To appear in: *Journal of Experimental Social Psychology*, vol. 97

Citation: Turpin, M. H., Walker, A. C., Fugelsang, J. A., Sorokowski, P., Grossmann, I., & Białek, M. (2021). The search for predictable moral partners: Predictability and moral (character) preferences. *Journal of Experimental Social Psychology*, *97*, 104196. https://doi.org/10.1016/j.jesp.2021.104196

The Search for Predictable Moral Partners: Predictability and Moral (Character) Preferences

Martin Harry Turpin[a,*], Alexander C. Walker[a,*], Jonathan A. Fugelsang[a], Piotr Sorokowski[b], Igor Grossmann[a], and Michał Białek[b]

[a] University of Waterloo
Department of Psychology

[b] University of Wroclaw
Institute of Psychology

Author Note

This research was supported by grants from The Natural Sciences and Engineering and Social Sciences and Humanities Research Councils of Canada.

Data from all studies has been made available at the following link: osf.io/xdepr

Correspondence concerning this article should be addressed to Martin Harry Turpin and Alexander C. Walker, Department of Psychology, University of Waterloo, Waterloo, ON, Canada N2L 3G1.

*Notes equal contribution from authors

Email: mhturpin@uwaterloo.ca and a24walke@uwaterloo.ca

Abstract

Across six studies (N = 1,988 US residents and 81 traditional people of Papua), participants judged agents acting in sacrificial moral dilemmas. Utilitarian agents, described as opting to sacrifice a single individual for the greater good, were perceived as less predictable and less moral than deontological agents whose inaction resulted in five people being harmed. These effects generalize to a non-Western sample of the Dani people, a traditional indigenous society of Papua, and persist when controlling for homophily and notions of behavioral typicality. Notably, deontological agents are no longer morally preferred when the actions of utilitarian agents are made to seem more predictable. Lastly, we find that peoples' lay theory of predictability is flexible and multi-faceted, but nevertheless understood and used holistically in assessing the moral character of others. On the basis of our findings, we propose that assessments of predictability play an important role when judging the morality of others.

Keywords: predictability, moral impressions, cooperation, utilitarian, deontology

The Search for Predictable Moral Partners: Predictability and Moral (Character) Preferences

Trade-offs are inevitable in life. There is very rarely a situation where all of one's goals can be simultaneously and completely satisfied. In many instances, sacrifices in one dimension (e.g., money, time, effort) are necessary for achieving a benefit in another (e.g., food, safety, health). Despite their necessity, the inevitability of trade-offs may be unintuitive to grasp and result in observers being harsh in their evaluations of decision makers. For example, following the 1986 Chernobyl nuclear disaster, radioactive clouds were carried in the direction of Moscow, and other densely populated areas in western Russia. Realising that the consequences of radioactive clouds raining over Moscow could cost millions of people their lives, a decision was made to intentionally seed the clouds with silver iodide. This seeding caused the clouds to release their radioactive payload over less populated areas in Belarus, harming the people there, while sparing a far greater number a life-threatening fate. Despite this decision leading to fewer lives being lost, it may be easy to view the decision as cold, immoral, and arbitrary. A similar story is told in health economics, where the condition of scarcity dictates that not everyone can have their health needs met. Decisions must be made as to how treatment is allocated and opting for the most utilitarian option often leaves people angry, even if it results in a greater number of lives saved (Tinghög & Västfjäll, 2018).

Perceptions of Utilitarians and Deontologists

Research in moral psychology and moral philosophy have often used sacrificial moral dilemmas (e.g., the trolley problem; Foot, 1967) as a way to illustrate and assess competing moral principles. Such dilemmas commonly describe a scenario in which a person has the choice of sacrificing a single individual in order to save the lives of many others (e.g., five people). Recent work has begun examining peoples' perceptions of individuals choosing the sacrificial

choice (utilitarian responders) as well as those allowing harm to befall multiple individuals as a result of a refusal to perform a sacrificial act (deontological responders). This work demonstrates that utilitarian responders are often judged to be less warm (Rom et al., 2017), less moral (Uhlmann et al., 2013), more a-moral (Kreps & Monin, 2014), more calculating and competent (Rom et al., 2017), and less trustworthy (Bostyn & Roets, 2017; Everett et al., 2016; Capraro et al., 2018) than those refusing to make the sacrificial choice, specifically in high-conflict dilemmas requiring a direct, as opposed to indirect, sacrificial act[1]. From a utilitarian perspective, such sacrificial actions are done in the service of an impartial process of benefit maximisation (Kahane et al., 2018). Nevertheless, outside observers may come to view utilitarians as mercenary, willing to commit acts of violence as long as some, potentially unknown, conditions are met.

Casual observers may often find it difficult to understand and accept the thought processes that underlie sacrificial decisions (e.g., trading lives for money in the present with the downstream goal of saving an even greater number of lives in the future). One possible reason for peoples' negative view of utilitarian decision-makers is that their actions may appear unpredictable. Although a person who is engaged in utilitarian calculation when making a moral decision has perfect access to all factors relevant to their choice, with their choice (to them) following naturally from impartial rational calculations, such internal calculations and moral reasoning are often inaccessible to third-party observers. Thus, when judging individuals choosing to perform a sacrificial act, it is the act itself that is most readily apparent. As such,

[1] Note that perceptions of utilitarians have been shown to be sensitive to whether their sacrificial choices take place within the context of high- or low-conflict moral dilemmas (Cushman et al., 2006; Sosa & Rios, 2019)

without the contextualizing information of a utilitarian responder's private calculations, the sacrificial actions taken by utilitarians may appear to be performed arbitrarily or whimsically. Conversely, the behavior of deontological agents may appear highly predictable, on account that it is guided by obeying simple and readily interpretable moral rules (e.g., thou shalt not kill). If people possess a moral preference for more predictable actors, as has been suggested by recent empirical work (Walker et al., 2021a), then to the extent that utilitarian responders appear less predictable so too should they appear less moral.

Morality as Cooperation

Cooperation is essential for human flourishing. Almost everything interesting the human species does is the product of multiple brains thinking, interacting, and coexisting together. The ability to access the thoughts and knowledge of others in a vast network of informational exchange transforms us from individuals toiling to eke out subsistence into consumers capable of accessing the collective effort of billions. The moral force of reciprocity underpins this productive cooperative system (Axelrod & Hamilton, 1981; Trivers, 1971). It must be possible to trust an individual to uphold their side of a bargain (at least partially) to benefit from an exchange. If distrust in others grows too great, if we become too uncertain about the thoughts and intentions of others, then the incentive to cooperate is lost and potential cooperators lose the advantage of having access to another's mind. The importance of trust has material consequences as evidenced by a strong association between societal trust and gross-domestic product across several societies (Balliet & Van Lange, 2013). There is good reason to believe that humans have developed a sense for morality for the purpose of ensuring cooperation among conspecifics (Greene, 2013, 2015; Haidt, 2012; Rai & Fiske, 2011; Tomasello & Vaish, 2013) as perhaps best reflected in the theory of "morality-as-cooperation" (Curry, 2016; Curry, Chesters, & Van Lissa,

2019; Curry, Mullins, & Whitehouse, 2019). If morality is fundamentally underpinned by the need to cooperate, then it follows that the ability to predict another's behavior should be paramount in determining moral character. Cooperation relies on the ability to predict what others will do if we choose to cooperate, yet only recently has the link between assessments of predictability and morality become a topic for investigation (Walker et al., 2021a).

A great deal of uncertainty exists when deciding whether to cooperate with others. A person's moral character is unclear when first encountering them. As such, one cannot be certain that the intention to cooperate is present in another's mind (Barrett et al., 2010; FeldmanHall & Shenhav, 2019; Vives & FeldmanHall, 2018). One way of reducing social uncertainty is to establish clear rules that everyone is expected to follow. If everyone is aware of the same rules or norms, then any given member of a society can generally be trusted to be a predictable cooperator. Humans as a species have appeared to adopt far more strategies to enforce cooperation than any other (Melis & Semmann, 2010). From an outside perspective, enforcing seemingly amoral norms (e.g., manner of dress) may seem superfluous or inefficient. However, the benefits in adopting this relatively inflexible system may be that adherence to such norms allows individuals to spend less effort in determining who can be trusted. Conversely, those who are caught violating societal norms—even if for an intended greater good—may very well prompt observers to consider *"what else might that person be willing to do?"*

The Present Research

The inspiration for the current work comes from the synthesis of three ideas. First, cooperation is immensely important for human flourishing (Pinker, 2012; Rand & Nowak, 2013; Ridley, 2010). Second, our sense of morality appears to have evolved for the purpose of supporting cooperation (Curry, 2016; Curry, Chesters, & Van Lissa, 2019; Curry, Mullins, &

Whitehouse, 2019). Lastly, the ability to predict the behavior of others is necessary for supporting productive cooperation (Barrett et al., 2010; FeldmanHall & Shenhav, 2019). Taking these ideas together, we hypothesize that the perceived predictability of an agent influences judgments of their moral character. In Study 1, we investigate the role of predictability in perceptions of moral character, specifically in the context of high-conflict moral dilemmas. Notably, we predict that agents choosing to perform sacrificial acts in high-conflict moral dilemmas will be perceived as less predictable and less moral compared to those who opt not to sacrifice.

Study 1

Method

Participants

A sample of 201 participants (48% Female; M_{age} = 36.74, SD_{age} = 11.69; 16% Asian-American; 11% African-American; 68% European-American; 4% mixed ethnicity) was recruited from Amazon Mechanical Turk and received $0.75 upon completion of a 7-minute online questionnaire. Participants were required to be residents of the United States and possess a Mechanical Turk HIT approval rate greater than or equal to 95% to be eligible to participate. We did not conduct an *a priori* power analysis, however, sensitivity power analysis indicated that with a sample of 201 participants, power of .80, and alpha at .05, we would be able to detect an effect size (Cohen's *d*) of 0.20 or greater for planned paired-samples *t*-tests. All experiments reported in the current study were pre-registered through Open Science Framework (see Open Practices). For all studies, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures used. Studies 1-5 received approval by a University of Waterloo Research Ethics Committee.

Materials

Study 1 featured two vignettes (Trolley and Bomb), both of which described a scenario in which a person had the choice of sacrificing a single individual in order to save the lives of five others. Below each vignette was a statement indicating whether a hypothetical person decided to sacrifice the single individual (Utilitarian actor) or let harm befall five strangers (Deontological actor; see Table 1). A full set of materials for all studies can be viewed in Part A of the supplementary materials.

Table 1

Study 1: Vignette Example

	Trolley Scenario
Vignette	A runaway trolley is heading down the tracks toward five workers who will all be killed if the trolley proceeds on its present course. Michael is on a footbridge over the tracks, in between the approaching trolley and the five workers. Next to him on this footbridge is a stranger who happens to be very large. The only way to save the lives of the five workers is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if Michael does this, but the five workers will be saved.
Deontological Actor	Michael decides not to push the very large stranger off the bridge.
Utilitarian Actor	Michael decides to push the very large stranger off the bridge.

Note. For each trial, one of two names for the hypothetical agent was randomly chosen.

Measures

Following the presentation of each vignette, participants were asked to judge the person who acted within the vignette on various moral dimensions, in line with recommendations that agent perception is the critical question for understanding moral psychology (Białek et al., 2019).

For each item, all measures were presented in a randomized order within a matrix table (see Figure 1).[2]

Moral Perception. Participants' moral perception of each actor was assessed by calculating the mean rating of four moral dimensions (i.e., Goodness, Morality, Peacefulness, and Empathy). Participants judged each actor on these dimensions using 7-point scales with labels surrounding the endpoints of each scale (i.e., Bad/Good, Immoral/Moral, Violent/Peaceful, Merciless/Empathetic). The reliability of this composite was calculated for both actor types and showed excellent reliability (Deontological actor: $\alpha = .91$, Utilitarian actor: $\alpha = .90$).

Predictability. Participants judged the predictability of each actor using a 7-point scale that ranged from "*Unpredictable*" to "*Predictable*."

Harm. Participants indicated how much harm they felt each actor had caused using a 7-point scale that ranged from "*Caused No Harm*" to "*Caused a Great Deal of Harm*."

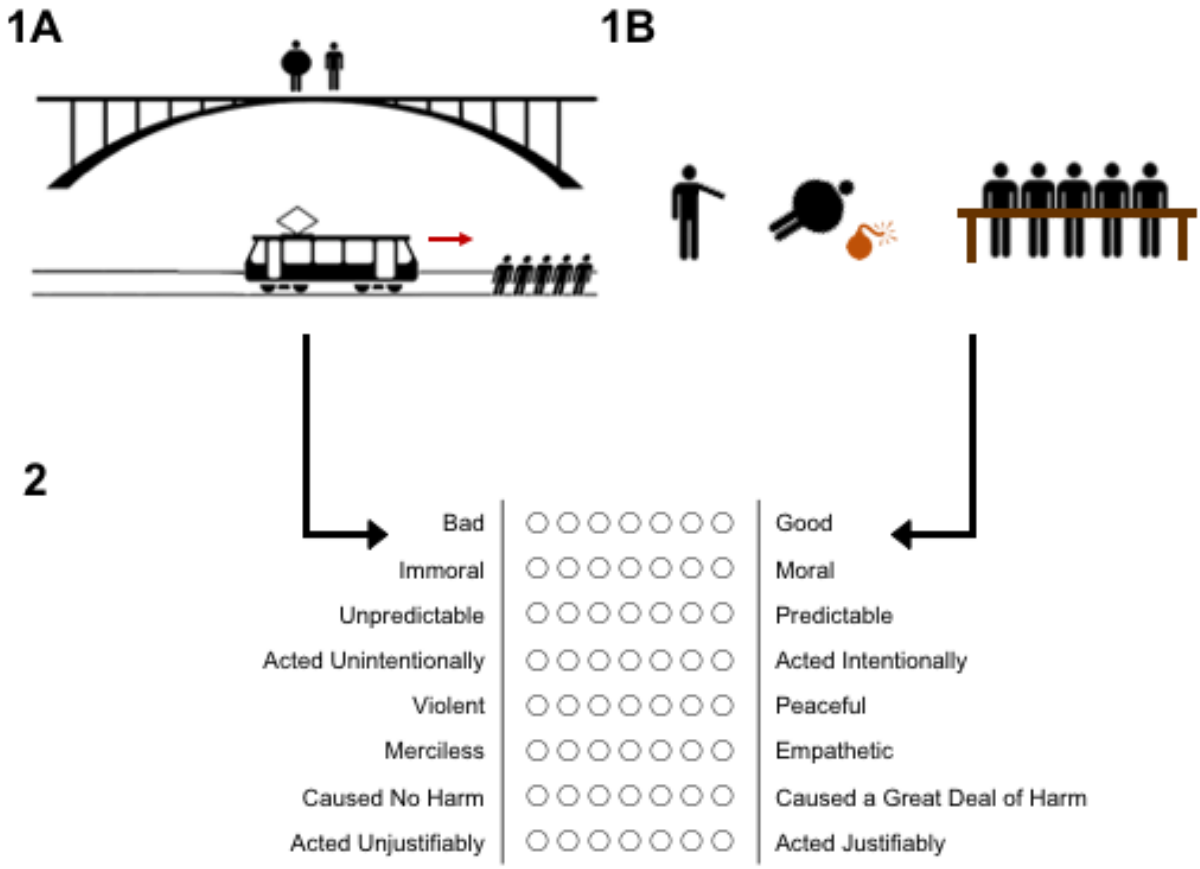
Intentionality. Participants judged how intentional they viewed the actions of each actor using a 7-point scale that ranged from "*Acted Unintentionally*" to "*Acted Intentionally*."

Justifiability. Participants judged the justifiability of each actor's actions using a 7-point scale that ranged from "*Acted Unjustifiably*" to "*Acted Justifiably*."

[2] Note that the moral dimensions of "Intentionality" and "Justifiability" were collected primarily for exploratory purposes. Analyses featuring these variables have largely been excluded from the main body of the manuscript yet can be viewed in Part B of the supplementary materials.

Figure 1

Study 1: Methodology Infographic



Note. Participants were presented with two vignettes (Trolley and Bomb) in a randomized order. One vignette described a hypothetical actor who decided to let harm befall five people in order to avoid sacrificing a single individual (1A) and the other described a hypothetical actor choosing to perform this sacrifice (1B). Participants' task was to evaluate each actor on eight moral dimensions presented in a randomized order within a matrix table (2).

Design and Procedure

Study 1 utilized a within-subjects design in which all participants were presented with vignettes featuring a runaway trolley and a bomb in a restaurant. Within one vignette (Trolley or Bomb), participants evaluated a hypothetical actor who decided to sacrifice a single individual in order to save five others (i.e., a utilitarian actor), and in the other, evaluated an actor who decided

to let harm come upon five people in order to avoid killing a single individual (i.e., a deontological actor; see Figure 1). Following participants' moral judgments of each actor, they concluded the experiment by responding to two demographic questions (i.e., age and gender).

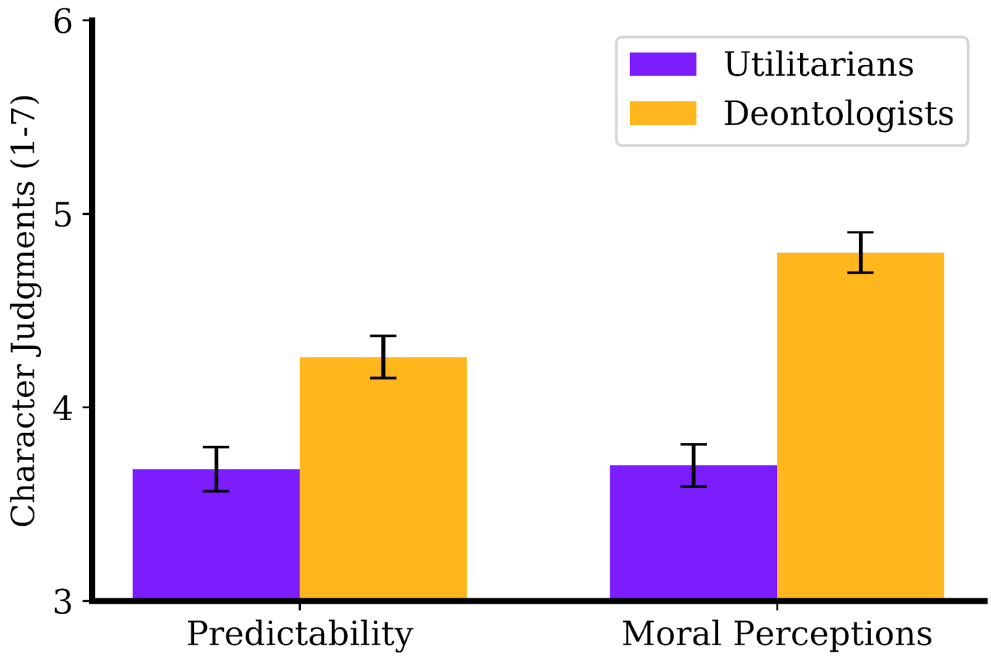
Results and Discussion

To assess our hypothesis that deontological actors would be judged as more predictable and more moral compared to utilitarian actors, we conducted paired-samples *t*-test[3] comparing judgments of predictability, moral character, and harm across deontological and utilitarian actors (see Figure 2). As predicted, deontological actors ($M = 4.26$, $SD = 1.54$) were judged to be more predictable than utilitarian actors ($M = 3.68$, $SD = 1.62$), $t(200) = 4.37$, $p < .001$, $d = 0.37$, 95% *CI* [0.17, 0.56]. Additionally, deontological actors ($M = 4.80$, $SD = 1.48$) were perceived as more moral compared to utilitarian actors ($M = 3.70$, $SD = 1.45$), $t(200) = 7.77$, $p < .001$, $d = 0.75$, 95% *CI* [0.55, 0.95]. Furthermore, despite the actions of utilitarian actors resulting in *fewer* lives lost, utilitarian actors were judged as causing *more* harm ($M = 5.10$, $SD = 1.44$) compared to deontological actors ($M = 4.05$, $SD = 1.89$), $t(200) = 5.99$, $p < .001$, $d = 0.62$, 95% *CI* [0.42, 0.82]. Lastly, we report the zero-order correlations between all key variables assessed in Study 1 (see Table 2). Most notably, participants' judgments of predictability were positively associated with their judgments of morality when judging both deontological and utilitarian actors.

[3] Across all studies, all post-hoc comparisons survive correction for multiple comparisons.

Figure 2

Study 1: Judgments of Predictability and Moral Perceptions



Note. Error bars represent +/- 1 standard error.

Table 2

Study 1: Zero-order Correlations

	1	2	3	4
1. Predictability	-	-.26***	-.09	.44***
2. Harm	-.18*	-	.19**	-.51***
3. Intentionality	-.13	.18*	-	-.04
4. Moral Perceptions	.38***	-.50***	-.14*	-

Note. Pearson correlations ($N = 201$). Bottom diagonal = Utilitarian actor. Top diagonal = Deontological actor. *** $p < .001$, ** $p < .01$, * $p < .05$.

Study 2

A possibility for why utilitarians are judged to be less predictable and moral compared to deontological agents is that the principles guiding deontologists' actions are clear and discoverable for observers (e.g., inflexibly follow rules), while the principles guiding utilitarian action may be less obvious (e.g., calculate subjective utility and act to maximize it)[4]. When reading a description of a utilitarian actor, people may think of a person acting whimsically or arbitrarily, deciding to sacrifice a person because they felt like it in the moment. It may therefore seem difficult to predict their behavior as there may appear to be no underlying principles guiding their actions. In contrast, a deontological actor may seem highly predictable as they appear to follow well-known moral rules. If an observer can be made to understand the principles guiding utilitarian action, they may find it easier to predict their future behaviors. Thus, agents that may have previously appeared as unpredictable violators of moral norms (e.g., do not kill) may—with transparent moral principles—appear as trustworthy cooperators, predictably causing harm when it helps achieve a greater good. As shown in Study 1, predictability judgments were positively associated with judgments of moral character. If explaining utilitarian principles results in utilitarians appearing more predictable, one may expect the gap in moral impressions between deontological and utilitarian actors to be reduced, if not eliminated. Similarly, if deontological agents can be made to appear to be acting capriciously, or "following their gut" rather than strictly adhering to moral rules, it may undermine their perceived predictability and consequently their perceived morality.

[4] A perfect utilitarian values every life equally, as in, they are impartially beneficent (Kahane et al., 2018). This feature of utilitarian thought is not a perfect correlate to the dimension we use in the current work which is that of instrumental harm. This adds a further layer of unpredictability as it is not possible to infer strictly from behavior the idiosyncratic weightings given to each life (see Cohen & Ahn, 2016).

Alternatively, it may be that it is the *behavior itself* that introduces an element of unpredictability such that, regardless of whether a sacrificial act is guided by stable moral principles, the decision to directly kill someone—even for a greater good—signals unpredictability. If this is the case, then explaining the underlying principles guiding utilitarian action should have no effect on perceptions of utilitarian agent's predictability nor morality.

Method

Participants

A sample of 901 US residents (50% Female; M_{age} = 38.51, SD_{age} = 12.71; 11% Asian-American; 9% African-American; 77% European-American; 2% mixed ethnicity) were recruited from Amazon Mechanical Turk using the same recruiting criteria as Study 1. Participants received $0.50 for completion of a 5-minute online questionnaire. Those who participated in Study 1 were restricted from participating in Study 2. We decided to significantly increase the size of our sample to accommodate switching to a between-subjects design. We once again did not conduct an *a priori* power analysis, however, a sensitivity power analysis indicated that with a sample of 901 participants, power of .80, and alpha at .05, we would be able to detect a minimum effect size of $\eta^2 = .010$ for planned factorial ANOVAs.

Materials

The vignettes used in Study 1 were once again used in Study 2. However, in Study 2, the described actions of each hypothetical actor were further explained by either a principled explanation, an emotional explanation, or no explanation (see Table 3).

Table 3

Study 2: Explanation examples

	Principled Explanation	Emotional Explanation	No Explanation
Deontological Actor	… because he believes one should never choose to kill another person regardless of the possible benefits	… because his gut feelings told him that it was the right thing to do	…
Utilitarian Actor	… because he believes one can choose to kill another person if the value of the lives saved outweigh the value of the lives sacrificed.	… because his gut feelings told him that it was the right thing to do.	…

Note. All text presented in this table appeared following a description of a hypothetical agent's action (e.g., "Michael decides to push the very large stranger off the bridge").

Measures

The measures used in Study 2 were identical to those used in Study 1. As in Study 1, we computed moral perception scores for each participant by computing the mean of their Good, Moral, Peaceful, and Empathetic judgments. Once again, the reliability of this composite was calculated ($\alpha = .92$) and demonstrated excellent reliability.

Design and Procedure

Study 2 utilized a between-subjects design for which each participant was randomly assigned to one of three explanation conditions (Principled explanation, Emotional explanation, or No explanation). All participants were presented with a single vignette in which they evaluated a person making a decision in a moral dilemma. Within each explanation condition,

participants had an equal chance of evaluating one of four possible vignettes (Scenario: Trolley or Bomb; Actor Type: Deontological or Utilitarian).

Results and Discussion

We assessed the degree to which explaining the underlying principles behind a moral act influences judgments of predictability and morality by comparing participants' ratings of predictability, moral perceptions, and harm using three, 3 (Explanation: Principled explanation, Emotional explanation, No explanation) x 2 (Actor Type: Deontological, Utilitarian) Factorial ANOVAs. We report the results of each ANOVA independently below. Additionally, we conducted correlational analyses to examine the zero-order associations between our key variables (see Table 4).

Table 4

Study 2: Zero-Order Correlations

	1	2	3	4
1. Predictability	-			
2. Harm	-.30***	-		
3. Intentionality	.03	.08*	-	
4. Moral Perceptions	.53***	-.59***	.02	-

Note. Pearson correlations (N = 901). *** $p < .001$, ** $p < .01$, * $p < .05$.

Predictability

We observed a main effect of explanation type, $F(2, 895) = 5.04$, $p = .007$, $\eta_p^2 = .011$, such that predictability ratings differed depending on whether a participant received a principled (M = 4.45, SD = 1.60), emotional (M = 4.17, SD = 1.65), or no explanation (M = 4.06, SD =

1.69) for an agent's action. However, this effect was small, suggesting that the principles (or lack of principles) guiding an agent's actions did not exert a large influence on assessments of their predictability (e.g., compared to the actions of agents). We also observed a main effect of actor type, $F(1, 895) = 93.85$, $p < .001$, $\eta_p^2 = .095$. Consistent with Study 1, deontological actors (M = 4.73 SD = 1.45) were perceived as more predictable compared to utilitarian actors (M = 3.72, SD = 1.69). No explanation by actor type interaction was detected, $F(2, 895) = 0.10$, *ns*.

Our hypothesis mainly focused on whether the lack of perceived predictability of utilitarian actors was due to ignorance of utilitarian principles. Therefore, we tested whether the predictability of utilitarian actors varied as a function of explanation type. The results of a follow-up One-Way ANOVA showed little evidence of a difference between explanation conditions for utilitarian actors, $F(2, 446) = 2.83$, $p = .060$, $\eta_p^2 = .013$. Thus, explaining utilitarian principles appeared to have no—or only a small—effect on their perceived predictability.

Moral perceptions

Consistent with Study 1, we observed a main effect of actor type, $F(1, 895) = 445.23$, $p < .001$, $\eta_p^2 = .332$, such that deontological actors (M = 5.42, SD = 1.20) were judged to be more moral than utilitarian actors (M = 3.58, SD = 1.42). No main effect of explanation type, $F(2, 895) = 1.87$, $p = .155$, $\eta_p^2 = .004$, nor an explanation type by actor type interaction was found, $F(2, 895) = 1.45$, $p = .236$, $\eta_p^2 = .003$. Therefore, whether a principled, emotional, or no explanation was given for an agent's action appeared to have no impact on the perceived morality of that agent.

Harm

In line with Study 1, we observed a main effect of actor type, $F(1, 895) = 211.56$, $p < .001$, $\eta_p^2 = .191$, such that utilitarian actors ($M = 5.35$, $SD = 1.44$) were judged to have caused more harm compared to deontological actors ($M = 3.68$, $SD = 1.95$), despite their actions resulting in fewer lives lost. The main effect of explanation type, $F(2, 895) = 0.69$, $p = .500$, $\eta_p^2 = .002$, and explanation type by actor type interaction were not significant, $F(2, 895) = 1.07$, $p = .345$, $\eta_p^2 = .002$.

Study 3

Explaining utilitarian principles had at most a small effect on how utilitarians were perceived with regards to their predictability and moral character. It may be the case that it is the sacrificial behaviors themselves that observers associate with unpredictability, rather than the principles which guide these actions. Most individuals, when presented with a high-conflict moral dilemma may find the option to directly kill another person unimaginable. As such, a person who chooses to perform this sacrificial act, even if for the greater good, may at once seem alien and unpredictable.

A common finding is that people are more willing to endorse sacrificial options in low- as opposed to high-conflict dilemmas (Bruers & Braeckman, 2014; Hauser et al., 2008; Waldmann & Dieterich, 2007). If the sacrificial option presented in high-conflict dilemmas (e.g., pushing a man onto trolley tracks) is unattractive to participants when making the decision themselves, a third party deciding to act in this manner may appear especially unpredictable. In contrast, sacrificial options presented in low-conflict dilemmas (e.g., pulling a track-switch) may seem obvious to observers and so a third party deciding to perform such actions may be less likely to signal unpredictability despite potentially being guided by the same utilitarian principles.

In Study 3, we added low-conflict moral dilemmas to test the hypothesis that utilitarian actors will be judged as more moral when the behavior used to satisfy their moral principles appears more predictable. This hypothesis follows naturally from two overarching principles which permeate this work. First, predictability and morality are positively linked, if it is a known finding that perceptions of morality change depending on the intensity of conflict (Cushman et al., 2006; Sosa & Rios, 2019), then we are naturally committed to hypothesizing that predictability will also display a reversal once examined. Second, we believe that for most people, directly murdering someone to meet a moral goal is far less obvious than initiating a mechanical process which indirectly results in the death of a person. There is precedent for this assumption as it has been demonstrated that cognitive load affects utilitarian, but not deontological responding in high-conflict moral dilemmas (Greene et al., 2008). While this does not guarantee that the deontic solution is necessarily more obvious, it suggests that the utilitarian response requires working memory, while the deontic choice does not. Additionally, peoples' greater endorsement of the sacrificial choice in low-conflict moral dilemmas (Bruers & Braeckman, 2014; Hauser et al., 2008; Waldmann & Dieterich, 2007) further suggests that the utilitarian response may be perceived as more predictable than the deontological response within this context. In line with this thinking, we predicted a reversal for participants' judgments of predictability and morality across low- and high-conflict dilemmas. That is, we hypothesized that participants would judge utilitarian actors as more predictable and more moral in low-conflict dilemmas and judge deontological actors as more predictable and moral in high-conflict dilemmas. Such a finding would highlight the importance of the predictability of an actor's behavior over the underlying moral philosophies informing such behaviors, when judging the moral character of others.

Method

Participants

A sample of 500 US residents was recruited from Amazon Mechanical Turk (9% Asian-American; 8% African-American; 79% European-American; 5% mixed ethnicity) using the same recruiting criteria as Studies 1 and 2. Participants received $0.50 upon completion of a 5-minute online questionnaire. Those who participated in Studies 1 or 2 were restricted from participating in Study 3. We excluded data from 145 participants who failed to pass two comprehension check questions, leaving data from 355 participants (43% Female; M_{age} = 39.03, SD_{age} = 12.14) to be analysed.[5] Sensitivity power analyses indicated that with a sample of 355 participants, power of .80, and alpha at .05, we would be able to detect a minimum effect size of η^2 = .017 for planned mixed factorial ANOVAs and d = 0.17 for follow-up paired samples t-tests.

Materials

Study 3 presented participants with either high- or low-conflict moral dilemmas depending on their randomly assigned condition. High-conflict dilemmas were identical to those used in Study 1 (see Table 1). Low-conflict dilemmas mirrored high-conflict dilemmas with the exception that the sacrificial option presented in low-conflict dilemmas involved an indirect (e.g., throwing a bomb on a patio near one individual) as opposed to direct harm (e.g., throwing a person on a bomb; see Table 5). As in Study 1, each vignette appeared with a statement indicating whether a hypothetical actor decided to sacrifice a single individual (Utilitarian actor) or let harm befall five strangers (Deontological actor).

[5] The results reported below do not depend on these exclusions. That is, the interpretation of all significance tests remains the same when analyzing our full sample.

Table 5

Study 3: Low-Conflict Vignette Example

	Trolley Scenario
Low-Conflict Vignette	A runaway trolley is heading down the tracks toward five workers who will all be killed if the trolley proceeds on its present course. Michael is standing at a switch. If he turns the switch the trolley will be diverted to a side track, but there is one stranger on this side track. The only way to save the lives of the five workers is to turn the switch and divert the trolley to the side track. The stranger will die if Michael does this, but the five workers will be saved.
Deontological Actor	Michael decides not to turn the switch.
Utilitarian Actor	Michael decides to turn the switch.

Note. For each trial, one of two names for the hypothetical agent was randomly chosen.

Measures

Study 3 used the same set of measures used in Studies 1 and 2 with the exception that we no longer asked participants to evaluate the justifiability of each agent's actions.

Design and Procedure

Study 3 used a mixed design in which conflict condition (high or low-conflict dilemmas) acted as a between-subjects factor and actor type (Deontological and Utilitarian) as a within-subjects factor. Participants were randomly assigned to either the high or low-conflict condition, for which they provided their moral judgments of hypothetical agents acting within either high or low-conflict moral dilemmas. All participants were presented with two vignettes (Trolley and Bomb) in which they evaluated a person who sacrificed a single individual to save five (Utilitarian actor) as well as a person whose inaction resulted in five individuals being harmed (Deontological actor). The order of both vignette type and actor type was counterbalanced. Following each moral evaluation, participants responded to a comprehension check item which assessed the degree to which they attended to the action described in the previous vignette.

Results and Discussion

To test whether a reversal in moral judgments would take place across high- and low-conflict dilemmas, we conducted 2 (Actor Type: Deontological vs. Utilitarian, within-subjects) x 2 (Conflict Type: High vs. Low, between-subjects) mixed factorial ANOVAs on judgments of predictability, moral perceptions, and harm. We report the results of each ANOVA independently below. Additionally, we conducted correlational analyses to examine the zero-order associations between our key variables within both High- and Low-Conflict conditions (see Tables 6 and 7).

Table 6

Study 3: Low-Conflict Condition Zero-Order Correlations

	1	2	3	4
1. Predictability	-	-.24**	.13	.41***
2. Harm	-.12	-	.30***	-.44***
3. Intentionality	.20**	.12	-	.04
4. Moral Perceptions	.50***	-.26***	.19*	-

Note. Pearson correlations (N = 177). Bottom diagonal = Utilitarian actor. Top diagonal = Deontological actor. *** $p < .001$, ** $p < .01$, * $p < .05$.

Table 7

Study 3: High-Conflict Condition Zero-Order Correlations

	1	2	3	4
1. Predictability	-	-.18*	.17*	.33***
2. Harm	-.14	-	-.26**	-.52***
3. Intentionality	-.06	.16*	-	.31***
4. Moral Perceptions	.46***	-.59***	-.06	-

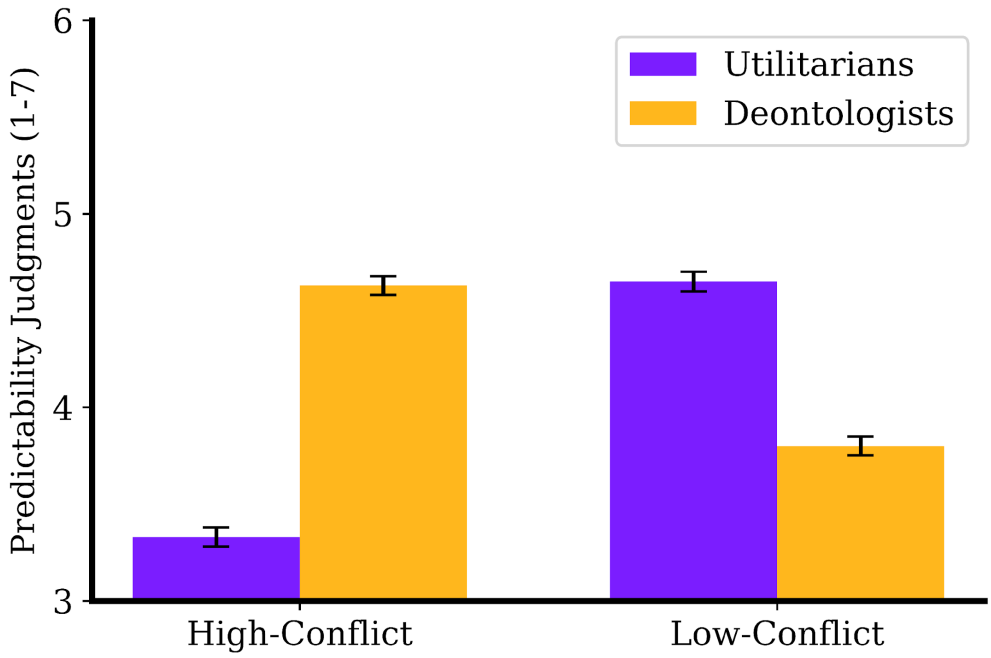
Note. Pearson correlations (N = 178). Bottom diagonal = Utilitarian actor. Top diagonal = Deontological actor. *** $p < .001$, ** $p < .01$, * $p < .05$.

Predictability

We observed a main effect of conflict type, $F(1, 353) = 5.88$, $p = .016$, $\eta_p^2 = .016$, such that agents in low-conflict dilemmas ($M = 4.23$, $SD = 1.36$) were judged to be more predictable than those in high-conflict dilemmas ($M = 3.98$, $SD = 1.34$). The main effect of actor type failed to reach statistical significance, $F(1, 353) = 3.78$, $p = .053$, $\eta_p^2 = .011$. The main effect of conflict type was qualified by a significant actor by conflict type interaction, $F(1, 353) = 84.14$, $p < .001$, $\eta_p^2 = .192$ (see Figure 3). Follow-up paired-samples t-tests showed a reversal in predictability judgments between utilitarian and deontological agents based on conflict type. Utilitarian agents ($M = 4.65$, $SD = 1.52$) were judged to be more predictable in low-conflict dilemmas than deontological agents ($M = 3.80$, $SD = 1.44$), $t(176) = 5.35$, $p < .001$, $d = 0.57$, 95% CI [0.36, 0.79], and less predictable ($M = 3.33$, $SD = 1.46$) than deontological agents ($M = 4.63$, $SD = 1.45$) in high-conflict dilemmas, $t(177) = 7.54$, $p < .001$, $d = 0.89$, 95% CI [0.67, 1.11].

Figure 3

Study 3: Judgments of Predictability



Note. Error bars represent +/- 1 standard error.

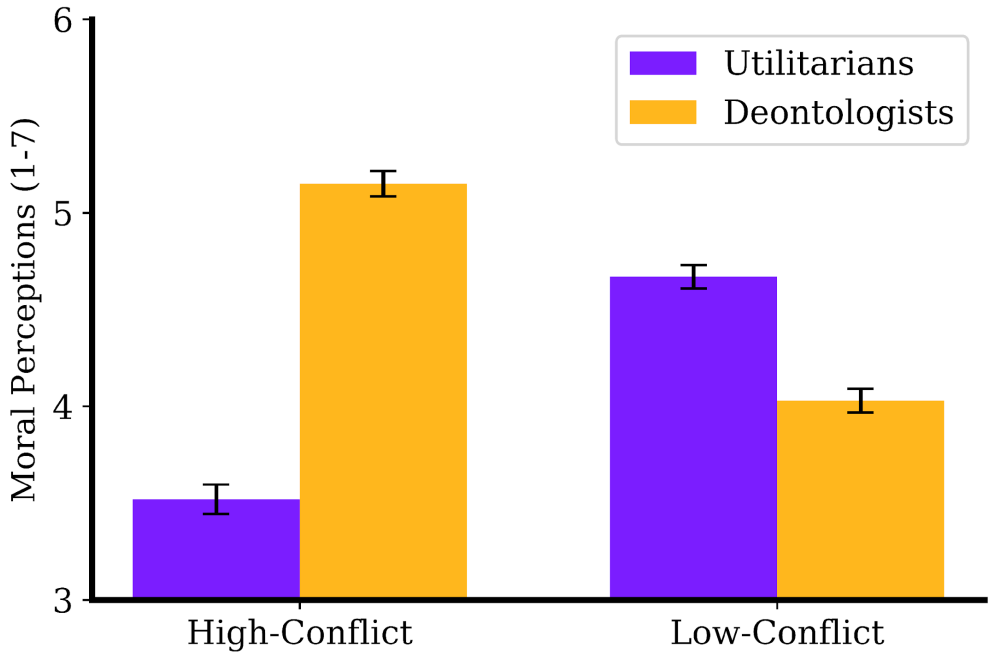
Moral perceptions

We observed a main effect of actor type, $F(1, 353) = 20.84$, $p < .001$, $\eta_p^2 = .056$, such that across high and low-conflict dilemmas, utilitarian actors ($M = 4.10$, $SD = 1.41$) were judged to be less moral compared to deontological actors ($M = 4.59$, $SD = 1.32$). The main effect of conflict type was not significant, $F(1, 353) = 0.04$, *ns*. The main effect of actor type was qualified by a significant actor by conflict type interaction, $F(1, 353) = 110.44$, $p < .001$, $\eta_p^2 = .238$ (see Figure 4). Follow-up paired-samples *t*-tests showed a reversal in moral perceptions between utilitarian and deontological agents based on conflict type. Utilitarian agents ($M = 4.67$, $SD = 1.14$) were perceived as more moral than deontological agents ($M = 4.03$, $SD = 1.17$) in

low-conflict dilemmas, $t(176) = 4.66$, $p < .001$, $d = 0.55$, 95% *CI* [0.34, 0.77], and less moral ($M = 3.52$, $SD = 1.43$) than deontological agents ($M = 5.15$, $SD = 1.22$) in high-conflict dilemmas, $t(177) = 9.79$, $p < .001$, $d = 1.23$, 95% *CI* [1.00, 1.45].

Figure 4

Study 3: Judgments of Moral Perceptions



Note. Error bars represent +/- 1 standard error.

Harm

We observed a main effect of actor type, $F(1, 353) = 23.03$, $p < .001$, $\eta_p^2 = .061$, such that utilitarian actors ($M = 4.83$, $SD = 1.44$) were judged to have caused more harm than deontological actors ($M = 4.18$, $SD = 2.04$) across high and low-conflict dilemmas. The main effect of conflict type was not significant, $F(1, 353) = 3.45$, $p = .064$, $\eta_p^2 = .010$. The main effect of actor type was qualified by a significant actor by conflict type interaction, $F(1, 353) = 51.76$, p

$< .001$, $\eta_p^2 = .128$. Unlike predictability and moral perceptions, no reversal took place for the amount of harm done by utilitarian and deontological agents. Instead, we did not find convincing evidence that utilitarians ($M = 4.45$, $SD = 1.41$) were perceived as causing less harm compared to deontological agents ($M = 4.77$, $SD = 1.89$) in low-conflict dilemmas, $t(176) = 1.82$, $p = .070$, $d = 0.19$, 95% *CI* [-0.02, 0.40]. Consistent with Studies 1 and 2, utilitarians ($M = 5.20$, $SD = 1.37$) were judged as causing more harm in high-conflict dilemmas than deontologists ($M = 3.58$, $SD = 2.01$), $t(177) = 7.97$, $p < .001$, $d = 0.94$, 95% *CI* [0.72, 1.16], despite their actions resulting in fewer lives lost.

Study 4

The results of Studies 1-3 are consistent with the idea that the perceived predictability of agents—specifically those acting within sacrificial moral dilemmas—influences judgments of their moral character. Nevertheless, what exactly participants imagined when judging the "predictability" of agents in Studies 1-3 remains unclear, as predictability as a common language word may evoke multiple concepts simultaneously[6] (e.g., behavioral consistency). To better understand the role of predictability in moral judgments it is necessary to find out what "predictability" means to participants. To address this question, we examined the extent to which participants' assessments of the predictability of utilitarian and deontological agents acting within high-conflict moral dilemmas were captured by various candidate predictability-related concepts.

[6] In fact, other work demonstrates how various factors (e.g., ambiguity) can impact the specific concepts brought to mind by a behavioral description, as well as bias moral judgments made on the basis of these descriptions (Fausey & Boroditsky, 2010; Walker et al., 2021b).

Method

Participants

A sample of 300 participants (41% Female; M_{age} = 38.77, SD_{age} = 11.47 8% Asian-American; 8% African-American; 75% European-American; 5% of Latin, Central, or South American origins; 2% Other) was recruited from Amazon Mechanical Turk. Each participant received $1.20 upon completion of an 8-minute online questionnaire. Participants were required to be residents of the United States and possess a Mechanical Turk HIT approval rate greater than or equal to 99%. In order to further ensure data quality, participants in Study 4 were required to correctly respond to two simple bot detection questions (see supplementary materials Part A). Sensitivity power analyses indicated that with a sample of 300 participants, power of .80, and alpha at .05, we would be able to detect a minimum effect size of r = .16 for correlational analyses and d = 0.19 for conducted paired samples t-tests.

Materials

The scenarios used in Study 4 were identical to those used in Study 1. That is, Study 4 featured two vignettes (Trolley and Bomb), both of which described a high-conflict moral dilemma in which a person had the choice of sacrificing a single individual to save the lives of five others. As in Studies 1-3, the actions of a hypothetical person were described below each vignette. Specifically, participants were presented with and judged a person who, within a high-conflict moral dilemma, sacrificed a single individual (Utilitarian actor) as well as a person who, within the same dilemma, let harm befall five strangers (Deontological actor; see Table 1).

Measures

As in Studies 1-3, participants were asked to judge hypothetical agents who acted within a high-conflict vignette on various dimensions. However, in order to assess the relation between judgments of various predictability-related concepts and judgments of predictability and morality, participants in Study 4 judged each agent on an expanded list of moral and non-moral dimensions. Once again, for each vignette, all measures were presented in a randomized order within a matrix table.

Predictability and Related Concepts. As in Studies 1-3, participants judged the predictability of each actor using a 7-point scale that ranged from "*Unpredictable*" to "*Predictable*." Additionally, in Study 4, participants also judged each agent on four additional dimensions, each of which represented a specific plausible concept which assessments of "predictability" may bring to mind. Specifically, participants judged the expected consistency of each actors' behavior ("*Expected to behave inconsistently*" to "*Expected to behave consistently*"), the intelligibility of their motivations ("*Hard to understand motivations"* to "*Easy to understand motivations*"), their reliability ("*Unreliable*" to "*Reliable*"), and the degree to which they were chaotic or methodical ("*Chaotic*" to "*Methodical*"). Mirroring judgments of predictability, each additional dimension was judged using a 7-point scale.

Moral Perceptions. Participants' moral perceptions of each actor presented in Study 4 was assessed by calculating the mean rating of five moral dimensions (i.e., Goodness, Morality, Peacefulness, Empathy, and Benevolence[7]). As in Studies 1-3, participants judged each actor on these dimensions using 7-point scales with labels surrounding the endpoints of each scale (i.e., Bad/Good, Immoral/Moral, Violent/Peaceful, Merciless/Empathetic, Threatening/Benevolent).

[7] "Benevolence" was included as part of our moral perceptions composite exclusively in Study 4 in order to match the number of moral dimensions presented (i.e., five) with that of our predictability and non-moral categories.

The reliability of this composite was calculated for both actor types and showed excellent reliability (Deontological actor: $\alpha = .91$, Utilitarian actor: $\alpha = .92$).

Non-Moral Dimensions. Participants also judged each actor on five additional dimensions (i.e., Status, Power, Confidence, Conventionality, and Wealth) designed to be at most weakly associated with the concepts of predictability and morality. These dimensions were included in order to dilute the salience of predictability and morality related terms as well as to allow for the strength of hypothesized associations (e.g., between predictability and predictability-related dimensions) to be compared with "baseline" associations featuring "unrelated" traits (e.g., the association between predictability and confidence). Consistent with all other measures, participants judged each actor on all five non-moral dimensions using 7-point scales with labels surrounding the endpoints of each scale (i.e., Low-status/High-status, Powerless/Powerful, Unassertive/Confident, Alternative/Conventional, Poor/Wealthy).

Design and Procedure

Study 4 featured an identical design and procedure to that of Study 1.

Results and Discussion

We examined the extent to which each of our hypothesized predictability-related concepts (i.e., consistency, intelligibility, reliability, and methodicalness) were associated with judgments of predictability (see Table 8). We pre-registered the intent to treat differences between correlation coefficients of .1, .2, and .3 as corresponding to small, medium, and large effect sizes of interest (Funder & Ozer, 2019). For judgments of utilitarian actors, we did not observe an effect size of interest when examining differences in correlation coefficients between judgments of predictability and predictability-related concepts. All predictability-related concepts demonstrated a strong association with judgments of predictability, $r(298) > .53$, $p <$

.001, suggesting that assessments of predictability, at least in this context, constitute a complex, multi-faceted idea that is difficult to distill down to a single concept (e.g., consistency of behavior). Next, evaluating judgments of deontological actors, we observed small differences of .13 or greater when comparing the correlation coefficient of predictability and consistency with the correlation coefficients of predictability and intelligibility, reliability, and methodicalness. Thus, in the context of judging deontological actors within high-conflict moral dilemmas, it appears that assessments of an agent's predictability are most strongly indexed by perceptions of the consistency of their behavior. Nevertheless, consistent with judgments of utilitarian actors, all predictability-related concepts demonstrated a strong association with judgments of predictability, $r(298) > .50$, $p < .001$. Thus, assessments of deontological actors' predictability, while most strongly indexed by notions of consistency, may still represent a complex and multi-faceted judgment.

Table 8

Study 4 Correlations: Moral Perceptions, Predictability, and Predictability-Related Concepts

	1	2	3	4	5	6
1. Moral Perceptions	-	.56***	.70***	.61***	.66***	.53***
2. Predictability	.55***	-	.67***	.51***	.54***	.51***
3. Consistency	.64***	.63***	-	.61***	.68***	.57***
4. Intelligibility	.60***	.54***	.61**	-	.56***	.52***
5. Reliability	.73***	.60***	.67***	.58***	-	.51***
6. Methodicalness	.65***	.60***	.64***	.58***	.64***	-

Note. Pearson correlations ($N = 300$). Bottom diagonal = Utilitarian actor. Top diagonal = Deontological actor. *** $p < .001$, ** $p < .01$, * $p < .05$.

Next, we assessed the relation between judgments of each predictability-related concept and judgments of predictability using simple linear regression. Specifically, we conducted two linear regressions (one for each actor type) predicting judgments of predictability using judgments of consistency, intelligibility, reliability, and methodicalness (see Tables 9A and 9B). Analyzing judgments of utilitarian actors, we find that an overall model regressing participants' predictability judgments on all predictability-related concepts was significant, $F(4, 295) = 71.80$, $p < .001$; $R^2 = 0.49$; and all predictability-related concepts explained variance in judgments of predictability (as shown in Table 9A), once again suggesting that assessments of predictability in this context are multi-faceted. Furthermore, analyzing judgments of deontological actors, $F(4, 295) = 70.50$, $p < .001$; $R^2 = 0.49$ (see Table 9B), we find that only judgments of consistency, $b = 0.47$, $p < .001$, and methodicalness, $b = 0.16$, $p < .001$, were significant predictors. Notably, consistent with our analyses featuring zero-order correlations, judgments of deontological actors' consistency were found to be the strongest predictor of judgments of their predictability. Thus, when judging people opting against the sacrificial choice in a high-conflict moral dilemma, judgments of an agent's predictability appear to be most strongly indexed by the perceived consistency of their behavior.

Table 9A

Regression analysis, judgments of predictability as predicted by all other candidate "predictability-related" concepts. Utilitarian agents.

Effect	*b*	*SE*	95% CI		*p*
Utilitarian agents – Predictability			*LL*	*UL*	
Intercept	0.19	0.21	-0.23	0.60	.377
Consistency	0.29	0.06	0.16	0.41	<.001
Reliability	0.21	0.06	0.08	0.33	.001
Intelligibility	0.11	0.05	0.01	0.22	.033
Methodicalness	0.20	0.05	0.08	0.29	<.001

Note. Total $N = 300$. *CI* = confidence interval; *LL* = lower limit; *UL* = upper limit, *b* = estimate of effect size.

Table 9B

Regression analysis, judgments of predictability as predicted by all other candidate "predictability-related" concepts. Deontological agents.

Effect	*b*	*SE*	95% CI		*p*
Deontological agents - Predictability			*LL*	*UL*	
Intercept	0.52	0.27	-0.00	1.04	.051
Consistency	0.47	0.06	0.34	0.60	<.001
Reliability	0.11	0.06	-0.01	0.23	.086
Intelligibility	0.09	0.05	-0.01	0.20	.092
Methodicalness	0.16	0.06	0.04	0.27	.007

Note. Total $N = 300$. *CI* = confidence interval; *LL* = lower limit; *UL* = upper limit *b* = estimate of effect size.

We also examined the extent to which judgments of non-moral dimensions (i.e., status, power, confidence, conventionality, and wealth), hypothesized to be at most weakly associated with judgments of predictability and morality, were correlated with assessments of predictability (see Table 10). First, for judgments of both utilitarian and deontological actors, judgments of conventionality were shown to be positively associated with judgments of predictability, $r(298) = .51$, $p < .001$, in both cases. These results suggest that conventionality may be better

categorized as a predictability-related concept, as opposed to being included as a non-moral dimension[8]. Nevertheless, correlations between judgments of all other non-moral dimensions and predictability were found to be at most small, $r(298) < .25$ (see Table 10), with the majority showing no significant association. Notably, comparisons between the strength of the associations between predictability and predictability-related dimensions and predictability and non-moral dimensions revealed large effects across all possible comparisons that excluded conventionality (which showed small to no effects). Thus, these analyses suggest that the predictability-related concepts featured in Study 4 were considerably more strongly associated with judgments of predictability than most non-moral dimensions.

Table 10

Study 4 Correlations: Moral Perceptions, Predictability, Non-Moral Dimensions

	1	2	3	4	5	6	7
1. Moral Perceptions	-	.56***	.18**	-.06***	.23***	.47***	-.08
2. Predictability	.55***	-	.11*	-.05	.10*	.51***	.02
3. Status	.34***	.24***	-	.26***	.33***	.12***	.44***
4. Power	.16*	-.02	.36***	-	.56***	-.21***	.30***
5. Confidence	.23***	.08*	.36***	.48***	-	-.04	.21***
6. Conventionality	.48***	.51***	.19***	.06	-.02	-	-.01
7. Wealth	.19***	.12*	.69***	.31***	.29***	.17**	-

Note. Pearson correlations ($N = 300$). Bottom diagonal = Utilitarian actor. Top diagonal = Deontological actor. *** $p < .001$, ** $p < .01$, * $p < .05$.

[8] It should be considered an oversight that "conventionality" was not a priori included as a predictability-related concept, as it clearly shares features with other concepts in that set.

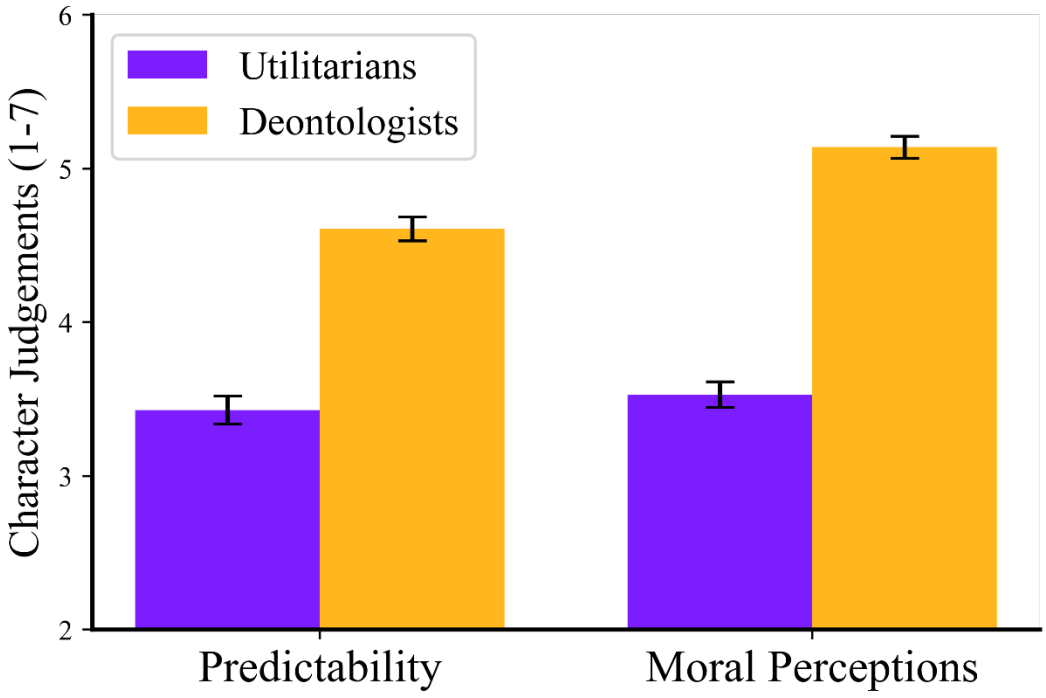
Finally, we assessed zero-order correlations between judgments of predictability-related and non-moral traits and judgments of moral perceptions (see Tables 8 and 10). These analyses revealed strong positive associations between all predictability-related concepts (i.e., consistency, intelligibility, reliability, and methodicalness) and moral perceptions across actor type, $r(298) > .52$, $p < .001$. Conversely, the associations between judgments of non-moral dimensions (i.e., status, power, confidence, conventionality, and wealth) and moral perceptions were mixed. For instance, judgments of conventionality were found to be positively associated with judgments of morality for utilitarian, $r(298) = .43$, $p < .001$, and deontological actors, $r(298) = .47$, $p < .001$. Conversely, a small correlation was observed between judgments of power and morality for utilitarian actors, $r(298) = .16$, $p = .004$, with this association disappearing for judgments of deontological actors, $r(298) = -.06$, $p = .277$. Moreover, comparing the strength of associations (with morality) between predictability-related and non-moral dimensions revealed many large differences. Therefore, as expected, judgments of consistency, intelligibility, reliability, and methodicalness shared considerably stronger associations with judgments of morality compared to judgments of non-moral dimensions (apart from conventionality, itself seemingly a predictability-related concept).

The design of Study 4 allowed us to once again assess whether deontological actors, choosing not to sacrifice a single individual in a high-conflict moral dilemma, are perceived to be more predictable and more moral compared to utilitarian actors choosing the sacrificial option within the same dilemma. As in Study 1, we conducted paired-samples *t*-tests comparing participants' judgments of predictability and moral character across deontological and utilitarian

actors (see Figure 5). Replicating the results of Studies 1-3, deontological actors were judged to be more predictable ($M = 4.61$, $SD = 1.46$) compared to utilitarian actors ($M = 3.43$, $SD = 1.59$), $t(298) = 8.68$, $p < .001$, $d = 0.51$, 95% CI [0.38, 0.62]. Also consistent with Studies 1-3, deontological actors were judged as more moral ($M = 5.14$, $SD = 1.22$) than utilitarian actors ($M = 3.53$, $SD = 1.35$), $t(298) = 13.49$, $p < .001$, $d = 0.78$, 95% CI [0.65, 0.91]. Thus, increasing confidence in predictability's role in moral judgments, we once again replicate the finding that people judge deontological actors as more predictable and more moral than utilitarian actors described as choosing the sacrificial choice in a high-conflict moral dilemma.

Figure 5

Study 4: Judgments of Predictability and Moral Perceptions



Note. Error bars represent +/- 1 standard error.

Study 5

An alternative explanation for the observed link between judgments of predictability and moral character relates to the role homophily (i.e., preferring others who are like oneself) may play when judging the character of others. It may be the case that, when judging agents acting within a sacrificial moral dilemma, people view agents choosing the same course of action that they themselves prefer (utilitarian or deontological) as both more predictable and more moral. From this perspective, assessments of predictability may guide judgments of moral character only to the extent that they capture participants' preferred course of action. Similarly, people may demonstrate a moral preference for agents choosing to perform what is believed to be the most "typical" action within a described moral scenario. For example, perceptions of deontological actors as both more predictable and moral (observed in Studies 1-4) may result from a majority of people endorsing the deontological option within high-conflict moral dilemmas (Greene et al., 2001; Greene, 2009; McGuire et al., 2009), resulting in deontological actors not only more frequently performing participants' preferred course of action but also the action that is more typical. If agents perceived as performing the more typical action are judged to be more predictable and more moral, then assessments of predictability may influence moral judgments only to the extent that they are associated with notions of typicality. In Study 5, we test these possibilities by asking participants' what they would do and what they believe most other people would do in the described high-conflict moral dilemmas. Overall, we hypothesize that the effect of predictability on perceptions of moral character will persist when accounting for participants' preferred course of action and the perceived typicality of an agent's behavior. Such findings would suggest that assessments of predictability play a unique role in judgments of moral character, specifically one that is not explained by notions of homophily or typicality.

Method

Participants

A sample of 300 US residents were recruited from Amazon Mechanical Turk (9% Asian-American; 11% African-American; 75% European-American; 4% Other) using the same recruiting criteria as Study 4. Participants received $1.20 upon completion of an 8-minute online questionnaire. We excluded data from 69 participants who failed to pass two comprehension check questions, leaving data from 231 participants (49 Female; M_{age} = 42.26, SD_{age} = 13.13) to be analysed. Sensitivity power analyses indicated that with a sample of 231 participants, power of .80, and alpha at .05, we would be able to detect a minimum effect size of r = .18 for correlation-based analyses and d = 0.22 for planned paired samples t-tests.

Materials

The scenarios featured in Study 5 were identical to those described in Studies 1 and 4.

Measures

Study 5 featured the same set of measures (i.e., moral perceptions, predictability, intentionality, and harm) as Study 3. As in Study 3, we computed moral perception scores for each participant by computing the mean of their Good, Moral, Peaceful, and Empathetic judgments. Once again, this composite showed good reliability (Utilitarian actor: α = .92, Deontological actor: α = .88). Furthermore, along with the aforementioned measures, Study 5 featured three additional measures, each of which were administered following character judgments of both utilitarian and deontological actors.

"What would you do?" (WWYD) Judgments. We assessed participants' preferred course of action in both Trolley and Bomb high-conflict vignettes by presenting each vignette[9] and asking participants "Which of the following best describes what you would do in this scenario?" Participants responded to this question by selecting one of the following five response options: I would definitely not [sacrificial choice]; I likely would not [sacrificial choice]; I would be indifferent/unsure; I likely would [sacrificial choice]; I would definitely [sacrificial choice]. The sacrificial choice described in these response options corresponded to the presented vignette (Trolley or Bomb) and were therefore described as either "push the very large stranger off the bridge" or "throw the very large stranger onto the bomb," depending on the vignette shown. Notably, participant responses were scored such that higher values were indicative of a greater certainty that one would make the sacrificial choice.

"What would others do?" (WWOD) Judgments. We also assessed participants' perceptions of what others would do in both Trolley and Bomb high-conflict vignettes. Participants were presented with these vignettes (as in WWYD judgments) and asked "If placed within this scenario, what percentage of people do you believe would [sacrificial choice]?" with the sacrificial choice again corresponding to the presented vignette (i.e., Trolley: "push the very large stranger off the bridge" and Bomb: "throw the very large stranger onto the bomb"). Responses to this question were elicited within a numerical entry text-box in which participants were instructed to "enter a number ranging from 0-100 representing the percentage of people that you believe would [sacrificial choice]."

[9] Note that these vignettes were slightly modified to remove any reference to a third-party acting within the vignette.

Oxford Utilitarianism Scale. The Oxford Utilitarianism Scale (OUS; Kahane et al., 2018) measures individuals' self-reported endorsement of utilitarian principles. This scale features two subscales, each of which is stated to measure a distinct dimension of utilitarian thinking: Instrumental Harm (5-items; measures peoples' willingness to commit harm for a greater moral purpose) and Impartial Beneficence (4-items; measures peoples' tendency to treat all lives as of equal moral value). Participants were presented with nine statements (e.g., Instrumental Harm: "It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people") and stated their level of agreement with each statement using a 7-point scale that ranged from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). Participants' mean agreement with the presented OUS statements was calculated for both OUS subscales. Higher scores indicated greater endorsement of utilitarian principles (specifically instrumental harm or impartial beneficence).

Design and Procedure

Study 5 utilized the same within-subjects design as Studies 1 and 4. The procedure of Study 5 also mirrored that of Studies 1 and 4, with the exception that participants in Study 5 responded to WWYD and WWOD judgments for both Trolley and Bomb vignettes following all character judgments of deontological and utilitarian actors. Furthermore, participants concluded Study 5 by completing nine OUS items and answering five demographic questions (e.g., age, sex, ethnicity, education, and household income).

Results and Discussion

We conducted two linear regression analyses (one for each actor type) predicting judgments of morality (i.e., moral perceptions) using predictability, WWYD, and WWOD

judgments (see Tables 11A and 11B). The overall models for both utilitarian, $F(3, 227) = 45.10$, $p < .001$, $R^2 = 0.37$, and deontological actors, $F(3, 227) = 43.80$, $p < .001$, $R^2 = 0.37$, were significant. The results of these analyses were inconsistent with the idea that the observed positive relation between judgments of predictability and morality in Studies 1-4 could be fully explained by participants viewing actors choosing their preferred course of action (or the one they perceived as more typical) as both more predictable and moral. That is, for both judgments of deontological and utilitarian actors, assessments of predictability were shown to be a significant predictor of judgments of morality, even when accounting for participants' WWYD and WWOD judgments, Deontologists: $b = 0.32$, $p < .001$; Utilitarians: $b = 0.30$, $p < .001$. Nevertheless, WWYD judgments were also found to be a significant predictor of moral perceptions, Deontologists: $b = -0.44$, $p < .001$; Utilitarians: $b = 0.45$, $p < .001$. That is, participants' tended to judge agents behaving in a manner consistent with their preferred course of action as more moral than those performing an alternative action. Conversely, WWOD judgments were not observed to be a significant predictor of moral perceptions when judging deontological, $b = -0.004$, $p = .134$, or utilitarian actors, $b = -0.002$, $p = .579$. Overall, the results of Study 5 further support the claim that judgments of an agent's predictability help guide judgments of their morality. Specifically, while participants' stating that they themselves would perform the sacrificial [non-sacrificial] choice tended to perceive utilitarian [deontological] actors as more moral, this moral preference could not account for the observed positive relation between judgments of an agent's predictability and morality. Thus, while homophily (i.e., preferring others who are like oneself) appears to play an important role when judging the morality of individuals acting within a moral dilemma, so too do assessments of predictability.

Table 11A

Regression analysis, judgments of moral perceptions as predicted by predictability judgments, what would you do judgments, and what would others do judgments. Utilitarian agents.

Effect	*b*	*SE*	95% CI		*p*
Utilitarian agents – Moral Perceptions			*LL*	*UL*	
Intercept	1.40	0.20	0.97	1.77	<.001
Predictability	0.30	0.05	0.20	0.41	<.001
WWYD	0.45	0.07	0.31	0.58	<.001
WWOD	0.01	0.01	-0.01	0.01	.579

Note. Total $N = 231$. *CI* = confidence interval; *LL* = lower limit; *UL* = upper limit *b* = estimate of effect size.

Table 11B

Regression analysis, judgments of moral perceptions as predicted by predictability judgments, what would you do judgments, and what would others do judgments. Deontological agents.

Effect	*b*	*SE*	95% CI		*p*
Deontological agents – Moral Perceptions			*LL*	*UL*	
Intercept	4.47	0.30	3.87	5.06	<.001
Predictability	0.32	0.05	0.21	0.41	<.001
WWYD	-0.44	0.06	-0.54	-0.33	<.001
WWOD	0.01	0.01	-0.00	0.01	.134

Note. Total $N = 231$. *CI* = confidence interval; *LL* = lower limit; *UL* = upper limit, *b* = estimate of effect size.

We also examined the zero-order correlations between several key variables of interest (see Table 12). Replicating the results of Studies 1-4, we observed a positive association between judgments of predictability and morality for judgments of both utilitarian, $r(229) = .47$, $p < .001$, and deontological actors, $r(229) = .43$, $p < .001$. Furthermore, we observed a positive association between WWYD judgments (with higher scores indicating greater certainty that one would themselves perform the sacrificial act) and moral judgments of utilitarian actors, $r(229) = .53$, $p < .001$. Likewise, we observed a negative association between WWYD judgments and moral judgments of deontological actors, $r(229) = -.50$, $p < .001$. An identical pattern of results was

observed for WWOD judgments, albeit with the observed correlation coefficients being reduced in magnitude for both utilitarian, $r(229) = .30$, $p < .001$, and deontological actors, $r(229) = -.19$, $p = .005$. Additionally, WWYD and WWOD judgments were both associated with the perceived predictability of utilitarian, WWYD: $r(229) = .35$, $p < .001$; WWOD: $r(229) = .27$, $p < .001$, and deontological actors, WWYD: $r(229) = -.22$, $p < .001$; WWOD: $r(229) = -.19$, $p = .003$. Thus, participants did tend to judge agents performing their preferred action, or the action they perceived to be most popular, as more predictable. Lastly, we observed that the endorsement of the utilitarian consistent principle of instrumental harm (as measured by the OUS) was positively associated with judgments of utilitarian actors as more predictable, $r(229) = .25$, $p < .001$, and more moral, $r(229) = .47$, $p < .001$. This was not the case for the principle of impartial beneficence, which was observed to share a small positive correlation with judgments of utilitarian actors' predictability, $r(229) = .14$, $p = .033$, and no association with judgments of these actors' morality, $r(229) = .09$, $p = .197$. Thus, while participants' beliefs regarding the acceptability of harming others for the greater good did appear to be predictive of their moral judgments of utilitarian (and deontological: $r(229) = -.37$, $p < .001$) actors, this was not the case for beliefs regarding the treatment of all lives as possessing equal moral value (i.e., impartial beneficence).

Table 12

Study 5 Correlations

	1	2	3	4	5	6	7	8
1. Moral Perceptions	-	.43***	-.47***	.29***	-.50***	-.19**	.02	-.37***
2. Predictability	.47***	-	-.21***	.15*	-.22***	-.19**	-.03	-.10
3. Harm	-.67***	-.32***	-	-.04***	.48***	.23***	.14*	.37***
4. Intentionality	-.07	-.16*	.09	-	-.08	.09	-.03	-.02
5. WWYD	.53***	.35***	-.43***	-.06	-	.45***	.06	.56
6. WWOD	.30***	.27***	-.20***	-.12	.45***	-	.09	.23
7. Impartial Beneficence	.09	.14*	-.05	-.11	.06	.09	-	.09
8. Instrumental Harm	.47***	.25***	-.39***	-.07	.56	.27	.09	-

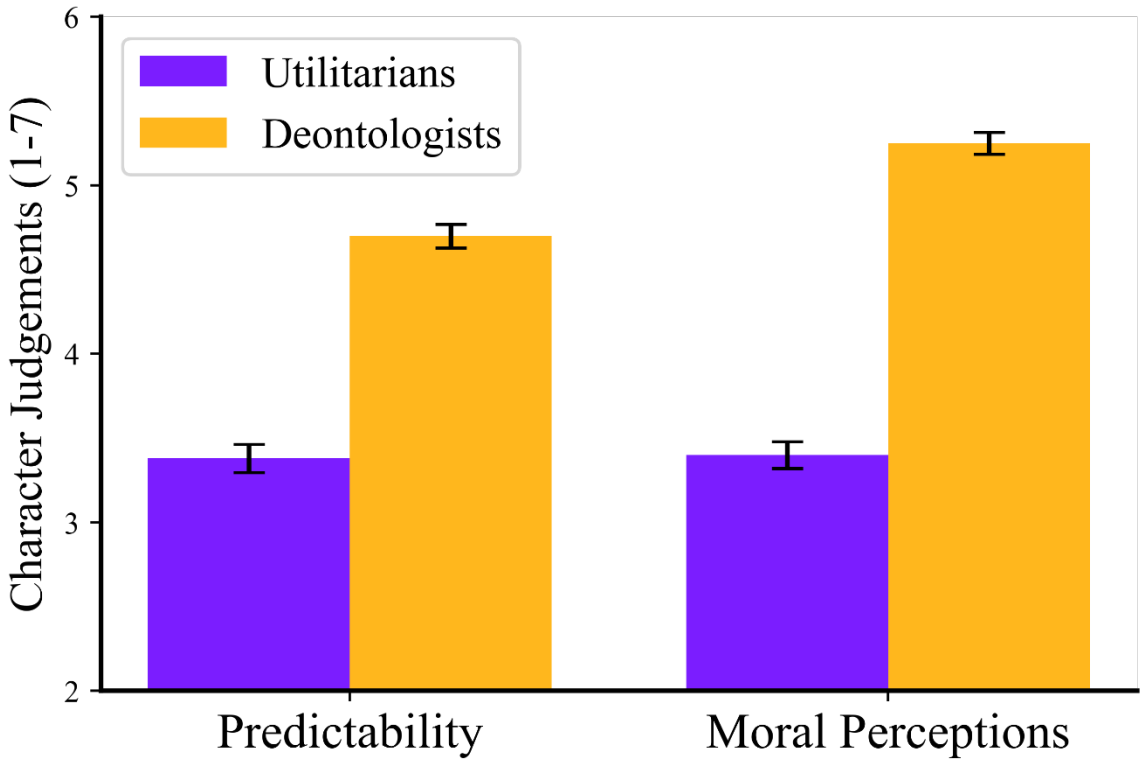
Note. Pearson correlations ($N = 231$). Bottom diagonal = Utilitarian actor. Top diagonal = Deontological actor. *** $p < .001$, ** $p < .01$, * $p < .05$.

The design of Study 5 allowed us to investigate whether agents described as performing a sacrificial act within a high-conflict moral dilemma (i.e., utilitarian actors) were once again judged by participants' to be less predictable and less moral than deontological actors (see Figure 6). Replicating the results of Studies 1-4, utilitarian actors were judged as less predictable ($M = 3.38$, $SD = 1.44$) and less moral ($M = 3.40$, $SD = 1.36$) than deontological actors (Predictability: $M = 4.70$, $SD = 1.21$; Moral Perceptions: $M = 5.25$, $SD = 1.13$), $t(230) = 9.55$, $p < .001$, $d = 0.62$, 95% *CI* [0.49, 0.77] and $t(230) = 13.65$, $p < .001$, $d = 0.90$, 95% *CI* [0.74, 1.05], respectively. Furthermore, despite the actions of utilitarian actors resulting in *fewer* lives lost, utilitarian actors were judged as having caused *more* harm ($M = 5.08$, $SD = 1.36$) compared to deontological actors ($M = 3.56$, $SD = 1.85$), $t(230) = 9.00$, $p < .001$, $d = 0.59$, 95% *CI* [0.45, 0.73]. Therefore,

the present findings provide additional support that, within high-conflict moral dilemmas, deontological actors are perceived as more predictable, more moral, and as causing less harm compared to utilitarian actors.

Figure 6

Study 5: Judgments of Predictability and Moral Perceptions



Study 6

So far, we have presented the results of five studies that suggest that perceptions of predictability influence assessments of moral character. These findings, however, could be limited to residents of Western societies who may place an unusually strong emphasis on predictability in their moral and social worlds. It is possible that those inhabiting non-Western societies may not attach a moral premium to predictable agents, as sociocultural factors have

been shown to alter moral judgments (Henrich et al., 2010; Nisbett et al., 2001). For example, recent work has demonstrated that the Yali people of Papua are more deontological compared to Western samples (Sorokowski et al., 2020), whereas those in Nicaragua are more utilitarian (Winking & Koster, 2020). If predictability is a fundamental aspect of moral judgment, then individuals across vastly different environmental circumstances should show a preference for predictable agents when judging moral character. To test this idea, we collected a sample from the Dani people, a small-scale, traditional, indigenous society of Western Papua. We hypothesize that, as in our Western samples, the Dani people will show a moral preference for predictable agents, with judgments of predictability being positively associated with judgments of morality.

Method

Participants

A sample of 81 adults (40% Female; M_{age} = 41.20, SD_{age} = 16.40) from the Dani traditional society of Papua participated in Study 6 in exchange for small gifts. A sensitivity power analysis indicated that with a sample of 81 participants, power of .80, and alpha at .05, we would be able to detect a minimum effect size of d = 0.63 for planned paired-samples t-tests. This study was approved by a University of Wroclaw ethics committee and by the chiefs of the local communities. The Dani people live in the central highlands of Western Papua (a province of Indonesia). This study was conducted in Baliem Valley in local villages around Wamena. Life in these villages is fundamentally different from life in an industrialized society. For example, there is no electricity, mobile coverage, running water, nor other modern amenities. Due to the remote location of their dwellings, the Dani can be described as a population with minor contact with Western culture (i.e., our participants' contact with members of other cultures is restricted to the few tourists visiting the town of Wamena).

Measures and Materials

Study 6 was conducted in collaboration with a local interpreter (from the Dani tribe), blind to the study's hypotheses. In a face-to-face interview, participants were described a modified version of the high-conflict Trolley dilemma used in Studies 1-5. To account for the Dani's lack of familiarity with trolleys and railways, we altered the high-conflict Trolley dilemma in a way that increased the ecological validity of the scenario for this sample (see Table 13). Within this vignette, participants were informed of the actions of a hypothetical agent who decided to sacrifice a single individual (Utilitarian actor) or let harm befall five strangers (Deontological actor). After being informed of this agent's action, participants were asked to assess the predictability (1 = *Very Unpredictable*; 2 = *Somewhat Unpredictable*; 3 = *Somewhat Predictable*; 4 = *Very Predictable*) and morality (1 = *Very Immoral*; 2 = *Somewhat Immoral*; 3 = *Somewhat Moral*; 4 = *Very Moral*) of the agent on a four-point Likert scale. In cases in which a participant was not answering or requested clarification the entire vignette was repeated.

Table 13

Study 6: Modified High-Conflict Vignette

	Trolley Scenario
Modified High-Conflict Vignette	Imagine [Local Name] is standing near a big collapsing tree on a hill. The tree is going to fall down. There are five people standing at the bottom of the hill. There is also a person standing half way up the hill. If [Name] does nothing, the tree will fall directly on the group of five people, causing their deaths. The only way to avoid the deaths of these people is to push the single person into the path of the tree so that it will be blocked. This will cause the single person to die but will save the other five people's lives.
Deontological Actor	[Name] decides not to push the single person into the path of the tree.
Utilitarian Actor	[Name] decides to push the single person into the path of the tree.

Note. This vignette was read to participants in the native language of the Dani people.

Design and Procedure

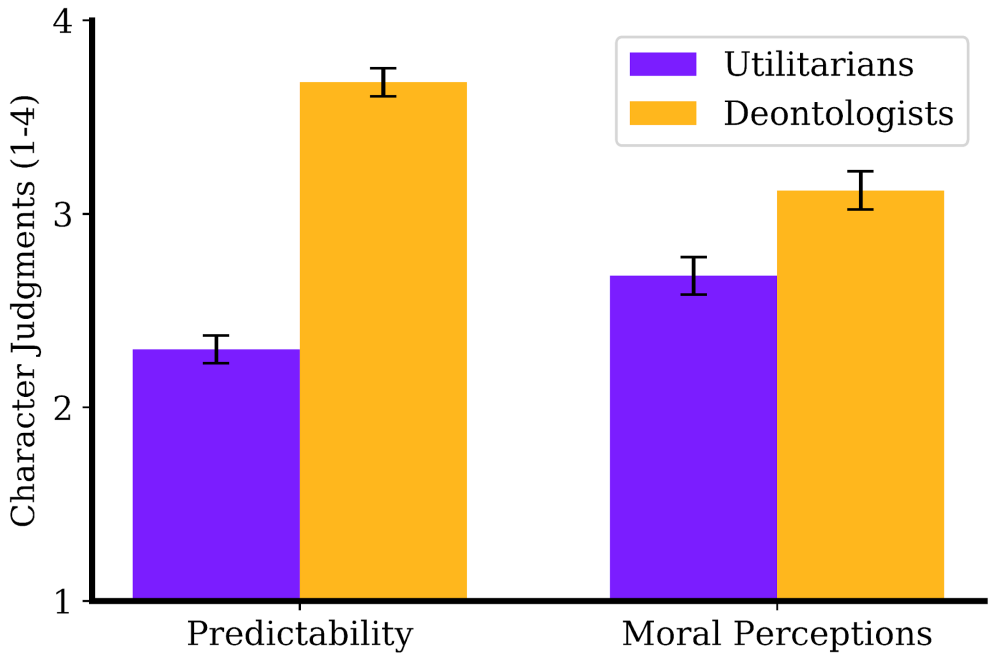
We used a between-subjects design in which each participant was described a vignette (in their native language) featuring either a utilitarian or deontological actor and were asked to judge the predictability and morality of the described agent.

Results

Consistent with our previous studies featuring US residents, participants perceived deontological actors as more predictable (M = 3.68, SD = 0.65) than utilitarian actors (M = 2.30, SD = 0.65), $t(79) = 9.59$, $p < .001$, $d = 2.12$, 95% CI [1.73, 2.50] (see Figure 7). Notably, deontological actors were also perceived as more moral (M = 3.12, SD = 0.87) than utilitarian actors (M = 2.68, SD = 0.89), $t(79) = 2.29$, $p = .025$, $d = 0.50$, 95% CI [0.18, 0.81]. As expected, judgments of predictability were positively associated with judgments of morality, $r(79) = .45$, $p < .001$. Overall, the present findings match those found in our Western samples, suggesting that the link between predictability and morality is not merely the result of the idiosyncratic tastes of an industrialized western population.

Figure 7

Study 6: Judgments of Predictability and Moral Perceptions



Note. Error bars represent +/- 1 standard error.

General Discussion

Across six studies, we establish how the moral impressions of agents acting within high- and low-conflict moral dilemmas are influenced by perceptions of their predictability. Regardless of the consequences of an agent's actions, and regardless of their violation of proscriptions against killing, participants in the current study consistently preferred the agent who they judged to be most predictable. That is, utilitarian actors opting to sacrifice an individual for the greater good were judged as more or less moral than a deontological actor refusing this sacrifice, depending on how predictable their actions appeared. We find that assessments of predictability are multi-faceted, strongly associated with judgments of an agent's consistency, reliability,

intelligibility, and methodicalness. However, we observe that assessments of predictability most strongly evoke judgments related to "consistency of behavior," particularly for judgments of deontological actors. Additionally, we show that participants' preferred course of action within the described moral dilemmas (i.e., WWYD judgments) are positively associated with judgments of predictability and morality. Nevertheless, assessments of predictability maintain a unique and non-trivial contribution to judgments of morality, even when controlling for participants' preferred moral decisions. Overall, we suggest that judgments of an agent's predictability inform judgments of their morality. Notably, we show that the observed moral preference for more predictable actors is not easily explained by a misunderstanding of utilitarian motivations, appeals to homophily, perceived action typicality, or a simple preference for inaction.

Do judgments of morality guide assessments of predictability, rather than the reverse? We believe this is the less parsimonious account for three reasons. First, mind perception has been argued to be the essence of morality (Gray et al., 2012; Schein & Gray, 2018). The attribution of moral character to an agent is downstream to first recognizing features of their mind (e.g., their agency or ability to experience pain). Predictability is one such mental feature related to mind perception that can be used to inform judgments of morality. Second, if the purpose of developing a sense of morality is for fostering cooperation (Curry, 2016), a natural prediction is that traits signalling that a person is likely to be a good cooperation partner[10] ought to influence whether they are judged to be a moral person, and not the reverse. Third, deciding whether a person is "moral" or "immoral" is complex and requires a specific understanding of their desires and mental states (Chakroff & Young, 2015; Cushman, 2015; Inbar et al., 2012). When making these mental state inferences, there is a degree of social uncertainty where one

[10] In fact, it has recently been demonstrated elsewhere that people explicitly endorse the importance of predictability when evaluating cooperation partners (Walker et al., 2021a).

cannot be sure of another's intentions (FeldmanHall & Shenhav, 2019; Vives & FeldmanHall, 2018). Only after gaining more information through various methods, such as observing a person's behavior, are we able to reduce uncertainty surrounding their desires and intentions. This reduction of social uncertainty affords us confidence in deciding whether or not to cooperate.

Traditionally, moral judgments have been thought to be a function of obedience to universal moral principles (deontology), or assessments of the consequences of a given action (utilitarianism; Bartels, 2008; Conway & Gawronski, 2013; Gawronski et al., 2017). When answering moral dilemmas, it is a consistent finding that people tend toward the sacrificial choice in low-conflict dilemmas and the non-sacrificial option in high-conflict dilemmas (Bruers & Braeckman, 2014; Hauser et al., 2008; Waldmann & Dieterich, 2007). If people are guided primarily by moral principles, such as utilitarianism or deontology, this is a puzzling finding, and a body of literature has emerged attempting to explain why such a reversal occurs (Greene et al., 2001; McGuire et al., 2009; Greene, 2009). However, if peoples' goal is instead to signal *behavioral* predictability (for instance, in order to signal one's value as a cooperation partner), this reversal makes sense. One take away from the current work is that those who take the sacrificial option in high-conflict dilemmas are perceived as unpredictable. Thus, it may be that individuals are sensitive to the signalling risk that endorsing a direct killing strategy presents and switch their responses in order to appear more predictable and more moral. Therefore, our data hint that such moral judgments may be guided more by a desire to signal predictability than considerations of moral principles. Though future work should explore the degree to which people are consciously aware of the signalling value of predictability.

The current study may also reveal one reason why utilitarians are disliked as cooperation partners and judged harshly by others evaluating their moral character (Bostyn & Roets, 2017; Everett et al., 2016; Uhlmann et al., 2013). That is, it is not sacrificial actions specifically that appear to drive peoples' dislike of utilitarians, nor a misunderstanding of their principles. Rather, it may be the extent to which sacrificial actions signal unpredictability that influences the harsh assessment of their character. When sacrificial actions are performed in a manner and context that is received as demonstrating greater predictability, as in the low-conflict moral dilemmas administered in Study 3, utilitarian actors are morally preferred over those who opt not to sacrifice.

The theory of morality-as-cooperation posits that a sense of morality emerged as a means to foster cooperation (Curry, 2016; Curry, Chesters, & Van Lissa, 2019; Curry, Mullins, & Whitehouse, 2019). As predictable behavior is necessary to form long lasting cooperative networks (especially between non-kin), this theory naturally accommodates the current findings. A person who signals unpredictability with their actions naturally makes a poor candidate for collaboration. If our moral sense emerged to facilitate cooperation, it makes sense for people to judge those they view as unpredictable as possessing an inferior moral character. One possible mechanism which facilitates predictability's influence on moral impressions may be that of social uncertainty, whereby unpredictable actions fail to reveal a person's intentions to cooperate or defect resulting in others avoiding interaction (FeldmanHall & Shenhav, 2019; Vives & FeldmanHall, 2018). Adding strength to evolutionary focused accounts, we observe a moral premium for predictability in both a Western (US residents) and indigenous non-Western society (adults from the Dani traditional society of Papua). These two types of societies are geographically, culturally, economically, and socially distinct, suggesting that individuals in both

societies may have inherited a preference for predictable agents as a foundational aspect of moral judgment from a common source. The results of Study 4 further support an account based on predictability's value as a trait for the purposes of cooperation. It seems that "predictability", as interpreted by participants, may most strongly index notions of consistent behavior as well as, to a lesser extent, notions of reliability, both of which are crucial traits for stable cooperative partnerships.

Cooperation is immensely important for human productivity, as being unable to predict the behavior of others globally would be disastrous for maintaining the incredible interconnectedness that characterises the human species. Moral impressions guide much of human decision making with respect to identifying potential cooperators (Goodwin et al., 2014; Graham & Haidt, 2010; Landy et al., 2016; Rand et al., 2012; Rand & Nowak, 2013). As such, understanding the factors that underlie these moral impressions is paramount in understanding the roots of cooperation. We propose that predictability is a key factor humans consider when judging the morality of others. Understanding the hugely powerful force of human cooperation requires an understanding of morality, and understanding the moral character of people we intend to cooperate with depends on being able to predict their behavior in the long term.

Open Practices

Preregistrations are viewable at the following links:

Study 1: osf.io/vhcbu/?view_only=54a3b5f1b7024d5384293dfce828c1be,

Study 2: osf.io/zps5y/?view_only=dafc468df1974638a80b74c3f99b6d7c,

Study 3: osf.io/8h6f2/?view_only=4ed25d5ae88d4677bb058c587f6d841e,

Study 4: osf.io/kqmeu/?view_only=2448d78f8ba24dc796569dce9a3c8732

Study 5: osf.io/njcgh/?view_only=5171e3cf4496479aa8099288f472ca40

Study 6: osf.io/3ukh8/?view_only=1c332808754d4d988e06b0743c316bfb

Data are available at: https://osf.io/xdepr/?view_only=42d4e21e416f4e2c997ed6c541324d13.

Materials are viewable in the Online Supplementary Materials.

References

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390-1396.

Balliet, D., & Van Lange, P. A. (2013). Trust, conflict, and cooperation: a meta-analysis. *Psychological Bulletin, 139*(5), 1090.

Barrett, H. C., Cosmides, L., & Tooby, J. (2010). Coevolution of cooperation, causal cognition and mindreading. *Communicative & Integrative Biology*, *3*(6), 522-524.

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, *108*(2), 381-417.

Białek, M., Turpin, M. H., & Fugelsang, J. A. (2019). What is the right question for moral psychology to answer? Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, *30*(9), 1383-1385.

Bostyn, D. H., & Roets, A. (2017). Trust, trolleys and social dilemmas: A replication study. *Journal of Experimental Psychology: General*, *146*(5), e1-e7.

Bruers, S., & Braeckman, J. (2014). A review and systematization of the trolley problem. *Philosophia*, *42*(2), 251-269.

Capraro V, Sippel J, Zhao B, Hornischer L, Savary M, Terzopoulou Z, et al. (2018) People making deontological judgments in the Trapdoor dilemma are perceived to be more prosocial in economic games than they actually are. *PLoS ONE 13*(10): e0205066.

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition*, *136*, 30-37.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, *104*(2), 216-235.

Curry, O. S. (2016). Morality as cooperation: A problem-centred approach. In *The evolution of morality*. T. K. Shackelford and R. D. Hansen, eds. Pp. 27-51. New York: Springer.

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of Morality-as-Cooperation in 60 societies. *Current Anthropology*, *6*0(1), 47-69.

Curry, O. S., Chesters, M. J., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality*, *78*, 106-124.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359-1381.

Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, 6, 97-103.

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

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772-787.

Fausey, C. M., & Boroditsky, L. (2010). Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic bulletin & review*, 17(5), 644-650.

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426-435.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 1-6.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156-168.

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343-376.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148-168.

Graham, J., & Haidt, J. (2010). Beyond beliefs: Religions bind individuals into moral communities. *Personality and Social Psychology Review*, *14*(1), 140-150.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101-124.

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

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144-1154.

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology*, *45*(3), 581-584.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greene, J. D. (2015). The rise of moral cognition. *Cognition*, *135*, 39-42.

Haidt, J. (2012). *The Righteous Mind: Why good people are divided by politics and religion*. New York: Pantheon.

Hauser, M., Young, L., & Cushman, F. (2008). Reviving Rawls' linguistic analogy: Operative principles and the causal structure of moral actions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology and Biology*. New York: Oxford.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61-83.

Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin, 38*(1), 52-62.

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

Kreps, T. A., & Monin, B. (2014). Core values versus common sense: Consequentialist views appear less rooted in morality. *Personality and Social Psychology Bulletin*, *40*(11), 1529-1542.

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart: The desirability of sociability and competence depends on morality. *Personality and Social Psychology Bulletin*, *42*(9), 1272-1290.

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, *45*(3), 577-580.

Melis, A. P., & Semmann, D. (2010). How is human cooperation different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), 2663-2674.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291-310.

Pinker, S. (2012). *The better angels of our nature: Why violence has declined*. New York: Penguin.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57-75.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427-430.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413-425.

Ridley, M. (2010). *The rational optimist: How prosperity evolves*. New York: HarperCollins.

Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology, 69*, 44–58.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32-70.

Sorokowski, P., Marczak, M., Misiak, M., & Białek, M. (2020). Trolley Dilemma in Papua. Yali horticulturalists refuse to pull the lever. *Psychonomic Bulletin & Review*, 1-6.

Sosa, N., & Rios, K. (2019). The utilitarian scientist: The humanization of scientists in moral dilemmas. *Journal of Experimental Social Psychology*, *84*, 103818.

Tinghög, G., & Västfjäll, D. (2018). Why people hate health economics–two psychological explanations. *LiU Working Papers in Economics*, *6*, 1-10.

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology*, *64*(1), 231-255.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35-57.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326-334.

Vives, M. L., & FeldmanHall, O. (2018). Tolerance to ambiguous uncertainty predicts prosocial behavior. *Nature Communications*, *9*(1), 1-9.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, *18*(3), 247-253.

Walker, A. C., Turpin, M. H., Fugelsang J. A., & Białek, M. (2021a). Better the two devils you know, than the one you don't: Predictability influences moral judgment. Retrieved from: psyarxiv.com/w4y8f. doi:10.31234/osf.io/w4y8f

Walker, A. C., Turpin, M. H., Meyers, E. A., Stolz, J. A., Fugelsang, J. A., & Koehler, D. J. (2021b). Controlling the narrative: Euphemistic language affects judgments of actions while avoiding perceptions of dishonesty. *Cognition*, *211*, 104633. https://doi.org/10.1016/j.cognition.2021.104633

Winking, J., & Koster, J. (2020). Small-Scale Utilitarianism: High acceptance of utilitarian solutions to trolley problems among a horticultural population. Retrieved from: https://psyarxiv.com/g6tc4/