Cognition 206 (2021) 104467

Contents lists available at ScienceDirect

Cognition

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

Principles of moral accounting: How our intuitive moral sense balances rights and wrongs

Samuel G.B. Johnson[a,b,c,*], Jaye Ahn[d]

[a] University of Warwick, Department of Psychology, United Kingdom of Great Britain and Northern Ireland
[b] University of Bath, School of Management, United Kingdom of Great Britain and Northern Ireland
[c] University College London, Centre for the Study of Decision-Making Uncertainty, United Kingdom of Great Britain and Northern Ireland
[d] University of Minnesota, Department of Psychology, United States of America

ARTICLE INFO

Keywords:
Moral judgment
Reputation
Intuitive ethics
Social cognition
Person perception

ABSTRACT

We are all saints and sinners: Some of our actions benefit others, while other actions lead to harm. How do people balance moral rights against moral wrongs when evaluating others' actions? Across 9 studies, we contrast the predictions of three conceptions of intuitive morality—outcome-based (utilitarian), act-based (deontologist), and person-based (virtue ethics) approaches. These experiments establish four principles: *Partial offsetting* (good acts can partly offset bad acts), *diminishing sensitivity* (the extent of the good act has minimal impact on its offsetting power), *temporal asymmetry* (good acts are more praiseworthy when they come after harms), and *act congruency* (good acts are more praiseworthy to the extent they offset a similar harm). These principles are difficult to square with utilitarian or deontological approaches, but sit well within person-based approaches to moral psychology. Inferences about personal character mediated many of these effects (Studies 1–4), explained differences across items and across individuals (Studies 5–6), and could be manipulated to produce downstream consequences on blame (Studies 7–9); however, there was some evidence for more modest roles of utilitarian and deontological processing too. These findings contribute to conversations about moral psychology and person perception, and may have policy and marketing implications.

1. Introduction

If you were to fly round-trip from NYC to LA, you would be responsible for emitting 1.3 tons of CO_2 into the atmosphere. This action imposes a cost on the environment and on society. But there is an easy way to neutralize these social costs—buying carbon offsets. In their most common form, the consumer contracts with a third-party to plant trees, which absorb CO_2 from the atmosphere and can thus neutralize any given amount of carbon. It turns out that planting 7 trees neutralizes approximately 1.3 tons of CO_2, and at current market prices this costs about $13. Many social and environmental scientists consider this a win–win, since this allows you to achieve whatever personal or economic benefits that motivated you to fly, while imposing zero net-cost on society and the planet.

But it is less clear how ordinary people, as opposed to policy wonks, think about carbon offsets. Anecdotally, many commentators seem to believe that they are ethically problematic. An op-ed in *The Guardian* characterized offsets as a way to "buy yourself a clean conscience by paying someone else to undo the harm you are causing" (Monbiot, 2006). Building on this argument, a parody website called cheatneutral.com even promised to offer "cheating offsets" to neutralize marital infidelity, boasting that their service "offsets your cheating by funding someone else to be faithful and NOT cheat" (quoted in May, 2007). Even though carbon offsets appear to be a good bargain for society from the utilitarian perspective of minimizing net harm, they may run up against deep psychological resistance.

The debate about carbon offsets is one example of *moral accounting*—how our intuitive morality balances harmful acts against beneficial acts. We are all saints as well as sinners; therefore, moral accounting is relevant to much human behavior. For example, a person might shirk off from work and make up for the shirking by working harder later on, might litter and then volunteer to pick up trash, or might discriminate against a black loan applicant and then make up for the discrimination by helping another applicant. This paper maps the principles governing moral accounting and tests the psychological mechanisms underlying these principles.

Many studies have looked at the behavioral effects of gaining moral credits or moral credentials on subsequent behavior. People *morally self-*

* Corresponding author at: University of Warwick, Department of Psychology, United Kingdom of Great Britain and Northern Ireland.
E-mail address: sam.g.b.johnson@warwick.ac.uk (S.G.B. Johnson).

https://doi.org/10.1016/j.cognition.2020.104467
Received 18 June 2020; Received in revised form 8 September 2020; Accepted 15 September 2020

license, becoming likelier to perform an immoral act after they or an in-group member perform an earlier positive act (Kouchaki, 2011; Merritt, Effron & Monin, 2010; Sachdeva et al., 2009). For example, after choosing a qualified female job candidate, people feel licensed to endorse gender stereotypes (Monin & Miller, 2001). Analogously, "virtuous" consumer behaviors (e.g., volunteering for community service) motivate "vice" behaviors (e.g., consuming luxury products) (Khan & Dhar, 2006). Although more research has looked at licensing behavior (performing good acts then bad acts), people are also known to engage in cleansing or redemption behavior (performing bad acts then good acts) (Tangney et al., 2007; Tetlock, 2003). For example, participants who relied on a "forbidden" (racially-tainted) base rate in setting insurance premiums later expressed greater interest in volunteering for race-related causes (Tetlock et al., 2000). Intriguingly, people sometimes act as though gaining moral credit and debits can have causal effects on future random outcomes (Callan et al., 2014), particularly when uncertainty is high and control is low (Converse et al., 2012).

But it is also critical to understand how others *judge* combinations of morally right and wrong actions. The study of praiseworthy and blameworthy acts have proceeded largely independently. Some research has *compared* moral judgments about blameworthy versus praiseworthy acts, documenting both symmetries (e.g., De Freitas & Johnson, 2018; Gray & Wegner, 2009; Siegel et al., 2017; Wiltermuth et al., 2010) and asymmetries (e.g., Bostyn & Roets, 2016; Guglielmo & Malle, 2019; Klein & Epley, 2014; Knobe, 2003; Pizarro et al., 2003), while other work has studied the ethicality of morally *ambiguous* acts that are not clearly blameworthy or praiseworthy (e.g., Everett et al., 2018; Levine et al., 2018; Levine & Schweitzer, 2014; Rottman et al., 2014). But the majority of this literature has theorized (separately) about the mechanisms underlying judgments about morally negative acts (e.g., Alicke, 1992; Baez et al., 2017; Cushman, 2008; Cushman et al., 2006; Graham et al., 2009; Guglielmo & Malle, 2017; Haidt et al., 1993; Inbar et al., 2012; Niemi & Young, 2016; Paxton et al., 2012; Schnall et al., 2008; Tannenbaum et al., 2011; Tetlock et al., 2000; Young & Saxe, 2011) or positive acts (e.g., Critcher & Dunning, 2011; Johnson, 2020; Johnson & Park, 2020; Lin-Healy & Small, 2013; Monin et al., 2008; Newman & Cain, 2014). Many of these articles propose detailed theories of how people assign praise or blame. But existing theory does not supply a ready account of how people evaluate *combinations* of praise and blame—a critical question if we are to understand how moral judgments of acts and persons unfold over time. We aim to fill this theoretical vacuum.

In addition to its theoretical value, it is practically useful to understand moral accounting. Moral decisions often depend on how we expect others to perceive our actions—people are aware that their (im)moral actions send signals to third-parties and therefore attend to those third parties' perceptions. For example, people conspicuously conserve resources: They are likelier to purchase "green" products when shopping in public rather than in private (Griskevicius et al., 2010). Moreover, moral signaling can sometimes lead to socially suboptimal behaviors: Since donations of time signal emotional investment more than donations of money, people with an affiliation goal express greater intention to donate time rather than money, even though people believe that such donations help fewer people (Johnson & Park, 2020). Since third-party moral judgments inform our predictions about how our actions will be perceived and therefore what actions we take, understanding these third-party judgments and their moderators can help to promote socially beneficial behaviors.

1.1. Moral accounting and theories of morality

In psychology and philosophy, the two dominant approaches are variants on *utilitarianism* (e.g., Bentham, 1907/1789; Mill, 1998/1861; Singer, 2011) and *deontology* (e.g., Aquinas, 2000/1274; Kant, 2002/1796; Nagel, 1979). Utilitarianism is outcome-centered, holding that our moral duty is to maximize positive consequences and minimize negative consequences. Deontology, in contrast, is act-centered, holding that our moral duty is to act according to moral laws. Although these approaches often agree, they sometimes diverge, as in moral dilemmas that involve instrumental harm—harming someone as a means to some greater end (Foot, 1967). Much of the theoretical and empirical discussion in moral psychology has concerned when, why, and how much these two factors—the outcome of an act versus the nature of the act itself—influence our moral judgments and decisions (e.g., Baron & Spranca, 1997; Bartels & Medin, 2007; Bartels & Pizarro, 2011; Conway & Gawronski, 2013; Côté et al., 2013; Greene et al., 2008; Kahane et al., 2015; Kahane et al., 2018; Paxton et al., 2012; Shenhav & Greene, 2010; Tetlock et al., 2000). At the risk of oversimplifying a complicated debate, it seems reasonably clear that both factors matter to most people, that their relative importance shifts across contexts, and that people do not adopt either a consistently utilitarian or deontological moral theory.

Yet, these approaches make quite different predictions about how moral accounting might work. According to utilitarianism, the net-benefit should drive judgments of blameworthiness: One would be morally blameworthy to the extent that one has caused more harm than good on balance and praiseworthy to the extent that one has caused more good than harm. This view is quite friendly to offsetting: Other things being equal, actions causing equal harm and benefit have zero net-harm and are equivalent to doing nothing at all. Different philosophical refinements of utilitarianism may very well give different verdicts. Whereas direct (or act) utilitarianism focuses on the immediate costs and benefits of actions, indirect utilitarianism allows agents to consider more far-flung consequences of their actions. For example, motive utilitarianism and rule utilitarianism account for the broader consequences of acting for particular reasons or in accordance with particular rules, respectively (Adams, 1976; Rawls, 1955; Singer, 1977). Since utilitarianism as understood in moral psychology is typically operationalized as direct or act utilitarianism, we stick with that operationalization here, while acknowledging that more sophisticated versions of utilitarianism may be flexible enough to accommodate many possible patterns of judgments.

According to deontology, some acts are wrong regardless of their consequences. Thus, it is wrong to perform forbidden actions as a means to some other end, even if that end itself is good. This view is much less friendly toward offsetting, which allows morally negative actions as long as they are balanced out by contravening positive outcomes. For acts that are viewed as forbidden, blame should differ little based on whether those acts are offset. As with utilitarianism, there are many philosophical refinements to deontology, with versions differing in where moral rules come from, the role of intention versus causation, whether actions are distinguished from omissions, the scope of actions that are supererogatory (permissible but not obligatory), and their relative emphasis on rights (Nozick, 1974; Quinn, 1989; Scheffler, 1982). Also as with utilitarianism, we operationalize deontology in a simple manner consistent with prior studies: That acts are blameworthy when they violate a moral norm and such acts are wrong regardless of their consequences (Baron & Spranca, 1997).

Both of these approaches, however, have been challenged by character-based approaches, which have a very old pedigree in philosophy (e.g., the *virtue ethics* of Aristotle, 1999/350 BCE and Hursthouse, 1999) but have only recently received attention in cognitive science (e.g., Goodwin et al., 2014; Uhlmann et al., 2015). On this view, morality is person-centered in the sense that it serves mainly to identify others who are likely to behave in cooperative and trustworthy ways in the future. Although utilitarianism and deontology benefit from their elegance and impressive philosophical pedigree, person-centered approaches benefit from their theoretical links with evolutionary biology, particularly the ideas of reciprocity, signaling, and reputation as key to the evolution of morality (e.g., Miller, 2007; Nowak & Sigmund, 2005; Silver & Shaw, 2018; Sperber & Baumard, 2012; Trivers, 1971). The core idea is that moral judgments such as blame serve to adaptively identify who one

should interact with in the future (reputation-tracking), which in turn motivates others to avoid blameworthy acts (reputation-management). Thus, when acts signal that a person has poor moral character, this triggers assignment of blame.

A number of empirical findings support character-based approaches, including the assignment of blame for harmless acts that seem to imply "wicked" desires (Inbar et al., 2012), people's computational facility at moral character evaluations relative to other equivalent information integration tasks (Johnson, Murphy, et al., 2019), outrage over inconsequential acts that are nonetheless diagnostic of character (Tannenbaum et al., 2011), and the outsized impact in praise judgments of the costs (Johnson, 2020) and emotional investment (Johnson & Park, 2020) signaled by charitable contributions rather than their effectiveness. Indeed, character inferences may be a key controlling factor that guides moral attention to both outcomes and actions; for example, character inferences moderate the relationship between consequences and blame (Siegel et al., 2017).

What would character-based approaches predict? Simply, combinations of positive and negative acts should be blameworthy to the extent that they provide negative evidence about a person's moral character or reputation. This provides a link between moral judgment and diagnostic or explanatory reasoning—acts are blameworthy when their best explanation implies negative underlying propensities that best explain those acts (see Johnson et al., 2020 and Lombrozo, 2016 for reviews of explanatory reasoning; see Gerstenberg et al., 2018 and Johnson et al., 2016 on the link between explanation and social cognition). Unlike utilitarianism and deontology, which provide some notion of how moral accounting would work based on first principles (notwithstanding the various refinements described above), character-based accounts of blame are inherently less theoretically constrained: They depend on auxiliary assumptions about how people evaluate moral character. To put some reins on these theories, we rely on prior research on person perception—the study of how people infer personality and character traits based on observed actions. Some of this prior work has looked at how positive and negative information is integrated into summary judgments such as liking (Anderson, 1965; Asch, 1946; Jones, 1990; Reeder & Brewer, 1979), allowing us to derive predictions about how moral accounting of blame might work on a character-based account.

1.2. Principles of moral accounting

In this article, we test four potential *principles of moral accounting* that might underlie how we judge combinations of rights and wrongs. These principles are motivated from person perception research, but in some cases, this past research does not uniquely determine the direction of the prediction. For this reason, our studies test both person perception and moral judgments to verify that these auxiliary hypotheses about character judgment hold for our stimulus set.

1.2.1. Principle 1. Partial offsetting: bad acts can be offset by comparable good acts, but only partially

For example, consider Betty, who litters, versus Anna, who litters and then volunteers to pick up an equivalent amount of trash. On balance, Anna has done no harm, since the world has the same amount of litter before and after this combination of actions. The partial offsetting principle makes two predictions. First, Anna should be blamed less than Betty, since Anna (but not Betty) offset the amount of harm by doing an equivalent amount of good. But second, Anna should not be perceived neutrally, but instead seen as somewhat blameworthy.

Although this principle has not been tested directly, a negativity bias has long been documented in impression formation, such that negative traits weigh more heavily on liking than do positive traits (Skowronski & Carlston, 1989), and people are more sensitive to situational factors when evaluating people who did positive rather than negative actions (Reeder & Spores, 1983). This makes good sense: Norm-violating actions are by their nature rarer and likelier to be diagnostic of underlying character. Of course, this negativity bias is also consistent with many related findings (Baumeister et al., 2001; Kahneman & Tversky, 1979), including some recent evidence that blame judgments, taken in isolation, tend to be more extreme than praise (Guglielmo & Malle, 2019). If character judgments are tightly linked with blame, as we propose, then the negativity bias implies that blame offsetting, if it occurs at all, should be partial, as negative actions receive greater weight in character evaluations than equivalently positive actions.

1.2.2. Principle 2. Diminishing sensitivity: moral judgments about offsetting are insensitive to the magnitude of the good act

Let's now compare Anna (remember, she littered and then picked up a similar amount of trash) versus Christine, who litters but then picks up *twice as much* trash as she littered. On balance, Christine has now done more good than bad and the world is a better place overall for her actions. The diminishing sensitivity principle says that Christine's greater benefit should not make her much less blameworthy than Anna; specifically, the difference in moral judgments for Anna versus Christine (offsetting the harm versus twice the harm) should be much smaller than between Anna and Betty (offsetting the harm versus not offsetting at all).

Character-based theories would predict this effect if diminishing sensitivity is also found in person perception. Indeed, such effects have been documented. For example, a single untrustworthy action requires a consistent series of many trustworthy actions before trust is restored because each successive good deed bears diminishing returns; indeed, in some situations trust may never be restored, such as if the initial untrustworthy action is accompanied by deception (Schweitzer et al., 2006). More broadly, negative impressions formed by a target person's extreme bad deeds require many good deeds before that person's reputation is restored, if ever (Birnbaum, 1973; Riskey & Birnbaum, 1974). Thus, while diminishing sensitivity has not been established for blame ascriptions for acts, it is known to occur in impression formation.

Although diminishing sensitivity has been found in other domains, it is not obvious whether it would apply to blame judgments. On the one hand, diminishing sensitivity is seen in many domains, such as valuation of public goods (Frederick & Fischhoff, 1998), risky choice (Kahneman & Tversky, 1979), and charitable giving (Slovic, 2007). Particularly relevant to the current theorizing, people are more sensitive to magnitudes for selfish rather than prosocial actions (Klein & Epley, 2014). For example, a person is seen as much warmer if they make a suggested $10 donation rather than donate nothing, but donating $20 instead of $10 does not buy additional perceived warmth.

But there is also reason to think we may not see it in the case described above. This is because people's sensitivity to magnitude along a dimension is tied to how *evaluable* that dimension is (Hsee, 1996). For example, when evaluating dictionaries one at a time, people pay little attention to the number of words (10,000 vs. 20,000) since this attribute is hard to understand out of context, but when evaluating these dictionaries side-by-side, people rely heavily on this attribute. In some cases, the amount of benefit may not be particularly evaluable (e.g., Johnson, 2020), but in this case it clearly is: Whether the actor offsets precisely the amount of harm is a natural reference-point, and the actor who offset twice their harm would be plainly producing twice as much benefit as an actor who offset their harm precisely. Thus, it is plausible we would only see diminishing sensitivity to benefits after the agent's net harm is neutral and might even see increasing sensitivity up to the neutral point. Indeed, the Klein and Epley (2014) study mentioned above is consistent with this: If making the suggested donation is the reference point, then people are *especially* sensitive to donations that bring the person up to that reference point. Given these contrasting predictions, it is important to measure character and blame judgments in the same study.

1.2.3. Principle 3. Temporal asymmetry: offsetting (a bad act followed by a good act) is more permissible than licensing (a good act followed by a bad act)

Now compare Anna (again, she littered and then picked up trash) versus Diane (who picked up trash and then littered). That is, Anna's actions look like moral cleansing, redemption, or offsetting (Tangney et al., 2007; Tetlock, 2003) whereas Diane's actions look more like moral self-licensing (Merritt et al., 2010). The temporal asymmetry principle says that Diane's licensing will be judged more harshly than Anna's offsetting, even though Anna and Diane did precisely the same set of things in different orders.

Person perception research here can motivate either prediction, which is one reason we test both person perception and moral judgments in our studies. On the one hand, classic literature points to the power of first impressions, often finding primacy effects in social judgment tasks (e.g., Anderson & Hubert, 1963). On the other hand, more recent work on hypocrisy points in the opposite direction (see Effron et al., 2018 for a review). People who act in opposition to their stated moral views tend to be judged more harshly when their avowal of a norm (a positive action) precedes its violation (a negative action) rather than the converse order (Barden et al., 2005). This is thought to occur because people are likelier to believe that a person's character has changed for the better when the norm avowal occurs after the violation, which explains why this asymmetry is larger for in-group rather than out-group members (Barden et al., 2014). Another example of a recency effect in person perception is the *end-of-life* bias, in which people's actions near the end of their lives receive far greater weight than their actions earlier in their lives when third-parties form summary judgments of their moral character (Newman et al., 2010), perhaps because the later actions are thought to be more revealing of the "true self." Given the mix of primacy and recency effects in the literature, we test both character and blame judgments to resolve this ambiguity.

1.2.4. Principle 4. Act congruency: moral judgments about offsetting depend on the match between the good and bad acts

Finally, consider again Anna (she littered and then picked up trash in the same area where she littered) versus Emma (who littered and picked up trash in a different area) versus Francine (who littered and mowed her neighbor's lawn). Even if these three offsetting acts are seen as equally beneficial in isolation, the act congruency principle says that people would nonetheless think that Anna is less blameworthy than Emma, who in turn is less blameworthy than Francine.

This principle is the most unique to the moral accounting framework, since it concerns the qualitative relationship between the harm and benefit: To what extent does "like offset like," or do our minds track a universal system of moral credits and debits? To our knowledge, the person perception literature contains no direct demonstrations of this, although there is related work on moral self-licensing. First, there is evidence of licensing both within-domain (e.g., hiring a minority applicant licenses expression of prejudiced attitudes; Monin & Miller, 2001) and cross-domain (e.g., eco-friendly behaviors license cheating in an unrelated task; Mazar & Zhong, 2010). Second, the mechanisms underlying these effects seem to differ (Effron & Monin, 2010). Within-domain licensing seems to occur because people accrue "moral credits" that they then feel licensed to "spend" on subsequent transgressions (Hollander, 1958; Nisan, 1991). In contrast, cross-domain licensing seems to occur mainly because people acquire "moral credentials" that they can integrate into their self-concept and which shapes the interpretation of, and can justify, subsequent behaviors (Monin & Miller, 2001). Third, when transgressions are blatant rather than ambiguous, within-domain is weaker than cross-domain licensing, and indeed may not occur at all, because within-domain transgressions trigger the perception of hypocrisy (Effron & Monin, 2010). All this suggests that *less* congruent acts would be more powerful offsets than more congruent acts—the opposite of the proposed principle.

Why might we nonetheless expect positive acts to better offset more congruent negative acts? One reason is that self-licensing and moral accounting take place at different time points. Whereas moral credentials and credits in self-licensing are evaluated after an initial positive act but before the negative act, moral accounts take account of both actions simultaneously. Thus, whereas highly congruent negative actions can feel hypocritical to an actor after having done a positive action, observers who get a broader sense of the overall picture may not interpret the sequence of actions in the same way, and indeed may view more similar acts as more redemptive as they can be more readily construed as expressions of remorse. This prediction has not been tested in person perception, so it is necessary to validate this assumption empirically in the current studies.

Another way to think about the act congruency principle is by analogy to mental accounting phenomena in consumer behavior (Thaler, 1985). The essence of mental accounting is that income, expenses, assets, and debts are segregated into different mental accounts, for instance based on income source, rather than mentally consolidating income streams as economists would recommend. These behaviors result from fundamental cognitive processes surrounding categorization (Henderson & Peterson, 1992) that apply equally to categorizing income streams and moral actions. Thus, analogous to traditional mental accounting, one might theorize that moral credits belong to different "moral accounts," such that a credit for a beneficial act can only be applied against a debit for a harmful act from the same category. This predicts the act congruency principle. Even though Francine might be thought praiseworthy for mowing her lawn in isolation, this does not help to clear the negative moral account for her littering. Francine has one moral account in the black and another in the red.

Why might a person with two neutral moral accounts be thought higher in moral character than a person with one moral account in the red and an offsetting moral account in the black? This follows directly from the same negativity bias in person perception that motivates the partial offsetting principle (Skowronski & Carlston, 1989). Even if the size of the moral credits and debits are equivalent, the debit looms larger than the credit, leading to overall negative character perception. Given the hypothesized link between character and blame, Anna (with her accounts nearly in balance) would be deemed less blameworthy than Francine (with a large account in the red and another in the black).

1.2.5. Predictions

Table 1 sets out the predictions made by utilitarian, deontological, and character-based approaches to moral judgment. On the most basic operationalization of utilitarianism, none of the four effects should occur, since moral credits and debits should be fully fungible. That is, positive and negative acts of equivalent harm and benefit should offset one another and be sensitive to the relative magnitudes. The temporal order and congruency of the actions do not influence overall utility and thus should not be incorporated into the moral calculus. As noted above, more sophisticated versions of utilitarianism might be devised to accommodate some of these effects, but on the most common operationalization used in psychology research utilitarianism has difficulty doing so.

On a basic operationalization of deontology, we would expect no offsetting at all, and therefore extreme magnitude insensitivity. Indeed, these predictions are core to how deontology and utilitarianism differ, since the very intuition deontology is trying to capture is that some actions cannot be permitted even for the greater good. Since deontology is act-based, it does not seem to predict effects of temporal order or the congruency between the harm and offset, since the act itself is held constant in all of these cases.

Finally, character-based approaches make conditional predictions. We assume, based on prior research in person perception, that effects of actions on character perception allow partial but not full offsetting and are insensitive to magnitude. We attempt to replicate these prior findings, and if we do, we would expect to find downstream effects on blame. As explained above, predictions about character perception for

Table 1
Predictions of different theoretical approaches for the four proposed principles of moral accounting.

Principle	Utilitarianism	Deontology	Character-based
1. ***Partial offsetting*****:** Bad acts can be offset by comparable good acts, but only partially.	No – zero net harm, so offset acts should not be blameworthy.	No – if the harm violated a moral rule, offsetting should not reduce blame.	Possibly – if beneficial act provides some character information, but not enough to override the initial harm.
2. ***Diminishing sensitivity*****:** Moral judgments about offsetting are insensitive to the magnitude of the good act.	No – net harm scales linearly with the amount of benefit.	Yes – increasing benefits do not override a rule violation.	Possibly – if increasingly beneficial acts reveal little about character.
3. ***Temporal asymmetry*****:** Offsetting (a bad act followed by a good act) is less blameworthy than licensing (a good act followed by a bad act).	No – net harm does not depend on temporal order.	No – temporal order does not affect rule violation.	Possibly – if offsetting implies remorse while licensing implies moral entitlement.
4. ***Act congruency*****:** Moral judgments about offsetting depend on the match between the good and bad acts.	No – net harm does not depend on match between harm and benefit.	No – rule was violated regardless of the kind of benefit.	Possibly – if higher congruency better keeps moral accounts balanced, reflecting better character.

temporal order and act congruency are unclear based on prior evidence, so we test the directions of these effects. If, as we intuited, we see recency effects and positive effects of congruency in character judgments, then we would also expect such effects for blame judgments.

1.3. The current studies

The central contributions of this article are (i) identifying and testing plausible principles of the moral accounting of blame; and (ii) examining whether a character-based account explains these blame judgments. In some cases, prior literature on person perception strongly constrains the plausible predictions of a character-based account (partial offsetting and diminishing sensitivity). In other cases, character-based accounts can make directional predictions only if we also establish whether these principles apply in character judgment (temporal asymmetry and act congruency). To be clear, character-based accounts of *blame* do not make predictions about how people evaluate moral *character*; that is the role of person perception theories, which is not our central theoretical contribution. We do provide arguments as to why we think these principles are plausible for character judgment, and speculate further in the General discussion. However, the prediction made by character-based accounts is simply *alignment* between character judgments and blame judgments.

Therefore, Studies 1–4 test Principles 1–4, respectively, documenting the key phenomena and anomalies underlying our propensity to mentally combine moral harms and benefits into overall judgments of blame or praise. To establish the compatibility (or lack thereof) of these principles with character-based approaches, Studies 2–4 measure character inferences as well as moral judgments.

Beyond compatibility between these theories and the proposed principles of moral accounting, Studies 5–9 test these theories directly. Studies 5 and 6 test the ability of utilitarian, deontological, and character-based approaches to explain, respectively, differences across the items used in Studies 1–4 and across individuals. Finally, Studies 7–9 test character-based approaches experimentally, by using several different manipulations of moral character to test for downstream consequences on blame.

2. Studies 1–4

We constructed a variety of vignettes involving various kinds of moral violations in which the consequences of harmful acts could be offset by an equivalently beneficial act. These vignettes were used to test the four principles summarized in Table 1.

2.1. Methods

2.1.1. Study 1

Participants read 10 vignettes, each describing a harmful act (see Table 2 for examples and Table A1 in the Supplementary Materials for all items in Study 1). Half of the vignettes appeared in the *Harm-Only* condition, so that the protagonist performed only the harmful act (e.g., littering 2 pounds of trash). The other half of the vignettes appeared in the *Offset* condition, so that the protagonist performed the harmful act as well as an equivalent act that serves to offset the harm from the harmful act (e.g., volunteering to pick up trash, collecting about 2 pounds). The assignment of vignette to condition was counterbalanced, with vignettes assigned pseudorandomly to the two counterbalancing conditions, and the order of the items was random.

The key dependent variable for each item was a *blame* judgment ("Overall, do you think [*protagonist*] deserves to be blamed or praised for these actions?") made on a scale from −5 to 5. For half of the participants, negative numbers corresponded to blame (−5 = "Very much blamed") and positive numbers to praise (5 = "Very much praised"), while the scale was reversed for the other half of the participants. For reporting results, scores were adjusted so that negative scores correspond to blame. Participants were also asked to report the overall *harm* of the action ("Considering the negative aspects (if any) of [*protagonist*]'s actions, how much harm do you think [*protagonist*] caused?") and the overall *benefit* of the action ("Considering the positive aspects (if any) of [*protagonist*]'s actions, how much good do you think [*protagonist*] caused?") on scales from 0 ("Not very much") to 10 ("Very much"). The order of the benefit and harm questions were counterbalanced across participants and always preceded the blame question. We do not analyze the benefit and harm judgments here, but instead use them to test hypotheses about variability across vignettes in Study 5.

2.1.2. Study 2

The method was based on Study 1, while measuring character and varying the magnitude of the offsetting benefit (see Table 2 for examples). For each item, participants were first told about the harmful act (e.g., littering 2 pounds of trash) and were then asked to judge the protagonist's moral character ("Given this information, how would you judge [*protagonist's*] moral character?") on a −5 ("Very bad") to 5 ("Very good") scale. On the next screen, participants were told about the offsetting benefit that the protagonist performed (e.g., picking up 2 pounds of trash) and asked to re-judge character ("Given this **new** information, how would you judge Taylor's moral character") on the same scale, as well as blame ("Considering [*protagonist's*] [*harm*] **and** their [*benefit*], do you think [*protagonist*] deserves to be blamed or praised for these actions overall?") on a scale from −5 ("Very much blamed") to 5 ("Very much praised"). On this screen, the harm was also re-presented with the new information presented in bold typeface.

Each vignette appeared in one of two benefit quantity conditions. In the *Single-Benefit* condition, the benefit was similar in magnitude to the harm (e.g., picking up 2 pounds of trash). In the *Double-Benefit* condition, the benefit was double this magnitude (e.g., picking up 4 pounds of trash). The vignettes always noted explicitly that the offset was of similar magnitude to the harm (e.g., "around 2 pounds of trash – the

Table 2
Sample stimuli from Studies 1–4.

Study 1	Offset		Harm-only
	Last week, Riley used around five pounds of non-renewable, plastic products, such as straws and plastic bags. Five pounds of plastic waste can be cleaned up for $9. Knowing this, Riley donates $9 to the Ocean Cleanup project to offset the amount of plastic they produced.		Last week, Riley used around five pounds of non-renewable, plastic products, such as straws and plastic bags.
Study 2	Single-offset		Double-offset
	Last week, Riley used around five pounds of non-renewable, plastic products, such as straws and plastic bags. Five pounds of plastic waste can be cleaned up for $9. Knowing this, Riley donates $9 to the Ocean Cleanup project to offset the amount of plastic they produced – the amount needed to offset the trash produced.		Last week, Riley used around five pounds of non-renewable, plastic products, such as straws and plastic bags. Five pounds of plastic waste can be cleaned up for $9. Knowing this, Riley donates $18 to the Ocean Cleanup project to more-than-offset the amount of plastic they produced – twice the amount needed to offset the trash produced.
Study 3	Offset		Licensing
	Last week, Riley used around five pounds of non-renewable, plastic products, such as straws and plastic bags. Five pounds of plastic waste can be cleaned up for $9. This week, Riley donates $9 to the Ocean Cleanup project, since this donation offsets last week's plastic consumption.		Last week, Riley donated $9 to the Ocean Cleanup project. Five pounds of plastic waste can be cleaned up for $9. This week, Riley uses around five pounds of non-renewable, plastic products, such as straws and plastic bags, since this plastic consumption was offset by last week's donation.
Study 4	High congruency	Medium congruency	Low congruency
	Last week, Riley used around five pounds of non-renewable, plastic products, such as straws and plastic bags, which can cause damage to the oceans. Later, Riley donates $9 to the Ocean Cleanup project, which helps clean up plastic in the ocean.	Last week, Riley used around five pounds of non-renewable, plastic products, such as straws and plastic bags, which can cause damage to the oceans. Later, Riley donates $9 to the Ocean Cleanup project, which helps clean up oil spills.	Last week, Riley used around five pounds of non-renewable, plastic products, such as straws and plastic bags, which can cause damage to the oceans. Later, Riley participates as a volunteer during election day.

amount they had littered") or double the magnitude ("around 4 pounds of trash – double the amount they had littered"). The procedure was otherwise the same as Study 1, with half of the vignettes assigned to each condition, counterbalanced across participants.

2.1.3. Study 3

The method was identical to Study 2, except that the vignettes each appeared in one of two temporal order conditions rather than quantity conditions (see Table 2 for examples). In the *Offset* condition, the harmful act preceded the beneficial act, as in Study 1 (e.g., littering and then picking up trash). Thus, the initial character judgment was made after knowing only about the harm, whereas the final character and blame judgments were made in light of both the harm and the benefit (the new information was bolded). The Offset condition was thus similar to the Single-Benefit condition of Study 2. In the *Licensing* condition, the beneficial act preceded the harmful act (e.g., picking up trash and then littering), as in the behaviors documented in moral self-licensing experiments (Monin & Miller, 2001). Thus, the initial character judgment was made after knowing only about the benefit, whereas the final judgments were made in light of both the benefit and the harm. The procedure was otherwise the same as Study 2, with half of the vignettes assigned to each condition, counterbalanced across participants.

2.1.4. Study 4

The method was identical to Study 2, except that the vignettes each appeared in one of three congruency conditions rather than quantity conditions (see Table 2 for examples). In the *High-Congruency* condition, the benefit directly counteracted the harm (e.g., picking up litter in the neighborhood where one had previously littered). In the *Medium-Congruency* condition, the beneficial act was in the same category as the High-Congruency condition, but had different beneficiaries or conserved a different resource (e.g., picking up litter in a different city). In the *Low-Congruency* condition, the beneficial act was completely unrelated to the harm (e.g., mowing the neighbor's lawn). The procedure was otherwise the same as Studies 2 and 3, with about one-third of the vignettes assigned to each condition, counterbalanced across participants.

We closely matched the stimuli between the High- and Medium-Congruency conditions so that the benefits are roughly equivalent (e.g., cleaning up carbon dioxide versus methane emissions). To ensure that the benefits in the Low-Congruency condition were perceived as equivalent to the High-and Medium-Congruency conditions, we conducted a pretest. We asked a separate group of participants (N = 50; 5 excluded) to rate each of 30 different benefits in terms of "how much good [*protagonist*] caused" on 0–10 scales. These 30 benefits included the 10 benefits from the Medium-Congruency condition, and 20 potential benefits to be used in the Low-Congruency condition. We selected items based on the results of this pretest, matching a Low-Congruency benefit as closely as possible to each item's Medium-Congruency condition. This resulted in very similar benefit scores across conditions (Ms = 5.28 vs. 5.30 for Medium- and Low-Congruency, respectively).

2.1.5. Participants

We recruited 100, 99, and 99 participants respectively for Studies 1–3 (with 2 conditions) and 150 participants for Study 4 (with 3 conditions). Participants in all studies were recruited and compensated through Amazon Mechanical Turk and were prevented from participating in multiple studies reported in this article.

Mechanical Turk samples are more diverse in socioeconomic status, education, and age compared to undergraduate samples; thus, we anticipate that the results reported in this article would generalize well across these variables. However, Turkers tend to be more politically liberal compared to the general American public, thus caution should be observed when generalizing results here that may be linked to broader political attitudes. (That said, we measure and statistically adjust for political orientation in Study 6 for the specific case of climate offsets.) We do not make any claims about the generality of these results beyond Western populations, but instead discuss broader cross-cultural issues in the General discussion.

The sample size was set a priori for all studies. Our planned sample sizes for Studies 1–3 and 4, respectively, suffice to detect with 90% power small- to medium-sized effects (d > 0.33 and 0.27 for N = 100 and 150, respectively) in the within-subjects designs. A larger sample size was planned for Study 4 because it divided items into 3 conditions rather than 2 conditions, requiring a larger number of participants to attain an equally precise estimate for an item analysis (Study 5).

We sought to further maximize statistical power by including data quality checks. After the main task for all studies, participants completed a series of check questions (testing recognition memory for the items) and were excluded from analysis if they answered more than one-third incorrectly (for Studies 1–4, Ns = 18, 17, 12, and 13, respectively) to avoid inattentive participants.

2.2. Results

Overall, the results support all four proposed principles of moral accounting—partial offsetting, diminishing sensitivity, temporal asymmetry, and act congruency. Means across all studies are shown in Figs. 1 and 2, and are broken down across items in Table A2 in the Supplementary Materials. All data are available through the Open Science Framework (https://bit.ly/2m9vCYT).

2.2.1. Study 1

Study 1 tested partial offsetting (Principle 1). This principle is composed of two claims—first, that an offset harm is perceived as less blameworthy than a non-offset harm, but second, that the offset harm is still perceived as more blameworthy rather than morally neutral.

We tested the first claim by contrasting Study 1's Harm-Only condition (the protagonist performed only a harmful act) against the Offset condition (the protagonist also performed a countervailing beneficial act). As shown in Fig. 1, blame judgments were significantly more negative in the Harm-Only than in the Offset condition. This difference

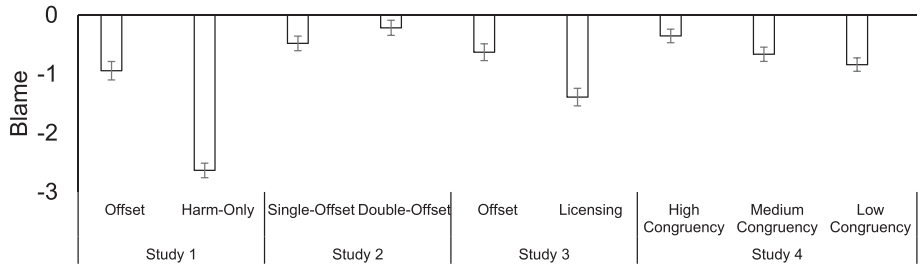

Fig. 1. Blame judgments across Studies 1–4.
Scale ranges from −5 to 5. Bars represent 1 SE.

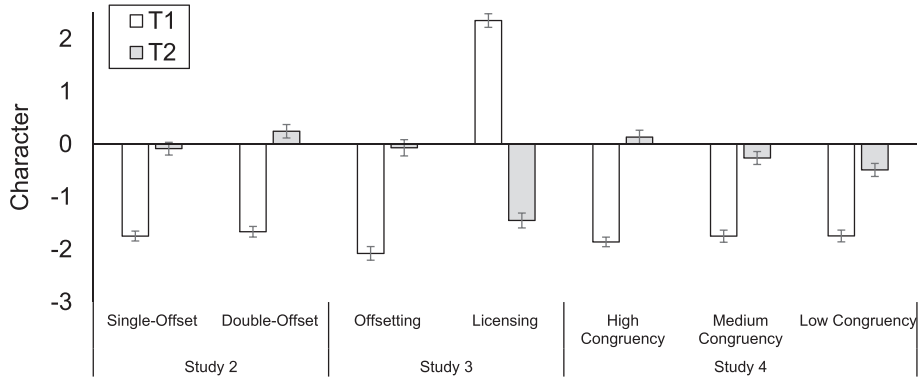

Fig. 2. Character judgments across Studies 2–4.
Scale ranges from −5 to 5. Bars represent 1 SE.

was statistically robust both across participants [$t(81) = 9.72$, $p < .001$, 95% CI_d[0.85,1.29]] and across items [$t(9) = 5.12$, $p < .001$, 95% CI_d[0.90,2.34]]. (We use 'CI $_d$' to refer to a CI on Cohen's *d* effect sizes, calculated by scaling the CI on a difference score by its standard deviation.) Thus, performing beneficial acts to offset harmful acts does result in less extreme blame judgments.

We tested the second claim by contrasting Study 1's Offset condition against the blame/praise scale's neutral midpoint of 0. These scores were more negative than the scale midpoint, significantly across participants [$t(81) = 6.02$, $p < .001$, 95% CI_d[0.45,0.88]] and marginally across items [$t(9) = 1.96$, $p = .082$, 95% CI_d[−0.10,1.33]]. Thus, even though protagonists are seen as *less* blameworthy when they perform beneficial acts to offset their harmful acts, such combinations are seen on balance as somewhat blameworthy. Overall, this supports the principle of partial offsetting: Offsets are not complete, but they are still fairly large in magnitude.

2.2.2. Study 2

Study 2 tested diminishing sensitivity (Principle 2), the notion that one receives diminishing degrees of moral credit as the amount of benefit increases. We tested this by comparing the Single-Benefit (the protagonist performed a benefit that exactly offset the harm) against the Double-Benefit condition (the protagonist performed a benefit that offset twice the harm). As shown in Fig. 1, blame judgments differed somewhat across conditions, reaching significance by subject [$t(81) = 2.25$, $p = .027$, 95% CI_d[0.03,0.47]] but not by item [$t(9) = 1.44$, $p = .18$, 95% CI_d[−0.26,1.17]].

However, compared to Study 1, this difference was much smaller ($d = 1.07$ vs. 0.25 by subject; $d = 1.62$ vs. 0.46 by item). Since the amount of benefit differed in equal steps between the Harm-Only, Single-Offset, and Double-Offset conditions, this result supports diminishing sensitivity: Increasing the benefit from zero to the size of the harm makes a large difference to blame, but increasing the benefit from the size of the harm to twice its size makes a much smaller difference.

We also could assess whether there is significant blame overall (relative to the scale midpoint) separately in the Single-Offset and Double-Offset conditions. Study 2 replicated partial offsetting, in that the Single-Benefit condition again produced negative blame judgments, which were significantly different from 0 when analyzed by subject [$t(81) = 3.89$, $p < .001$, 95% CI_d[0.21,0.65]], though not by item [$t(9) = 0.73$, $p = .48$, 95% CI_d[−0.48,0.95]]. Of particular interest, blame judgments were still negative even in the Double-Offset condition, though not significantly so [$t(81) = 1.69$, $p = .095$, 95% CI_d [−0.03,0.41] by subject; $t(9) = 0.41$, $p = .69$, 95% CI_d[−0.59,0.85] by item]. That is, even when the protagonist offset the harm by a factor of two, the act was still not seen as praiseworthy.

Although diminishing sensitivity is clearly incompatible with utilitarianism—since twice as much good was accomplished in the Double-Benefit than the Single-Benefit condition, despite minimal differences in blame—it could be compatible with both deontological and character-based approaches. We test deontology in Studies 5 and 6, but in the meantime we examine the dynamics of character inferences to see whether the results are empirically consistent with a person-based account. That is, we can test the auxiliary hypotheses of character-based accounts summarized in Table 1.

Recall that Study 2 measured character judgments twice—after the protagonist did the harmful action (Time-1), and after they did the beneficial action (which either offset the harm by a factor of one or two; Time-2). Thus, if character judgments explain diminishing sensitivity we would expect a large difference between the Time-1 and Time-2 judgments in the Single-Benefit condition. But we would expect only a modest difference between the Time-2 judgments in the Single-Benefit and Double-Benefit conditions, even though the difference in benefit is the same in these two contrasts.

This is exactly what we found, as shown in Fig. 2. In the Single-Benefit condition, participants believed the protagonist had dramatically better moral character at the Time-2 than the Time-1 judgment, after the protagonist had offset their harm [$t(81) = 11.95$, $p < .001$, 95% CI_d[1.10,1.54] by subject; $t(9) = 5.45$, $p < .001$, 95% CI_d [1.01,2.44]]. But the difference in Time-2 character judgments between the Single- and Double-Benefits conditions was, though statistically significant by subject, comparatively modest in magnitude [$t(81) = 2.72$, $p = .008$, 95% CI_d[0.08,0.52] by subject; $t(9) = 1.40$, $p = .19$, 95% CI_d[−0.27,1.16] by item]. Thus, the pattern of blame judgments—exhibiting highly decreasing sensitivity to increasing amounts of benefit—is mirrored in participants' character inferences. Participants did not think that a person willing to offset their harm was much less virtuous than a person willing to offset their harm by a factor of two.

Moreover, Study 2 uncovered mediation patterns consistent with a person-based account. Using the MEMORE package for mediation with repeated measures (Montoya & Hayes, 2017), the Time-2 character judgments mediate the effect of Single- vs. Double-Benefit condition on blame judgments [$b = 0.26$, $SE = 0.09$, 95% CI[0.07,0.43]].

One could read these results as supporting person-based accounts for diminishing sensitivity (since increasing levels of benefit result in decreasing informational returns for character) but not partial offsetting (since Time-2 character judgments are near the zero-point in the Single-Benefit condition, despite negative blame judgments). However, Study 2 did not include a Time-0 character judgment before the harm occurred, so we do not know whether the Time-2 judgment is indeed more negative than a hypothetical Time-0 judgment. Yet, we do measure Time-0 judgments in Study 7, and (to preview our findings) we find there that when only neutral information about moral character is provided, character judgments tend to be rather positive ($M = 1.60$; see Fig. 3). Thus, the Study 2 results are consistent with a character-based

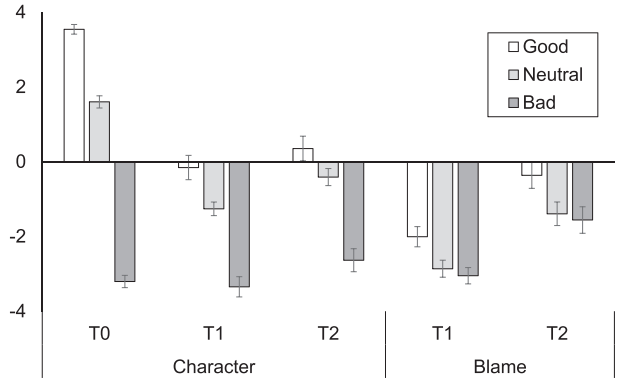

Fig. 3. Character and blame judgments in Study 7.
Scales range from −5 to 5. Bars represent 1 SE.

account of partial offsetting, since the beneficial act does not successfully restore character back to default (pre-harm) levels.

2.2.3. Study 3

Study 3 tested temporal asymmetry (Principle 3), the notion that pairs of harmful and beneficial acts are seen as less blameworthy when the harm occurs before, rather than after, the benefit. We tested this by comparing the Offsetting condition (the harmful act preceded the beneficial act) against the Licensing condition (the harmful act occurred after the beneficial act). As shown in Fig. 1, blame judgments were more negative in the Licensing than in the Offsetting condition [$t(86) = 5.63$, $p < .001$, 95% CI_d[0.39,0.82] by subject; $t(9) = 4.51$, $p = .001$, 95% CI_d[0.71,2.14] by item]. This supports temporal asymmetry (Principle 3). Moreover, Study 3 again replicates partial offsetting (Principle 1), as blame judgments were negative whether offsetting or licensing.

These results are clearly incompatible with utilitarianism, since the total harm and benefit are equal regardless of their temporal order. They are also hard to explain on a deontological account, since the same violations of moral rules occurred in either order. But they could be compatible with character-based moral judgment. To explain this pattern, the Time-2 character judgments would need to be more negative in the Licensing condition than in the Offset condition. For example, this pattern of character judgments would be consistent with the inference that people who offset harms with later beneficial acts are doing so out of guilt or shame, whereas people who use earlier beneficial acts to license later harms are doing so out of a sense of entitlement.

As shown in Fig. 2, Time-2 character judgments were indeed more negative in the Licensing than in the Offset condition [$t(86) = 8.00$, $p < .001$, 95% CI_d[0.64,1.07] by subject; $t(9) = 5.88$, $p < .001$, 95% CI_d[1.14,2.57]]. This occurred because the second act was perceived as much more diagnostic of character when it was a harmful rather than beneficial act, as manifested in the much larger differences between the Time-1 and Time-2 character judgments in the Licensing versus the Offset condition (Fig. 2). Moreover, as in Study 2, the Time-2 character judgments mediated the effect of condition (Offset vs. Licensing) on blame [$b = 0.85$, $SE = 0.11$, 95% CI[0.64,1.07]].

2.2.4. Study 4

Finally, Study 4 tested act congruency (Principle 4), the notion that pairs of harmful and beneficial acts are seen as less blameworthy to the extent that the beneficial act is seen as directly counteracting the harm. We tested this by comparing the High-Congruency condition (in which the benefit directly offset the harm) against the Medium-Congruency condition (in which the benefit was of a similar kind to the harm, but did not directly offset it) and the Low-Congruency condition (in which the benefit was dissimilar to the harm).

As shown in Fig. 1, blame judgments became increasingly harsh as the offset became less congruent with the harm: Blame was more negative in the Medium Congruency than in the High Congruency condition [$t(136) = 2.27$, $p = .025$, 95% CI_d[0.02,0.36] by subject; $t(9) = 2.93$, $p = .017$, 95% CI_d[0.21,1.64] by item]. However, the trend for more negative blame judgments in the Low Congruency than in the Medium Congruency condition did not reach significance [$t(136) = 1.28$, $p = .20$, 95% CI_d[−0.06,0.28] by subject; $t(9) = 0.37$, $p = .71$, 95% CI_d[−0.60,0.83] by item].

These results are incompatible with utilitarianism, since we equated the total benefit across the three conditions, and with deontology, since the same moral rule was violated in each condition. But they could be consistent with character-based accounts, if people believe that people have different "moral accounts," such that having moral accounts in the red signals poor moral character and more congruent offsets better offset an associated harm. In that case, the less congruent the offset is, the less the offset ameliorates perceived character, and the more blame is assigned.

Indeed, the Time-2 character judgments were significantly more negative for the Medium- than for the High-Congruency condition [$t(136) = 2.74$, $p = .007$, 95% CI_d[0.07,0.40] by subject, $t(9) = 2.92$, $p = .017$, 95% CI_d[0.21,1.64]], corresponding to the large increase in blame between these conditions. Moreover, Time-2 character judgments mediated the effect of Medium- vs. High-Congruency on blame [$b = 0.33$, $SE = 0.12$, 95% CI[0.10,0.56]].

Given the smaller and less statistically robust increase in blame between the Medium and Low Congruency conditions, character-based accounts would predict a smaller decrement in Time-2 character judgments across these conditions—which is exactly what was observed, with this difference reaching marginal significance by subject [$t(136) = 1.69$, $p = .093$, 95% CI_d[−0.02,0.31]] but not by item [$t(9) = 0.74$, $p = .48$, 95% CI_d[−0.48,0.95]].

2.3. Discussion

The results support the four proposed principles of moral accounting—partial offsetting, diminishing sensitivity, temporal asymmetry, and act congruency. These principles conflict with utilitarianism, and temporal asymmetry and act congruency are also hard to explain on deontological accounts. However, the observed patterns of character judgments confirm the auxiliary hypotheses made by a character-based account (Table 1) and in several cases mediated the effects on blame. Thus, these results support both our empirical framework and our proposed theoretical explanation for it.

3. Study 5

Whereas Studies 1–4 tested how well the four *principles* of moral accounting are explained by utilitarianism, deontology, and character-based moral judgment, Study 5 examined how well these accounts explained variability across the ten *vignettes*. Study 5A measured beliefs about deontology for the Harm-Only version of each vignette, while Study 5B measured inferences about character for both the Harm-Only and Offset versions. These predictors were combined with judgments of harm and benefit from Study 1 (testing utilitarian accounts) to predict blame.

3.1. Method

The method of Study 5A was broadly similar to Studies 1–4, except that (i) participants saw only the Harm-Only version of each vignette from Study 1 (see Table 2), and (ii) instead of judging blame or character, they completed a series of ten questions adapted from Baron and Spranca (1997) (see Part B of the Supplementary Materials), intended to measure the extent to which people think about each scenario deontologically. This scale was composed of five subscales, measuring

Quantity Insensitivity (e.g., "It is equally wrong for some of this to happen as for twice as much to happen"), *Agent Relativity* (e.g., "I mainly have an obligation to stop this only if I am personally involved"), *Objectivity* (e.g., "People have an obligation to stop this even if they think they do not"), *Anger* (e.g., "Thinking about this bothers me"), and *Trade-off Denial* (e.g., "In the real world, there is nothing we can gain by this happening"); one item from each subscale was reverse-coded. Participants were also asked to rate the harmfulness of each act ("This causes a large amount of harm"). Each question was answered on a scale from −5 ("Strongly disagree") to 5 ("Strongly agree"). The order of the vignettes was random, as was the order of the 11 questions for each vignette.

Study 5B followed the same procedure as Study 5A, with two changes. First, each participant saw both the Harm-Only and Offset versions of each vignette from Study 1, with these versions contained in separate blocks, with the order of the vignettes randomized and the order of the blocks counterbalanced. Second, rather than the deontology scale, participants made a series of eight character judgments (in a random order) for each vignette, asking participants to "judge [*protagonist*] on the following traits," with measurements of perceived honesty ("trustworthy," "dishonest"), justice ("fair," "unjust"), kindness ("kind", "mean"), and responsibility ("prudent", "irresponsible"), on scales from −5 ("Not at all") to 5 ("Very much"); the latter item from each pair was reverse-coded. These dimensions were adapted from cross-cultural work on character virtues (Dahlsgaard et al., 2005).

We recruited 99 and 100 participants, respectively, for Studies 5A and 5B. We excluded participants using the same criterion as Studies 1–4 (*Ns* = 23 and 15, respectively).

3.2. Results

Overall, variability in deontology across items had little predictive power for blame judgments (providing little support for act-based processes), whereas harm and character had relatively consistent effects (supporting outcome-based and person-based processes).

We analyze the deontology and character scales themselves in detail in Part B of the Supplementary Materials. Here, we simply note that the deontology scale had good reliability with the agent relativity subscale removed [$\alpha = 0.91$], and the character scale had excellent reliability without any deletions [$\alpha = 0.99$]. The means for deontology (Study 5A), character (Study 5B), and net harm (Study 1) for each item are given in Table A3 in the Supplementary Materials.

For our main analyses, we first used multiple regression to test correlates of blame in the Harm-Only condition (Study 1) at the item level. We entered three variables as predictors of blame (each centered at its mean and scaled by its standard deviation). First, according to deontological approaches, the extent to which an act is perceived as deontological should predict blame judgments (i.e., should have a significantly negative coefficient, since negative numbers correspond to higher blame). However, deontology scores from Study 5A were not a significant predictor of blame [$b = 0.25$, $SE = 0.27$, $p = .39$]. Second, according to utilitarianism, the net harm (total harm – total benefit) should predict blame (with a negative coefficient). Using the difference between perceived harm and benefit judgments from Study 1, this variable did significantly predict blame [$b = -1.09$, $SE = 0.38$, $p = .028$]. Finally, according to person-based approaches, character judgments (positive traits – negative traits) should predict blame (with a positive coefficient). Averaging separately the positive and negative trait judgments in Study 5B and taking the difference for each item, these character judgments significantly predicted blame [$b = 0.73$, $SE = 0.23$, $p = .021$]. Overall, this model explained the vast majority of the variance in blame [$R^2 = 0.98$]. Thus, these results support outcome- and person-based processes, but not act-based processes, as important drivers of blame.

This last analysis looked at how people assign blame to acts that are merely harmful. But our main interest in this article is examining blame when harmful acts are bundled with beneficial, offsetting acts. Thus, we sought to model the *difference scores* between the Offset and Harm-Only conditions of Study 1, as a measure of how much an act could be offset. These scores too we predicted from three variables (again, centered at their means and scaled by their standard deviations). First, deontology assumes that an act is offsetable to the extent that it does not violate deontological rules. The deontology scores from Study 5A did not predict the difference scores [$b < 0.01$, $SE = 0.10$, $p = .96$], contradicting this prediction. Second, outcome-based approaches assume that an act is offsetable to the extent that the offset produces net benefits relative to the harm. Thus, we calculated the difference scores in net harm (harm – benefit) between the Offset and Harm-Only conditions of Study 1. These scores were marginally significantly predictors of differences in blame across conditions [$b = -0.48$, $SE = 0.20$, $p = .057$]. Finally, person-based accounts assume that an act is offsetable to the extent that the offset rehabilitates the person's moral character. Thus, we calculated the difference scores in character between the Offset and Harm-Only conditions of Study 5B. These scores significantly predicted differences in blame [$b = 0.57$, $SE = 0.20$, $p = .026$]. This model explained almost all the variance in blame difference scores [$R^2 = 0.95$]. Thus, the results for blame reduction in light of offsetting run parallel to the results for the harmful acts alone—blame reduction due to offsetting is predicted by improvements in net harm and in perceived character, but not the deontological characteristics of the act itself.

3.3. Discussion

These results add to Studies 1–4 in further supporting character-based accounts, not only in explaining the basic principles of moral accounting, but also in explaining variability across different kinds of harms and offsets. Study 5 also demonstrated attention to net harm, but not to violations of deontological principles (see Bartels & Medin, 2007 for related results).

4. Study 6

Whereas Study 5 sought to explain variability across scenarios, Study 6 looked at variability across *individuals*. We focused on carbon offsets as an especially important test case, given potential policy and marketing implications for attitudes toward individuals who purchase offsets. Specifically, we test the extent to which (i) vignette-specific differences in deontological construals and perceptions of harm, (ii) trait differences in utilitarianism and deontology, and (iii) attitudes toward the environment predict judgments about people purchasing carbon offsets.

4.1. Method

Participants each completed a series of measures in a fixed order. First, participants reported their political attitudes toward a number of environmental policies, including carbon offsets, cap-and-trade, carbon fines, and phasing-out of coal power. For each policy, participants read a brief description and were asked the extent to which they support the policy on a scale from −5 ("Strongly disagree") to 5 ("Strongly agree").

Second, participants responded to two offsetting vignettes, similar to the Offset condition of Study 1. One vignette was identical to the Plane scenario used in Study 1 (see Table 2). The other vignette was a parallel scenario with a company (rather than individual) paying to offset carbon emissions. For each vignette, participants completed first the Study 5A measures (i.e., the deontology scale) in a random order, and then the Study 5B measures (i.e., character judgments) in a random order. Since participants were given the Offset (rather than Harm-Only) version of the vignettes, the deontology scale was prefaced by

"Concerning the [*harm*], to what extent do you agree or disagree with each of the following statements?" Each vignette appeared on its own page in a counterbalanced order.

Third, participants completed two scales, measuring beliefs in utilitarianism and deontology (in a counterbalanced order, with items randomized within each scale). The utilitarianism measure was the Oxford Utilitarianism Scale (Kahane et al., 2018), itself composed of two subscales. The five items on the *impartial beneficence* subscale [$\alpha = 0.78$] measured impartial concern for the good of humanity (e.g., "It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal"), while the four items on the *instrumental harm* subscale [$\alpha = 0.78$] measured willingness to sacrifice others to the greater good (e.g., "It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people"). The deontology measure was the deontology subscale of Robinson's (2012) consequentialism scale [$\alpha = 0.78$], testing people's rigidity in the face of moral rules (e.g., "Some rules and laws are universal and are binding no matter the circumstances you find yourself in"). Two filler items were included but not analyzed. Items on both scales were judged from –5 ("Strongly disagree") to 5 ("Strongly agree").

Fourth, participants completed an environmental concern scale (Albrecht et al., 1982), itself composed of three four-item subscales—*balance of nature* ($\alpha = 0.86$; e.g., "Humans must live in harmony with nature in order to survive"), *limits to growth* ($\alpha = 0.77$; e.g., "The earth is like a spaceship with only limited room and resources"), and *man over nature* ($\alpha = 0.79$; e.g., "Plants and animals exist primarily to be used by humans" [reverse-coded]). Items were presented in a random order, on the same response scale used for the utilitarianism and deontology scales. For our main analyses, we average across all 12 items [$\alpha = 0.89$].

Finally, participants reported standard demographic information (e.g., age and gender) as well as political party orientation and religiosity on 1 to 9 scales.

We recruited 400 participants for Study 6. We excluded participants if they failed to meet either of two criteria ($N = 76$). First, we excluded participants if they failed a series of memory check questions similar to those used in Studies 1–5. Second, we excluded participants if they failed any of the three attention checks (e.g., "Please choose –3") embedded within the utilitarianism, deontology, and environmental concern scales.

4.2. Results

The two utilitarianism subscales and the deontology scale were correlated weakly, suggesting three distinct psychological traits are being captured. Consistent with previous work (Kahane et al., 2018), the correlation between utilitarianism subscales was small ($r = 0.12$). Interestingly, the correlations between these subscales and the deontology scale were not only fairly weak, but of opposite signs for the two subscales [$r = 0.32$ with impartial beneficence and $r = -0.28$ with instrumental harm]. These findings are consistent with a number of studies challenging the idea that utilitarianism and deontology are psychological opposites (Bartels & Medin, 2007; Bartels & Pizarro, 2011; Kahane et al., 2018). Given the independence of these factors, we analyze them separately as predictors.

For our main analysis, we fit a series of multiple regressions predicting blame (Table 3). Since the models for each of the two vignettes revealed similar patterns, we average blame judgments across the two vignettes. All predictors were centered at their means and scaled by their standard deviations.

For Model 1, we entered our two vignette-specific scales—measuring deontological beliefs and character inferences about each vignette—as predictors, averaging across the two vignettes. (For the deontology scale, we used the same four sub-scales as in Study 5A.) As shown in Table 3, these factors independently predicted blame judgments. More deontological conceptualizations of the actions led to more blame [$b = -0.17$, $p = .029$], and more positive inferences about character led to less blame [$b = 1.25$, $p < .001$], with the latter effect about 7 times larger than the former (since all predictors were scaled). Thus, these results support a modest role for deontological judgment and a large role for character inferences. Overall, this model explained about half of the variance in blame. The results are similar in Model 4 when all other covariates are included.

For Model 2, we entered our three general measures of utilitarianism and deontology—the impartial beneficence and instrumental harm subscales of the Oxford Utilitarianism Scale (Kahane et al., 2018) and the deontology scale (Robinson, 2012). These scales did not predict blame judgments, undercutting the idea that stable, generalized moral theories underlie individual differences in blame, as opposed to the situation-specific construals and inferences shown in Model 1 (as well as Study 5) to be highly predictive. The results are similar in Model 4 when all other covariates are included.

For Model 3, we tested three potential predictors of attitudes toward those who purchase carbon offsets. It seems plausible that one's attitudes about environmental policies related to offsetting, such as cap-and-trade or carbon fines, could predict blame judgments for individuals who perform offsets. Perhaps surprisingly, environmental policy support (averaging across the four policies) does not predict blame judgments [$b = 0.18$, $p = .13$], although support for carbon offsets in specific does somewhat predict (less negative) blame [$b = 0.24$, $p = .023$] in an alternate model. The environmental concern scale did not significantly predict blame in Model 3 [$b = 0.16$, $p = .18$], although it did in Model 4 when other covariates were included [$b = 0.21$, $p = .010$]. Specifically, participants higher in environmental concern were more favorable toward people who purchased carbon offsets, even though those individuals had also polluted; this result is consistent with Polonsky et al. (2012). Likewise, participants who were more closely affiliated with the Democratic party, and therefore likely higher in environmental concern on average, had more favorable blame judgments [$b = 0.33$, $p = .004$]. These results seem more consistent with utilitarianism rather than deontology, since pro-environmental individuals are likely to see environmental transgressions as more harmful (and therefore offsets as more beneficial) as well as more norm-violating (and therefore offsets as less acceptable). Since people higher in environmental concern are more favorable toward purchasers of offsets, the former effect appears to outweigh the latter.

4.3. Discussion

As in Studies 1–5, the effect of person-based processes predominated, explaining far more variance in blame than any other factor. Yet, Study 6 found some support for deontological judgment as a contributing factor, since individuals who construed the acts more deontologically issued more blame toward carbon offset purchasers, and for utilitarian judgment, since individuals higher in environmental concern issued less blame. But these effects appear to be specific to how the situation is construed, rather than broader traits, since trait utilitarianism and deontology scales had little predictive power.

5. Study 7–9

In our previous studies, we have measured moral character, finding that it explains anomalies in how people judge combinations of harms and benefits (Studies 1–4), differences across scenarios (Study 5), and differences across individuals (Study 6). In Studies 7–9, we aim instead to *manipulate* perceived character. On person-based accounts of moral accounting, these manipulations should influence blame judgments even when the action is held constant. We do this in three different ways. Study 7 directly manipulates moral character (Johnson, 2018; Nadler, 2012; Nisan & Horenczyk, 1990). Study 8 manipulates framing

Table 3
Predictors of blame in Study 6.

Predictor	Model 1	Model 2	Model 3	Model 4
Vignette-specific deontology	−0.17 (0.08)*			−0.26 (0.08)**
Vignette-specific character	1.25 (0.08)***			1.20 (0.08)***
Impartial beneficence scale		0.01 (0.11)		0.03 (0.08)
Instrumental Harm Scale		0.18 (0.11)		0.12 (0.08)
Deontology Scale		0.04 (0.11)		0.06 (0.08)
Environmental policy support			0.18 (0.12)	0.12 (0.08)
Environmental Concern Scale			0.16 (0.12)	0.22 (0.08)*
Political party (high = democrat)			0.33 (0.11)**	0.24 (0.08)**
R^2	0.54	< 0.01	0.03	0.57

Entries are b coefficients (SEs), with predictors centered at their means and scaled by their standard deviations. More negative blame judgments indicate higher blame.
° $p < .10$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

(abstract versus concrete), which is known to influence the acceptability of psychological explanations for behavior, such as character-based explanations in terms of personality (Kim et al., 2017; see also Trope & Liberman, 2010). Study 9 manipulates the protagonist's emotional states (e.g., regret), which people use as a cue to moral character (Barasch et al., 2014). In each of these cases, we anticipated that these manipulations would have downstream consequences for the blameworthiness of combinations of harmful and beneficial acts.

5.1. Method

5.1.1. Study 7

The study used a 3 (Character: Good, Bad, or Neutral) × 2 (Dependent Measure: Character or Blame) design, with the character manipulation within-subjects and the dependent measure varied between-subjects. To keep the length of the study reasonable, we chose three vignettes, selected to be maximally distinct from one another (Carbon Emissions, Discrimination, and Fraud). For each vignette, we constructed three versions, based on the Offset condition from Study 1. These versions added a paragraph about the protagonist's biography, either suggesting good moral character, bad moral character, or providing no character information. See Table 4 for sample stimuli.

Participants completed a total of three items, seeing each vignette and each condition once, counterbalanced using a Latin square. For each item, participants first read the character's biography (Good, Bad, or Neutral). Participants in Character judgment condition (but not the Blame judgment condition) then judged the protagonist's character on the same scale used in Study 5B—the Time-0 judgment. On the next screen, participants read about the harm (character information was repeated at the top of the screen), and then made either character judgments (on the same scale) or blame judgments ("Considering [*protagonist's action*], do you think [*protagonist*] deserves to be blamed or praised for this action?")—the Time-1 judgment. Finally, participants read about the offset (character and harm information was repeated at the top of the screen), and then made either character or blame judgments ("Considering [*protagonist's action and their offset*] …")—the Time-2 judgment.

5.1.2. Study 8

The study used a 2 (Concrete vs. Abstract) × 2 (Offset vs. Harm-Only) within-subjects design. We constructed four versions of each of the vignettes used in Study 7, using a manipulation and design similar to Kim et al. (2016, 2017). The Concrete versions were the same as those used in Study 1, prefaced with "Consider the following case." The Abstract versions consisted of a general description of the concrete behavior, prefaced by "Consider the following kind of case." Thus, the abstract version omitted any specific details about the protagonist or their actions, while describing the general sort of harm and offset they had done. Examples are given in Table 4.

Participants saw all four versions of each vignette, which were blocked by condition. In the first half of the experiment, participants would see either the Concrete/Harm-Only and Abstract/Offset conditions, or the Concrete/Offset and Abstract/Harm-Only conditions, with the remaining conditions in the second half of the experiment. This prevented participants from directly comparing two versions of the same item that varied only in abstractness or in offsetting. Within each block, items were presented in a random order. For each item, participants rated moral character using the scale from Study 5B, with the scale items in a random order, followed by blame ("Do you think [*protagonist/these protagonists*] deserve to be blamed or praised for these actions?").

5.1.3. Study 9

The study used a three-condition within-subjects design, manipulating the protagonist's emotional reaction to their harm (Guilt, Shame, or No-Remorse). We constructed three versions of each of three vignettes (Carbon Emissions, Discrimination, and Fraud), based on the Offset condition from Study 1. In all conditions, the protagonist committed a harm and then offset it, but an additional paragraph was added between the descriptions of the harm and offset, explaining why the protagonist did the offsetting act. This paragraph either invoked privately-directed guilt, publicly-directed shame, or neither, as illustrated in Table 4. We primarily expected that people would derive more positive character inferences toward those whose offsets were emotionally motivated (Barasch et al., 2014). However, we also thought that privately-directed guilt may be a more effective driver of character inferences than publicly-directed shame, since the former is internally rather than externally focused (Stearns & Parrott, 2012) and may therefore be seen as signaling a more stable underlying trait.

Participants completed a total of three items, seeing each vignette once and each of the conditions once, counterbalanced using a Latin square. For each item, participants judged moral character and blame, on the scales used in Studies 2–4, as well as a prediction of the protagonist's propensity to do the harm again (e.g., "How likely do you think it is that Morgan will make denials based on race again in the future?") on a scale from −5 ("Not likely at all") to 5 ("Very likely"). Predictions were reverse-coded for analysis for consistency with the other measures.

5.1.4. Participants

We recruited 200, 101, and 200 participants, respectively, for Studies 7–9. Relative to Studies 1–5, we used a larger sample size for Study 7 because the dependent variable was varied between-subjects, and for Study 9 because we anticipated a relatively small effect size (if

Table 4
Sample stimuli from Studies 7–9.

Study 7	Good character	Bad character	Neutral character
	For the past few years, Taylor has been a regular volunteer at the city's homeless shelter. Taylor often makes donations to charity organizations that support domestic abuse shelters. Taylor also enjoys participating as a volunteer to organize charity fundraisers. Recently, Taylor took a plane ride across the U.S., emitting carbon dioxide. Later, Taylor donated money to plant the number of trees required to absorb these emissions, since this donation offsets the carbon from the plane ride.	Taylor has been cheating on their significant other for the past year. This is not the first time Taylor has cheated in a relationship and plans on continuing the affair. Taylor also frequently lies to their relatives and close friends. Recently, Taylor took a plane ride across the U.S., emitting carbon dioxide. Later, Taylor donated money to plant the number of trees required to absorb these emissions, since this donation offsets the carbon from the plane ride.	For the past few years, Taylor has been a regular customer at a coffee shop near their house. Taylor picks up a cup of coffee on their way to work in the morning. Recently, Taylor took a plane ride across the U.S., emitting carbon dioxide. Later, Taylor donated money to plant the number of trees required to absorb these emissions, since this donation offsets the carbon from the plane ride.

Study 8	Concrete/offset	Concrete/harm-only
	Consider the following case: Recently, Taylor took a plane ride across the U.S., emitting carbon dioxide. Later, Taylor donated money to plant the number of trees required to absorb these emissions, since this donation offsets the carbon from the plane ride.	Consider the following case: Recently, Jay took a plane ride across the U.S., emitting carbon dioxide.
	Abstract/offset	Abstract/harm-only
	Consider the following kind of case: There are cases when travelers take plane rides across the U.S., emitting carbon dioxide. Later in these cases, these travelers donate money to plant the number of trees required to absorb these emissions, since this donation offsets the carbon from the plane ride.	Consider the following kind of case: There are cases when travelers take plane rides across the U.S., emitting carbon dioxide.

Study 9	Guilt	Shame	No-remorse
	Recently, Taylor took a plane ride from San Francisco to New York, emitting 0.45 metric tons of carbon dioxide. Taylor knows that it is wrong to emit large amounts of carbon dioxide and felt guilty afterwards because Taylor does not want to be a bad person. It takes around 3 trees to absorb this amount of carbon dioxide. Later, Taylor donates money to plant 3 trees, since this donation offsets the carbon from the plane ride.	Recently, Taylor took a plane ride from San Francisco to New York, emitting 0.45 metric tons of carbon dioxide. Taylor knows that it is wrong to emit large amounts of carbon dioxide and felt ashamed afterwards because Taylor does not want to be seen as a bad person. It takes around 3 trees to absorb this amount of carbon dioxide. Later, Taylor donates money to plant 3 trees, since this donation offsets the carbon from the plane ride.	Recently, Taylor took a plane ride from San Francisco to New York, emitting 0.45 metric tons of carbon dioxide. Taylor knows that it is wrong to emit large amounts of carbon dioxide but did not feel bad about it afterwards. It takes around 3 trees to absorb this amount of carbon dioxide. Later, Taylor donates money to plant 3 trees, since this donation offsets the carbon from the plane ride.

any) for the difference between the guilt and shame conditions. This larger sample can detect with 90% power small effects ($d > 0.23$) in Study 9's within-subjects design. We excluded participants using the same criterion as Studies 1–5 (Ns = 17, 19, and 21, respectively).

5.2. Results

Overall, the results support all three moderators of character—character information, abstractness versus concreteness, and emotional displays—which in all cases had downstream effects on blame. The means are shown in Figs. 3–5.

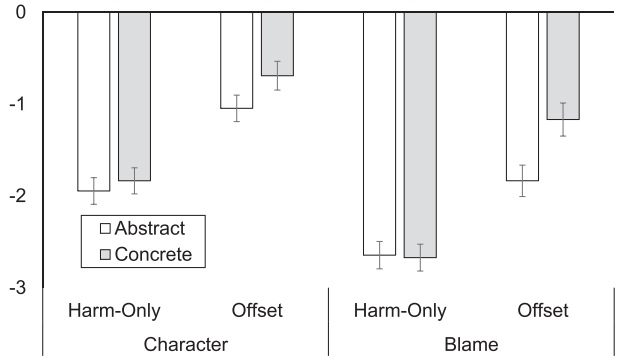

Fig. 4. Character and blame judgments in Study 8. Scales range from −5 to 5. Bars represent 1 SE.

Fig. 5. Character judgments, blame judgments, and predictions in Study 9. Scales range from −5 to 5. Bars represent 1 SE. The prediction scale was reversed for comparability to the other scales, so that negative scores correspond to a higher chance of repeating the blameworthy act.

5.2.1. Study 7

Looking first at the Time-0 character judgments in Fig. 3, we can see that our character manipulation was successful, as participants viewed the protagonists as having more positive character in the Good than in the Bad character condition [$t(89) = 26.98$, $p < .001$, 95% CI_d [2.63,3.05]]; both conditions also differed significantly from the Neutral character condition [$t(89) = 10.90$, $p < .001$, 95% CI_d[0.94,1.36] for Good vs. Neutral and $t(89) = 20.92$, $p < .001$, 95% CI_d[2.00,2.41] for Bad vs. Neutral]. (As noted above in conjunction with Study 2, Time-0 character judgments were positive even in the Neutral condition, suggesting that people's baseline character inferences are positive.) Moreover, although these differences diminished at later time points, the Good and Bad conditions continued to significantly differ at Time-1 [$t(89) = 8.09$, $p < .001$, 95% CI_d[0.64,1.06]] and Time-2 [$t(89) = 7.45$, $p < .001$, 95% CI_d[0.57,0.99]]. Thus, our character manipulation successfully influenced character judgments even after participants had learned additional information about the protagonists' specific harmful and beneficial acts.

Given these character inferences on the part of the participants judging character, we would also expect the participants judging blameworthiness to make correspondingly harsher blame judgments in light of more negative global perceptions of character. This was indeed observed: Blame was significantly harsher in the Bad than in the Good character condition, both at Time-1 when only the harm had occurred [$t(92) = 2.76$, $p = .007$, 95% CI_d[0.08,0.49]] and at Time-2 after the offset had also occurred [$t(92) = 2.38$, $p = .019$, 95% CI_d[0.04,0.45]].

Although the differences between the Good and Bad character conditions strongly support the effect of character inferences on blame, one aspect of these results that appears more puzzling is the comparison of these judgments to the Neutral condition. Character judgments are more similar between the Good and Neutral conditions than between the Neutral and Bad conditions, which seems to reflect a default assumption that most people are morally good. However, this did not translate into a comparable asymmetry in blame judgments, which were actually more similar between the Bad and Neutral conditions than between the Good and Neutral conditions. We suspect that this last result is due to a methodological artifact—blame judgments in the Neutral condition may be as extreme as they could plausibly be given that some of the infractions (e.g., emitting carbon by flying) were relatively minor, leaving relatively little room for them to become more negative. Thus, we do not think these shifting asymmetries reflect deep psychological processes.

5.2.2. Study 8

As shown in Fig. 4, character judgments were indeed more extreme (i.e., negative) in the Concrete than in the Abstract condition when the protagonist offset the harm [$t(81) = 3.35$, $p = .001$, 95% CI_d [0.15,0.59]], but not when the protagonist only committed the harm [$t(81) = 1.12$, $p = .26$, 95% CI_d[−0.10,0.34]]. Thus, our manipulation of character was successful for the Offset condition but not for the Harm-Only condition. We speculate that this difference between the Harm-Only and Offset conditions may have occurred because the motivation for offsetting is often more ambiguous than the motivation for harm, allowing more room for concrete context to influence character inferences.

Blame judgments showed a closely aligned pattern, consistent with the idea that character inferences drive blame. Corresponding to the significant difference in character judgments, blame was more extreme for the Concrete than the Abstract version of the Offset condition [$t(81) = 2.93$, $p = .004$, 95% CI_d[0.10,0.54]]. Moreover, these effects on blame were mediated by character judgments [$b = 0.21$, $SE = 0.09$, 95% CI[0.06,0.41]]. There was no difference in blame between the Concrete and Abstract versions of the Harm-Only condition [$t(81) = -0.31$, $p = .75$, 95% CI_d[−0.25,0.19]], as befits the corresponding null effect on character judgments. Thus, these results support our character-based account of blame judgments for moral offsets.

5.2.3. Study 9

Contrary to expectations, there were no significant differences between the Shame and Guilt conditions for any of our measures [$ts < 0.80$, $ps > 0.40$]. This is somewhat surprising given previous results finding more negative character inferences in light of public-facing (reputation-oriented) shame versus private-facing guilt (Stearns & Parrott, 2012), but it is not itself inconsistent with our theorizing about character and blame. For subsequent analyses, we collapse across the Shame and Guilt conditions (referring to these as the Remorse conditions).

As shown in Fig. 5, remorseful offsetters were deemed better in moral character [$t(178) = 2.83$, $p = .005$, 95% CI_d[0.06,0.36]], less blameworthy [$t(178) = 2.37$, $p = .019$, 95% CI_d[0.03,0.32]], and less likely to perform the harmful act again [$t(178) = 5.29$, $p < .001$, 95% CI_d[0.25,0.54]], compared to the non-remorseful offsetters. Remarkably, even protagonists who had performed offsetting actions precisely because they felt guilt or shame were deemed relatively

blameworthy, and do not seem to be any less blameworthy than the protagonists of Studies 1–4. This may be because people assume by default that offsets are motivated by shame or guilt (or both) and specifying this explicitly may add little beyond participants' default assumptions.

Character judgments mediated the effect of remorse on blame [$b = 0.74$, $SE = 0.26$, 95% CI[0.25,1.25]] as in previous experiments. There was also marginal evidence that character judgments mediated the effect of remorse on predictions [$b = 0.10$, $SE = 0.08$, 95% CI [-0.01,0.29]]; this is consistent with our adaptive explanation for why character judgments predominate in moral judgment—that global representations of moral character serve to identify individuals in the social environment who are likely to be cooperative and trustworthy in the future.

5.3. Discussion

Studies 7–9 identified factors that shift blame by influencing character judgments while holding the action itself constant. In Study 7, providing explicit information about a person's character influenced character judgments in one group of participants and caused parallel shifts in blame for offsetting actions in a separate group. In Study 8, concrete (vs. abstract) framing led to more negative character inferences about moral agents, but only when the actions were offset; this interactive effect of offsetting and framing translated into blame judgments. And in Study 9, participants inferred less negative character on the part of moral agents whose offsetting actions were motivated by shame or guilt (which did not differ from another); this once again translated into blame.

These studies add to our empirical case for the centrality of person-based moral judgment in moral accounting. Studies 1–4 could establish the compatibility of character-based moral judgment with the principles of moral accounting and provided mediation evidence, these studies did not manipulate character directly, and Studies 5 and 6 looked at differences across items and individuals. Although these studies help to demonstrate the broad applicability and explanatory power of character inferences, these designs are less well-suited to establishing strong causal claims. Demonstrating that three different manipulations of moral character have the expected consequences for blame goes a long way toward strengthening those causal claims.

6. General discussion

Much of our behavior is tinged with shades of morality. How third-parties judge those behaviors has numerous social consequences: People judged as behaving immorally can be socially ostracized, less interpersonally attractive, and less able to take advantage of win–win agreements. Indeed, our desire to avoid ignominy and maintain our moral reputations motivates much of our social behavior. But on the other hand, moral judgment is subject to a variety of heuristics and biases that appear to violate normative moral theories and lead to inconsistency (Bartels et al., 2015; Sunstein, 2005). Despite the dominating influence of moral judgment in everyday social cognition, little is known about how judgments of individual acts scale up into broader judgments about sequences of actions, such as moral offsetting (a morally bad act motivates a subsequent morally good act) or self-licensing (a morally good act motivates a subsequent morally bad act). We need a theory of *moral accounting*—how rights and wrongs add up in moral judgment.

This article has sought to document several key principles guiding moral accounting and to explain how these principles arise from broader notions. Across a wide range of moralized harms, Studies 1–4 found that people view offsets as impartial—a bad act followed by a good act of comparable magnitude is not seen as morally neutral, but as somewhat blameworthy, albeit less blameworthy than if one had not done the good act at all. Blame is relatively insensitive to the magnitude of a moral offset, viewing acts that offset double the harm they had done as only a little less blameworthy than acts that offset the same amount of harm they had done. The temporal order of the actions influences judgments, such that bad acts preceding good acts are seen as less blameworthy than the converse. And the similarity of the harm and benefit influences the extent to which the benefit can offset the harm, suggesting that people do not keep a single "moral account" for each individual, but instead keep track of harms separately and apply a moral benefit against the same kind of harm.

6.1. Accounting for moral accounting

To what extent are these principles compatible with the various approaches to morality? Although neither utilitarianism nor deontology nor character-based approaches were originally formulated with the purpose of explaining how people evaluate combinations of actions, we can attempt to extend their logic to understand what these theories would most plausibly say about our principles. Doing so is useful to maintain continuity with existing debates and to spark future research.

6.1.1. Utilitarianism

According to utilitarianism, we should evaluate actions by adding up their costs and benefits. Since fully offset actions have no net cost or benefit, the most basic version of utilitarianism should consider such actions morally neutral. Thus, partial offsetting and diminishing sensitivity seem quite inconsistent with the most fundamental principles of utilitarianism. Still, a defender of utilitarianism might point out that offsetting, though partial, was fairly large in magnitude, and argue that a more sophisticated version of utilitarianism can account for these findings. One approach is to consider additional costs, such as the potential third-party effects of seeing someone flying; if such third-parties are likelier to fly too, then this increases *total* CO_2 emissions even if one's *personal* emissions are offset. On the other hand, making such a move seems to require that comparable *benefits* also be considered, such as raising awareness of carbon offsets, which may *further* decrease CO_2.

Another possibility would be to add in not just the concrete harms but also the emotional effects on victims, such as loss aversion. Since the displeasure of having $10 stolen is greater than the pleasure of a $10 gift, then it is only the *financial* harm that if offset, not the emotional harm. Indeed, prospect theory (Kahneman & Tversky, 1979) suggests both partial offsetting (due to loss aversion) *and* diminishing sensitivity (due to the shape of the value function). That said, this approach can only go so far, since many of our items (including our environmental items) concern harms that do not plausibly have an emotional component. Since CO_2 emissions only affect the Earth's temperature in the aggregate, there is little role for third-parties' emotional reactions to individuals' emissions.

While partial offsetting and diminishing sensitivity demonstrate at least *some* sensitivity to magnitudes, which is the hallmark of utilitarianism, temporal asymmetry and act congruency are much more difficult to explain on a utilitarian account. Clearly, the net harm is equated between two identical actions performed in opposite orders, and utilitarianism aims to trade-off different kinds of harms against one another, making congruency irrelevant. Perhaps some variant on rule utilitarianism could be made to account for these findings—for instance, the world might be better if people adopted an obligation to offset and refrained from licensing (temporal asymmetry) and a principle of offsetting their own specific actions (act congruency). However, one begins to worry that such versions of utilitarianism become so flexible that they can account for any moral intuition at all.

6.1.2. Deontology

According to deontology, actions are blameworthy when they break moral norms. Relative to utilitarianism, deontology seems to have more difficulty explaining offsetting. It is not clear that harms can be offset at all by beneficial actions, since the norm remains broken—thus partial

offsetting seems at first blush to contradict deontology. Diminishing sensitivity *is* consistent with deontology, and indeed is taken as a hallmark of protected values or deontological thinking (Baron & Spranca, 1997); the only complaint one might lodge here is that sensitivity does not diminish *enough*.

The case of temporal asymmetry is perhaps the most interesting to consider. Deontology does have the resources to consider an actor's motivations, since motivations are relevant to understanding whether a norm has been broken. This is why deontological theories include notions such as the "doctrine of double effect" (Mangan, 1949), which excuses harms only if they were unintended side-effects of a beneficial action, rather than an intentional means to bring about that beneficial action. Indeed, one of the conditions for this doctrine to hold is precisely that the good outweigh the harm (Mangan, 1949), indicating that deontology has some access to trade-offs. While this doctrine does not apply in cases of offsetting (since the harm was neither a means nor a side-effect), it is relevant because it suggests that whether an action is norm-violating depends partly on the actor's motivations. Since temporal asymmetry and act congruency probably result in part from inferences about why the actor is taking the second act (redemption or entitlement), perhaps some refinement of deontology could be constructed to account for these principles.

6.1.3. Character-based accounts

Whereas utilitarianism and deontology could potentially be refined to account for some of the principles observed here, character-based accounts naturally account for all of them. According to such accounts, blame judgments flow from judgments of moral character, which themselves follow a set of inferential principles that sit outside the theory. For this reason, we rely on existing research in person perception (e.g., Reeder & Brewer, 1979) and on empirical tests of these inferential principles to make clear directional predictions about blame judgments. In all cases, character and blame were closely aligned. By measuring both blame and character judgments in our studies, we were able to provide evidence for these auxiliary principles governing how people infer moral character from combinations of actions. Thus, although this article's central contribution is in presenting a theory of how people judge blame for combinations of moral and immoral acts, it also demonstrates several phenomena of moral character inference which are, to our knowledge, novel.

Several other pieces of evidence also supported the character-based account. Studies 5 and 6 found that character judgments are key drivers of differences in blame across items and across individuals. Studies 7–9 demonstrated that various experimental manipulations of moral character, holding constant the action itself, led to shifts in blame. This is not to say that people totally ignore outcomes and violations of moral rules. For instance, Studies 5 and 6 both found evidence for the influence of outcomes, and Study 6 found some evidence for the influence of perceived rule violations. Yet, character judgments seem to overwhelm these other forces, perhaps because rule-violations and outcomes are themselves inputs into character judgments. This would be consistent with evidence that people are more sensitive to outcomes in assigning blame to agents known to have poor, versus exemplary, moral character (Siegel et al., 2017).

Where do character judgments come from? Two kinds of processes are relevant. First, character inferences likely follow some of the same principles of *abductive inference*—seeking explanations from observations—as many other mental processes, such as causal reasoning and theory-of-mind (Johnson et al., 2020; Lombrozo, 2016). That is, beliefs about moral character are *hypotheses* about how best to explain necessarily limited *data* about an individual's behavior, which in turn license future predictions. Second, these processes were likely shaped by evolutionary forces that put a premium on survival-relevant information. It has long been acknowledged that people prize and carefully track social information, such as their reputation for cooperation and trustworthiness (e.g., Sperber & Baumard, 2012). Given our impressive capacity for abductive inference and obsessive interest in social reputation, it is natural that character inferences would be central to our thoughts about others.

Why would our inferential and social capacities endow us with the specific character-inference mechanisms we found here? Here, we must speculate and future research will be revealing. But we suspect that the basic intuitions behind the character inference principles are straightforward. Someone who is capable of doing something bad maintains that capability even if they also do something good (hence, offsetting is only partial). A person who does a lot of good may not be that much more trustworthy or honest than someone who does a moderate amount of good (hence, diminishing sensitivity to increasingly large offsets). Someone who feels entitled to do something bad because they have done something good is probably more likely to maintain that immoral motivation than someone who felt remorseful after doing something bad (hence, licensing evidences worse character than offsetting). And someone who does not keep their moral accounts in the black (e.g., by helping someone other than the victim of their harmful act) may be a less desirable social partner than someone who keeps all their moral ducks in a row (hence, the congruency of the bad and good acts contributes to moral evaluations). These strike us as rational, if fallible, general rules about character inference that probably yield reasonably good predictions in realistic cases, yet further research is needed to fully understand their basis. More broadly, we think that understanding character inference is an exciting direction for research at the intersection of person perception, abductive inference, and evolutionary psychology.

6.2. Limitations

Although the converging evidence for these principles across different kinds of harms and different methodological constraints suggests both internal validity and generalizability across harms, these studies do have limitations that might be addressed in future work. Perhaps most significantly, our participants were from the United States and skewed politically liberal, and moral intuitions are known to differ globally across cultures and between political tribes within a particular culture (Graham et al., 2011). Still, attentiveness to harm (more than other "moral foundations" such as in-group loyalty) is relatively consistent across cultures and political orientations; since the current studies focus on harm, these results may be more likely to generalize across cultures. Moreover, beliefs in karma—an apparent downstream consequence of moral accounting—appear to be prevalent across cultures. Perhaps surprisingly, more is probably known in psychology about the perceived causal effects of karma in Western populations (e.g., Banerjee & Bloom, 2014; Callan et al., 2014; Converse et al., 2012) who often explicitly disavow the causal force of karma, than in non-Western populations that are likelier to acknowledge karmic forces explicitly (White et al., 2017). Thus, studying these effects cross-culturally may be useful for building a broader understanding of how basic psychological forces interact with culture to produce specific belief systems.

A related limitation is that we focused on harm as opposed to other kinds of moralized actions, such as violations of loyalty, fairness, or purity. This is in part because theories such as utilitarianism (and to some extent deontology) have little to say about such considerations, making it more difficult to generate predictions from these theories. By focusing on harm, we challenge those perspectives on their own turf. Moreover, harm is particularly ubiquitous in moral psychology, although researchers disagree about whether this is harm is used as a post hoc rationalization to justify moralizing harmless acts (Haidt, 2001) or because people perceive a wide variety of actions as harmful (Schein & Gray, 2018). Still, to the extent that people endorse harmless actions as moral violations, a complete theory of moral accounting must also explain how people add up judgments of multiple actions that adhere to versus violate these principles. More broadly, the issue of how people

evaluate trade-offs among different kinds of violations (e.g., fairness versus harm) strikes us as a ripe topic for research.

Finally, we must acknowledge limitations on the formulation and empirics of the theories of blame we rely on in this article. At a general level, utilitarianism, deontology, and character-based theories all originated in normative ethics and were not intended to have the level of precision required to make specific predictions in this tasks. At best we can derive what a basic version of each theory implies, leaving open the possibility of future refinements. Even our preferred character-based theory has some inherent flexibility because it depends for its predictions on a broader theory of character inference. Thus, even as the empirical findings in this article are clear, their theoretical interpretation can be subject to debate. A related issue is that deontology and utilitarianism tend to predict null effects, whereas character-based accounts predict positive effects. In future theory testing, it would be useful to derive additional predictions where the converse is true.

6.3. Future research

In raising a new question about how our intuitive morality combines benefits and harms, we cannot conceivably address all possible aspects of this topic. Indeed, one exciting aspect of this research is its potential to inspire future research programs. Here are four possible directions.

First, some of our studies identified as central the mental states of the actors, such as the finding that experiencing emotions such as shame or guilt mitigates the blameworthiness of combinations of harmful and beneficial actions (Study 9). Many other studies have also found that mental states are relevant to moral ascriptions (e.g., Barasch et al., 2014; Cushman, 2008), with a particularly interesting set of findings suggesting that people can be blamed for the mental states *themselves* (Cusimano & Goodwin, 2019; Inbar et al., 2012). But much more needs to be done to understand how people infer moral character from mental states. In our own studies, for example, it remains unclear whether the temporal asymmetry principle is explained by the belief that the most recent action is most diagnostic of character or whether it is tied specifically to the motivation for that action. For instance, is the action seen as morally better if it is specifically motivated by the desire to offset an earlier harm or if it is spontaneous? The former plausibly suggests a lower capacity for future harm (since it implies remorse), while the latter plausibly suggests a greater capacity for future benefits (since it does not require a specific antecedent). As another variant, does it make one's moral character better or worse if one plans *in advance of the harmful action* to make up for it later? Both understanding these specific questions related to our moral accounting principles and the broader relationship between mental-state ascriptions and character judgments are important directions for future work.

Second, we tested these principles individually but never in combination due to the large number of cells that would require. However, such tests could be informative about theory and yield interesting interactions. For example, when we test the act congruency principle, we found that more congruent offsets (occurring after the harms) are seen as more offsetting. But we earlier pointed to some literature finding that highly congruent *licensing* is seen as more hypocritical and therefore occurs less often (Effron & Monin, 2010). Thus, it is possible that the act congruency principle would be diminished in magnitude or even reversed if we examined harms that occurred *after* rather than before the benefits they offset. Since hypocrisy is likely to be one component of moral character judgments, this prediction can fit squarely within character-based accounts, but empirical testing is needed to examine both the character inferences and any downstream consequences on blame.

Third, although we consider a reasonably wide range of harms—from relatively minor violations such as littering to fairly extreme behaviors such as fraud—we do not consider moral violations at the extremes of behavior, such as murder. It is unclear whether standard principles of moral accounting apply at such extremes, or whether instead a sufficiently red moral account leads to moral bankruptcy from which one cannot recover. It is particularly interesting to consider whether different kinds of moral judgments are especially likely to decouple (Cushman, 2008; Martin & Cushman, 2016) at the extremes. One arguable depiction of such a decoupling can be seen in the film *Sympathy for Mr. Vengeance* (2002, dir. Park Chan-wook). See footnote for a plot summary with spoilers (or alternatively, stop reading and watch this excellent movie!).[1] Even though the sum of the character's actions were seen as diagnostic of good moral character, the bad aspects of his actions were so extreme that it was felt that he must be punished. One explanation for such decoupling is that character judgments and punishment have different evolved functions: character is about reputation tracking (Sperber & Baumard, 2012) while punishment is about norm enforcement (Fehr & Fischbacher, 2004). Future research could examine this issue by varying the extremity of both positive and negative behaviors while measuring multiple judgments, including character, blame, and punishment. If punishment is primarily motivated by norm enforcement, such judgments may look more deontological.

Finally, people may endorse lay theories that differ from their actual patterns of moral judgments. Utilitarianism is a particularly valuable framework for justifying moral intuitions, so much so that people often confabulate utilitarian, harm-based justifications for harmless acts that nonetheless feel wrong ("moral dumbfounding"; Haidt, 2001). Thus, if people are intuitive virtue ethicists but hold lay utilitarian theories, they may seek reasons why the harms of offset actions outweigh the benefits even in cases where this transparently does not apply. More generally, lay theories of morality are important for understanding moral discourse and would be a useful object for future study.

6.4. Practical implications

Although the study of moral accounting is in its infancy, we think this topic is well-positioned to generate actionable policy and marketing advice. As we noted above, many policy-makers and social scientists view carbon offsets as one partial solution to curbing carbon emissions, but this solution can only be effective if consumers are actually motivated to undo their harmful actions through offsets. Consumers do value both moral and financial costs in their decision-making and trade them off against one another (e.g., Johnson, Zhang, Keil, 2019; Smith, 1990). Thus, any way to improve the moral calculus at the margin is likely to lead to greater uptake of offsets.

The current studies offer several stylized facts that may be of use to marketers of consumer offsets such as carbon offsets. First, the partial offsetting principle says that offsetting one's harm directly is, while better than not offsetting at all, still less morally praiseworthy than abstaining from the harm altogether. Meanwhile, the diminishing sensitivity principle says that further increasing the offset (e.g., double one's offset) may barely get one back to the morally neutral point. But sometimes consumer offsets are less expensive than one might expect—e.g., around $13 to offset a round-trip transcontinental flight. Thus, offering customers to offset double their harm may be worth it for some customers to approach moral neutrality. Second, the temporal asymmetry principle says that moral evaluations are more favorable when beneficial actions are taken after—rather than before—the harm,

[1] Factory worker Ryu loses his job at a time when he desperately needs money to pay for his sister's kidney transplant. Ryu kidnaps his former employer's daughter Yu-sun for ransom, treating her very kindly and with every intention of returning her. Yu-sun dies in a freak accident, and her father Dong-jin vows to avenge her death. When he finally locates Ryu, he understands that Ryu made hard choices in an impossible situation—"I know you're a good guy"—but still carries out his revenge—"But you know why I have to kill you?" Despite everything, Dong-jin acknowledges Ryu's good character based on the sum-total of his actions and his situation. Yet he feels he has a duty to levy punishment.

suggesting that consumers may be achieve better moral closure by purchasing offsets after their flights. Third, the act congruency principle says that moral evaluations are more favorable when the benefit closely resembles the harm, suggesting that marketers should emphasize the tight match between the harm and the benefit. Finally, and more broadly, methods to generate more positive character inferences from offsetting actions (e.g., by manipulating abstract vs. concrete framing as in Study 8 or by cueing emotional signaling as in Study 9) may help to enhance the benefit of offsets. More broadly, ethical consumerism is increasing in popularity as society becomes richer and consumers are increasingly motivated to account for the wider consequences of their choices. Thus, one could imagine other forms of offsets, such as for plastic consumption, gaining a following and helping to raise funds to restore or conserve the environment.

Although environmental offsets may be among the most pressing applications of moral accounting, we suspect that this research has a much wider range of policy and marketing implications. Three examples: first, people often make moral judgments about companies (e.g., Chernev & Blair, 2015), which in turn impact their purchasing decisions (Smith, 1990). Thus, companies can benefit from understanding moral principles in devising strategies to recover from public relations disasters, and society can benefit as companies do more to satisfy their social stakeholders. Second, legal decision-making depends on questions of fact and law, but often on moral intuitions too. Thus, as jurors and other legal stakeholders track the overall blameworthiness of a person's actions, the framing of those actions may shape legal decision-making. Third, citizens and policy-makers often make moral judgments about the behavior of other nations. Indeed, debates about whether other countries have behaved immorally shape public discourse in the lead-up to wars. It may not be an exaggeration to say that how people add up moral rights and wrongs, aggregated across society, can be a matter of war and peace.

CRediT authorship contribution statement

SGBJ and JA conceptualized the project, designed the studies, and analyzed the data. JA collected the data under the supervision of SGBJ. SGBJ wrote the manuscript, with critical revisions from JA.

Acknowledgements

We thank Frank Keil for his support, the Yale Cognition & Development Lab for useful comments, and the University of Bath School of Management for funding.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104467.

References

Adams, R. M. (1976). Motive utilitarianism. *Journal of Philosophy, 73*, 467–481.
Albrecht, D., Bultena, G., Hoiberg, E., & Nowak, P. (1982). Measuring environmental concern: The new environmental paradigm scale. *Journal of Environmental Education, 13*, 39–43.
Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*, 368–378.
Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology, 70*, 394–400.
Anderson, N. H., & Hubert, S. (1963). Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Verbal Behavior, 2*, 379–391.
Aquinas, T. (2000/1274). *Summa theologica* (Fathers of the Dominican Province, Trans.) Notre Dame, IN: Ave Maria Press.
Aristotle (1999/350 BCE). Nicomachean ethics (T. Irwin, Trans.). Indianapolis, IN: Hackett.
Asch, S. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258–290.
Baez, S., Herrera, E., García, A. M., Manes, F., Young, L., & Ibáñez, A. (2017). Outcome-oriented moral evaluation in terrorists. *Nature Human Behaviour, 1*, 0118.
Banerjee, K., & Bloom, P. (2014). Why did this happen to me? Religious believers' and non-believers' teleological reasoning about life events. *Cognition, 133*, 277–303.
Barasch, A., Levine, E. E., Berman, J. Z., & Small, D. A. (2014). Selfish or selfless? On the signal value of emotion in altruistic behavior. *Journal of Personality and Social Psychology, 107*, 393–413.
Barden, J., Rucker, D. D., & Petty, R. E. (2005). "Saying one thing and doing another": Examining the impact of event order on hypocrisy judgments of others. *Personality and Social Psychology Bulletin, 31*, 1463–1474.
Barden, J., Rucker, D. D., Petty, R. E., & Rios, K. (2014). Order of actions mitigates hypocrisy judgments for ingroup more than outgroup members. *Group Processes & Intergroup Relations, 17*, 590–601.
Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes, 70*, 1–16.
Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2015). Moral judgment and decision-making. In G. Keren, & G. Wu (Eds.). *Blackwell handbook of judgment and decision-making* (pp. 478–515). New York, NY: Wiley.
Bartels, D. M., & Medin, D. L. (2007). Are morally motivated decision makers insensitive to the consequences of their choices? *Psychological Science, 18*, 24–28.
Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*, 154–161.
Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323–370.
Bentham, J. (1907/1789). *An introduction to the principles of morals and legislation.* Oxford, UK: Clarendon.
Birnbaum, M. H. (1973). Morality judgment: Test of an averaging model with differential weights. *Journal of Experimental Psychology, 99*, 395–399.
Bostyn, D. H., & Roets, A. (2016). The morality of action: The asymmetry between judgments of praise and blame in the action–omission effect. *Journal of Experimental Social Psychology, 63*, 19–25.
Callan, M. J., Sutton, R. M., Harvey, A. J., & Dawtry, R. J. (2014). Immanent justice reasoning: Theory, research, and current directions. In J. M. O., & M. P. Zanna (Vol. Eds.), *Advances in experimental social psychology. Vol. 49. Advances in experimental social psychology* (pp. 105–161). London, UK: Academic Press.
Chernev, A., & Blair, S. (2015). Doing well by doing good: The benevolent halo of corporate social responsibility. *Journal of Consumer Research, 41*, 1412–1425.
Converse, B. A., Risen, J. L., & Carter, T. J. (2012). Investing in karma: When wanting promotes helping. *Psychologcial Science, 23*, 923–930.
Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology, 104*, 216–235.
Côté, S., Piff, P. K., & Willer, R. (2013). For whom do the ends justify the means? Social class and utilitarian moral judgment. *Journal of Personality and Social Psychology, 104*, 490–503.
Critcher, C. R., & Dunning, D. (2011). No good deed goes unquestioned: Cynical reconstruals maintain belief in the power of self-interest. *Journal of Experimental Social Psychology, 47*, 1207–1213.
Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*, 353–380.
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*, 1082–1089.
Cusimano, C., & Goodwin, G. P. (2019). Lay beliefs about the controllability of everyday mental states. *Journal of Experimental Psychology: General, 148*, 1701–1732.
Dahlsgaard, K., Peterson, C., & Seligman, M. E. P. (2005). Shared virtue: The convergence of valued human strengths across culture and history. *Review of General Psychology, 9*, 203–213.
De Freitas, J., & Johnson, S. G. B. (2018). Optimality bias in moral judgment. *Journal of Experimental Social Psychology, 79*, 149–163.
Effron, D. A., & Monin, B. (2010). Letting people off the hook: When do good deeds excuse transgressions? *Personality and Social Psychology Bulletin, 36*, 1618–1634.
Effron, D. A., O'Connor, K., Leroy, H., & Lucas, B. J. (2018). From inconsistency to hypocrisy: When does "saying one thing but doing another" invite condemnation? *Research in Organizational Behavior, 38*, 61–75.
Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology, 79*, 200–216.
Fehr, E., & Fishbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*, 63–87.
Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*.
Frederick, S., & Fischhoff, B. (1998). Scope (in)sensitivity in elicited valuations. *Risk Decision and Policy, 3*, 109–123.
Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition, 177*, 122–141.
Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*, 148–168.
Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*, 1029–1046.
Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101*, 366–385.
Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*, 505–520.
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*,

1144–1154.
Griskevicius, V., Tybur, J. M., & van den Bergh, B. (2010). Going green to be seen: Status, reputation, and conspicuous conservation. *Journal of Personality and Social Psychology, 98*, 392–404.
Guglielmo, S., & Malle, B. F. (2017). Information-acquisition processes in moral judgments of blame. *Personality and Social Psychology Bulletin, 43*, 957–971.
Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLoS ONE, 14*, Article e0213544.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.
Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*, 613–628.
Henderson, P. M., & Peterson, R. A. (1992). Mental accounting and categorization. *Organizational Behavior and Human Decision Processes, 51*, 92–117.
Hollander, E. P. (1958). Conformity, status, and idiosyncrasy credit. *Psychological Review, 65*, 117–127.
Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 67*, 247–257.
Hursthouse, R. (1999). *On virtue ethics.* Oxford, UK: Oxford University Press.
Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefitting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin, 38*, 52–62.
Johnson, S. G. B. (2020). Dimensions of altruism: Do evaluations of charitable behavior track prosocial benefit or personal sacrifice? Retrieved from SSRN https://papers.ssrn.com/abstract=3277444.
Johnson, S. G. B., Bilovich, A., & Tuckett, D. (2020). Conviction narrative theory: A theory of choice under radical uncertainty. Retrieved from PsyArXiv https://psyarxiv.com/urc96/.
Johnson, S. G. B., Kim, H. S., & Keil, F. C. (2016). Explanatory biases in social categorization. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.). *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 776–781). Austin, TX: Cognitive Science Society.
Johnson, S. G. B., Murphy, G. L., Rodrigues, M., & Keil, F. C. (2019). Predictions from uncertain moral character. *Proceedings of the 40th annual conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
Johnson, S. G. B., & Park, S. Y. (2020). Moral signaling through donations of money and time. Retrieved from SSRN https://papers.ssrn.com/abstract=3343284.
Johnson, S. G. B., Zhang, J., & Keil, F. C. (2019). Consumers' beliefs about the effects of trade. Retrieved from SSRN https://papers.ssrn.com/abstract=3376248.
Jones, E. E. (1990). *Interpersonal perception.* New York, NY: Freeman.
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*, 131–164.
Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). "Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.
Kant, I. (2002). *Groundwork for the metaphysics of morals* (A. Zweig, Trans.). New York, NY: Oxford University Press (Original work published 1796).
Khan, U., & Dhar, R. (2006). Licensing effect in consumer choice. *Journal of Marketing Research, 43*, 259–266.
Kim, N. S., Ahn, W., Johnson, S. G. B., & Knobe, J. (2016). The influence of framing on clinicians' judgments of the biological basis of behaviors. *Journal of Experimental Psychology: Applied, 22*, 39–47.
Kim, N. S., Johnson, S. G. B., Ahn, W., & Knobe, J. (2017). The effect of abstract versus concrete framing on judgments of biological and psychological bases of behavior. *Cognitive Research: Principles and Implications,* 2.
Klein, N., & Epley, N. (2014). The topography of generosity: Asymmetric evaluations of prosocial actions. *Journal of Experimental Psychology: General, 143*, 2366–2379.
Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*, 190–194.
Kouchaki, M. (2011). Vicarious moral licensing: The influence of others' past moral actions on moral behavior. *Journal of Personality and Social Psychology, 101*, 702–715.
Levine, E. E., Hart, J., Moore, K., Rubin, E., Yadav, K., & Halpern, S. (2018). The surprising costs of silence: Asymmetric preferences for prosocial lies of commission and omission. *Journal of Personality and Social Psychology, 114*, 29–51.
Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Personality and Social Psychology, 53*, 107–117.
Lin-Healy, F., & Small, D. A. (2013). Nice guys finish last and guys in last are nice: The clash between doing well and doing good. *Social Psychological and Personality Science, 4*, 692–698.
Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences, 20*, 748–759.
Mangan, J. (1949). An historial analysis of the principle of double effect. *Theological Studies, 10*, 41–61.
Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition, 147*, 133–143.
May, P. (2007, May 14). Offset your infidelity? Observations on ethical cheating. Retrieved from New Statesman http://www.newstatesman.com.
Mazar, N., & Zhong, C. (2010). Do green products make us better people? *Psychological Science, 21*, 494–498.
Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass, 4*, 344–357.
Mill, J. S. (1998/1861). *Utilitarianism.* Oxford, UK: Oxford University Press.
Miller, G. F. (2007). Sexual selection for moral virtues. *Quarterly Review of Biology, 82*, 97–125.
Monbiot, G. (2006, October 18). Paying for our sins. Retrieved from The Guardian http://www.theguardian.com.
Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Presonality and Social Psychology, 81*, 33–43.
Monin, B., Sawyer, P. J., & Marquez, M. J. (2008). The rejection of moral rebels: Resenting those who do the right thing. *Journal of Personality and Social Psychology, 95*, 76–93.
Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods, 22*, 6–27.
Nadler, J. (2012). Blaming as a social process: The influence of character and moral emotion on blame. *Law and Contemporary Problems, 75*, 1–31.
Nagel, T. (1979). *The view from nowhere.* New York, NY: Oxford University Press.
Newman, G. E., & Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological Science, 25*, 648–655.
Newman, G. E., Lockhart, K. L., & Keil, F. C. (2010). "End-of-life" biases in moral evaluations of others. *Cognition, 115*, 343–349.
Niemi, L., & Young, L. (2016). When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin, 42*, 1227–1242.
Nisan, M. (1991). The moral balance model: Theory and research extending our understanding of moral choice and deviation. In W. M. Kurtines, & J. L. Gewirtz (Eds.). *Handbook of moral behavior and development* (pp. 213–249). Hillsdale, NJ: Erlbaum.
Nisan, M., & Horenczyk, G. (1990). Moral balance: The effect of prior behaviour on decision in moral conflict. *British Journal of Social Psychology, 29*, 29–42.
Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*, 1291–1298.
Nozick, R. (1974). *Anarchy, state, and utopia.* New York, NY: Basic Books.
Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*, 163–177.
Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14*, 267–272.
Polonsky, M. J., Vocino, A., Grau, S. L., Garma, R., & Ferdous, A. S. (2012). The impact of general and carbon-related environmental knowledge on attitudes and behaviour of US consumers. *Journal of Marketing Management, 28*, 238–263.
Quinn, W. S. (1989). Actions, intentions, and consequences: The doctrine of doing and allowing. *Philosophical Review, 98*, 287–312.
Rawls, J. (1955). Two concepts of rules. *Philosophical Review, 64*, 3–32.
Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86*, 61–79.
Reeder, G. D., & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social Psychology, 44*, 736–745.
Riskey, D. R., & Birnbaum, M. H. (1974). Compensatory effects in moral judgment: Two rights don't make up for a wrong. *Journal of Experimental Psychology, 103*, 171–173.
Robinson, J. S. (2012). *The consequentialist scale: Elucidating the role of deontological and utilitarian beliefs in moral judgments.* Unpublished masters thesisToronto, ON: University of Toronto.
Rottman, J., Kelemen, D., & Young, L. (2014). Taining the soul: Purity concerns predict moral judgments of suicide. *Cognition, 130*, 217–226.
Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science, 20*, 523–528.
Scheffler, S. (1982). *The rejection of consequentialism.* Oxford, UK: Oxford University Press.
Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*, 32–70.
Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34*, 1096–1109.
Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes, 101*, 1–19.
Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron, 67*, 667–677.
Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition, 167*, 201–211.
Silver, I., & Shaw, A. (2018). Pint-sized public relations: Developing reputation management. *Trends in Cognitive Science, 22*, 277–279.
Singer, P. (1977). Actual consequence utilitarianism. *Mind, 86*, 67–77.
Singer, P. (2011). *Practical ethics* (3rd ed.). New York, NY: Cambridge University Press.
Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*, 131–142.
Slovic, P. (2007). "If I look at the amss I will never act": Psychic numbing and genocide. *Judgment and Decision making, 2*, 79–95.
Smith, N. C. (1990). *Morality and the market: Consumer pressure for corporate accountability.* London, UK: Routledge.
Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language, 27*, 495–518.
Stearns, D. C., & Parrott, W. G. (2012). When feeling bad makes you look good: Guilt, shame, and person perception. *Cognition & Emotion, 3*, 407–430.
Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences, 28*, 531–542.
Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*, 345–372.
Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology, 47*, 1249–1254.
Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology, 78*, 853–870.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences, 7*, 320–324.
Thaler, R. (1985). Mental accounting and consumer choice. *Marketing Science, 4*, 177–266.
Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*, 35–57.
Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*, 440–463.
Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*, 72–81.
White, C., Baimel, A., & Norenzayan, A. (2017). What are the causes and consequences of belief in karma? *Religion, Brain, and Behavior, 7*, 339–342.
Wiltermuth, S. S., Monin, B., & Chow, R. M. (2010). The orthogonality of praise and condemnation in moral judgment. *Social Psychological and Personality Science, 1*, 302–310.
Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*, 202–214.