfy one moral value while neglecting others entirely (14, 15). Consider, by contrast, the following dilemma from warfare. It shares many properties with trolley dilemmas, without forcing an extreme judgment.

> Two countries, A and B, have been at war for years (you are not a citizen of either country). The war was initiated by the rulers of B, against the will of the civilian population. Recently, the military equilibrium has broken, and it is certain that A will win. The question is how, when, and at what cost.
>
> Country A has two strategies available: attacking the opposing army with conventional weapons and bombing the civilian population. They could use one, the other, or a combination of both. Bombing would demoralize country B: The more civilians are killed, the sooner B will surrender, and the fewer soldiers will

Author contributions: R.A.G., M.T.B., D.S., and L.C. designed research; R.A.G., M.T.B., and D.S. performed research; R.A.G., M.T.B., and L.C. analyzed data; and R.A.G. and L.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

See online for related content such as Commentaries.

[1]R.A.G. and M.T.B. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: rguzman@udd.cl or cosmides@ucsb.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2214005119/-/DCSupplemental.

Published October 10, 2022.

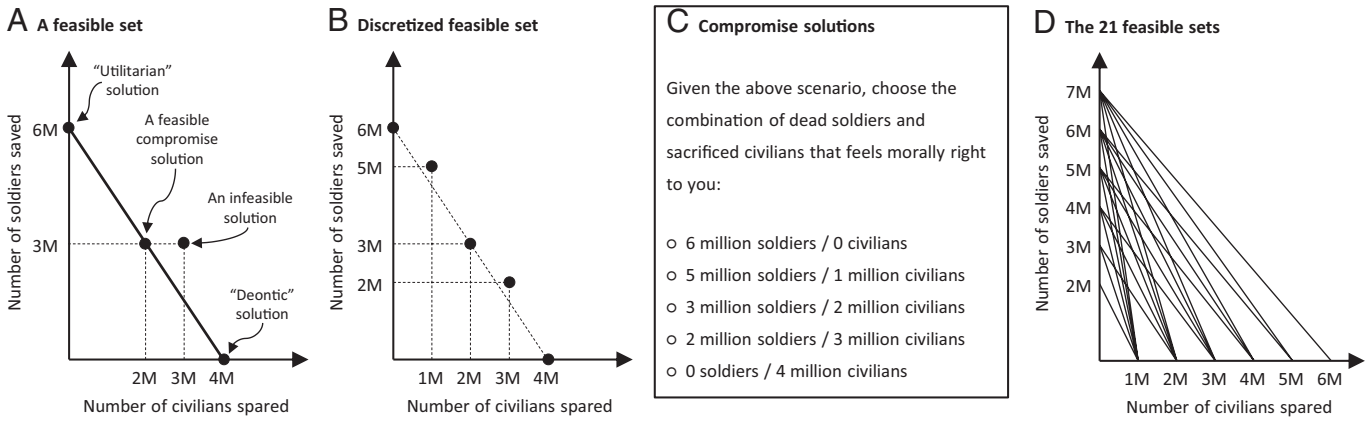

Fig. 1. (*A*) Feasible set for the introductory example. Solutions are expressed in terms of "moral goods": the lives of civilians and soldiers. (*B*) Discretized feasible set corresponding to that scenario. (*C*) Alternatives presented to subjects, corresponding to the solutions in the discretized feasible set (e.g., the alternative "0 sacrificed civilians and 6 million dead soldiers" corresponds to the feasible solution "4 million civilians spared and 0 soldiers saved"). Alternatives are expressed in terms of deaths, which are "moral bads." (*D*) Feasible sets for the 21 scenarios of the war dilemma. Subjects responded to all 21 scenarios, for each willingness frame.

> die—about half from both sides, all forcibly drafted. Conventional fighting will minimize civilian casualties but maximize lives lost (all soldiers).
>
> More precisely: If country A chooses not to bomb country B, then 6 million soldiers will die, but almost no civilians. If 4 million civilians are sacrificed in the bombings, B will surrender immediately, and almost no soldiers will die. And, if A chooses an intermediate solution, for every four civilians sacrificed, approximately six fewer soldiers will die.
>
> How should country A end the war? What do you feel is morally right?

In moral psychology, "sacrifice 4 million civilians" is typically interpreted as a utilitarian response because it saves the most lives (all soldiers). The cost is inflicting maximum harm on civilians. "Do not sacrifice any civilians" is typically interpreted as a deontic response because it respects the principle of not harming bystanders (the civilians). The cost is maximizing the death toll. Both are extreme judgments, because they satisfy one moral value fully but the other not at all. Compromise judgments strike a balance between competing moral values by partially satisfying both. An intermediate solution, such as "sacrifice x civilians to save y soldiers," is a compromise judgment: It spares some (but not all) civilians while saving more (but not the most) lives.*

Each judgment is associated with a solution to this dilemma, expressed as a "bundle" of two moral goods: a number of civilians spared, denoted by c, and a number of soldiers saved, denoted by s. All solutions with $c, s \geq 0$ are conceivable, but not all are available. There is a feasibility constraint, defined by two conditions stated in the dilemma: $s = 6 \text{ million} - 1.5c$, and $c \leq 4$ million. Fig. 1*A* represents these conditions graphically; all points on the line are feasible solutions for this dilemma. This feasible set includes the two extreme solutions—utilitarian and deontic—as well as all points in between: intermediate solutions, such as "2 million civilians spared and 3 million soldiers saved." Solutions that fall above or below the line are not available to be chosen: You might prefer to spare 3 million civilians and 3 million soldiers, for example, but the feasible set does not include this solution.

*Extreme vs. compromise judgments would be called corner vs. interior solutions in rational choice theory. Later, we will refer to maximizing a "rightness function" rather than a utility function, because "utility" is too easily confused with utilitarian in the moral sense. Maximizing a rightness function can result in a compromise or deontic judgment—it does not imply one is maximizing a "utilitarian" moral value, such as saving the most lives.

The lifeways of our hunter-gatherer ancestors routinely created moral dilemmas (16–20). For many, a compromise judgment would have promoted fitness better than an extreme one. These situations should have selected for a cognitive system that is well designed for making trade-offs between conflicting moral values. The adaptive function of this moral trade-off system (MTS) would be to identify, among the available solutions, one that is most right. These judgments would inform (but not determine) behavior.

The Moral Tradeoff System

Four Design Features. To accomplish its adaptive function well, an MTS requires features designed to solve four adaptive problems.

Feature 1. The MTS should be able to produce the full spectrum of judgments: extreme ones and compromises.

Consider this dilemma. A forager fished all day, but his luck was bad. He returns to camp with a catch too small to feed his children and sick brother. The forager's neighbor has been smoking fish for her grandchildren's visit. He asks her for some, but she gives him far less than he requested. The forager feels it would be wrong to steal additional fish from his neighbor, but it would also be wrong to neglect his sick brother. Should he steal from her? If he does, how many fish should he take? The more he steals, the sooner his brother will recover, but the greater the harm to his neighbor.

The solution the forager experiences as most right could be to steal none, some, or all of her fish. The MTS should be capable of delivering any of those answers.

Feature 2. Judgments should vary with incentives. These are variables that determine which solutions to the dilemma are feasible (that is, available to be chosen).

The solutions available to the forager depend on variables such as how much each fish he steals would improve his brother's health and harm his neighbor. He may feel he should steal all of her fish if that would deliver his brother from the verge of death. If his brother has a cold, and some extra food would hasten his recovery by only a day, the forager may feel he should refrain from stealing altogether. If some extra food would significantly hasten his brother's recovery without harming his neighbor too much, the forager may feel that he should steal some, but not all, of her fish.

Feature 3. Judgments should vary with morally relevant variables, such as willingness, fairness, reciprocity, entitlement, merit, and

honesty. The social cognitive systems activated by a dilemma determine which variables are morally relevant to its resolution.

Reciprocity, for instance, is morally relevant to the cognitive system that regulates social exchange (6). When the forager recalls that he helped his neighbor generously last week but she has been stingy in return, that cognitive system may infer that he is entitled to more fish than she gave him. A different "rulebook" regulates altruism toward kin: The forager will feel duty bound to care for his sick brother, even if his brother never reciprocates favors. The fact that the forager's catch was small due to bad luck, not laziness, will be morally relevant to the cognitive system that regulates risk pool sharing: If the dilemma activates that system, it will deliver the intuition that he deserves help from his fellow fishers (21).

Feature 4. The MTS should be able to weigh conflicting moral values and choose a solution that is most right, delivering that solution as an intuitive moral judgment.

The forager has a dilemma because his obligation to help his brother conflicts with his obligation to not harm his neighbor. His MTS should be able to strike a balance between these obligations, and choose the solution he feels is most right. Suppose he feels it would be right to steal 3 of her 10 remaining fish—a solution that would reduce (but not minimize) his brother's recovery time, without severely harming his neighbor. This intuitive moral judgment will serve as input to decisions about how to behave (22, 23), along with prudential factors (e.g., risk of retribution) and temptations (e.g., the extra calories he could consume by stealing more than three fish).

How an MTS Should Work. These four adaptive problems have parallels in rational choice theory, which formally analyzes trade-off decisions. We repurposed tools from rational choice for moral trade-offs, and used this "theory of the computation" [*sensu* Marr (24)] to develop a cognitive model of how an MTS should work.

Considering a decision problem activates one or more social cognitive systems. Each system reads the situation, makes its own inferences, and creates a morally laden mental representation of the situation. If the different representations are contradictory, the MTS is activated.

We propose the MTS is composed of three subsystems: MV, FS, and MAX.

The MV subsystem integrates moral values. It receives the morally laden representations as inputs, assigns weights to the conflicting moral goods, and constructs a rightness function on the fly. A rightness function is a temporary mental representation (a type of data structure), specific to the situation at hand. The function maps all solutions that the mind can conceive of onto a level of rightness.[†] When comparing two or more solutions, the one with the highest level of rightness will feel most right.

A rightness function can be denoted as follows:

$$v(\mathbf{x}, \boldsymbol{\beta}) : X \rightarrow V \subseteq \mathbb{R},$$

where v is the function; $\mathbf{x}$ is a conceivable solution; X is the set of conceivable solutions; V is the set of levels of rightness (which are real numbers); and $\boldsymbol{\beta}$ is a vector of situation-specific parameters computed by MV, by operating on the morally laden representations of the situation. The function gives rise to a rightness order.

[†]Because a rightness function ranks all conceivable solutions to a dilemma, many different scenarios—each representing different harms and benefits to the parties involved—can be evaluated in advance of discovering which one pertains when a decision must be made. This feature enables prospective thinking, as well as on-the-spot judgments.

If $v(\mathbf{x}_1, \boldsymbol{\beta}) > v(\mathbf{x}_2, \boldsymbol{\beta})$, $\mathbf{x}_1$ feels more right than $\mathbf{x}_2$.
If $v(\mathbf{x}_1, \boldsymbol{\beta}) = v(\mathbf{x}_2, \boldsymbol{\beta})$, $\mathbf{x}_1$ and $\mathbf{x}_2$ feel equally right.

The parameters in $\boldsymbol{\beta}$ modulate the way in which the rightness function ranks solutions. It includes the weights assigned to the two moral goods: the lives of civilians and soldiers in the war dilemma, and the welfare of his brother and neighbor in the forager's dilemma. Context can affect these parameters: If the civilians had supported the war, their lives would probably weigh less. If the neighbor had been generous, her welfare would probably weigh more.

The MV subsystem and the cognitive systems that inform it are universal, but they are calibrated by the individual's experiences. So, faced with the same situation, different people might generate different parameter values. In consequence, their moral judgments may differ.

Working in parallel to MV, the FS (for feasible set) subsystem constructs a feasible set that captures the incentives of the dilemma. A feasible set is a mental representation of the solutions that the MTS perceives as available.

Subsequently, the MAX subsystem maximizes the rightness function, given the constraints posed by the feasible set. The result is an optimal solution $\mathbf{x}^*$. MAX outputs the intuitive judgment "solution $\mathbf{x}^*$ is the most right."

When this process ends, the rightness function and feasible set are erased or fade from memory.

MV, FS, and MAX operate nonconsciously: Their computations are not performed via deliberative reasoning. The MTS operates like the visual system: Its final products are objects of awareness, but its computations are not. Just as we "see" objects, we "feel" that some options are more right than others. Like the sight of an object, the feeling "solution $\mathbf{x}^*$ is most right" is a representation that can be read by many downstream systems, including those for deciding how to behave (25).

Even though the MTS operates nonconsciously, conscious deliberations can play a role in judgment. Arguments and reflection can change which social cognitive systems are activated by a dilemma and, therefore, the representations MV uses to compute $\boldsymbol{\beta}$. This can affect intuitive judgments.

A cognitive system with this architecture is capable of producing compromise judgments (feature 1). It responds to incentives and morally relevant variables (features 2 and 3). And it assigns weights to competing moral goods, which allows it to choose a feasible solution that is most right (feature 4).

Rightness functions are a type of utility function, and maximizing a utility function produces choices that comply with the axioms of rational choice (26, 27). Alternative theories of moral judgment do not predict compromise judgments that respect the axioms of rational choice. Theories that attribute judgment to adaptive specializations are silent on these issues but consistent with them (3, 4, 8, 28–38). But rigid heuristics, inflexible emotions, or deliberative reasoning cannot explain rational judgments that include compromises. Indeed, a well-known dual-process model—one that invokes both inflexible emotions and deliberative reasoning—makes opposing predictions.

A Competing Theory: Greene's Model

Is it right to stop a runaway trolley by pushing a bystander onto the tracks, sacrificing his life to prevent the trolley from running over five workers? A prohibition against inflicting harm says no; this is considered a deontic judgment. A principle of maximizing aggregate welfare says yes; this is considered a utilitarian judgment. If making consistent deductions from a single normative principle

is the standard for rationality (14, 15, 39), then human moral judgment falls short: Many people flip-flop between deontic and utilitarian judgments when the consequences of a trolley problem remain the same but other features of the situation change (14, 40, 41).

Moral flip-flopping is usually explained by a dual-process model (42). System 1 is composed of emotions, heuristics, and inferences, which produce moral intuitions. System 2 performs deliberative reasoning. Judgments can flip because the intuitions produced by system 1 can preempt, interfere with, or bias judgments reached by reasoning. These models vary greatly (3, 14, 36–38), but most assume that system 1 computations are automatic, nonconscious, effortless, and fast, whereas system 2 computations are controlled, conscious, effortful, and slow.

Greene's dual-process model (14) makes three additional claims: 1) Emotions produce inflexible responses; 2) flexibility—responding to context by considering multiple factors—requires deliberative reasoning; and 3) utilitarian judgments are produced by reasoning, whereas deontic judgments are produced by emotions.

According to this model, we experience the trolley problem as a dilemma because "two [dissociable psychological] processes yield different answers to the same question" [ref. 15, p. 269]. When considering whether sacrificing the bystander to save five people is morally permissible, System 2, operating on emotionally neutral representations of the situation, does a cost–benefit analysis. Not pushing saves one life (the bystander), whereas pushing saves four (five workers minus one bystander), so System 2 outputs the utilitarian judgment: "Push the bystander." But the prospect of pushing the bystander activates an "alarm bell" emotion, which issues an inflexible internal command: DO NOT HARM. That command translates into the deontic judgment: "Do not kill this innocent bystander!"

According to Greene and colleagues (14, 15), this prohibition against inflicting harm is "nonnegotiable": It cannot be weighed against other values. Commands issued by the alarm bell are, by design, difficult to override, because their adaptive function is to prevent actions that disrupt cooperative relationships (14, 20).

Cushman and Greene (15) argue that nonnegotiability makes moral conflict fundamentally different from motivational conflict in other domains. "When motivational systems conflict, this conflict can often be negotiated by weighing the preferences against each other" (p. 276). "Intractable dilemmas arise when psychological systems produce outputs that are. . .non-negotiable because their outputs are processed *as absolute demands, rather than fungible preferences* [emphasis added]." As a result, the trolley problem is experienced as intractable: A deontic judgment will *feel* right, a utilitarian judgment will *seem* right, and subtle changes in context will cause people to flip between these two extremes (14, 40, 43). But no middle ground will ever feel right or seem right, because there is no psychological machinery for negotiating a compromise judgment.

Acknowledging that the nonnegotiability hypothesis is speculative, Cushman and Greene (15) say that "putting it to empirical test is an important matter for further research." Here we test it—along with five novel MTS predictions.

Testing the Nonnegotiability Hypothesis

Testing the nonnegotiability hypothesis requires a sacrificial moral dilemma that permits compromise judgments. The trolley problem and other standard dilemmas cannot be used, because they force subjects to choose extreme responses (e.g., push or do not push).

We used the war dilemma because it satisfies both requirements. It is sacrificial because increasing the number of survivors entails sacrificing bystanders: the civilians. It also permits compromise judgments.

The war dilemma has two additional features tailored to intensify a nonnegotiable demand. First, the dilemma has three of four characteristics, each of which makes inflicting harm less morally acceptable to people (41): The harm is inflicted on others, not self; it is instrumental, not a side-effect; and harming civilians is avoidable, not inevitable. Second, history tells us that the war dilemma would activate an alarm bell, if one exists. Bombing civilian targets caused moral outrage in World War II, leading to a revision of the Geneva Accords: They permit attacks on military targets, but they prohibit deliberate or indiscriminate attacks on civilians.

To illustrate how Greene's dual-process model would handle the war dilemma, suppose that 6 million soldiers are at risk, and that each civilian sacrificed saves three soldiers.

If the prospect of sacrificing one bystander activates an alarm bell emotion, as Greene maintains, then the prospect of killing a civilian should too. The prohibition issued by this alarm—"Do not bomb civilians!"—should be experienced as a nonnegotiable demand. Because intermediate solutions kill civilians, they will not satisfy that demand. Only the deontic judgment, "spare all civilians (2 million) and save zero soldiers" should feel right.

System 2 would determine that sacrificing all 2 million civilians will maximize the number of survivors. Therefore, system 2 would conclude that "spare zero civilians and save all 6 million soldiers" is more right than other solutions.

Systems 1 and 2 would issue competing judgments for the war dilemma. But there is no psychological machinery that can weigh these moral preferences against each other to produce a compromise, because the prohibition against killing civilians is nonnegotiable. So people will always opt for an extreme solution, even when intermediate solutions are available. They will never make compromise judgments.

Contrasting predictions follow from the MTS model.

Testing for a Moral Tradeoff System

The war dilemma is also designed to test features of the MTS. Warfare is a domain that was relevant during human evolution (19, 20) and activates multiple moral intuitions (3, 13, 44–46). Minimizing loss of life in warfare is often seen as a moral good; so is sparing innocent lives. Across societies, from hunter-gatherers to nation states, people make moral distinctions between warriors and those they protect—family and friends ancestrally, civilians now (44–46). So the evolved systems activated by the war dilemma are likely to produce conflicting moral intuitions, thereby activating the MTS.

Compromise Moral Judgments. A well-designed MTS should be able to produce extreme and compromise judgments (feature 1).

Prediction 1. Compromise judgments will be common for the war dilemma.

Support for this prediction is evidence in favor of the MTS model and against the nonnegotiability hypothesis.

Response to Incentives. Judgments should respond to incentives (feature 2). We tested this feature by systematically varying the dilemma's incentives over a succession of scenarios. These incentives are given by two parameters: S, the number of soldiers at risk of death, and S/C, the number of soldiers saved for each civilian sacrificed. In all scenarios, $S/C > 1$, so that the number

of deaths is always minimized when C civilians are sacrificed. The 21 feasible sets for these scenarios are depicted in Fig. 1*D*.

Prediction 2. Subjects will respond to incentives by changing their judgments.

The "cost-effectiveness" of sacrificing one life has received attention in studies with standard dilemmas: Utilitarian judgments were more likely when sacrificing one life saved larger numbers of people. This makes sense on both evolutionary and philosophical grounds (23, 47).

Conjecture 1. Subjects will say that a larger percentage of soldiers should be saved if S/C is greater, all else being equal.

Response to Morally Relevant Variables. Morally laden representations of the situation serve as input to the MTS. When a change in situation alters these representations, the MTS should be capable of responding with different judgments (feature 3).

In war, people's willingness to participate is morally relevant to assigning responsibility, blame, and honor. The introductory example said the people at risk were unwilling participants: Civilians had "opposed the war," and soldiers "were forcibly drafted." If one of these variables in the vignette is changed to "civilians supported the war" or "soldiers volunteered," the MV subsystem may construct a different rightness function. Maximizing a different rightness function is likely to shift judgments.

To study this effect, we created three different frames for the dilemma—BU: both unwilling, the baseline; CW: civilians willing, a variant in which the civilians supported the war (but soldiers remain unwilling); and SW: soldiers willing, a variant in which the soldiers volunteered (but civilians remain unwilling).

Prediction 3. Between frames, subjects will make different judgments.

Each subject responded to the 21 scenarios twice, once to the BU frame and once to either CW or SW.

Moral Coherence. As a check that subjects are responding to willingness because of its *moral* relevance, we assessed whether their judgments shifted coherently with this variable. In the war dilemma, the logic of consent suggests that equal or greater harm should befall willing than unwilling participants (47), as follows.

Conjecture 2. Holding S and S/C constant: Taking BU as a baseline, subjects will say an equal or greater number of civilians should be sacrificed in CW, and an equal or lesser number of civilians should be sacrificed in SW.

Judgments Will Respect the Axioms of Rational Choice. Rational choice models assume that an agent has a preference order. This allows the agent to compare any options it can conceive of. Preference orders vary with context, but they do not change when incentives change. Different agents can have different preferences, so, when facing the same problem, they might make different choices. Here, the MTS is an agent, and its preferences are moral: They rank solutions in terms of rightness.

Preference orders have several properties. Consistency states that the statements "$\mathbf{x}$ is at least as good as $\mathbf{y}$" and "$\mathbf{y}$ is better than $\mathbf{x}$" cannot both be true at the same time. Transitivity states that, if "$\mathbf{x}$ is at least as good as $\mathbf{y}$" and "$\mathbf{y}$ is at least as good as $\mathbf{z}$," then "$\mathbf{x}$ is at least as good as $\mathbf{z}$." Rational choice theory assumes that, among all feasible options, the agent will choose the one that it most prefers, or one of the most preferred if there is a tie.

Revealed preference methods can be used to assess the consistency of a sequence of choices, as defined by the axioms of rational choice theory. Broadly speaking, the experimental strategy is to present a person with a sequence of problems that are identical in all respects except for the incentives. If she reveals, through her choices, that "$\mathbf{x}$ is at least as good as $\mathbf{y}$," she will not reveal through later choices that "$\mathbf{y}$ is better than $\mathbf{x}$." Each violation of this logical requirement is an inconsistency. The fewer the inconsistencies, the more rational the person's choices.

We used a revealed preference method to test whether judgments are made by maximizing a rightness function.

Prediction 4. Within each frame, a subject will reveal few inconsistencies.

If prediction 4 is confirmed even for subjects who chose compromises, that is strong evidence for an MTS and against the nonnegotiability hypothesis, which must view compromises as random errors.

Rightness Functions Are Temporary. If rightness functions are temporary mental representations, then the MTS can construct a different one for each frame, resulting in different judgments. We tested this hypothesis by examining subjects who changed their judgments with willingness.

Prediction 5. Subjects who make different judgments between frames will reveal few inconsistencies within each frame.

Empirical Investigation

1,745 subjects participated in two conditions each (order counterbalanced): one with the BU frame and another with a variant—either CW ($n = 845$) or SW ($n = 900$).

Each condition consisted of 21 scenarios of the war dilemma (order randomized) with the same frame but different incentives. Across scenarios, S varied from 2 million to 7 million deaths, C varied from 1 million to 6 million deaths, and S/C varied from 1.17 to 7 soldiers saved per civilian sacrificed. The values of S/C cover the range typical of trolley problems. Fig. 1*D* depicts the feasible sets of the 21 scenarios. All include extreme solutions (deontic and utilitarian). The scenarios were repeated across frames, so every subject encountered each scenario twice.

A scenario offers several alternatives for ending a war. An alternative consists of a number of civilians sacrificed and a number of soldiers killed. Fig. 1*C* shows the alternatives presented to the subjects in one scenario, while Fig. 1*B* depicts the discretized feasible set. Alternatives are bundles of moral bads (deaths), whereas the feasible solutions are bundles of moral goods (lives). To simplify choices, we rounded lives to the nearest million. This resulted in two to seven alternatives per scenario. For comparison to past studies, six scenarios offer only extreme solutions. Examples are shown in Table 1 (see *SI Appendix*, Table S1 for all 21).

Since our hypothesis is about moral intuitions (not the ability to reason from a philosophical principle), subjects were encouraged to answer "what you feel is morally right, which may or may not be the same as what you think is morally right." The full text of the instrument is provided in *SI Appendix*.

Results and Discussion

Compromise Judgments Were Common. Prediction 1 was confirmed: 71% of subjects made compromise judgments in at least one condition. Fig. 2*A* shows the percent who made compromises in each frame. We will call these subjects "compromisers." The figure also shows the percent of subjects who made the same extreme judgment 21 times and the percent who flip-flopped between extreme judgments. We will call these subjects "extreme responders" and "flip-floppers."

Table 1. Four scenarios of the dilemma

No. soldiers at risk (S)	No. soldiers saved for each civilian sacrificed (S/C)	Alternatives (civilians sacrificed, soldiers dead)	Feasible set (civilians spared, soldiers saved)
4	2.00	(0, 4) (1, 2) (2, 0)	(2, 0) (1, 2) (0, 4)
6	2.00	(0, 6) (1, 4) (2, 2) (3, 0)	(3, 0) (2, 2) (1, 4) (0, 6)
7	3.50	(0, 7) (1, 4) (2, 0)	(2, 0) (1, 3) (0, 7)
7	1.75	(0, 7) (1, 6) (2, 4) (3, 2) (4, 0)	(4, 0) (3, 1) (2, 3) (1, 5) (0, 7)

Note: All quantities in millions of lives.

The nonnegotiability hypothesis predicts that subjects always intend to make an extreme judgment, because they lack the cognitive capacity to negotiate deontic and utilitarian values. They could, however, make compromise judgments occasionally, due to "trembling hand" mistakes: clicking on an option other than intended by accident (due to a slip of the hand or lapse of attention). This could happen in the 15 scenarios with intermediate solutions.

As Fig. 2*B* shows, the distribution of compromise judgments is incompatible with a preference for extreme judgments, peppered with an occasional mistake. If hand trembling were the correct explanation, the "compromised" category would have been crowded with subjects who made one or two compromise judgments. But it was not. In every frame, 85% of subjects who chose an intermediate alternative made 3 to 15 compromise judgments (*SI Appendix*, Fig. S1), 8.8 on average (SD 4.7). When intermediate alternatives were available, compromisers chose them 58% of the time.

The prevalence of compromise judgments rules out the nonnegotiability hypothesis. It is evidence for a system that can resolve dilemmas with compromises (feature 1).

Subjects Responded to Incentives. Prediction 2 and conjecture 1 were also confirmed: The percent of soldiers that subjects felt should be saved increased with S/C. The effect is shown in Fig. 3*A*.

Six scenarios forced a choice between a deontic and a utilitarian solution, as in the standard design for moral dilemma experiments. When S/C increased, the percent of subjects choosing the utilitarian alternative increased, showing that incentives matter even when subjects are forced to make an extreme judgment, as shown in Fig. 3*B*.

These results show that the cognitive system in charge of moral judgment can respond to incentives (feature 2).

Subjects Responded Coherently to Willingness. Moral coherence is visible in the subjects' average responses. Fig. 3*A* shows, for each scenario, the average percent of soldiers that subjects felt it was right to save in CW, BU, and SW. For every scenario, the percent to be saved was highest when civilians were willing, intermediate when both were unwilling, and lowest when soldiers were willing. The proportion of extreme responders reveals the same response pattern: "Utilitarians" were most common in CW and least common in SW; the reverse was true for "deontics" (Fig. 2*A*). These results support prediction 3 and conjecture 2.

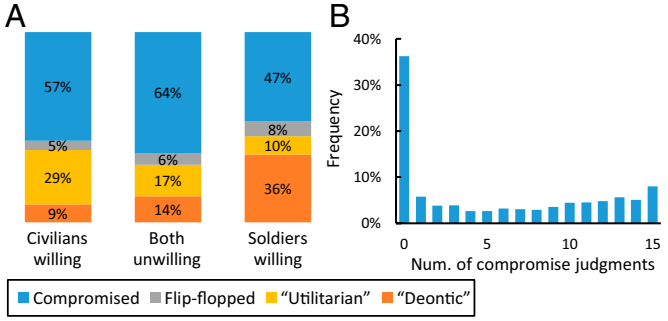

Fig. 2. (*A*) The majority of subjects made compromise judgments. (*B*) Eighty-five percent of subjects who compromised made three or more compromise judgments. BU is pictured.

Willingness Changes the Subjective Value of Lives. What weights did subjects assign to civilian and soldiers' lives, and how did those weights change when willingness changed? To find out, we fitted the rightness function of a "representative agent": one whose response to each scenario is the average of the subjects' responses.

Fig. 3*A* depicts the representative agent's 63 answers (21 per frame). Its rightness function matched a constant elasticity of substitution utility function (*SI Appendix* explains the estimation procedure). The function can be written as follows:

$$v(\mathbf{x}, \boldsymbol{\beta}) = \alpha^{\frac{1}{\sigma}} c^{\frac{\sigma-1}{\sigma}} + (1-\alpha)^{\frac{1}{\sigma}} s^{\frac{\sigma-1}{\sigma}}$$

where $\mathbf{x} = (c, s)$ is a solution, c and s are the numbers of surviving civilians and soldiers, and $\boldsymbol{\beta} = (\alpha, \sigma)$ is a vector of parameters. The value of both parameters can differ across the three frames, because frames differ in two morally relevant variables: the willingness of civilians and the willingness of soldiers. The estimated values of α and σ give us a central tendency of subjects' moral preferences; they are shown in Table 2.

Parameter $\alpha \in [0, 1]$ represents the weight of civilians in moral value, whereas $1 - \alpha$ represents the weight of soldiers. If $\alpha = 0.5$, the agent cares equally about both types of people; if $\alpha > 0.5$, it cares more about civilians than soldiers. As expected, it valued civilians most highly when soldiers were willing, at an intermediate level—but still more highly than soldiers—when both were unwilling, and equally when civilians were willing.

Parameter $\sigma \geq 0$ is the elasticity of substitution: It regulates the sensitivity of the agent's responses to changes in S/C. The elasticity of substitution did not vary with willingness. The high value of σ (approximately two) means that the representative agent is highly sensitive to incentives.

Coherence of Individual Subjects. We counted coherence violations for each individual as follows: Let (c_{ij}, s_{ij}) be the solution chosen by the subject in scenario j of frame i, where $j = 1, \ldots, 21$ and $i \in \{\mathrm{BU}, \mathrm{CW}, \mathrm{SW}\}$. A subject violated coherence in scenario j if $c_{\mathrm{CW}j} > c_{\mathrm{BU}j}$ (a choice that entails the death of more unwilling than willing civilians) or if $s_{\mathrm{SW}j} > s_{\mathrm{BU}j}$ (a choice that entails the death of more unwilling than willing soldiers). By definition, each subject could violate coherence up to 21 times.

We could ask whether subjects' responses were more coherent than expected by chance. This benchmark is too lax, however, because a subject that responds to incentives but does not maximize rightness would violate coherence less often than expected by chance alone. So we created a tougher benchmark: the judgments of simulated agents that respond to incentives exactly like subjects

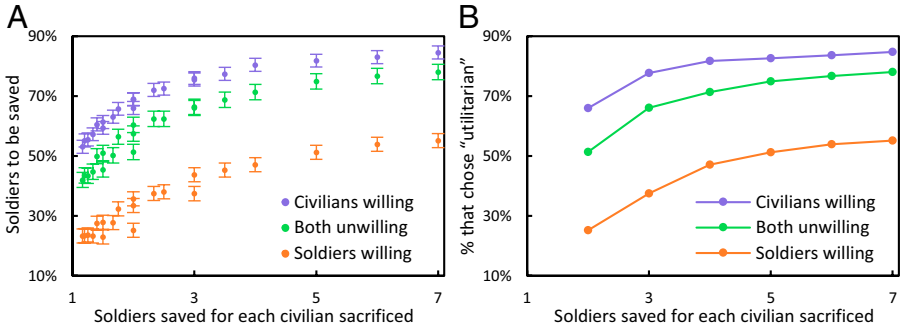

Fig. 3. Subjects responded to the human cost of saving lives; the *x* axes represent S/C. (*A*) Percent of soldiers saved in each of 21 scenarios, with 95% CIs. As S varied, approximately the same proportion of soldiers were saved for each value of S/C. (*B*) All subjects responded to six scenarios in which they were forced to choose one of two extreme options; percent "utilitarian" responses for those scenarios.

did on average. We call this benchmark "incentives + chance." The 21 responses of a simulated agent were generated by independently resampling subjects' choices for each separate scenario in a given frame. One million agents were created for BU and CW, and 1 million for BU and SW.

Subjects outperformed the incentives + chance benchmark by a wide margin (*SI Appendix*, Fig. S2); 69% of subjects never violated coherence, compared to less than 1% of simulated agents.

The 367 subjects who always made the same extreme judgment in both frames cannot violate coherence. The other 1,378 subjects could. Coherence was high for these subjects: 62% made zero violations, and 81% made coherent judgments in 90 to 100% of the 21 scenarios. Responding to incentives does not, by itself, produce these high levels of moral coherence: Only 20% of the incentives + chance agents were coherent in 90 to 100% of scenarios. See *SI Appendix*, Fig. S2 for subjects in CW/BU and SW/BU separately.

The average and individual-level results converge. Seventy-eight percent of subjects made different judgments in their two frames, confirming prediction 3. The cognitive system producing judgments responded to a morally relevant variable (feature 3), and did so coherently, confirming conjecture 2.

Measuring Moral Rationality. Judgments made by maximizing a well-behaved rightness function will respect GARP—the generalized axiom of revealed preferences (26, 27).‡ (See *SI Appendix* for a detailed explanation.) This mathematical fact allows for a crisp test of moral rationality.

We used the number of GARP violations made by each subject as a measure of irrationality. A subject makes a GARP violation by revealing an inconsistency of the form "$\mathbf{x}$ feels more right than $\mathbf{y}$" and "$\mathbf{y}$ feels at least as right as $\mathbf{x}$." Revelations can be direct or indirect. Direct revelations involve two or more solutions that are in the same feasible set. Indirect revelations are inferred from transitivity chains that link three or more options that are not all in the same feasible set.

Fig. 4 illustrates a simple type of GARP violation involving two consecutive scenarios. Respecting GARP becomes increasingly difficult as an agent faces more scenarios, due to the indirect revelation mechanism operating within and across them. Simulations show that, for the 21 scenarios, a subject can make up to 152 violations.

A perfectly functioning MTS will produce zero GARP violations. If it resides in a subject who occasionally makes a trembling hand mistake, then her judgments may fail to respect GARP fully. The fewer the GARP violations, the more evidence that a subject's judgments were made by maximizing a well-behaved rightness function.

Making few GARP violations cannot be accomplished by chance: Random responders make a median of 61 violations. Responding to incentives is also insufficient to appear rational: Incentives + chance agents make a median of 45 violations.

Subjects Made Rational Judgments. Subjects outperformed both benchmarks, exhibiting high levels of rationality. The percent of subjects who never violated GARP was 71% in BU, 77% in CW, and 78% in SW. The corresponding figures were 0%, 1%, and 0% for the incentives + chance agents, and 1% for random responders. Moreover, 88% of subjects made no violations in at

Table 2. Parameters of representative rightness function

Condition	α	99% CI	σ	99% CI
Civilians willing	0.49	[0.47, 0.51]	1.98	[1.88, 2.08]
Both unwilling	0.61	[0.59, 0.63]	1.99	[1.89, 2.08]
Soldiers willing	0.80	[0.78, 0.81]	1.88	[1.78, 1.97]

Results from a nonlinear least squares regression. For details, see the *SI Appendix*.

‡Well behavedness is an auxiliary assumption. It consists of three mathematical properties: continuity, nonsatiation, and convex indifference curves. This assumption biases the test against the MTS hypothesis: Judgments produced by maximizing a function with different properties may satisfy the axioms of rational choice (48), yet violate GARP.

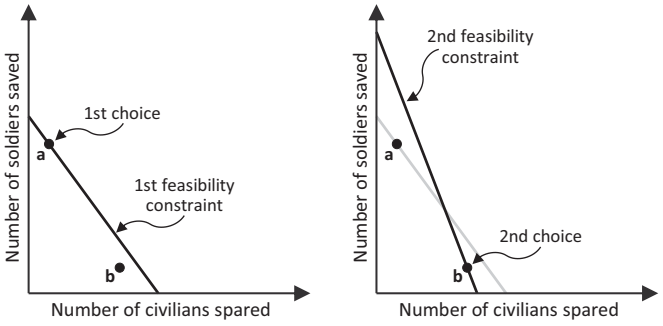

Fig. 4. A simple type of GARP violation can occur when a subject faces two consecutive scenarios with intersecting feasibility constraints. Suppose the subject chooses **a** from the first feasibility constraint. This choice reveals that **a** feels more right than every solution below the constraint, including **b**. We can infer this even though **a** and **b** are not in the same feasible set (i.e., **b** was not available to be chosen), and **a** is not Pareto superior to **b**. This inference follows from the well-behavedness assumption (the proof is in *SI Appendix*). Next, the subject chooses **b** from the second feasibility constraint. This choice reveals that **b** feels more right than all the solutions below that constraint, including **a**. We can therefore infer that **b** feels more right than **a**, even though **a** is not feasible in the second scenario. These two inferences are contradictory: They imply that **a** feels both more right and less right than **b**. This inconsistency is a GARP violation.

Table 3. Rationality in the BU frame

Individuals	Never violated GARP, %	Made seven or fewer violations, %	Median no. of violations	90th percentile	95th percentile
All subjects	71	94	0	5	10
Compromisers	59	92	0	7	15
Flip-floppers	53	88	0	9	17
Benchmarks					
Random responders	∼0	1	61	90	97
Incentives + chance	∼0	5	45	78	86
Rational + 1 tremble	39	76	1	15	23
Dual + compromises	39	62	3	26	35

least one of their two frames. In each frame, 94% of subjects made seven or fewer violations, whereas <10% of incentives + chance agents did.

For detecting a cognitive system that maximizes rightness, zero violations is too strict a standard. The hypothesized cognitive system resides in a human, whose hand could tremble. So we compared subjects to simulated agents that are perfectly rational but make a single, random mistake. This *rational + 1 tremble* benchmark was created by resampling the subset of subjects with zero GARP violations in a given frame. In each iteration, we randomly changed one of the 21 choices of the drawn subject, to create one mistake per rational agent. We created 2 million agents for BU, 1 million for CW, and 1 million for SW.

For rational + 1 tremble agents, the median number of GARP violations was low: one in BU, two in CW, and one in SW. About 38% remained perfectly consistent, and ∼77% made seven or fewer violations. Yet subjects scored better on all of these measures, as Table 3 and *SI Appendix*, Tables S2 and S3 show.

Compromisers Provide a Critical Test. Subjects who respect GARP are expected on the MTS model. Is the same true of the dual-process model?

Extreme responders (who always respect GARP) can be explained by both models. The dual-process model will produce a deontic response profile if an alarm bell emotion preempts reasoning every time, and a utilitarian profile if reasoning always prevails. An MTS can also produce these profiles. A rightness function that assigns zero weight to soldiers $[v(c, s) = c]$ will produce a deontic profile; one that assigns zero weight to civilians $[v(c, s) = s]$ will produce a utilitarian profile.[§]

Both models can also explain GARP-respecting flip-floppers, but only if we make the charitable assumption that a dual process somehow responds to incentives. The dual-process model will produce them if and only if emotion trumps reason when S/C is below a threshold. An MTS will produce them if it constructs a linear rightness function, with positive weights assigned to both soldiers and civilians $[v(c, s) = \alpha c + (1 - \alpha)s$, where $0 < \alpha < 1]$. When S/C exceeds a threshold, the optimal solution is utilitarian; when S/C is below that threshold, the optimal solution is deontic (see *SI Appendix* for the proof).

Most subjects were compromisers, however. Their rationality—or lack thereof—provides a critical test between the MTS hypothesis and the dual-process model.

The MTS hypothesis predicts that compromisers will be rational: They will make few, if any, GARP violations.[¶] The dual-process model does not predict that compromisers will exist, let alone be rational. It can attribute their existence to hand trembles, of course. But compromises made by mistake create many GARP violations, as we show below.

Compromisers Made Rational Judgments. Most compromisers never violated GARP: 59% in BU, 64% in CW, and 58% in SW. In every condition, ∼90% of compromisers made seven or fewer violations. By contrast, ∼0% of incentives + chance agents made zero violations, and less than 10% made seven or fewer violations.

The compromisers even outperformed the rational + 1 tremble agents. In all frames, the percent of compromisers with zero violations was higher by at least 20 points. The percent of compromisers who made seven or fewer violations was also higher: ∼90% of them compared to 77% of rational + 1 tremble agents. See Table 3 and *SI Appendix*, Tables S2 and S3.

Of the 1,237 compromisers, 85% made zero violations in at least one frame; 81%, if we exclude cases in which the same extreme option was chosen 21 times (49). Even if their hand trembled in one condition, their flawlessly rational performance in the other evidences a cognitive system that maximizes rightness.

The rationality of compromisers confirms prediction 4.

A fine-grained analysis of performance as a function of difficulty supports the optimization hypothesis even more strongly. How difficult it is to respect GARP varies with the number of compromises an agent makes. The median number of violations made by random responders is an indicator of difficulty. As Fig. 5*A* shows, difficulty depends on the number of compromises made. It is high for all numbers of compromises, and has an inverted U shape.

Difficulty thus defined should not matter for an MTS. When the input to MAX is a well-behaved rightness function, its optimization algorithm will identify the most right solution in a scenario. That process does not depend on how many compromises MAX produced in previous scenarios. Respecting GARP is a cost-free byproduct of optimization. Effortful, conscious reasoning is unnecessary to remain consistent across scenarios.

Difficulty would matter, however, if compromise judgments were produced by nonoptimizing algorithms: heuristics, inflexible emotions, or deliberative reasoning. Respecting GARP is not a byproduct of their design, so avoiding violations would require a backward-looking algorithm. Each new judgment would have to be made factoring in all the previous ones. It follows that GARP violations should increase with difficulty. A heuristic that responds to incentives without looking backward is a case in point: GARP violations increased with difficulty for the incentives + chance agents.

As Fig. 5*A* and *SI Appendix*, Fig. S3 show, compromisers made highly rational judgments regardless of the number of compromises they made: Their median number of violations was zero in 87% of cases, and two at most. Remarkably, compromisers outperformed the rational + 1 tremble agents, the most exacting

[§]Assigning zero weight is only one of many ways the MTS can produce an extreme profile.

[¶]An MTS can produce compromises by maximizing a rightness function with strictly convex indifference curves, for example; see *SI Appendix*.

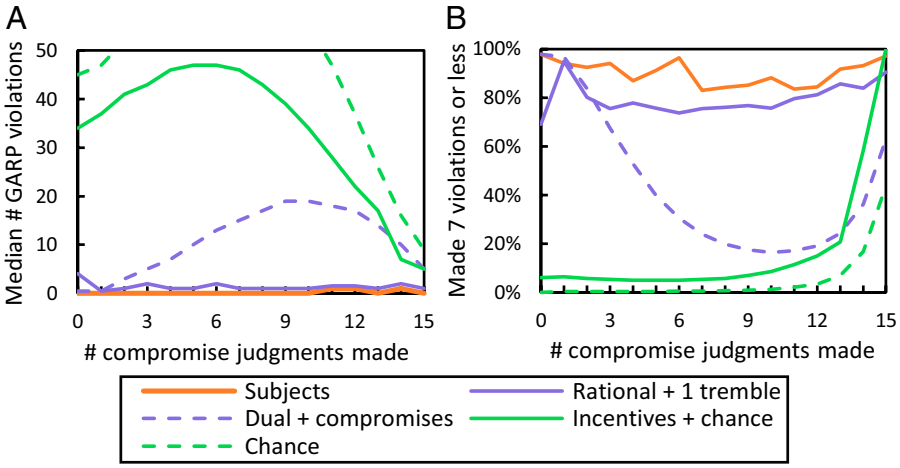

Fig. 5. Subjects made rational responses, no matter how many compromise judgments they made (BU). (*A*) The median number of GARP violations as a function of number of compromises made; (*B*) percent making few violations (seven or fewer). Regardless of difficulty (the curve for chance), the median number of GARP violations by compromisers was approximately zero, and the percent making few violations was high.

benchmark. Respecting GARP across all levels of difficulty is a signature of optimization.

A Dual Process Cannot Create Rational Compromisers. The dual-process model does not specify how the two processes would respond to incentives, so it makes no specific predictions about GARP violations. We can, however, bracket the violations expected between two bounds. Random flip-flopping sets an upper bound: Since S/C and S vary unpredictably across scenarios, an alarm bell could go off unpredictably in response. This would create a median of 45 violations, not the median of 0 found for subjects. A lower bound is set by extreme judgments that are rational, with compromises as hand trembles.

To quantify this lower bound, we created a *dual* + k *compromises* benchmark: simulated agents created by resampling the extreme responders and GARP-respecting flip-floppers in a given frame. For each agent, we randomly created k compromises (range: 0 to 15). For comparability, we made the distribution of compromises for the bootstrapped samples the same as for subjects. These distributions are depicted in Fig. 2*B* and *SI Appendix*, Fig. S1. We created 2 million agents for BU, 1 million for CW, and 1 million for SW.

The performance profile for dual + k compromise agents has an inverted U shape: They make more violations when the task is more difficult (Fig. 5*A* and *SI Appendix*, Fig. S3). This shape is the fingerprint of systems that do not optimize. It contrasts with the performance profile for subjects as a function of k, which is flat at zero violations.

We also examined the percent of subjects and dual + k compromise agents who made seven or fewer violations as a function of k. The results are presented in Fig. 5*B* and *SI Appendix*, Fig. S3. At the highest level of difficulty ($k = 6, 7$; median = 63 violations), over 85% of subjects made seven or fewer violations (BU: 90%; CW: 92%; SW: 86%), whereas a minority of dual + k compromise agents did (BU: 27%; CW:21%; SW: 37%). Subjects were far more rational than expected from a dual process + mistakes model.

The prevalence of rational compromisers—ones who respected GARP even when it was most difficult to do so—is strong evidence of a cognitive system that weighs different moral values and chooses a most right solution (feature 4).

Moral Judgments Were Intuitive. Deliberative reasoning cannot explain GARP-respecting compromisers. It would require constructing and maintaining a consistent preference order that expands as each new judgment is made. In this experiment, your working memory would have to hold a growing record of up to 435 revealed preference relations. This is because the 21 scenarios include 29 different solutions, and each solution can be in a preference relationship with itself and all other solutions. To avoid violating GARP, you would have to check each feasible solution in a scenario for possible violations, choose one that is consistent with all previously revealed preferences, and then update the record accordingly. The cognitive load of this task makes it intractable. The only way to ensure no GARP violations through deliberative reasoning is to make the same extreme judgment 21 times.[#]

An intuitive optimization process can produce rational compromisers; System 2 reasoning cannot.

Rightness Functions Are Temporary Representations. Judgments that respect GARP in both frames, yet vary across frames, are evidence that rightness functions are constructed on the fly. To test this, we examined the 78% of subjects whose judgments varied with willingness. These subjects were rational within each frame: Most never violated GARP (BU: 63%; CW: 72%; SW: 71%), and 92% made seven or fewer violations. Indeed, most of them were rational in both frames (zero violations: 49%; seven or fewer: 86%), confirming prediction 5. This indicates that their MTS had constructed, and maximized, different rightness functions in each willingness frame. The initial function persisted while subjects made (rational) judgments in the first frame they faced, but it was replaced by a different function when they faced the second frame.

Conclusion

The results of the war dilemma revealed a previously unknown cognitive competence: a moral tradeoff system. It is composed of three subsystems: MV, FS, and MAX. MV integrates conflicting values into a rightness function: a temporary representation that maps each conceivable solution onto a level of rightness. In parallel, FS constructs a feasible set: a temporary

[#]Economists participating in GARP studies know this. Harbaugh et al. (49) could not use economists as a control group, because many were more concerned with appearing rational than with choosing the options that they preferred. They avoided intermediate options entirely, confessing afterward that they chose the same extreme option every time to prevent "embarrassing" GARP violations.

representation of the subset of conceivable solutions that are perceived as available. MAX uses these representations to identify a feasible solution with the maximum level of rightness, delivering it as an intuitive judgment.

How do we know that the MTS works like this?

The results support every prediction this model makes, including unique ones. But the most decisive evidence is that the vast majority of subjects were rational in the sense of GARP: They made few GARP violations. They were rational no matter what mix of extreme and compromise judgments they made, and remained rational across frames—even when their judgments changed with the willingness of soldiers or civilians. This is the signature of a process that maximizes rightness.

We considered five alternative hypotheses; none could explain the results—especially rational judgments by compromisers. Responding randomly produces many GARP violations; so does responding to incentives without maximizing rightness. Three inflexible rules—always choose deontic, always choose utilitarian, and flip-flop at a threshold—can produce extreme judgments that respect GARP. But no inflexible rule can produce GARP-respecting compromises. This includes the inflexible command, DO NOT HARM, issued by System 1 of Greene's dual-process model (14). Yet more than 70% of subjects made compromises, and when they did, over 90% made rational judgments (zero to seven violations; the maximum possible number is 152).

The dual-process model proposed by Greene and colleagues (14, 15) makes several unique predictions, all contradicted by the data.

A straightforward version of their model predicts that the war dilemma will always elicit extreme judgments, because its conflicting values cannot be "negotiated by weighing preferences against each other" (15). This rules out compromise judgments, because they result from weighing moral preferences: Compromises strike a balance between conflicting values by partially satisfying both. It follows that compromise judgments will be infrequent and unsystematic, because they are noise. The fact that compromises were both frequent and rational contradicts these predictions. Indeed, compromise judgments that respect GARP are evidence of a rightness function that assigns positive weights to conflicting values (in this case, the lives of civilians and soldiers). The MTS "negotiates" the conflicting values by assigning a rightness level to each conceivable solution, including the intermediate ones.

A version of the dual-process model that jettisons the nonnegotiability hypothesis is refuted as well. In that model, flexibility—the ability to respond to context by integrating multiple considerations—requires conscious, deliberative reasoning. This implies that only System 2 could produce compromise judgments. But this version of the model also fails to explain the prevalence of rational compromisers. If intermediate solutions are chosen, deliberative reasoning cannot prevent GARP violations across 21 scenarios with intersecting feasibility constraints (49). The cognitive load of the task makes it impossible, not only for Greene's model, but for any model that seeks to attribute the results to deliberative reasoning.

An original feature of the MTS hypothesis is that rightness functions and feasible sets are temporary mental representations, constructed on the fly for a specific dilemma. We know that the MTS can construct new rightness functions on the fly because most subjects changed their judgments when the willingness frame was different but the scenarios were the same. We know that the MTS can construct new feasible sets on the fly because, within a frame, most subjects responded to different scenarios by changing their judgments. This flexibility necessitates three subsystems: two subsystems that construct the temporary representations, and a third subsystem that uses them to identify an optimal solution.

These findings open many questions. How does MV integrate conflicting values? Does it have a library of functional forms to draw on, with free parameters calculated on the fly (50)? Does FS represent the available options as a set, or in summary form—analogous to the "budget lines" in Fig. 1*D*?

Finally, the three-subsystem architecture proposed here may provide a useful template for psychologists and social scientists studying choice in domains outside moral psychology. An architecture like this could be present in many other cognitive systems, each specialized for a different domain of choice.

Materials and Methods

Study procedures were approved by the University of California, Santa Barbara Institutional Review Board. Participants gave fully informed consent before answering the survey. The survey was implemented in Qualtrics. US adults were recruited through Amazon MTurk in 2016. Subjects were paid $1. A session lasted ∼15 min. Criteria for inclusion were set in advance. The dataset for analysis was all subjects who completed the survey and correctly answered two attention checks and an English language comprehension question. $N = 1,745$ met these criteria: 54% female; mean age = 36 y (SD 12 y), range 18 y to 87 y. Subjects were randomly assigned to the different treatments. See *SI Appendix* for full text of the instrument. Data are available at Open Science Framework (OSF; https://osf.io/kd34j/).

Data, Materials, and Software Availability. Excel spreadsheet and codes have been deposited in OSF (https://osf.io/kd34j/) (51).

ACKNOWLEDGMENTS. We thank Ryan Oprea, Carlos Rodríguez-Sickert, Víctor Lima, John Tooby, and two anonymous reviewers for their advice and encouragement. Research was funded by John Templeton Foundation Grant 29468 (L.C.).

Author affiliations: [a]Centro de Investigación en Complejidad Social, Facultad de Gobierno, Universidad del Desarrollo, Santiago 7610658, Chile; [b]Department of Psychology, University of Montreal, Montreal, QC, Canada H3C 3J7; [c]Oklahoma Center for Evolutionary Analysis, Department of Psychology, Oklahoma State University, Stillwater, OK 74078-3064; and [d]Center for Evolutionary Psychology, Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA 93106-9660

1. A. P. Fiske, *Structures of Social Life: The Four Elementary Forms of Human Relations* (Free, 1991).
2. D. B. Bugental, Acquisition of the algorithms of social life: A domain-based approach. *Psychol. Bull.* **126**, 187–219 (2000).
3. J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (Pantheon, 2012).
4. L. Cosmides, R. A. Guzmán, J. Tooby, "The evolution of moral cognition" in *The Routledge Handbook of Moral Epistemology*, A. Zimmerman, K. Jones, M. Timmons, Eds. (Routledge, New York, 2019), pp. 174–228.
5. D. Lieberman, J. Tooby, L. Cosmides, The architecture of human kin detection. *Nature* **445**, 727–731 (2007).
6. L. Cosmides, J. Tooby, "Can a general deontic logic capture the facts of human moral reasoning?" in *Moral Psychology*, W. Sinnott-Armstrong, C. B. Miller, Eds. (The MIT Press, 2008), vol. 1, pp. 53–119.
7. L. Cosmides, H. C. Barrett, J. Tooby, Adaptive specializations, social exchange, and the evolution of human intelligence. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9007–9014 (2010).
8. N. Baumard, *The Origins of Fairness: How Evolution Explains Our Moral Nature* (Oxford University Press, New York, 2016).
9. H. Kaplan *et al.*, Food sharing among ache foragers: Tests of explanatory hypotheses. *Curr. Anthropol.* **26**, 223–246 (1985).
10. H. S. Kaplan, E. Schniter, V. L. Smith, B. J. Wilson, Risk and the evolution of human exchange. *Proc. Biol. Sci.* **279**, 2930–2935 (2012).
11. L. Aarøe, M. B. Petersen, Crowding out culture: Scandinavians and Americans agree on social welfare in the face of deservingness cues. *J. Polit.* **76**, 684–697 (2014).
12. A. W. Delton, L. Cosmides, M. Guemo, T. E. Robertson, J. Tooby, The psychosemantics of free riding: Dissecting the architecture of a moral concept. *J. Pers. Soc. Psychol.* **102**, 1252–1270 (2012).
13. J. Tooby, L. Cosmides, "Groups in mind: The coalitional roots of war and morality" in *Human Morality and Sociality: Evolutionary and Comparative Perspectives*, H. Høgh-Olesen, Ed. (Palgrave Macmillan, 2010), pp. 91–234.
14. J. Greene, "The secret joke of Kant's soul" in *Moral Psychology*, W. Sinnott-Armstrong, Ed. (MIT Press, 2008), vol. 3, pp. 35–80.
15. F. Cushman, J. D. Greene, Finding faults: How moral dilemmas illuminate cognitive structure. *Soc. Neurosci.* **7**, 269–279 (2012).

16. R. L. Kelly, *The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways* (Smithsonian Institution Press, 1995).
17. H. Kaplan, K. Hill, J. Lancaster, A. M. Hurtado, A theory of human life history evolution: Diet, intelligence, and longevity. *Evol. Anthropol.* **9**, 156–185 (2000).
18. A. V. Jaeggi, M. Gurven, Natural cooperators: Food sharing in humans and other primates. *Evol. Anthropol.* **22**, 186–195 (2013).
19. L. H. Keeley, *War before Civilization* (Oxford University Press, 1997).
20. R. Wrangham, *The Goodness Paradox: The Strange Relationship between Virtue and Violence in Human Evolution* (Pantheon, 2019).
21. M. B. Petersen, Social welfare as small-scale help: Evolutionary psychology and the deservingness heuristic. *Am. J. Pol. Sci.* **56**, 1–16 (2012).
22. D. Sznycer *et al.*, Shame closely tracks the threat of devaluation by others, even across cultures. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2625–2630 (2016).
23. S. Tassy, O. Oullier, J. Mancini, B. Wicker, Discrepancies between judgment and choice of action in moral dilemmas. *Front. Psychol.* **4**, 250 (2013).
24. D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, 1982).
25. J. Tooby, L. Cosmides, A. Sell, D. Lieberman, D. Sznycer, "Internal regulatory variables and the design of human motivation: A computational and evolutionary approach" in *Handbook of Approach and Avoidance Motivation*, A. J. Elliot, Ed. (Psychology, 2008), pp. 251–271.
26. H. R. Varian, The nonparametric approach to demand analysis. *Econometrica* **50**, 945–973 (1982).
27. J. C. Cox, On testing the utility hypothesis. *Econ. J. (Lond.)* **107**, 1054–1078 (1997).
28. D. Lieberman, C. Patrick, *Objection: Disgust, Morality, and the Law* (Oxford University Press, 2018).
29. G. Kahane, On the wrong track: Process and content in moral psychology. *Mind Lang.* **27**, 519–545 (2012).
30. G. Kahane, "Intuitive and counterintuitive morality" in *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics*, J. D'Arms, D. Jacobson, Eds. (Oxford University Press, New York, 2014), pp. 9–39.
31. S. Nichols, R. Mallon, Moral dilemmas and moral rules. *Cognition* **100**, 530–542 (2006).
32. M. J. Crockett, Models of morality. *Trends Cogn. Sci.* **17**, 363–366 (2013).
33. C. Helion, K. N. Ochsner, The role of emotion regulation in moral judgment. *Neuroethics* **11**, 297–308 (2016).
34. C. Hu, X. Jiang, An emotion regulation role of ventromedial prefrontal cortex in moral judgment. *Front. Hum. Neurosci.* **8**, 873 (2014).
35. J. Mikhail, Universal moral grammar: Theory, evidence and the future. *Trends Cogn. Sci.* **11**, 143–152 (2007).
36. J. May, V. Kumar, "Moral reasoning and emotion" in *The Routledge Handbook of Moral Epistemology*, A. Zimmerman, K. Jones, M. Timmons, Eds. (Routledge, New York, 2019), pp. 139–156.
37. M. Białek, W. DeNeys, Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgm. Decis. Mak.* **12**, 148–167 (2017).
38. S. Bretz, R. Sun, Two models of moral judgment. *Cogn. Sci. (Hauppauge)* **42**, 4–37 (2018).
39. J. Doris, S. Stich, J. Phillips, L. Walmsley, "Moral psychology: Empirical approaches" in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta., Ed. (Spring edition, 2020). https://plato.stanford.edu/archives/spr2020/entries/moral-psych-emp/. Accessed 1 July 2021.
40. J. D. Greene *et al.*, Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition* **111**, 364–371 (2009).
41. J. F. Christensen, A. Flexas, M. Calabrese, N. K. Gut, A. Gomila, Moral judgment reloaded: A moral dilemma validation study. *Front. Psychol.* **5**, 607 (2014).
42. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York, 2011).
43. A. Shenhav, J. D. Greene, Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex. *J. Neurosci.* **34**, 4741–4749 (2014).
44. S. French, *The Code of the Warrior: Exploring Warrior Values Past and Present* (Rowman & Littlefield, Lanham, 2003).
45. A. P. Fiske, T. S. Rai, *Virtuous Violence: Hurting and Killing to Create, Sustain, End, and Honor Social Relationships* (Cambridge University Press, New York, 2014).
46. H. M. Watkins, The morality of war: A review and research agenda. *Perspect. Psychol. Sci.* **15**, 231–249 (2020).
47. M. J. Lewis Petrinovich, Patricia O'Neill, An empirical study of moral intuitions: Toward an evolutionary ethics. *J. Pers. Soc. Psychol.* **64**, 467–478 (1993).
48. M. A. Diaye, M. W. Urdanivia, Violation of the transitivity axiom may explain why, in empirical studies, a significant number of subjects violate GARP. *J. Math. Psychol.* **53**, 586–592 (2009).
49. W. T. Harbaugh, K. Krause, T. R. Berry, GARP for kids: On the development of rational choice behavior. *Am. Econ. Rev.* **91**, 1539–1545 (2001).
50. J. Andreoni, J. Miller, Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* **70**, 737–753 (2002).
51. R. A. Guzmán, M. T. Barbato, D. Sznycer, L. Cosmides. Moral Tradeoff System, Open Science Framework. OSF. https://osf.io/kd34j/. Deposited 5 September 2022.

Supplementary information for

A moral tradeoff system produces intuitive judgments that are rational, coherent, and strike a balance between conflicting moral values

Ricardo Andrés Guzmán[a,*], María Teresa Barbato[a],
Daniel Sznycer[b], and Leda Cosmides[c,*]

[a]Centro de Investigación en Complejidad Social, Universidad del Desarrollo, Santiago.
[b]Department of Psychology, University of Montreal and Oklahoma Center for Evolutionary Analysis, Department of Psychology, Oklahoma State University, Stillwater, OK 74078-3064.
[c]Center for Evolutionary Psychology, University of California, Santa Barbara.
[*]Corresponding authors. E-mail: rguzman@udd.cl, cosmides@ucsb.edu.

September 2, 2022

Contents

1 Additional figures and tables

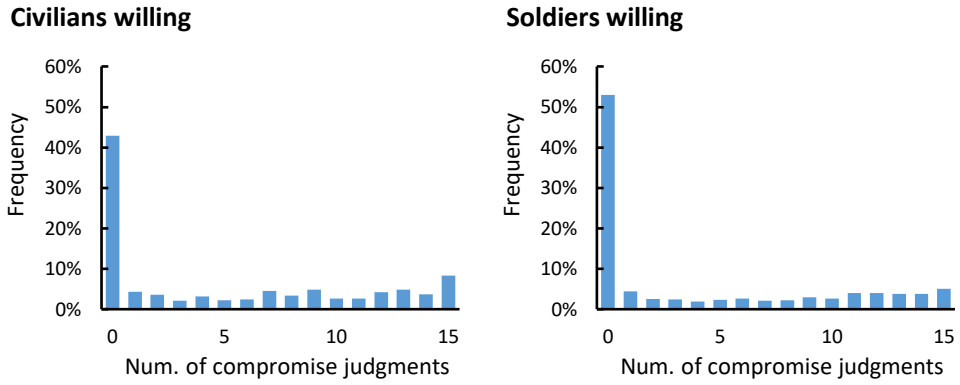

Figure S1. Frequency of compromise judgments made in the civilians willing and soldiers willing conditions.

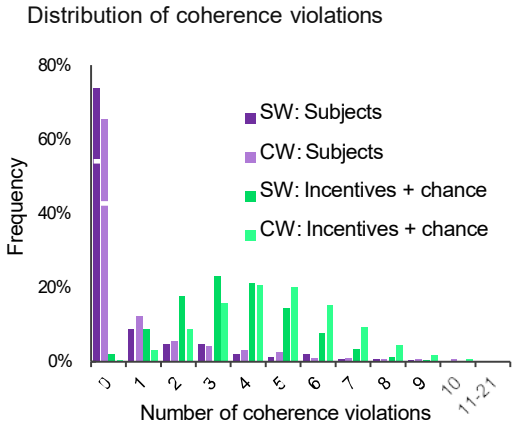

Figure S2. Number of coherence violations made by subjects and simulated agents. Most subjects responded coherently to changes in *willingness*. (Sections above the white segments represent subjects who always made the same extreme judgment in both conditions.

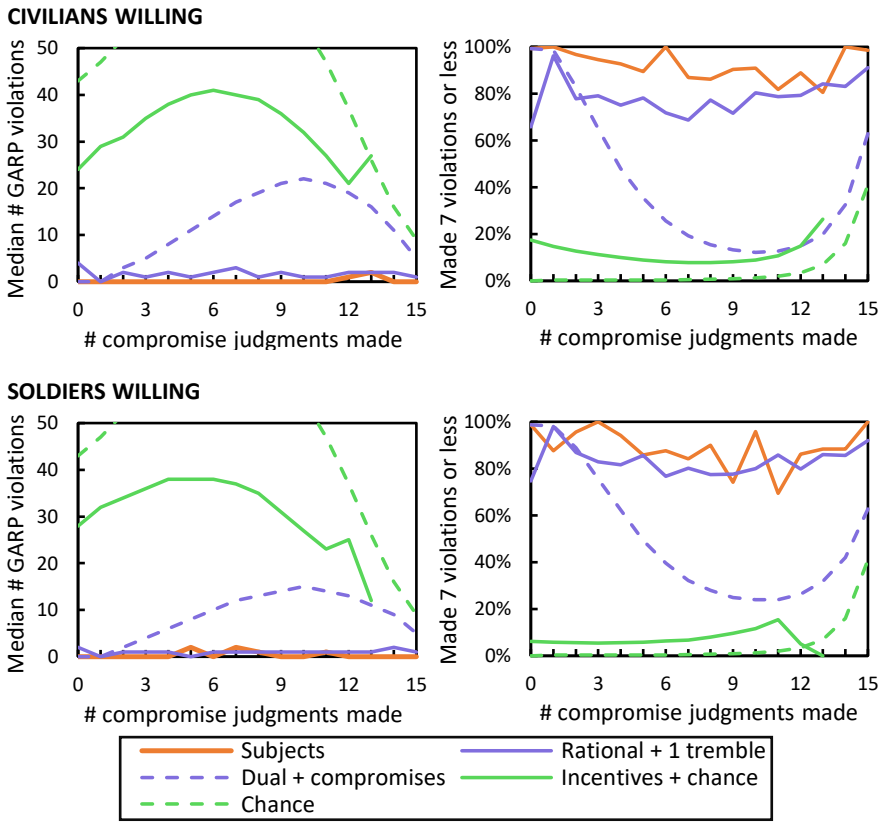

Figure S3. Performance of subjects and simulated agents in the war dilemma.

Table S1. Scenarios of the war dilemma

Scenario #	Soldiers at risk (S)	Soldiers saved for each civilian sacrificed (S / C)	Alternatives (civil. sacrificed, sold. dead)	Feasible set (civil. spared, sold. saved)	Pareto inferior solutions (civil. spared, sold. saved)
1	2	2.00	(0, 2) (1, 0)	(1, 0) (0, 2)	(0, 0) (0, 1)
2	3	3.00	(0, 3) (1, 0)	(1, 0) (0, 3)	(0, 0) (0, 1) (0, 2)
3	3	1.50	(0, 3) (1, 2) (2, 0)	(2, 0) (1, 1) (0, 3)	(0, 0) (1, 0) (0, 1) (0, 2)
4	4	4.00	(0, 4) (1, 0)	(1, 0) (0, 4)	(0, 0) (0, 1) (0, 2) (0, 3)
5	4	2.00	(0, 4) (1, 2) (2, 0)	(2, 0) (1, 2) (0, 4)	(0, 0) (1, 0) (0, 1) (1, 1) (0, 2) (0, 3)
6	4	1.33	(0, 4) (1, 3) (2, 2) (3, 0)	(3, 0) (2, 1) (1, 2) (0, 4)	(0, 0) (1, 0) (2, 0) (0, 1) (1, 1) (0, 2) (0, 3)
7	5	5.00	(0, 5) (1, 0)	(1, 0) (0, 5)	(0, 0) (0, 1) (0, 2) (0, 3) (0, 4)
8	5	2.50	(0, 5) (1, 3) (2, 0)	(2, 0) (1, 2) (0, 5)	(0, 0) (1, 0) (0, 1) (1, 1) (0, 2) (0, 3) (0, 4)
9	5	1.67	(0, 5) (1, 4) (2, 2) (3, 0)	(3, 0) (2, 1) (1, 3) (0, 5)	(0, 0) (1, 0) (2, 0) (0, 1) (1, 1) (0, 2) (1, 2) (0, 3) (0, 4)
10	5	1.25	(0, 5) (1, 4) (2, 3) (3, 2) (4, 0)	(4, 0) (3, 1) (2, 2) (1, 3) (0, 5)	(0, 0) (1, 0) (2, 0) (3, 0) (0, 1) (1, 1) (2, 1) (0, 2) (1, 2) (0, 3) (0, 4)
11	6	6.00	(0, 6) (1, 0)	(1, 0) (0, 6)	(0, 0) (0, 1) (0, 2) (0, 3) (0, 4) (0, 5)
12	6	3.00	(0, 6) (1, 3) (2,0)	(2, 0) (1, 3) (0, 6)	(0, 0) (1, 0) (0, 1) (1, 1) (0, 2) (1, 2) (0, 3) (0, 4) (0, 5)
13	6	2.00	(0, 6) (1, 4) (2,2) (3,0)	(3, 0) (2, 2) (1, 4) (0, 6)	(0, 0) (1, 0) (2, 0) (0, 1) (1, 1) (2, 1) (0, 2) (1, 2) (0, 3) (1, 3) (0, 4) (0, 5)
14	6	1.50	(0, 6) (1, 5) (2,3) (3,2) (4,0)	(4, 0) (3, 1) (2, 3) (1, 4) (0, 6)	(0, 0) (1, 0) (2, 0) (3, 0) (0, 1) (1, 1) (2, 1) (0, 2) (1, 2) (2, 2) (0, 3) (1, 3) (0, 4) (0, 5)
15	6	1.20	(0, 6) (1, 5) (2, 4) (3, 3) (4, 2) (5, 0)	(5, 0) (4, 1) (3, 2) (2, 3) (1, 4) (0, 6)	(0, 0) (1, 0) (2, 0) (3, 0) (4, 0) (0, 1) (1, 1) (2, 1) (3, 1) (0, 2) (1, 2) (2, 2) (0, 3) (1, 3) (0, 4) (0, 5)
16	7	7.00	(0, 7) (1, 0)	(1, 0) (0, 7)	(0, 0) (0, 1) (0, 2) (0, 3) (0, 4) (0, 5) (0, 6)
17	7	3.50	(0, 7) (1, 4) (2, 0)	(2, 0) (1, 3) (0, 7)	(0, 0) (1, 0) (0, 1) (1, 1) (0, 2) (1, 2) (0, 3) (0, 4) (0, 5) (0, 6)
18	7	2.33	(0, 7) (1, 5) (2, 3) (3, 0)	(3, 0) (2, 2) (1, 4) (0, 7)	(0, 0) (1, 0) (2, 0) (0, 1) (1, 1) (2, 1) (0, 2) (1, 2) (0, 3) (1, 3) (0, 4) (0, 5) (0, 6)
19	7	1.75	(0, 7) (1, 6) (2, 4) (3, 2) (4, 0)	(4, 0) (3, 1) (2, 3) (1, 5) (0, 7)	(0, 0) (1, 0) (2, 0) (3, 0) (0, 1) (1, 1) (2, 1) (0, 2) (1, 2) (2, 2) (0, 3) (1, 3) (0, 4) (1, 4) (0, 5) (0, 6)
20	7	1.40	(0, 7) (1, 6) (2, 5) (3, 3) (4, 2) (5, 0)	(5, 0) (4, 1) (3, 2) (2, 4) (1, 5) (0, 7)	(0, 0) (1, 0) (2, 0) (3, 0) (4, 0) (0, 1) (1, 1) (2, 1) (3, 1) (0, 2) (1, 2) (2, 2) (0, 3) (1, 3) (2, 3) (0, 4) (1, 4) (0, 5) (0, 6)
21	7	1.17	(0, 7) (1, 6) (2, 5) (3, 4) (4, 3) (5, 2) (6, 0)	(6, 0) (5, 1) (4, 2) (3, 3) (2, 4) (1, 5) (0, 7)	(0, 0) (1, 0) (2, 0) (3, 0) (4, 0) (5, 0) (0, 1) (1, 1) (2, 1) (3, 1) (4, 1) (0, 2) (1, 2) (2, 2) (3, 2) (0, 3) (1, 3) (2, 3) (0, 4) (1, 4) (0, 5) (0, 6)

Note: All quantities in millions of lives. In a scenario, each alternative presented to subjects corresponds to a solution in the feasible set. Each solution is an ordered pair (c, s), where c is the number of civilians spared and s is the number of soldiers saved. The corresponding alternative is the ordered pair ($C - c$, $S - s$), where $C - c$ is the number of civilians sacrificed and $S - s$ is the number of soldiers dead. The "Pareto inferior solutions" are inferior to *at least one* feasible solution. There are only 29 different solutions, and almost all of them appear in more than one scenario, either as a feasible solution or a Pareto inferior solution.

Table S2. Rationality in the civilians willing condition.

Individuals	Never violated GARP	Made 7 viols. or less	Median num. of violations	90th Percentile	95th Percentile
All subjects	78%	95%	0	3	7
Compromisers	64%	92%	0	6	15
Flip-floppers	67%	93%	0	6	15
Benchmarks					
Random responders	~0%	1%	61	90	97
Incentives + chance	1%	9%	38	74	82
Rational + 1 tremble	39%	73%	2	16	24
Dual + compromises	47%	65%	2	27	35

Table S3. Rationality in the soldiers willing condition

Individuals	Never violated GARP	Made 7 viols. or less	Median num. of violations	90th Percentile	95th Percentile
All subjects	77%	94%	0	3	8
Compromisers	58%	88%	0	8	14
Flip-floppers	54%	94%	0	6	13
Benchmarks					
Random responders	~0%	1%	61	90	97
Incentives + chance	~0%	6%	36	65	73
Rational + 1 tremble	36%	82%	1	11	16
Dual + compromises	54%	75%	0	19	27

2 Written description of the war dilemma

Here we will present the written description of the dilemma, as presented to subjects. Yellow highlight indicates phrases that varied between the three conditions and four treatments (a treatment consists of the baseline condition and one of two variants, shown to subjects in that order or in reverse order).

2.1 Both unwilling followed by civilians willing

Two foreign countries, A and B, have been at war for several years (you are not a citizen of either of these two countries). The war was initiated by the rulers of country B, against the will of the civilian population. The war has been bloody: Millions have died during the conflict, which so far had been deadlocked. Recently, the military equilibrium has broken, and it is now certain that Country A will win the war sooner or later. The question is how, when, and at what cost. Country A has two strategies available. Country A could use one, the other, or a combination of both. The first strategy is to attack the opposing army with conventional weapons, preventing civilian casualties almost completely. If Country A applies this strategy, the war will continue for some time (perhaps years). The delay in the end of the war will cause the deaths of a great number of soldiers of both sides. Of the soldiers that die, about half will be from country A and half from country B. Nearly all are young soldiers who were forced to join the army against their will, and are desperate to return to their families. The second strategy available to Country A is to bomb cities of Country B, killing civilians (who opposed the war from the beginning) and almost no soldiers. This strategy would demoralize Country B and force it to surrender quickly. The war would end soon. There is a third approach. Country A could bring the war to an end by using both strategies, resulting in the deaths of some civilians and some soldiers. The more civilians are sacrificed (killed) during the bombings, the sooner Country B will surrender, and the fewer soldiers will die on the battlefield. How should Country A end the war?

[Before answering the scenarios of the next condition(civilians willing), the subjects saw the following message.]

Now imagine a war in which everything is the same as before, except for the original attitude of the civilians of the aggressor country toward the war. The situation is as follows: The war was initiated by the rulers of country B, with the support

of the civilian population. As in the previous case, the armies of both countries are made up of young soldiers who were forced to join the army against their will, and are desperate to return to their families. Remember that Country A could bring the war to an end in three ways:

1. By attacking the opposing army with conventional weapons, in which case many soldiers would die, but no civilians would die.

2. By bombing cities, in which case many civilians would die, but no soldiers.

3. By a combination of both strategies, in which case some soldiers and some civilians would die.

Also remember that of the soldiers that die, about half will be from country A and half from country B. How should Country A end the war?

2.2 Both unwilling followed by soldiers willing

Two foreign countries, A and B, have been at war for several years (you are not a citizen of either of these two countries). The war was initiated by the rulers of country B, against the will of the civilian population. The war has been bloody: Millions have died during the conflict, which so far had been deadlocked. Recently, the military equilibrium has broken, and it is now certain that Country A will win the war sooner or later. The question is how, when, and at what cost. Country A has two strategies available. Country A could use one, the other, or a combination of both. The first strategy is to attack the opposing army with conventional weapons, preventing civilian casualties almost completely. If Country A applies this strategy, the war will continue for some time (perhaps years). The delay in the end of the war will cause the deaths of a great number of soldiers of both sides. Of the soldiers that die, about half will be from country A and half from country B. Nearly all are young soldiers who were forced to join the army against their will, and are desperate to return to their families. The second strategy available to Country A is to bomb cities of Country B, killing civilians (who opposed the war from the beginning) and almost no soldiers. This strategy would demoralize Country B and force it to surrender quickly. The war would end soon. There is a third approach. Country A could bring the war to an end by using both strategies, resulting in the deaths of some civilians and some soldiers. The more civilians are sacrificed (killed) during the bombings, the sooner Country B

will surrender, and the fewer soldiers will die on the battlefield. How should Country A end the war?

[Before answering the scenarios of the next condition (soldiers willing), the subjects saw the following message.]

Now imagine a war in which everything is the same as before, except for the motivation of the soldiers. The situation is as follows: The armies of both countries are made up of young soldiers who volunteered, and are willing to fight for their country. As in the previous case, the civilian population of country B, the aggressor, did not support the war. Remember that Country A could bring the war to an end in three ways:

1. By attacking the opposing army with conventional weapons, in which case many soldiers would die, but no civilians would die.

2. By bombing cities, in which case many civilians would die, but no soldiers.

3. By a combination of both strategies, in which case some soldiers and some civilians would die.

Also remember that of the soldiers that die, about half will be from country A and half from country B. How should Country A end the war?

2.3 Civilians willing followed by both unwilling

Two foreign countries, A and B, have been at war for several years (you are not a citizen of either of these two countries). The war was initiated by the rulers of country B, with the support of the civilian population. The war has been bloody: Millions have died during the conflict, which so far had been deadlocked. Recently, the military equilibrium has broken, and it is now certain that Country A will win the war sooner or later. The question is how, when, and at what cost. Country A has two strategies available. Country A could use one, the other, or a combination of both. The first strategy is to attack the opposing army with conventional weapons, preventing civilian casualties almost completely. If Country A applies this strategy, the war will continue for some time (perhaps years). The delay in the end of the war will cause the deaths of a great number of soldiers of both sides. Of the soldiers that die, about half will be from country A and half from country B. Nearly all are young soldiers who were forced to join the army against their will, and are desperate

to return to their families. The second strategy available to Country A is to bomb cities of Country B, killing civilians (who supported the war from the beginning) and almost no soldiers. This strategy would demoralize Country B and force it to surrender quickly. The war would end soon. There is a third approach. Country A could bring the war to an end by using both strategies, resulting in the deaths of some civilians and some soldiers. The more civilians are sacrificed (killed) during the bombings, the sooner Country B will surrender, and the fewer soldiers will die on the battlefield. How should Country A end the war?

[Before answering the scenarios of the next condition (both unwilling), the subjects saw the following message.]

Now imagine a war in which everything is the same as before, except for the original attitude of the civilians of Country B (the aggressor) toward the war. The situation is as follows: The war was initiated by the rulers of country B, against the will of the civilian population. As in the previous case, the armies of both countries are made up of young soldiers who were forced to join the army against their will, and are desperate to return to their families. Remember that Country A could bring the war to an end in three ways:

1. By attacking the opposing army with conventional weapons, in which case many soldiers would die, but no civilians would die.

2. By bombing cities, in which case many civilians would die, but no soldiers.

3. By a combination of both strategies, in which case some soldiers and some civilians would die.

Also remember that of the soldiers that die, about half will be from country A and half from country B. How should Country A end the war?

2.4 Soldiers willing followed by both unwilling

Two foreign countries, A and B, have been at war for several years (you are not a citizen of either of these two countries). The war was initiated by the rulers of country B, against the will of the civilian population. The war has been bloody: Millions have died during the conflict, which so far had been deadlocked. Recently, the military equilibrium has broken, and it is now certain that Country A will win the war sooner or later. The question is how, when, and at what cost. Country A has two strategies

available. Country A could use one, the other, or a combination of both. The first strategy is to attack the opposing army with conventional weapons, preventing civilian casualties almost completely. If Country A applies this strategy, the war will continue for some time (perhaps years). The delay in the end of the war will cause the deaths of a great number of soldiers of both sides. Of the soldiers that die, about half will be from country A and half from country B. Nearly all are young soldiers who volunteered, and are willing to fight for their country. The second strategy available to Country A is to bomb cities of Country B, killing civilians (who opposed the war from the beginning) and almost no soldiers. This strategy would demoralize Country B and force it to surrender quickly. The war would end soon. There is a third approach. Country A could bring the war to an end by using both strategies, resulting in the deaths of some civilians and some soldiers. The more civilians are sacrificed (killed) during the bombings, the sooner Country B will surrender, and the fewer soldiers will die on the battlefield. How should Country A end the war?

[Before answering the scenarios of the next condition (both unwilling), the subjects saw the following message.]

Now imagine a war in which everything is the same as before, except for the motivation of the soldiers. The situation is as follows: The armies of both countries are made up of young soldiers who were forced to join the army against their will, and are desperate to return to their families. As in the previous case, the civilian population of country B, the aggressor, did not support the war. Remember that Country A could bring the war to an end in three ways:

1. By attacking the opposing army with conventional weapons, in which case many soldiers would die, but no civilians would die.

2. By bombing cities, in which case many civilians would die, but no soldiers.

3. By a combination of both strategies, in which case some soldiers and some civilians would die.

Also remember that of the soldiers that die, about half will be from country A and half from country B. How should Country A end the war?

3 Scenarios of the war dilemma

In each condition, subjects were given 21 scenarios of the war dilemma, summarized in table S1. A scenario is a multiple-choice question that offers a minimum of two and a maximum of seven alternatives for ending a war. An alternative consists of x civilians sacrificed and y soldiers killed. The number of civilians that could be sacrificed ranges from zero to C, and the number of soldiers that could be killed ranges from S to zero, where $S > C$. The values of C and S vary from scenario to scenario.

The more civilians are sacrificed in a given scenario, the fewer soldiers will die. Each civilian death saves $S/C > 1$ soldiers, approximately. It follows that total deaths are minimized when C civilians are sacrificed (the maximum possible number). Parameters S and S/C specify the incentives of the dilemma.[1] Saving one soldier requires sacrificing C/S civilians; S/C is the reciprocal of the relative price of saving a soldier—it expresses how many soldiers will be saved by sacrificing a single civilian.[2] C, on the other hand, is analogous to what economists would call "income": It is the number of civilian lives that are available to exchange for soldiers' lives.

Figure S4 shows the description of a scenario presented to subjects. Total deaths in this scenario range from $C = 4$ million (all civilians) to $S = 6$ million (all soldiers).

Alternatives were presented to subjects expressed in terms of bads (human deaths). To analyze subjects' responses using a revealed preference method, we re-expressed the alternatives as their corresponding goods (human lives).

We use the term "solution" to denote an outcome of the war expressed in terms of lives. Solution (c, s) is a bundle of c civilian lives and s soldiers' lives.[3] If an alternative is "x civilians sacrificed and y soldiers killed," its corresponding feasible solution is $c = C - x$ civilians spared and $s = S - y$ soldiers saved. For example, the fourth alternative in Fig. S4—3 million sacrificed civilians and 2 million dead soldiers—corresponds to solution $(1, 4)$; that is, 1 million civilians spared and 4 million soldiers saved.

[1] In economics, the incentives would be the relative price of two goods and the income available to spend on them.

[2] The reciprocal of price is called "cost-effectiveness" in economics. The greater S/C is, the more lives are saved by sacrificing one civilian.

[3] "Bundle" is a term of art in microeconomics. A bundle is an n-dimensional vector or n-tuple containing non-negative quantities of n goods.

1. If no civilians are sacrificed, 6 million soldiers will die on the battlefield.
2. To end the war and save all the soldiers, 4 million civilians would have to be sacrificed.
3. For every 4 civilians sacrificed during the bombings, 6 soldiers less will die on the battlefield, approximately.

Remember that half of the soldiers who die are from each country, and that you are not a citizen of either country.

Given the above scenario, choose the combination of dead soldiers and sacrificed civilians that feels morally right to you.

- ○ 6 million soldiers / 0 civilians
- ○ 5 million soldiers / 1 million civilians
- ○ 3 million soldiers / 2 million civilians
- ○ 2 million soldiers / 3 million civilians
- ○ 0 soldiers / 4 million civilians

Figure S4. Text of a scenario presented to subjects.

4 Moral rationality in the war dilemma

In this section we present the foundations of rational choice theory, focusing on its application to moral psychology and the war dilemma. For those readers interested in a more general treatment of the subject, we recommend Jehle & Reny's microeconomics textbook [1].

4.1 Rational choice theory in a nutshell

Rational choice theory is a formalization of folk-psychological theories of intentional behavior [2]. It aims to explain and predict the choices of decision makers who act purposefully and react to incentives.

The core of a rational choice model is *the agent.* She is an abstract representation of a real decision maker. Depending on the decision problem, the agent may represent a person, a household, a company, a political party, an army, or a non-human animal, to mention a few applications. In our application, the agent represents the moral tradeoff system.

A rational choice model has the following elements:

1. **A set of conceivable options**: This set contains all options that the agent "can conceive of," whether feasible or not at the moment of choice. Options are problem-specific, but do not depend on the prevailing incentives.

2. **A preference order**: For any two conceivable options $\mathbf{x}$ and $\mathbf{y}$, the agent can tell whether $\mathbf{x}$ is "at least as good" as $\mathbf{y}$, according to her preferences. This is written as $\mathbf{x} \succsim \mathbf{y}$. This expression is also read as "$\mathbf{x}$ is weakly preferred to $\mathbf{y}$."

3. **A feasible set**: This set contains the options that the agent perceives as available to solve her problem. A feasible set is a proper subset of the set of conceivable options.

4. **The optimization assumption**: Among all feasible options, the agent will choose one of the most preferred (there may be a tie for first place).

This minimal set of assumptions about preferences, feasible options, and optimization constitute the "axioms of rational choice."

4.2 *Homo economicus*

The axioms of rational choice theory are usually lumped together with certain auxiliary assumptions about the *content* of preferences. These assumptions include selfish preferences, love of leisure, expected utility, and geometric or exponential discounting of future utility. As a group, these assumptions loosely define the preferences of a *Homo economicus*.

Contrary to popular belief, the empirical validity of *Homo economicus* preferences is not essential to the survival of rational choice theory. If the auxiliary assumptions fail in a particular application, they can be replaced by better ones without having to revise the axioms of the theory. Rational choice theory is not about the content of an agent's preferences; it is about how the agent makes choices given whatever preferences she has.

4.3 What are preferences?

The technical definition of preference is loosely related to its colloquial meaning.

In common parlance, "to prefer" means "to like." This is a folk psychological concept that predicts a wide range of mental states, physiological reactions, and behaviors. If you tell me that you like Snickers better than Twizzlers, I will think you enjoy eating Snickers more than eating Twizzlers. I will think that you salivate more at the sight of a Snickers. I will anticipate that, given a choice between a Snickers and a Twizzlers, you will pick the Snickers.

In rational choice theory, by contrast, the word "preference" is a label attached to a mathematical object: a binary relation between the elements of a set (the set of conceivable options).

The connection between preferences (in their technical sense) and the cognitive mechanisms of decision making is mostly ignored by researchers in the field. Most are agnostic about the psychological reality of preferences: They argue that it is irrelevant whether preferences are real or fictitious. In their view, all that matters is that decision makers behave *as if* they had preferences and optimized their choices accordingly [3, 4].

This nihilistic ontology of preferences is an element of "predictionism," the prevailing epistemology of rational choice theory. Predictionism is an idiosyncratic variant of instrumentalism. It states that rational choice models are neither true nor false;

only useful or useless for predicting the choices of real decision makers [5]. Other predictions of the theory—including its axioms, which trivially predict themselves—are dismissed as immaterial. The arbitrary dismissal of some predictions separates predictionism from standard forms of instrumentalism, which maintain that all predictions of a theory—including its axioms—must be under constant empirical scrutiny.

Sweeping axioms under the rug may seem like cheating, but it's the only workable epistemology when decision makers are households or firms. Families and firms make purposeful decisions, factoring in the incentives they face, but they lack minds capable of optimization: Decisions in both types of organizations arise from internal negotiations between their members, each of whom has his own individual agenda. In the absence of a theory of internal negotiations, economists recourse to the best existing alternative, which is rational choice theory. Their hope is that, when viewed from the outside, the negotiation process approximates an optimization process, at least at the aggregate or "market" level.

But when the decision maker is a person, the claim that he can make optimal choices without a cognitive system that does the optimizing is less plausible.[4]

4.4 Moral preferences

In rational choice models, preferences can be about any kind of goods: consumer goods and services, leisure time, financial assets, prestige, mates—anything that is scarce and people desire. When rational choice theory is applied to moral psychology, preferences are about moral goods, such as human lives, improving people's welfare, or fulfilling obligations.

Regarding notation, the statement "option $\mathbf{x}$ is weakly preferred to option $\mathbf{y}$" is interpreted as "solution $\mathbf{x}$ is felt to be at least as right as solution $\mathbf{y}$." Likewise, the statement "option $\mathbf{x}$ is strictly preferred to option $\mathbf{y}$" is interpreted as "solution $\mathbf{x}$ is felt to be more right than solution $\mathbf{y}$."

Unlike orthodox rational choice theorists, who take an agnostic stance on the psychological reality of preferences, we commit to a realist ontology: Moral preferences exist temporarily in memory, represented as rightness functions.

[4]Of course, no one consciously optimizes his choices given an explicit preference order of all conceivable options; but this does not invalidate the theory. The mind is known to carry out very complex algorithms in real time (as in language and vision) without our being aware of it.

4.5 Properties of a preference order

A preference order has the following properties:

1. It is **total**: $\mathbf{x} \succsim \mathbf{y}$, or $\mathbf{y} \succsim \mathbf{x}$, or both. This means that all options can be compared with each other.

2. It is **antisymmetric**: If $\mathbf{x} \succsim \mathbf{y}$ and $\mathbf{y} \succsim \mathbf{x}$, then the agent is "indifferent" between $\mathbf{x}$ and $\mathbf{y}$. In notation, $\mathbf{x} \sim \mathbf{y}$.

3. It is **reflexive**; that is, $\mathbf{x} \succsim \mathbf{x}$. In words, every option is weakly preferred to itself (it is at least as good as itself).

4. It is **transitive**: $\mathbf{x} \succsim \mathbf{y}$ and $\mathbf{y} \succsim \mathbf{z}$ implies $\mathbf{x} \succsim \mathbf{z}$.

Transitivity is a well-known and generally uncontroversial assumption. It is usually explained with apples and oranges: "If an apple is preferred to a banana, and a banana is preferred to an orange, then an apple is preferred to an orange." The example is accurate, but leaves the wrong impression that transitivity is trivial, when in fact the opposite is true. Being rational or "consistent" (complying with the axioms of rational choice at all times) is a dauntingly complex task when options are not single goods (an apple, a banana, an orange) but bundles of goods (k apples, m bananas, and n oranges). The complexities will be apparent in section 4.18, where we will explain GARP in the context of the war dilemma.

4.6 Indifference and strict preference

A preference order creates as a by-product an indifference relation and a strict preference relation.

Indifference relation: $\mathbf{x} \sim \mathbf{y}$ if and only if $\mathbf{x} \succsim \mathbf{y}$ and $\mathbf{y} \succsim \mathbf{x}$

Strict preference relation: $\mathbf{x} \succ \mathbf{y}$ if and only if $\mathbf{y} \not\succsim \mathbf{x}$.

Both relations are transitive.

In the context of moral judgment, the statement "$\mathbf{x}$ is indifferent to $\mathbf{y}$" means "$\mathbf{x}$ is felt to be as right as $\mathbf{y}$." Likewise, the statement "$\mathbf{x}$ is strictly preferred to $\mathbf{y}$" means "$\mathbf{x}$ is felt to be more right than $\mathbf{y}$."

4.7 Utility functions

In some decision problems, the set of conceivable options is potentially infinite and boundless. The war dilemma is a case in point (imagine two billion soldiers saved and one billion civilians spared; now imagine one billion + 1).[5]

The question naturally arises as to how the mind can store and manipulate an order of preference over infinitely many options. This seems like a computational impossibility, but rational choice theorists have found a workaround: An infinite preference order can be represented by a continuous real-valued "utility function," provided that the order satisfies certain mathematical conditions.[6]

A utility function is a finite mathematical object that represents, in compact form, a preference order for all solutions that the mind can conceive of. To each conceivable solution, a utility function assigns a level of utility: a real number that serves to compare options in terms of their subjective "goodness." If two options have the same utility, they are considered equally good. Otherwise, the option with the higher utility is strictly preferred.

Formally, a utility function is written as follows:

$$u(\mathbf{x}) : X \to V \subseteq \mathbb{R},$$

where u is the function, $\mathbf{x}$ is a conceivable option, X is the set of conceivable options, and V contains the values that utility can take (which are real numbers). Like the preference order it represents, a utility function is problem-specific. But it does not depend on which options are available: The function is invariant with respect to the content of the feasible set.

A utility function gives rise to a preference order in the following manner:

If $u(\mathbf{x}) \geq u(\mathbf{y})$, then $\mathbf{x} \succsim \mathbf{y}$.

If $u(\mathbf{x}) = u(\mathbf{y})$, then $\mathbf{x} \sim \mathbf{y}$.

[5]It is an open question whether *all* quantities that are conceivable formally (in the context of mathematics and a linguistic number system; e.g., "one billion + 1") can be represented by a cognitive system that constructs preference orders (see [6]). The minds of humans and other animals represent quantities, but the format of these representations and the limits of what each computational system can represent are active areas of cognitive research.

[6]See [1, section 1.2] for an in-depth discussion of preference orders and their connection to utility functions.

If $u(\mathbf{x}) > u(\mathbf{y})$, then $\mathbf{x} \succ \mathbf{y}$.

It is tempting to associate utility with pleasure, happiness, satisfaction, or welfare; but the analogy is improper. Utility, as a mathematical object, has no intensity or degree: It tells you whether one option is better than another, not how much better it is.

When a preference order has certain mathematical properties, the corresponding utility function is continuous and doubly-differentiable. Continuity and double differentiability are convenient for modeling, because they allow the use of calculus, which simplifies constrained maximization.

4.8 Rightness functions are utility functions

In abstract terms, a rightness function can be written as follows:

$$v(\mathbf{x}, \quad) : X \to V \subseteq \mathbb{R},$$

where v is the function; $\mathbf{x}$ is a conceivable solution; X is the set of conceivable solutions; V contains the values that rightness can take (which are real numbers); and is a vector of situation-specific parameters computed by the MV subsystem, by operating on morally-laden representations of the situation. (See the main text for a model of the moral tradeoff system; MV is the component that constructs rightness functions.)

The rightness function gives rise to a rightness order of conceivable solutions, as follows:

If $v(\mathbf{x}, \quad) = v(\mathbf{y}, \quad)$, then both solutions are felt to be equally right.

If $v(\mathbf{x}, \quad) > v(\mathbf{y}, \quad)$, then $\mathbf{x}$ is felt to be more right than $\mathbf{y}$.

The parameters in modulate the way in which the rightness function ranks solutions. For example, in the war dilemma, contains the relative weights assigned to the lives of civilians and soldiers, plus other parameters having to do with elasticity of substitution and/or other aspects of preferences.

Context can affect the parameters values. For example, if the civilians had supported the war, rather than opposed it, civilian lives would probably weigh less. Likewise, if the soldiers had volunteered instead of being forcibly drafted, soldiers' lives would probably weigh less.

4.9 Testing for consistency requires auxiliary assumptions

One can always rationalize a sequence of moral judgments by saying that the subject is indifferent among all conceivable solutions, and hence all her choices were arbitrary. More formally, the rightness function $v(\mathbf{x}) = k$, where k is a constant, rationalizes any sequence of moral judgments. It follows that the moral tradeoff system hypothesis cannot be falsified in isolation (an inconvenience that extends to rational choice theory in general). Making auxiliary assumptions about the shape of the rightness function is an inescapable requirement.

4.10 Well-behavedness

In our empirical analysis, we made the auxiliary assumption that rightness functions are "well-behaved." Contrary to what the term suggests, well-behavedness is not an ethical standard that a rightness function can instantiate. Rather, well-behavedness is a set of mathematical properties that some utility functions satisfy: continuity, nonsatiation, and convex indifferent curves.[7] The meaning of these properties is mathematically involved, but it has an intuitive geometric interpretation.

Figure S5 depicts two rightness functions viewed from above: One is well-behaved, while the other is not. The points on the upper-right quadrant of the Cartesian plane (axes included) represent the set of conceivable solutions. Point (c, s) is a solution involving c civilians spared and s soldiers saved. Elevation at a given point is the rightness of the corresponding solution. All solutions that lie on the same contour line are felt equally right. The arrows indicate the direction in which rightness increases.

A well-behaved rightness function (such as the one depicted in panel *a*) resembles a mountain with no peak and no backside. It is nonsatiated: Starting from any solution (any point on the Cartesian plane), one can increase rightness by increasing c, s, or both quantities (mathematically speaking, the function is strictly increasing). The contour lines of a well-behaved rightness function are "indifference curves." This means that all solutions that are equally right lie on the same contour line. Viewed from the origin, well-behaved indifference curves are convex (though not necessarily strictly convex). In jargon, this means that there is "substitution" between goods.

[7] A more general form of well-behavedness assumes local nonsatiation, a less stringent mathematical requirement on the structure of the preference order.

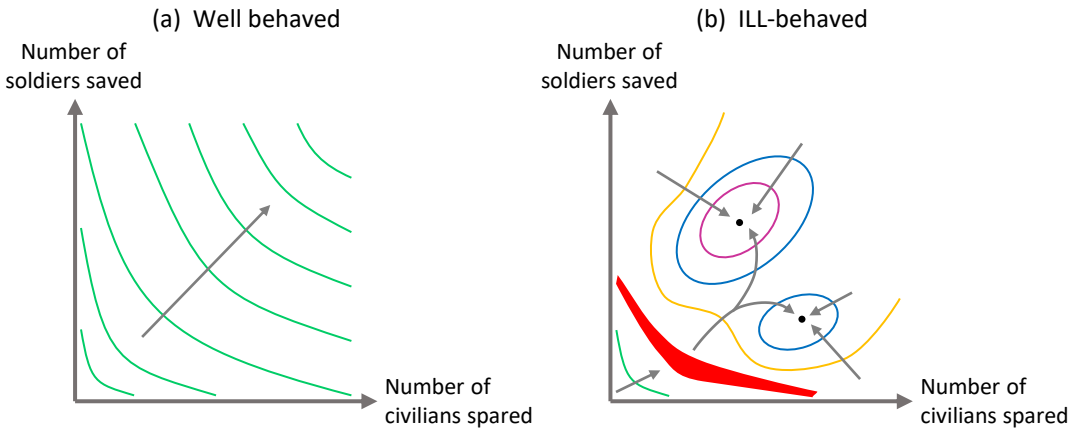

Figure S5. Two rightness functions viewed from above.

The rightness function depicted in panel *b* misbehaves in various ways. It has two peaks (marked with black dots), where it is satiated. It has more than one contour line with the same level of moral rightness (such as the blue ones), which are thus not indifference curves. Its surface has a flat "indifference zone" (highlighted in red), where the function has a constant value. And some indifference curves are not convex (for example, the orange one).

We consider nonsatiation and convex indifference curves to be psychologically plausible assumptions for the war dilemma.

Nonsatiation formalizes two intuitions: Given the alternative of saving more soldiers without increasing civilian deaths, a typical subject will take that alternative. Likewise, given the alternative of saving more civilians without increasing soldiers' deaths, a typical subject will take that alternative.

Convex indifference curves, on the other hand, formalize an intuition about the substitutability of soldiers' and civilian lives: For a subject to be indifferent between a series of alternative solutions, each additional civilian death must be compensated by saving the lives of a non-decreasing number of additional soldiers. In other words, measured in terms of civilian lives, each additional soldier life is worth the same or less than the previous one.

It should be emphasized that we do not posit that all rightness functions are well-behaved. It just seemed likely to us that, in this particular experiment, most subjects would exhibit well-behaved rightness functions.

The experimental results confirmed our guess.

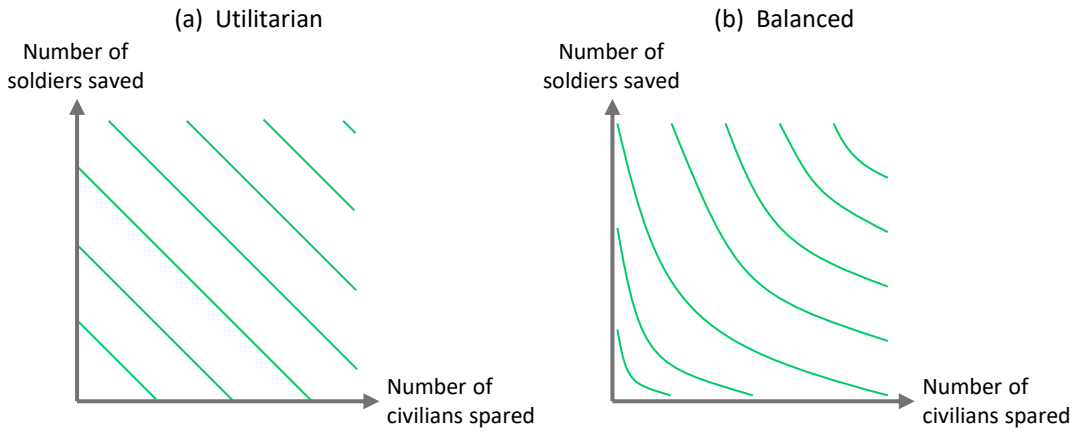

Figure S6. Two types of rightness functions.

4.11 Utilitarian and balanced rightness functions

Figure S6 depicts two rightness functions. Both are well-behaved.

The utilitarian value function (depicted in panel *a*) is given by $v(c, s) = c + s$. According to this formula, only the total number of survivors matters from a moral point of view. The function's indifference curves are straight lines: They have the form $s = r - c$, where parameter $r > 0$ is the level of rightness of that curve. Straight indifference curves are convex, but not strictly convex.

To illustrate one kind of rightness function that can produce compromise judgments, consider the "balanced" function in panel *b* (short for "striking a balance"). It has indifference curves that bend smoothly inwards. In mathematical terms, they are strictly convex to the origin. Strict convexity implies that, for the subject to remain indifferent among a series of alternative solutions, each additional civilian death must be compensated by the lives of an increasing number of additional soldiers. Put differently, each additional surviving soldier is worth less to the subject than the previous one in terms of civilian lives.

4.12 Feasible solutions to the war dilemma

The feasible set or "feasibility constraint" of a scenario is given by

$$F = \{(c, s) : s = -(S/C)c + S, \text{ where } c, s \geq 0\}.$$

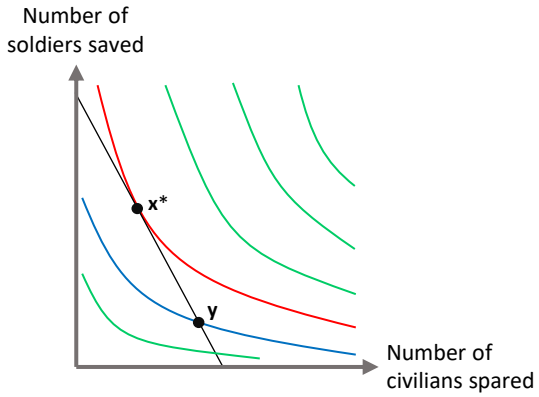

Figure S7. A balanced rightness function and a linear feasibility constraint. The optimal solution $\mathbf{x}^*$ is at the point of tangency between the constraint and the red indifference curve.

where S and S/C are parameters, and $S > C > 0$. These parameters constitute the incentives of the dilemma. S is the number of soldiers at risk of death. S/C is the efficacy of bombing: the number of soldiers saved per civilian sacrificed.

Figure S7 provides an example.

4.13 Compromises judgments as rightness maximizing choices

Figure S7 depicts a balanced rightness function. The diagonal line segment represents a feasibility constraint; imagine that all points on the line segment are available to be chosen.

Since this balanced rightness function has strictly convex indifference curves that approach, but do not intersect, the axes, the optimal solution $\mathbf{x}^* = (c^*, s^*)$ is unique and lies at the point of tangency between the feasibility constraint and an indifference curve. In consequence, the optimal solution is intermediate. More formally, when all points on the feasibility constraint are available and the rightness function is balanced, it is always the case that $c^*, s^* > 0$: That is, the optimal solution will always be a compromise.[8]

[8]For the war dilemma, feasibility constraints were discretized to the nearest million, so not all points on the line segment were available to be chosen. In this case, a person with these indifference curves might choose a mixture of compromise and extreme solutions. Imagine, counterfactually, that the feasible set included every point on the line connecting $(0, S)$ and $(C, 0)$. The optimal solution would be the point of tangency, and intermediate. But let's say the number of civilians spared at the optimum is closer to 0 than to 1M (i.e., anywhere from 1 to 499,000); when the feasible set is

As a proof, consider solution $\mathbf{y}$, which lies at the intersection between the feasibility constraint and the blue indifference curve. Observe that solutions on the blue curve have a lower level of rightness than those on the red curve. It follows that $\mathbf{y}$ is less right than $\mathbf{x}^*$. We conclude that $\mathbf{y}$, and any feasible solution other than $\mathbf{x}^*$, cannot be optimal.

In contrast, a subject who maximizes a utilitarian rightness function [i,e., $v(c, s) = c + s$] will choose the solution with the greatest possible number of survivors. In the war dilemma, the number of survivors is maximized when all soldiers are saved and no civilians are spared (from a maximum possible number of C). That is, $\mathbf{x}^* = (0, S)$.

4.14 Deontic moral values

Deontic values cannot be represented by a well-behaved rightness function, but they can be represented by a rule: "Prefer a solution that maximizes the number of civilians spared. Among solutions that maximize the number of civilians spared, prefer one that maximizes the number of soldiers saved."[9] The moral tradeoff system could construct this rule on the fly and store it temporarily in memory, like a computer stores an active program in RAM (but this is speculative).

4.15 Rational moral flip-flopping

Linear rightness functions can cause moral flip-flopping.

Consider the following example:

$$v(c, s) = \alpha c + (1 - \alpha)s,$$

where $0 < \alpha < 1$. And recall that the feasibility constraint of the war dilemma is

$$s = -\frac{S}{C}c + S,$$

where $0 \leq c \leq C$ and $0 \leq s \leq S$. This constraint implies that, if $c = 0$, then $s = S$. It also implies that, if $c = C$, then, $s = 0$.

discretized to the nearest million, an extreme solution that spares no civilians might be closer to that point than the intermediate solutions offered. A mix of compromise and extreme judgments can be produced by indifference curves with other properties as well (e.g., ones with a similar shape that do intersect the axes).

[9]Deontic values are a form of "lexicographic preference," which are notorious for not being representable by a continuous utility function.

The maximization problem can be easily solved by replacing s in the rightness function with the feasibility constraint. This reduces to:

$$v(c) = \alpha c + (1-\alpha)\left(-\frac{S}{C}c + S\right).$$

Moreover, grouping terms, we obtain:

$$v(c) = (1-\alpha)\left(\beta - \frac{S}{C}\right)c + (1-\alpha)S,$$

where $\beta = \alpha/(1-\alpha)$. Observe that β is an increasing function of α. Also, $\beta \in (0, \infty)$.

The solution to this optimization problem will depend on the sign of the coefficient accompanying c. There are three cases:

1. If $S/C > \beta$, then the coefficient will be negative, and so the optimal solution will be to minimize the value of c. Therefore, $c^* = 0$ and $s^* = S$. This means that the subject will issue a "utilitarian" judgment.

2. If $S/C < \beta$, then the coefficient will be positive, so the optimal solution will be to maximize the value of c. Therefore, $c^* = C$ and $s^* = 0$. This means that the subject will issue a "deontic" judgment.

3. If $S/C = \beta$, then the coefficient will be zero. Therefore, all feasible values of (c, s) will maximize rightness, and so the model will make no prediction. This degenerate case will occur with zero probability if S/C is a continuous random variable.

It follows that the subject will exhibit a flip-flopping response pattern, in response to changes in the incentives of the dilemma. Note that linear rightness functions are one of many types of rightness functions that can cause a flip-flopping response profile.

4.16 Revealed preferences, hand trembles, and inconsistency

As a subject makes choices, she gradually reveals preference relations, which may or may not match her true preference relations.

The simplest type of revelation occurs when the feasible set contains two options: $\mathbf{x}$ and $\mathbf{y}$. If the subject chooses $\mathbf{x}$, then "$\mathbf{x}$ is revealed weakly preferred to $\mathbf{y}$." Otherwise, we presume, the subject would have chosen $\mathbf{y}$. We denote this inference "$\mathbf{x} \succsim^R$

$\mathbf{y}$." Superscript R distinguishes revealed preferences (inferred from choices) from the subject's true, underlying preferences.

If a subject abides by the axioms of rational choice, without ever making a mistake, her revealed preferences will be accurate. Perfect rationality, in the sense of consistency, guarantees that $\mathbf{x} \succsim^R \mathbf{y}$ implies $\mathbf{x} \succsim \mathbf{y}$, $\mathbf{x} \sim^R \mathbf{y}$ implies $\mathbf{x} \sim \mathbf{y}$, and $\mathbf{x} \succ^R \mathbf{y}$ implies $\mathbf{x} \succ \mathbf{y}$.

A human, however, occasionally makes "trembling hand mistakes": unintended choices caused by clumsiness, distraction, and other forms of noise. Because of trembling hand mistakes, some revealed preference relations may be incorrect. For instance, an unintended choice could reveal $\mathbf{x} \succsim^R \mathbf{y}$, whereas in reality $\mathbf{y} \succ^R \mathbf{x}$. Contradictory revelations of this sort are called *inconsistencies*. A special case follows logically: "$\mathbf{x} \succ^R \mathbf{x}$" is an inconsistency, because $\mathbf{x} \succsim^R \mathbf{x}$ by definition.

A subject's revealed preference order grows as she makes more choices, but it can never fully encompass a boundless set of conceivable solutions (unlike true preference orders, revealed preference orders are not necessarily total). Even if a subject always makes consistent choices, her true preference order never becomes fully known.

4.17 Preference inference rules

The preference elicitation procedure that we used in this study is based on GARP. It exploits three inference rules.

The first inference rule derives from the optimization assumption:

> **Rule 1:** If a subject chooses $\mathbf{x}$, she reveals that she weakly prefers $\mathbf{x}$ to every other option in the feasible set (she reveals $\mathbf{x} \succsim^R \mathbf{y}$ for each feasible option $\mathbf{y}$).

This rule produces directly revealed preference relations; that is, relations that link options that are in the same feasible set.

The second rule derives from the transitivity assumption. It links three options in a revealed preference chain:

> **Rule 2:** $\mathbf{x} \succsim^R \mathbf{y}$ and $\mathbf{y} \succ^R \mathbf{z}$ implies $\mathbf{x} \succ^R \mathbf{z}$.

Options $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ need not all be in the same feasible set. If they are, the preference revelation is direct. Otherwise the preference revelation is indirect.

The third rule derives from the nonsatiation assumption:

> **Rule 3:** If $\mathbf{x}$ is Pareto superior to $\mathbf{y}$, then $\mathbf{x} \succ^R \mathbf{y}$.

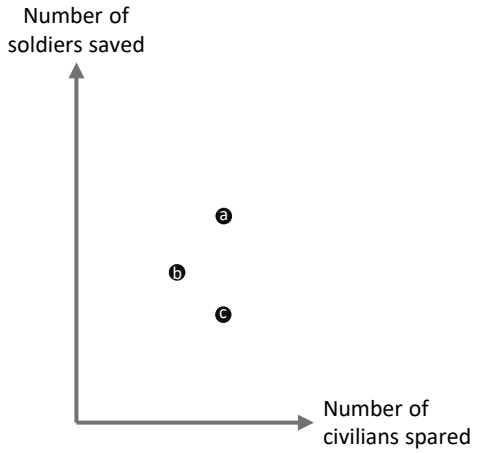

Figure S8. Solution **a** is Pareto superior to **b** and **c**. Therefore, **a** is revealed to be strictly preferred to both solutions.

This rule takes the preference for Pareto superior options as *a priori* knowledge.

In the war dilemma, solution **x** is Pareto superior to solution **y** if and only if three conditions are met: (1) More soldiers are saved and/or more civilians are spared in **x** than in **y**. (2) The number of civilians spared in **x** is not less than in **y**. (3) The number of soldiers saved in **x** is not less than in **y**.

By way of example, consider solutions **a**, **b**, and **c** depicted in figure S8. **a** is Pareto superior to **b**, because **a** saves more soldiers and spares more civilians than **b**. For this reason, we assume that the subject strictly prefers **a** to **b** (i.e., **a** is felt more right than **b**). Moreover, **a** is Pareto superior to **c**, because **a** and **c** spare the same number of civilian lives, but **a** saves more soldiers. Therefore, rule 3 leads us to conclude that the subject strictly prefers **a** to both **b** and **c**.

On the other hand, **b** is not Pareto superior to **c**, and **c** is not Pareto superior to **b**, because **b** saves more soldiers but spares fewer civilians than **c**. Therefore, we cannot tell *a priori* whether the subject is indifferent between **b** and **c**, or strictly prefers one solution to the other.

In addition to the three inference rules, the procedure assumes $\mathbf{x} \succsim^{R} \mathbf{x}$. Intuitively speaking, it is taken for granted that each option is at least as good as itself.

4.18 The generalized axiom of revealed preferences

Here we will derive GARP for the particular case of the war dilemma, using the three inference rules presented in section 4.17.

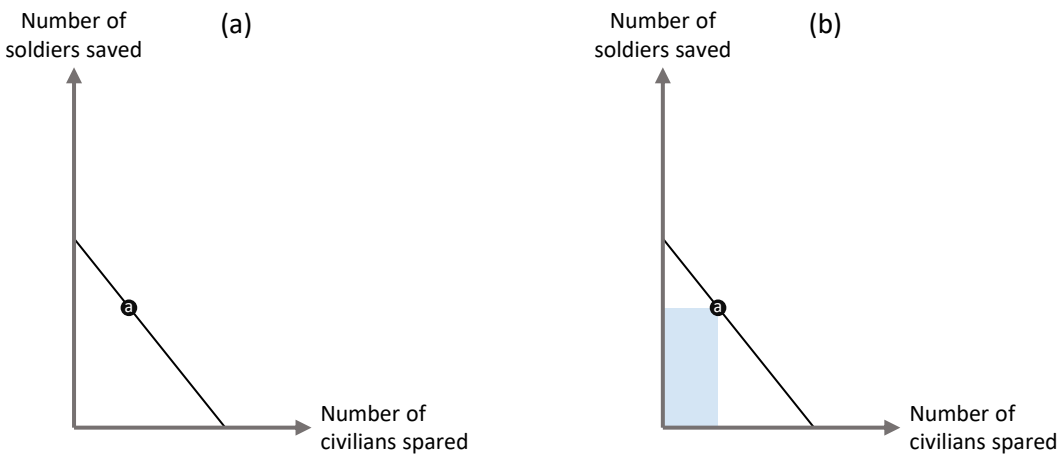

Figure S9. (a) A scenario of the war dilemma. The subject chooses solution **a**. (b) For each solution **x** in the light blue rectangle, **a** is Pareto superior to **x**. It follows that $\mathbf{a} \succ \mathbf{x}$ for each **x** in the rectangle (inference rule 3).

Two preliminary notes on the well-behavedness assumption: First, well-behavedness is sufficient but more stringent than necessary for GARP. Second, ill-behavedness does not invalidate the moral tradeoff system hypothesis, because ill-behaved rightness functions lead to rational judgments, though not necessarily of the GARP-respecting kind. Depending on their choices, subjects who exhibit ill-behaved rightness functions may be misclassified by GARP as inconsistent, a possibility that biases the test against the hypothesis.

We begin by analyzing the simplest case: a subject facing two scenarios.

First, the subject faces the scenario depicted in figure S9a. She chooses solution **a** from the feasibility constraint. This choice reveals that she weakly prefers **a** to every other solution on the constraint, and strictly prefers **a** to every solution below the constraint. This inference is explained step by step in figures S9b to S10b.

As a corollary, we obtain a fourth inference rule:

> **Rule 4:** If the subject chooses solution **y** from a feasibility constraint, then, for each solution **x** below the constraint, $\mathbf{y} \succ^R \mathbf{x}$.

Next, the subject faces the scenario depicted in figure S11a. This scenario has a different feasibility constraint, with a steeper slope and a higher intercept. The steeper slope (a higher value of S/C) means that the death of one civilian saves a greater number of soldiers. The higher intercept (a higher value of S) means that more soldiers are at risk of death.

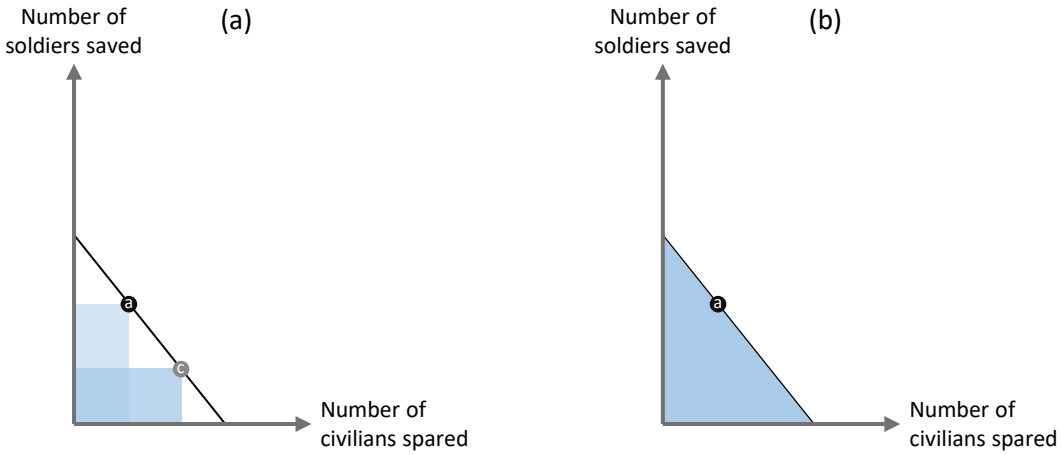

Figure S10. (a) Solution $\mathbf{c}$ was feasible, but the subject did not choose it. It follows that $\mathbf{a} \succsim^R \mathbf{c}$ (inference rule 1). Moreover, for each solution $\mathbf{x}$ in the darker blue rectangle, $\mathbf{c}$ is Pareto superior to $\mathbf{x}$; hence, $\mathbf{c} \succ^R \mathbf{x}$ (inference rule 3). From $\mathbf{a} \succsim^R \mathbf{c}$ and $\mathbf{c} \succ^R \mathbf{x}$, it follows that $\mathbf{a} \succ^R \mathbf{x}$ for each $\mathbf{x}$ in the darker blue rectangle (inference rule 2). (b) The previous argument extends to all other unchosen solutions on the feasibility constraint. Each of these solutions has a corresponding blue rectangle of Pareto inferior solutions. The rectangles overlap, creating a blue triangle below the constraint. We conclude that, for each solution $\mathbf{x}$ below the feasibility constraint, $\mathbf{a} \succ^R \mathbf{x}$.

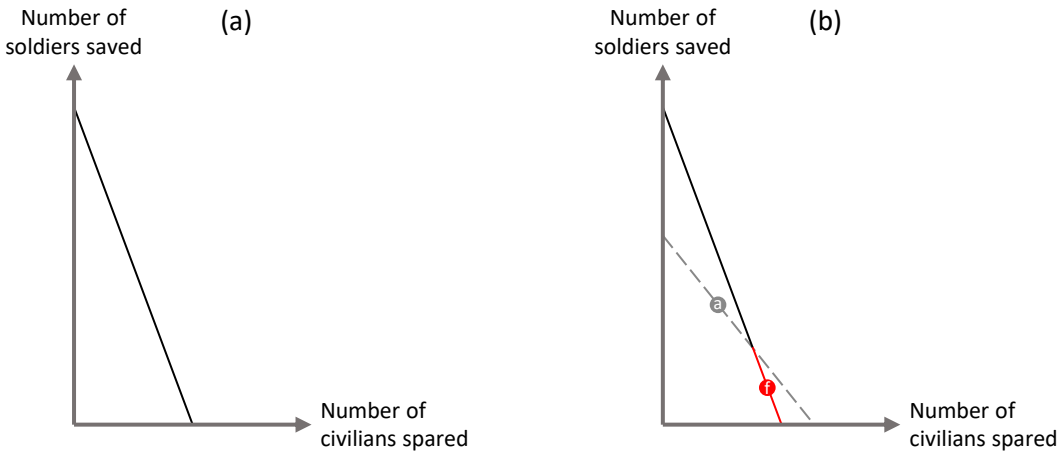

Figure S11. (a) The subject faces a second scenario. (b) GARP predicts that a morally rational subject would not choose a solution on the red segment. To see why, suppose that she chooses $\mathbf{f}$. This choice reveals that $\mathbf{f} \succ^R \mathbf{a}$, because $\mathbf{a}$ is below the current feasibility constraint (inference rule 4). On the other hand, $\mathbf{f}$ is below the previous feasibility constraint (the dashed line). It follows that $\mathbf{a} \succ^R \mathbf{f}$ (inference rule 4), which contradicts $\mathbf{f} \succ^R \mathbf{a}$. This inconsistency is a GARP violation. Note: There is a special case in which $\mathbf{a}$ is at the intersection of the two constraints. In that case, $\mathbf{f} \succsim^R \mathbf{a}$, because $\mathbf{a}$ is an unchosen solution on the second constraint (inference rule 1). This revelation also contradicts $\mathbf{a} \succ^R \mathbf{f}$.

What solutions could the subject rationally choose?

Figure S12 shows the feasibility constraint divided into two segments, one green and one red. GARP predicts that a morally rational subject would choose a solution on the green segment. We prove this graphically in figure S12.

There are also cases in which the second choice will not cause inconsistencies or "GARP violations," regardless of what the subject chose first. This is proved in figure S13.

More generally, GARP allows for arbitrarily long sequences of scenarios. The following inference rules link n solutions, within and across scenarios, in chains of revealed preference relations:

> **Rule 5a:** If $\mathbf{x}_1 \succsim^R \mathbf{x}_2 \succsim^R \ldots \succsim^R \mathbf{x}_n$ and $\mathbf{x}_n \succsim^R \mathbf{x}_{n+1}$, then $\mathbf{x}_1 \succsim^R \mathbf{x}_2 \succsim^R \ldots \succsim^R \mathbf{x}_n \succsim^R \mathbf{x}_{n+1}$.
>
> **Rule 5b:** If $\mathbf{x}_1 \succsim^R \mathbf{x}_2 \succsim^R \ldots \succsim^R \mathbf{x}_n$ and $\mathbf{x}_n \succ^R \mathbf{x}_{n+1}$, then $\mathbf{x}_1 \succsim^R \mathbf{x}_2 \succsim^R \ldots \succsim^R \mathbf{x}_n \succ^R \mathbf{x}_{n+1}$.

These rules are generalizations of rule 2.

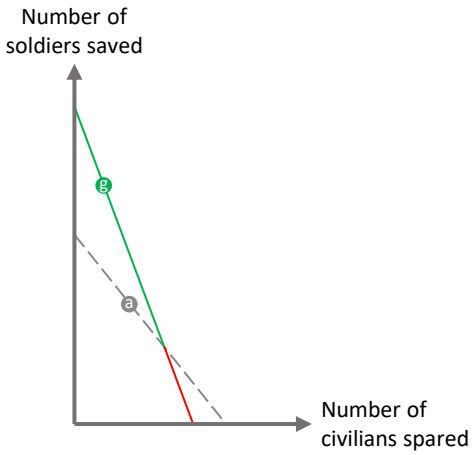

Figure S12. A morally rational subject would choose a solution on the green segment. To see why, suppose she chooses $\mathbf{g}$. This reveals that $\mathbf{g} \succ^R \mathbf{a}$, because $\mathbf{a}$ is below the second constraint (inference rule 4). On the other hand, because $\mathbf{g}$ is above the first constraint (the dashed line) it was not available to be chosen when the subject chose $\mathbf{a}$ from the first feasibility constraint. So there are no reasons to infer $\mathbf{a} \succsim^R \mathbf{g}$ (rules 1 or 4 cannot be used). Hence, no inconsistency is found.

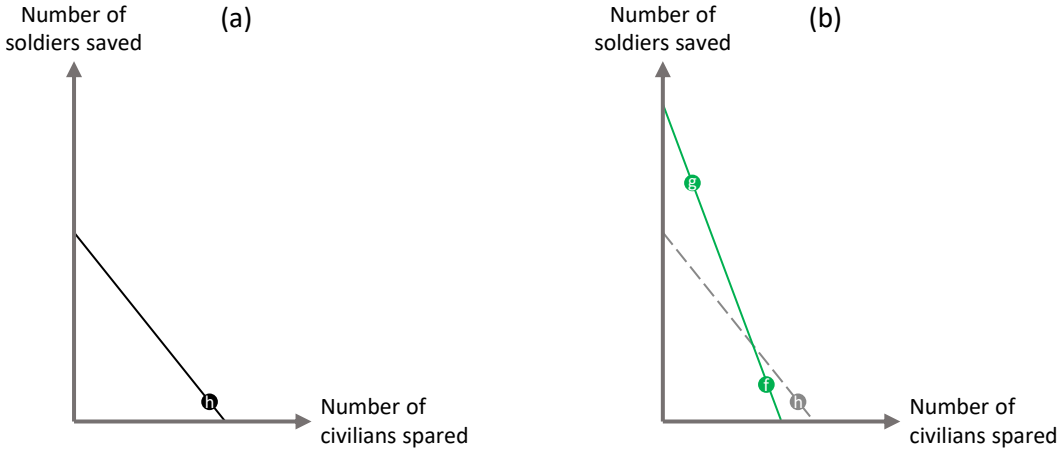

Figure S13. (a) A different subject chooses $\mathbf{h}$ from the first feasibility constraint. (b) In the second scenario, whatever choice he makes (such as $\mathbf{f}$ or $\mathbf{g}$) would be morally rational. This is because $\mathbf{h}$ is to the right of the second constraint, so the subject's second choice does not reveal a preference relation between $\mathbf{h}$ and any choice on the second constraint (rules 1 or 4 cannot be used).

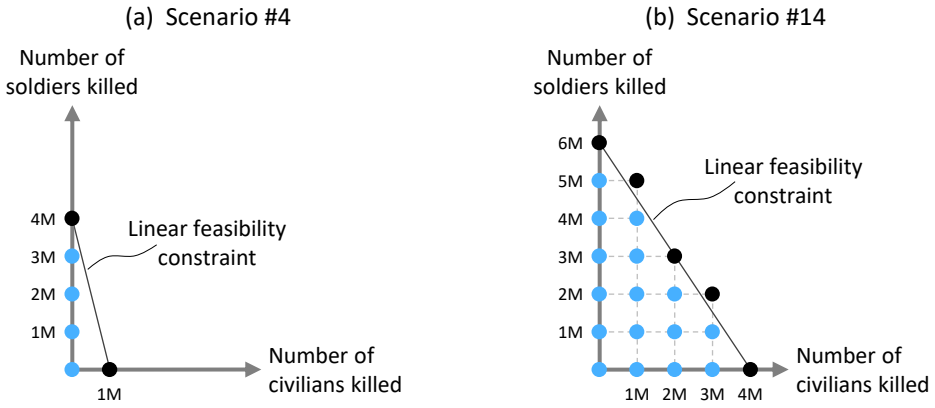

Figure S14. Black markers represent discretized feasible solutions. Blue markers represent solutions that are Pareto inferior to at least one feasible solution. (a) $S = 4$ and $S/C = 4$. This scenario offers two extreme solutions. (b) $S = 6$ and $S/C = 1.5$. This scenario offers two extreme and three intermediate solutions.

Due to rules 5a and 5b, a single choice may reveal a large number of preference relations. These relations may link solutions that never appear in the same feasibility constraint, or even in the same scenario. Cascades of GARP violations can occur as a consequence.

4.19 Discretized solutions

Counting GARP violations requires discretizing the solution space. A discretized feasible set approximates a linear feasibility constraint, such as the ones shown in figure S14. Solutions in the discretized feasible set fall approximately on the line (they are rounded to the nearest million). A Pareto inferior solution is inferior to at least one member of the feasible set.

4.20 Counting GARP violations

To count GARP violations, we used the same method as Andreoni & Miller [7].

Consider, for example, a subject that faces scenarios 13, 17, and 19. Table S4 shows the feasible set and Pareto inferior solutions of these scenarios.

First, the subject faces scenario 13 and chooses solution $(2, 2)$. Since $(1, 4)$ also appears in the feasible set, her choice reveals $(2, 2) \succsim^R (1, 4)$ (inference rule 1). Moreover, since $(1, 4)$ is Pareto superior to $(1, 3)$, we assume that $(1, 4) \succ^R (1, 3)$ (inference

Table S4. Three scenarios of the war dilemma

Scenario #	Soldiers at risk (S)	Soldiers saved for each civilian sacrificed (S / C)	Alternatives (civil. sacrificed, sold. dead)	Feasible set (civil. spared, sold. saved)	Pareto inferior solutions (civil. spared, sold. saved)
13	6	2.00	(0, 6) (1, 4) (2,2) (3,0)	(3, 0) (2, 2) (1, 4) (0, 6)	(0, 0) (1, 0) (2, 0) (0, 1) (1, 1) (2, 1) (0, 2) (1, 2) (0, 3) (1, 3) (0, 4) (0, 5)
17	7	3.50	(0, 7) (1, 4) (2, 0)	(2, 0) (1, 3) (0, 7)	(0, 0) (1, 0) (0, 1) (1, 1) (0, 2) (1, 2) (0, 3) (0, 4) (0, 5) (0, 6)
19	7	1.75	(0, 7) (1, 6) (2, 4) (3, 2) (4, 0)	(4, 0) (3, 1) (2, 3) (1, 5) (0, 7)	(0, 0) (1, 0) (2, 0) (3, 0) (0, 1) (1, 1) (2, 1) (0, 2) (1, 2) (2, 2) (0, 3) (1, 3) (0, 4) (1, 4) (0, 5) (0, 6)

Note: All quantities in millions of lives.

rule 3). Applying rule 5b to both revelations, we infer:

$$(2,2) \succsim^R (1,4) \succ^R (1,3).$$

Next, the subject faces scenario 17 and chooses solution $(1,3)$. Since $(0,7)$ also appears in the feasible set, her choice reveals $(1,3) \succsim^R (0,7)$ (inference rule 1); and we previously inferred that $(2,2) \succsim^R (1,4) \succ^R (1,3)$. Applying rule 5a to both revelations, we infer:

$$(2,2) \succsim^R (1,4) \succ^R (1,3) \succsim^R (0,7). \tag{1}$$

Lastly, the subject faces scenario 19 and chooses solution $(0,7)$. Since $(2,3)$ also appears in the feasible set, her choice reveals $(0,7) \succsim^R (2,3)$. Moreover, since $(2,3)$ is Pareto superior to $(2,2)$, we assume that $(2,3) \succ^R (2,2)$. Applying rule 5b to both revelations, we infer:

$$(0,7) \succsim^R (2,3) \succ^R (2,2) \tag{2}$$

Using inference rules 5a and 5b we can link preference chains (1) and (2), as follows:

$$(2,2) \succsim^R (1,4) \succ^R (1,3) \succsim^R (0,7) \succsim^R (2,3) \succ^R (2,2).$$

We have thus found a "preference cycle."

The above preference cycle entails that, for each pair of solutions $\mathbf{x}$ and $\mathbf{y}$ in the set $\{(2,2),(1,4),(1,3),(0,7),(2,3)\}$, it is the case that "$\mathbf{x} \succ^R \mathbf{y}$ and $\mathbf{y} \succsim^R \mathbf{x}$" (which includes as a particular case "$\mathbf{x} \succ^R \mathbf{x}$"). Each of these contradictions is a GARP violation. Since there are five solutions in the cycle, the number of GARP violations created by the cycle is $\binom{5}{2} + 5 = 15$.

Note that the preference cycle that we have detected is one of several caused by the subject's third choice. You can find the remaining cycles using the same procedure that we have illustrated here.

5 The representative agent's rightness function

We conjectured that the representative agent maximizes a CES rightness function:

$$u((c,s),\beta(k)) = \alpha(k)^{\frac{1}{\sigma(k)}} c^{\frac{\sigma(k)-1}{\sigma(k)}} + [1-\alpha(k)^{\frac{1}{\sigma(k)}}] s^{\frac{\sigma(k)-1}{\sigma(k)}},$$

where c and s are the numbers of surviving civilians and soldiers, $k = \{\text{SW}, \text{BU}, \text{CW}\}$ represents the frame, and $\beta(k) = [\alpha(k), \sigma(k)]$ is a vector of parameters whose values can be frame-dependent.

Parameter α is the weight of civilian lives on moral rightness. Possible values of α range from 0 to 1, where $\alpha > 0.5$ indicates more weight on the lives of civilians relative to soldiers, and $\alpha = 0.5$ indicates equal weight.

Parameter $\sigma > 0$ is the elasticity of substitution between civilian and soldiers' lives. Elasticity of substitution determines the degree of curvature of the indifference curves. It has a behavioral implication: The higher the value of σ, the more sensitive the agent's responses are to changes in S/C.

We estimated the representative agent's rightness function indirectly, by fitting her optimal response function (recall that her response in each scenario is the average for all subjects). A optimal response function gives the number of soldiers to be saved, for given values of S, S/C, and k (the frame). In logarithms,

$$ln(s^*) = ln\frac{[1-\alpha(k)]S}{1-\alpha(k)+\alpha(k)(S/C)^{1-\sigma(k)}} + \varepsilon,$$

where s^* is the optimal response, and ε is an error term.

We performed a nonlinear least-squares regression with $n = 21 \times 3 = 63$ observations (21 obs. per condition). Its results are reported in Table S5.

The estimated values of α are significantly different in each condition. The representative agent puts the most weight on civilian lives ($\alpha = 0.80$) when they opposed the war but soldiers were willing to fight for their country. When civilians and soldiers were both unwilling participants, the agent also put more weight on the lives of civilians ($\alpha = 0.61$), but less than in the previous case. Only when civilians supported the war and soldiers were unwilling draftees, did the agent weight civilians and soldiers similarly ($\alpha = 0.49$).

Table S5. Nonlinear least-squares regression of the representative agent's optimal response function

				Num. obs.	63	
				R-squared	0.9975	
				Adj. R-squared	0.9972	
				Root MSE	0.0569	
				Res. dev.	-186.61	
ln(*s*)	Coef.	Robust std. err	t-stat.	P-value	[99% conf. inter.]	
α(SW)	0.80	0.0064	124.23	0.000	0.78	0.81
α(BU)	0.61	0.0076	79.99	0.000	0.59	0.63
α(CW)	0.49	0.0062	79.37	0.000	0.47	0.51
σ(SW)	1.88	0.0369	50.86	0.000	1.78	1.97
σ(BU)	1.99	0.0367	54.11	0.000	1.89	2.08
σ(CW)	1.98	0.0384	51.53	0.000	1.88	2.08

References

[1] Geoffrey A. Jehle and Philip J. Reny. *Advanced Microeconomic Theory*. 3rd. New York: Pearson, 2011. ISBN: 9780273731917.

[2] Alexander Rosenberg. *Economics: Mathematical Politics or Science of Diminishing Returns*. Chicago: University of Chicago Press, 1992. ISBN: 9780226727233.

[3] Milton Friedman. "The methodology of positive economics". In: *Essays in Positive Economics*. Chicago: University of Chicago Press, 1953, pp. 3–43. ISBN: 0226264033.

[4] Faruk Gul and Wolfgang Pesendorfer. "The case for mindless economics". In: *The Foundations of Positive and Normative Economics: A Handbook*. Ed. by Andrew Caplin and Andrew Schotter. New York: Oxford University Press, 2008, pp. 3–39. ISBN: 9780195328318.

[5] Daniel M. Hausman. "Economic methodology in a nutshell". In: *Journal of Economic Perspectives* 3.2 (1989), pp. 115–127. DOI: 10.1257/jep.3.2.115.

[6] Christina Boyce-Jacino et al. "Large numbers cause magnitude neglect: The case of government expenditures". In: *Proceedings of the National Academy of Sciences* 119.28 (2022), e2203037119. DOI: 10.1073/pnas.220303711.

[7] James Andreoni and John Miller. “Giving according to GARP: An experimental test of the consistency of preferences for altruism”. In: *Econometrica* 70.2 (2002), pp. 737–753. DOI: 10.1111/1468-0262.00302.