Hypocritical blame is associated with reduced prosocial motivation

Luis Sebastian Contreras-Huerta[a,b,c,d*], Hongbo Yu[e], Annayah M. B. Prosser[b,f], Patricia L. Lockwood[b,c,g,h], Molly J. Crockett[i,j,1] and Matthew A.J. Apps[,b,c,g,h1*]

[a] Center for Social and Cognitive Neuroscience (CSCN), School of Psychology, Universidad Adolfo Ibáñez, Viña del Mar, Chile

[b] Department of Experimental Psychology, University of Oxford, Oxford, Oxford OX1 3PH, UK

[c] Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham B15 2TT, UK

[d] Center of Social Conflict and Cohesion Studies, Santiago, Chile

[e] Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara CA 93106, USA.

[f] Marketing, Business and Society Division, School of Management, University of Bath, BA2 7AY, UK

[g] Institute for Mental Health, School of Psychology, University of Birmingham, Birmingham B15 2TT, UK

[h] Christ Church, University of Oxford, Oxford OX1 1DP, UK

[i] Department of Psychology, Princeton University, Princeton NJ 08540, USA

[j] University Center for Human Values, Princeton University

***Correspondence**: Luis Sebastian Contreras-Huerta, sebastian.contreras@uai.cl; Matthew A.J. Apps, m.a.j.apps@bham.ac.uk.

[1] Shared senior co-authorship

Abstract

People often act hypocritically. One form of hypocrisy occurs when people blame others for transgressing moral principles – such as profiting from harming others – that they themselves have violated in the past. However, the psychological processes associated with this hypocritical blame are largely unknown. One possibility is that hypocritical blame is related to the costs of being prosocial, such that a person could have the intention to help but might not be willing to put in the effort. Here, we test whether a measure of hypocritical blame that quantifies the discrepancy between how willing people are to profit from another's harm, and how much they blame somebody else for similarly profiting, is associated with a task measuring how willing someone is to choose and then exert physical effort to benefit themselves or an anonymous other. Results revealed that hypocritical blame is associated with reduced prosocial motivation specifically, and not with how willing people are to exert effort for their own benefit. This effect was found in both a reduced willingness to choose to be prosocial and for energising prosocial acts. This suggests that the discrepancy between moral standards and actions is related to the willingness to overcome the costs of being prosocial, with some people being simply unwilling to exert the effort required to live up to their moral principles.

Key words: Hypocritical blame, prosocial motivation, effort-based decision-making, harm aversion, moral behaviour.

Introduction

Moral principles often guide how people act and are reinforced in society through the judgements we make of others. Judging someone as morally virtuous or vicious, depending how much they deviate from moral principles, serves to perpetuate these principles as norms. It has often been assumed that people guide both their moral decisions and their moral judgements according to the same standards. However, this is often not the case[1–5]. In fact, it is common that people show, at least to some degree, discrepancies between their actions and how harshly they judge others. For example, consider a politician who publicly condemns violations to Covid-19 restrictions as selfish, but then violates those same rules when it is in their interest. There is a clear discrepancy between the politician's moral judgements and their moral actions. This discrepancy has been identified by philosophers, psychologists, and the general public as a form of moral hypocrisy: hypocritical blame[1,4,6–9].

Philosophical accounts have proposed that such hypocrisy could be due to having stricter moral standards for others, but more lax ones for ourselves[10,11]. This could imply deceptive intentions - actively putting on a moral appearance to hide one's immoral reality[12,13]. Indeed hypocritical blamers are considered untrustworthy and are morally condemned[9,14–16]. However, a recent experimental study[9] showed that underlying hypocritical blame there might be authentic moral standards. Participants made moral decisions about whether to inflict pain on another person in exchange for profit. A week after their decisions, and unbeknown to the participants, they judged how morally blameworthy similar decisions in the same task were. Hypocritical blame could then be estimated for each participant as the degree of discrepancy between their decisions in the first session and their moral judgements of similar behaviours. Using this measure, it was found that hypocritical blame was related to feelings of conflict and guilt during moral decision making. This suggests that, at least in some hypocritical blamers, failing to live up their moral standards is not necessarily linked to deceptive intentions but to a weakness of the will (*akrasia*[16–18]).

One potential factor that might underlie a failure to live up to moral standards is how motivated the person is to incur costs to themselves for the benefit of others[19–21]. An important cost people have to incur in everyday life is effort, and many prosocial acts require us to decide whether we are willing to exert that effort for others. For instance, holding the door open for a stranger is a relatively small effort. But, typically, people are averse to effort[22,23]. Given equal rewards that can be obtained, people will choose a course of action that is less effortful[24]. This effort aversion is exacerbated for prosocial acts – people readily perform low effort prosocial tasks, but are much less likely to do something highly effortful if it benefits another person compared to when the effort benefits themselves[25–29]. Previous work suggests that if the costs to behave morally increase, people fail to behave according to their moral principles[19,21]. Thus, even though hypocritical blamers could genuinely share the moral principles they use to judge and blame others, aversion to the costs of performing helpful actions could overpower their good intentions[2,21,30], leading to a failure to live up to their own genuine moral standards[6,17–19]. Such an account would predict that people who show a high degree of hypocritical blame will be those who are more averse to exerting effort to benefit others. However, whether people's levels of hypocritical blame are associated with people's aversion to exerting effort for another's benefit is unclear.

Here, we test the association between hypocritical blame and motivation to effortfully help, We operationalised individual differences in hypocritical blame as the discrepancy between decisions in a moral task and the blame assigned on similar decisions made by others[9] (**Fig. 1A**). Participants traded-off profit against electric shocks that were delivered to a stranger. Using computational modelling, we calculated the probability of each participant to harm others for profit based on their decisions in the task. At least a week later, participants completed a different task where they witnessed the same trials again, but this time they rated how blameworthy harmful decisions were. We calculated a hypocritical blame index as the product between probability of harm and their judgement towards the same actions, such that higher values indexed more judgement for actions that participants would perform themselves.

People's willingness to exert effort to benefit others – prosocial motivation – was measured using a prosocial effort task, in which participants traded-off different levels of physical effort in exchange for different magnitudes of reward received by either participants themselves or another unknown person[25–27] (**Fig. 1B**). If participants decided to work in order to maximise the rewards for either themselves or the other person, they had to exert physical effort in the form of squeezing a handle at the appropriate level for that trial. Thus, the prosocial effort task allowed us to measure two aspects of motivation: people's decisions to exert effort and their energisation of effortful actions.

Having these measures, two alternative hypotheses could be tested for how hypocritical blame is linked to both motivational aspects. Firstly, if hypocritical blame is in part explained by a reduced willingness to exert effort in general, rather than just for prosocial acts, it would be expected that hypocritical blamers show less willingness to exert effort regardless of who the beneficiary may be, self or other. In contrast, if hypocritical blame is associated only with people's willingness to act prosocially, then it will be associated with people's aversion to effort only when another person benefits. These two hypotheses can be tested in terms of both people's choices of whether to exert effort, and in terms of how much force – or energy – they exert into the effortful acts they have chosen to undertake.

We show that hypocritical blame is associated with effort aversion, specifically for prosocial acts. Those higher in hypocritical blame were less willing to choose to exert effort for another's benefit, and exerted less force into those actions when they chose to perform them. These results suggest motivation may be a key factor that leads to hypocritical blame.

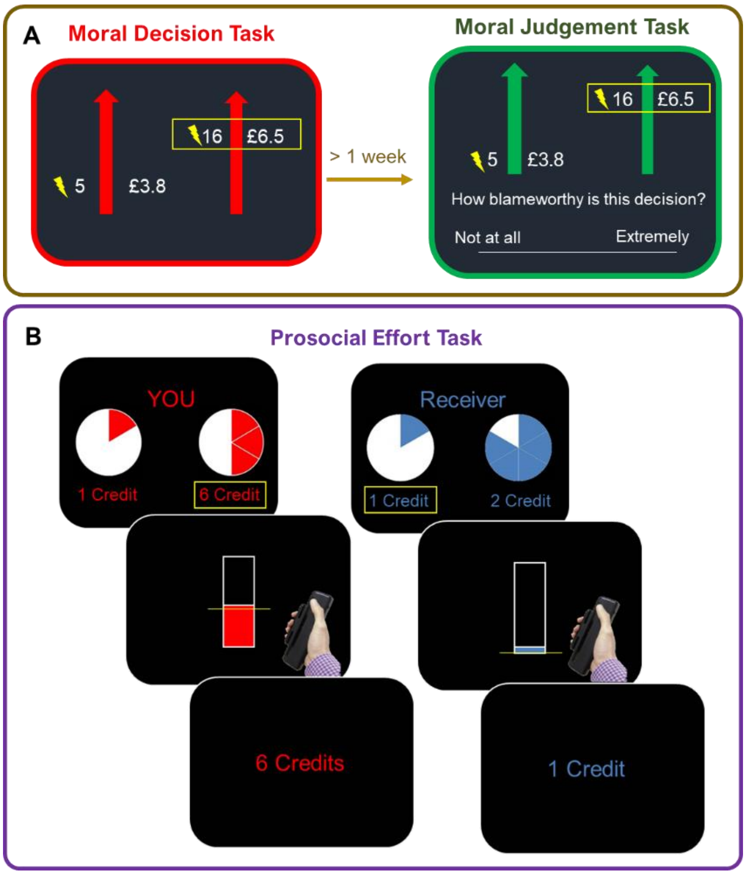

Figure 1. Behavioural measures. A. Hypocritical blame measure. Participants completed two tasks separated for at least a week - (i) a moral decision task, where participants traded-off profit against pain delivered to another person, choosing between a helpful (less profit and electric shocks) and a harmful option (more money and shocks), and (ii) a moral judgement task, where participants judged how blameworthy a series of decisions were when the harmful options were chosen. Unknown by participants, this set was the same trials they completed a week earlier. Hypocritical blame index was calculates using behaviour in these two tasks. B. Prosocial motivation measure. In the prosocial effort task, participants chose between a rest, low reward option, and a work, higher reward offer. If the work option was chosen, the participant has to exert force thresholded to their own maximum voluntary contraction (MVC). For the work offer, different combinations of the five levels of effort (30-70% MVC) and reward (2-10 credits) were presented. Crucially, in half of the trials, rewards were delivered to the participant themselves, while in the other half it was given to an anonymous stranger.

Results

Hypocritical blame is highly prevalent and varies across participants

The hypocritical blame data and analysis are the same that appeared in Yu et al., 2022[9]. Participants (n = 61) first completed a harm aversion task measuring their moral decisions, where people traded-off money they can earn against electric shocks, choosing between a more harmful (but more profitable) option or a less harmful option[31,32] (**Fig. 1A**). On half the trials, the harm –number of electric shocks – was delivered to oneself, and on the other half, to an anonymous person. This task allowed us to quantify the degree to which someone would themselves profit from harm being delivered to another person. Importantly, at least one week later, participants also completed a moral judgment task. In this task they witnessed the same options from the harm aversion task that they had performed before, but now they would rate how blameworthy those decisions were, between not at all to extremely blameworthy. Crucially, the harmful option was always chosen, and this task only had trials where shocks were delivered to someone else. Unbeknown by participants, this was the same set of trials they performed a week earlier. Thus, this task measured how much they blamed someone for choosing to profit from someone else's harm.

Using these tasks, a measure of hypocritical blame was quantified[9] by comparing (i) the participants' likelihood of making a harmful decision on the harm aversion task on the trials where the shocks were received by the other person (calculated through a computational model, see **Materials and Methods**), with (ii) the blame they assigned for a harmful decision in the same trial. Hypocritical blame was defined by the sum of the blame assigned on each trial, weighted by the participants' own likelihood to harm on the same trial. In this way, hypocritical blame was the discrepancy between the probability to harm in each trial and how harsh participants blame others for the same action. Thus, participants who assign a high level of blame on trials where they themselves are likely to choose the harmful option have a high hypocrisy score. In contrast, participants who almost never assign blame on the trials where they themselves are likely to choose the harmful option have a low hypocrisy score. With this operationalisation of hypocritical blame, 97% of participants displayed at least some level of hypocrisy, with a wide range of individual variation in the degree in which it is manifested, following a normal distribution (M±SD = 12.9±7.6; Kolmogorov-Smirnov normality test, **Fig. 2A**).

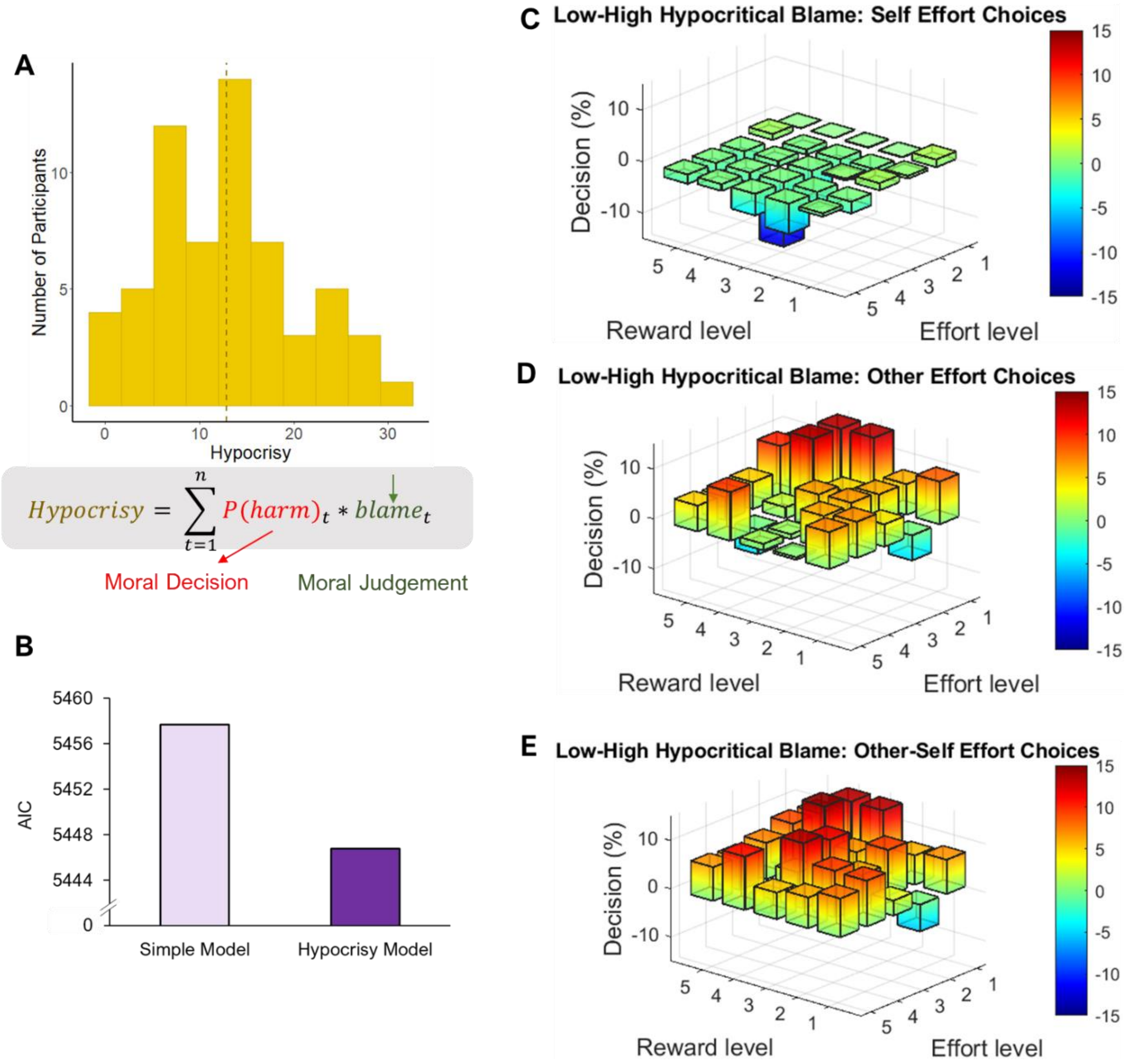

Figure 2. Hypocritical blame interacts with effort, reward and beneficiary in the prosocial effort task. A. Hypocritical blame was formalised as the sum of the trial-by-trial blame in the judgement task weighted by the probability to harm computed from their moral decisions. Hypocritical blame index followed a normal distribution. B. Adding hypocritical blame, in addition to effort, reward and beneficiary, as an independent variable into a model predicting decisions to work improves model fitting according to AIC values (Akaike Information Criterion, y-axis)**.** C-E. 3D plots on percentage of choices to work versus rest in the prosocial effort task across different effort and reward levels in low compared to high hypocritical blamers (median split). C. Self trials. Participants who have high and low scores in hypocritical blame show similar patterns of decisions in self trials across different effort and reward levels. D. Other trials. Participants who score low versus high in hypocritical blame are more motivated to work for others across many of the effort and reward combinations, and especially when high reward is obtained by low effort. E. Differences between low and high hypocritical blamers in choices to work for others vs self. People with a high hypocritical blame score are more biased to work for self than other, especially in trials where high reward is obtained by low effort. Note: Median split dividing participants in high and low hypocritical blamers was only used for illustrative purposes. All statistical analyses used hypocrisy as a continuous variable.

Hypocritical blame is linked to prosocial motivation

In the same experimental session where the moral judgement task was performed, participants completed the prosocial effort task, where they decided between a low reward, rest option, and a variable high reward, high effort option (five levels for effort and reward respectively, **Fig. 1B**). Importantly, on half of the trials the reward was for participants themselves, while in the other half the money was received by an anonymous receiver. This person was a different receiver relative to the harm aversion task, to avoid the potential influence of reciprocity (see **Material and Methods** and **Supplementary Methods** for details). Using this task, we tested whether hypocritical blame was linked to motivation to work to benefit others and/or self.

We used a linear mixed effects model (MM) predicting decisions to work versus rest with predictors of level of effort, magnitude of reward, and beneficiary of the reward, together with their potential interactions. In doing so we could examine whether it was specifically effort, rather than reward, that was associated with hypocritical blame. These models had a random intercept for each subject, and random effects for effort and reward levels, as previous studies have shown that sensitivity to this information vary across participants[25,29]. First, we tested whether adding hypocritical blame as a variable into this MM improved model fitting, demonstrating a better prediction of choices to exert effort by including the blame measure. We found that a model containing effort, reward, beneficiary and hypocritical blame outperformed a model without the latter variable (AIC: simple model= 5457.7; hypocrisy model = 5446.8, **Fig. 2B**). A loglikelihood ratio test confirmed that the hypocrisy model significantly improved the model fit ($\chi^2_{diff} = 28.93$, $df_{diff} = 9$, $p < 0.001$) suggesting that hypocritical blame is important to predict decisions to exert effort for reward.

Within the winning model (for detailed results see **Supplementary Table S1**), a four-way interaction was found between hypocritical blame, effort, reward and beneficiary ($b = 0.27$, $SEM = 0.1$, $z = 2.7$, $p < 0.007$). Qualitative examination suggested this effect was driven by high levels of hypocritical blame being associated with a reduced willingness to exert effort for reward on other trials, but not in self trials. There was no difference between high and low hypocritical blamers for the self condition in decisions to work (**Fig. 2C**), but people who had higher scores in hypocritical blame were less willing to work for others across almost all reward and effort levels (**Fig. 2D**). Comparing the willingness to choose to exert effort for self vs other in high vs low hypocritical blame (**Fig. 2E**), shows that hypocritical blamers are less prosocially motivated in general, and especially so in trials where little effort is required.

We performed a series of post-hoc analyses to test the qualitative description above. Thus, for each reward and effort combination, we tested whether the slope of hypocritical blame predicting decisions to work were significantly different between self and other trials. We found that differences between self and other occurred mainly at lower effort levels, especially at mid to high level of rewards (see **Supplementary Table S2** and **Supplementary Figure S1**). Crucially, as the effort level increases, differences between self and other begin to lose statistical significance. Indeed, at the highest effort level, there were no significant differences between self and other across reward magnitudes. Likewise, no significant differences were found in the effect of hypocritical blame on decisions at the lowest reward magnitude, independent from effort. These results suggest that the reduction of motivation to help others in highly hypocritical people occurs especially when the effort costs are low-to-moderate, and others' benefit moderate-to-high. Thus, higher levels of hypocritical blame are associated with a reduced incentivisation by rewards and a higher sensitivity to effort costs, specifically when another person will benefit.

Consistent with the above results, hypocritical blame also showed a significant three-way interaction with beneficiary and effort ($b = 0.27$, $SEM = 0.11$, $z = 2.5$, $p < 0.02$) and a two-way interaction with beneficiary ($b = -0.52$, $SEM = 0.11$, $z = -4.54$, $p < 0.001$), but not a main effect by its own. These results support the four-way interaction described above - as participants scored higher in hypocritical blame, they worked less for others compared with self (**Fig. 3**). Finally, we found similar effects revealed by previous studies showing that participants were less willingness to work for others than for self, especially in high effort trials, and in trials with low rewards on offer (see **Supplementary Table S1**).

Hypocritical blame is associated with reduced energisation of actions when working for others.

Hypocritical blamers exert less force when invigorating actions that benefit others.

Even after people make a decision to exert effort, they might not always put as much energy into those actions for others compared to self. Next, we tested whether hypocritical blame was linked to how much force participants exerted into actions when helping others. We examined the "normalised force" for each trial in the prosocial effort task – calculated as the area under the curve (AUC) of the force normalised to a participants maximum AUC across all trials. We first tested whether including hypocritical blame into the model improved model fitting. Similar to what we found looking at decisions to work, a model that included hypocritical blame outperformed a simpler model with only effort, reward, beneficiary and their interactions as independent variables (AIC, simple model = -17179, hypocrisy model = -17193). This was later confirmed by a loglikelihood test that revealed a significant model fitting improvement by the hypocrisy model ($\chi^2_{diff} = 27.59$, $df_{diff} = 7$, $p < 0.001$).

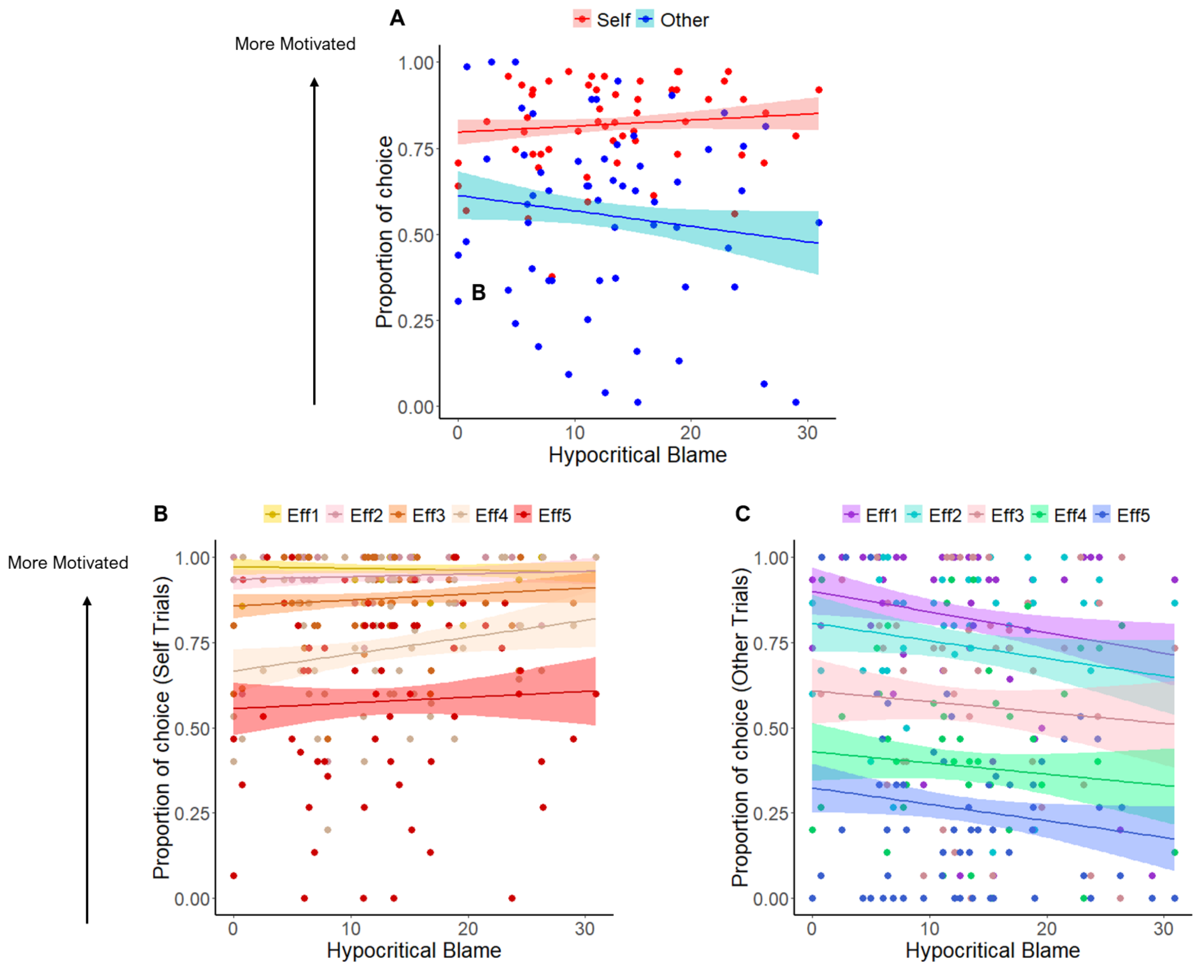

Figure 3. Hypocritical blame is associated with less willingness to work for others compared to self. A. Y-axis shows proportion of choosing the work versus the rest option. Participants get less motivated to work for others compared to self as they have higher scores in hypocritical blame B-C. Effects of hypocritical blame on choices according to effort and beneficiary. Hypocritical blame is associated with less motivation for others across effort levels (C) compared to self (B). Shaded areas show the 70% confidence interval around the slopes. Individual points show the score of each participant for each condition. Eff = Effort Level.

An interaction between hypocritical blame and beneficiary was found (b = -0.01, SEM = 0.002, t = -3.94, p < 0.001), indicating that as participants are more hypocritical, the difference in force exertion between self and other increased (**Fig. 4A**), such that they exerted less force for actions that benefitted others relative to self than participants low in hypocritical blame. Thus, the discrepancy between moral judgements and actions is associated with higher gaps between self and other action invigoration. Furthermore, this model revealed main effects of effort (high effort levels require more force exertion), reward (participants exerted more force when the reward was high), and beneficiary, together with an effort x beneficiary interaction, showing that participants generally exerted less force for others than for self, especially in high effort trials, replicating previous studies (see **Supplementary Table S3,** and **Supplementary Figure S2**).

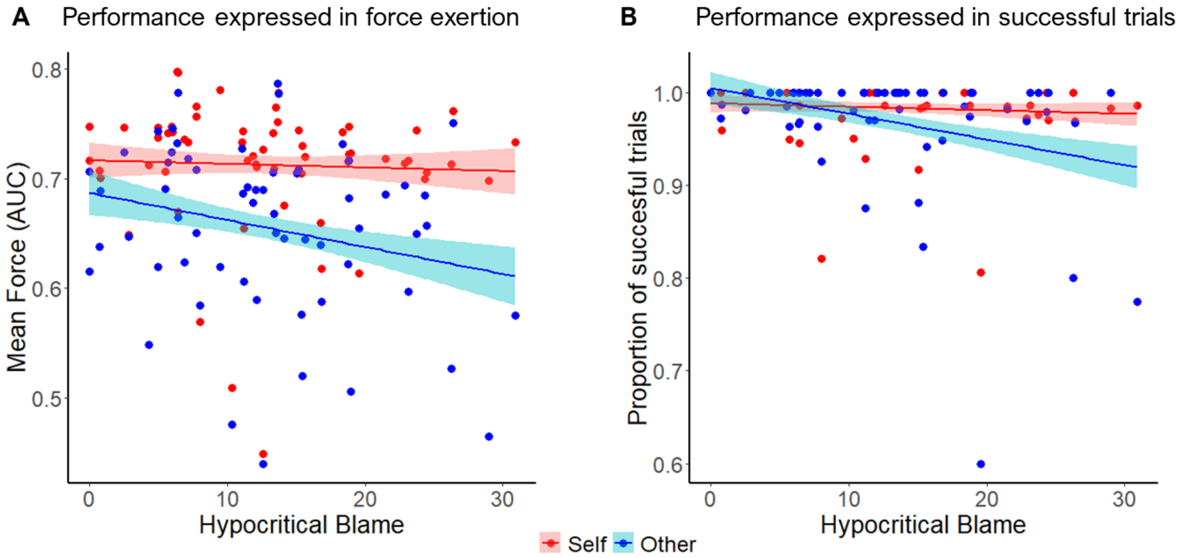

Figure 4. Hypocritical blame is linked to lower performance in actions that benefit others compared to self. A. Hypocritical blame is associated with less force exerted for others compared with self. Participants who scored high in hypocritical blame showed more differences in how they exerted force for others compared to self when similar effort was required, suggesting more superficiality in their prosocial actions. Y-axis depicts the mean area under the curve (AUC) during the 3 seconds force period normalised to participants maximum level of force exerted across trials. B. Hypocritical blame is associated with less success in trials benefiting others. As people score higher in hypocritical blame, they fail more in trials where the benefit is received by others compared with themselves. Y-axis depicts the proportion of successful vs failed trials. Shaded areas show the 70% confidence interval around the slopes. Individual points show the score of each participant for each condition.

Since participants, regardless of hypocritical blame, chose to benefit others much more when the effort required was low compared to high, the above effect could be accounted by a reduced sampling of trials at the higher level of effort in the other condition. That is, there are less high effort trials for other in people who are highly hypocritical. To confirm this was not driving the effect, we examined if removing the higher effort trials still revealed an effect of hypocritical blame. Consequently, we removed consecutively the three highest effort levels from the model testing each time whether the effect was still present. The hypocritical blame x beneficiary interaction proved to be a robust effect and it was present even if the highest effort level (70% of maximum voluntary contraction, MVC; t = -4.1, p < 0.001), as well as the two highest effort level (60-70% MVC; t = -2.75, p < 0.006), and the three highest effort levels (50-70% MVC; t = -2.17, p < 0.04). Thus, the people who were highest in hypocritical blame simply exerted less force for others, even when they had chosen to do so.

Hypocritical blame is linked to higher failure to exert the required force when performing prosocial actions

As effort levels are set as a percentage of participants' MVC in the prosocial effort task, all effort levels should be attainable, and failure indicates a reduction in the motivation to exert the required effort even when they freely chose to undertake it. In fact, success rates were very high overall in the prosocial effort task. Participants on average obtained the reward on offer in 98.3% of the self trials (SEM = 0.5%), and in 96.9% of the other trials (SEM = 0.8%). Is hypocritical blame related to failure to succeed at the required effort level? The above results on force exertion suggests that this could be the case. In order to test whether hypocritical blame modulated the probability to succeed in a given trial, two different models were built using the same approach described above, i.e. a simple model against a hypocrisy model, but now predicting the binary variable 'success'. Note, however, that the results revealed in this analysis could be driven only by a number of participants in only a number of trials due to the high success rates of this task. The following results should be therefore interpreted with caution.

The hypocrisy model improved model fitting relative to the simple model according to the AIC (simple model = 953.7, hypocrisy model = 949.1). A loglikelihood test confirmed that the hypocrisy model significantly outperformed the simple model ($\chi^2_{diff} = 18.61$, $df_{diff} = 7$, $p < 0.01$). Crucially, this model revealed that hypocritical blame interacted with beneficiary (b = -0.74, SEM = 0.3, z = -2.48, p < 0.02, **Fig. 4B**), such that people who scored high in hypocritical blame failed more in other trials compared with people who scored low, showing bigger differences in success between self and other. These results indicate that participants with high scores in hypocritical blame not only exerted less force for others than for themselves for similar actions, but also, they did so to the extent that more often failed to energise the required force to obtain the reward on offer. Finally, the hypocrisy model revealed main effects of effort, indicating that people failed more as trials get harder, and beneficiary, showing that people failed more in other compared to self trials (**see Supplementary Table S4**).

Given that hypocritical blame modulated success in other trials, it could be argued that its effect on force exertion might have been driven only by those trials where participants exerted such a low amount of force that they did not achieve the required effort goal. However, this seems unlikely, since the effect of hypocritical blame on force remained even if highest effort trials (50-70% of participants' MVC), which are more likely for participants to fail, are removed from the model. Despite this, and to add robustness to these results, we performed a final analysis to test the effects of hypocritical blame on normalised force building again a simple and a hypocrisy model on trial-by-trial force exertion but this time considering only trials where participants succeeded. Again, the hypocrisy model significantly improved model fitting relative to the simple one (AIC: simple model = -18858, hypocrisy model = -18865, χ^2_{diff} = 20.39, df_{diff} = 7, $p < 0.005$). Crucially, the hypocritical blame x beneficiary interaction was maintained (b = -0.003, SEM = 0.001, t = -2.18, $p < 0.03$), indicating that people who scored high in hypocritical blame exerted less force for others across trials. A three-way interaction between hypocritical blame, beneficiary and effort was also significant in this model with only successful trials (b = 0.004, SEM = 0.001, t = 2.64, $p < 0.01$), suggesting that the modulation of hypocritical blame on the force exerted for self and other varied across different effort levels. This effect is likely to be triggered by the removal of the failed trials from the analysis, which corresponds to the manifestation of hypocritical blamers' low motivation in force exertion especially at the higher effort levels. Thus, while hypocritical blame effects on lower effort levels remained in this model, its effects in the harder ones were diminished by eliminating these failed trials.

Discussion

Every day people blame others for actions that they judge as morally wrong, denoting in this way their moral standards, but people may often also fail to live up to those moral standards themselves. We tested the possibility that this hypocritical blame could be related to a reduced motivation to overcome the costs of being moral and doing prosocial acts that benefit others[19]. Here, we investigated this hypothesis using decision-making tasks measuring hypocritical blame and prosocial effort motivation within the same participants, testing for their association. This approach revealed that people with higher levels of hypocritical blame, i.e. wider gaps between moral judgements and actions, are (i) less willing to put in effort to benefit others relative to self, and (ii) exerted less force when performing effortful actions to benefit others compared to their own. These results suggest that hypocritical blame might be related to a dissociation between moral standards manifested in blame judgements, and the motivation to overcome the costs to act accordingly.

This study revealed that hypocritical blame is associated with less willingness to incur effort costs specifically for prosocial actions, but not for self-benefitting acts. This aligns with the fact that the motivation to put in effort for ourselves may be fundamentally different that choosing to help others, which may be more of a moral choice. Choosing whether to exert effort to obtain rewards is a goal-directed behavioural problem that even non-social animals must consider. People vary in their willingness to work in the absence of any social context. However, prosocial effort is morally relevant. Doing something to benefit someone else requires a personal cost[33] without an immediate or well defined benefit to oneself. For people high in hypocritical blame, while they are very willing to judge others actions, overcoming the personal cost of exerting effort may be too high. This could be due to a difference between self and other interests in moral hypocrites[34,35]. This is consistent with evidence suggesting that low hypocritical people might be more self-driven by the welfare of others, valuing self and other benefits more equally[35], in contrast with people who act morally only to conform to the social norm[10,36,37]. Indeed, previous work has shown that people scoring high on moral responsibility are not necessarily less hypocritical, but those who score high in self-motivated moral intentions display more alignment between moral principles and actions[2,10,35–38]. The current results suggest that this reduced prosocial motivation might be due to high sensitivity to effort costs and reduced incentivisation by rewards when benefitting others.

Strikingly, hypocritical blamers were less willing to work prosocially especially in trials where the effort cost was low to moderate and the benefit high to moderate. At first, this result seems counterintuitive, since it is in the high effort levels where people in general show more reluctance to incur in prosocial effort[25–29], and hypocritical blame could have just augmented this effect. However, people have consistently shown high motivation to work in easy trials regardless of the beneficiary, a ceiling effect that could impact on the current results. Thus, as hypocritical blamers also display low motivation in these trials, the difference in this spectrum of the design is more pronounced compared with participants low in hypocritical blame. In fact, we did not see effects of hypocritical blame on decisions in the highest effort levels between self and other - people were equally demotivated to work for others regardless of hypocrisy. Furthermore, the effects of hypocritical blame were evident at higher levels of rewards, especially in combination with low effort costs. Taken together, these results suggest that the effort sensitivity of hypocritical people is manifested in social situations where most people would be willing to work to benefit others. Hypocritical blamers might reduce self-costs, even if those costs are small, especially in contexts where there is not a clear moral norm for behaviour like the prosocial effort task - i.e. not putting in effort to help others is not as socially prohibitive as harming others for profit[21,39–42].

Intriguingly, hypocritical blame in the current study was not only associated with less willingness to choose to work for others, but also, when they decided to do so, they energised less the prosocial actions compared with self-benefiting behaviour ones. Indeed, sometimes people high in hypocritical blame failed to achieve the required force to obtain the money for others due to their poorer performance. These results suggest that hypocritical blamers’ interest towards others could also be *superficial* – they may appear prosocial in their willingness to help on the surface, but the effort they invest to benefit others suggests otherwise. Consequently, hypocrites might display moral principles mainly through their judgements and desires, but not through actual actions. When faced with the opportunity to skip the self-cost, they would do so, although in appearance they could maintain their moral status[2,38].

Research on goal-directed behaviour can offer an explanatory framework for interpreting these results. In reward processing, two components can be identified[43,44] - a hedonist impact of pleasure given by the reward, a '*liking'* component; and the incentive saliency of the reward, a '*wanting'* component, that drives the agent towards a reward-seeking behaviour that leads to a willingness to exert effort. Even though these two components are related to each other, they can, in principle, be dissociated and linked distinct neural mechanisms. The current results could be interpreted as hypocritical blamers 'liking' the idea of being moral – they share moral principles that lead to high moral standard reflected in their judgements and desires, but they fail in 'wanting' to be moral, and thus do not overcome self-costs to benefit others.

The present results, especially the superficiality of the hypocritical prosocial decisions, could be interpreted as supporting a claim that hypocritical blame is a deceptive behaviour. That is people might want to display moral standards that they have no intention of meeting themselves. However, this does not necessarily have to be the case[6,17,18]. First participants, regardless of hypocritical blame, decided to help others in a high number of trials, including ones with high effort costs. Second, in the vast majority of the trials, participants, regardless of hypocritical blame, achieved the force required to obtain the reward on offer. Finally, failing to meet their prosocial and moral standards might trigger guilt and frustration in the hypocritical blamer due to their moral failures[6,17,18,45,46]. Indeed, according to previous results[9], hypocritical blamers can feel conflict and guilt when they fail to behave according to their moral principles. Taken together, therefore, it seems more appropriate to consider hypocritical blame as the extreme of a continuum, associated with reduced motivation to overcome self-costs to benefit others. People with higher levels of hypocritical blame still worked for others, only that they did so to a lesser degree than the lower extreme, as they are more sensitive to their self-interest when benefitting others.

Future research could extend this work and control for some limitations of the current study. Firstly, it could test directly whether moral hypocrisy is associated with deceptive or genuine moral intentions. Secondly, it could also examine whether the current results could be extended to different scenarios, where the cost to be overcome by participants is other than effort, where hypocrisy is manifested as other types of moral discrepancy (e.g. attitudes/actions, judgements for self/other actions), where judgments and actions are made publicly vs privately, and where hypocrisy is embedded in real social contexts[38,47,48]. Furthermore, future research could shed light on what personality traits could modulate the motivational aspects of hypocritical blame[2,11,35,36,46,49–51]. Thus, the present results open a research opportunity for future investigations to illuminate the psychological aspects of moral hypocrisy.

In summary, here we tested whether hypocritical blame, a discrepancy between moral judgements and actions, is associated with lower prosocial motivation in the form of a reduced willingness to exert effort for others' benefit. Our results highlight that those that hypocritically blame others, may do so because they are too sensitive to the costs to be moral. These results open an investigative door into a motivational, goal-directed component of moral hypocrisy.

Materials and Methods

Participants

All protocols were approved by the ethics committee of the University of Oxford (R50262/RE001). The participants used in this study were part of the sample used by Yu et al., 2022[9] (n = 62). One of these 62 participants never chose to help others in the prosocial effort task, meaning an absence of sufficient variance of force data, crucial for hypothesis testing. Therefore, 61 participants were included in the analyses reported here (age M = 22.6, SD = 3.8, 34 females). All participants gave written informed consent and were financially compensated for their time.

General Procedure

Participants took part of a multi-stage, multi-task study researching social decision-making. As part of this study, participants had two experimental sessions. First, participants completed the harm aversion task. Second, and at least one week later (range = 7–74 days, median = 13 days), participants attended to a behavioural session, where they completed the moral judgement task and the prosocial effort task (see Yu et al., 2022[9] for details about the general procedure).

First Experimental Session

Harm Aversion Task

Before completing the harm aversion task, participants completed a pain thresholding procedure[31,32,52] aiming to familiarise them with the painful shocks to be traded-off in the decision task, and to identify the physical intensity according to each participant's subjective pain scale and match it across the sample. Next, participants went through a role assignment procedure where they were mock-assigned the role of Deciders while a confederate was assigned the role of Receiver, following a well-established protocol[25,31] (**see Supplementary Methods**).

After these procedures, participants received instructions about the task and completed some practice trials. They were told that one trial was going to be randomly selected and implemented at the end of the session, and they were ensured that their decisions would be completely anonymous and confidential. Participants then completed the harm aversion task, measuring their moral behaviour[31,32,52]. In this task, participants traded-off a certain number of electric shocks against financial profit. Crucially, in half of the trials participants believed that the electric shocks were delivered to another unknown person, the Receiver, while in the other half shocks were delivered to participants themselves, the Deciders. Money, on the other hand, was always obtained by the participants/Deciders.

In every trial, participants decided between two options: a harmful option, associated with more shocks for more profit, and a helpful option, associated with less shocks but in exchange of less money. 72 trials were included per condition (generated using the criteria described in Crockett et al., 2017[32]), plus four catch up trials (harmful options were associated with lower profit), making a total of 76 trials. In half of these trials, the harmful option was on the left side of the screen, while in the other half they were on the right. These 76 trials were duplicated to have the same trials for self and other conditions. These trials were displayed in two blocks of 76 trials to each participant, with each block equal amount of self and other trials. Four sets of 76 trials were produced following this protocol, and participants were assigned randomly one of these sets. This trial generation protocol ensured that number of shocks were decorrelated from the amount of profit ($|r| < 0.07$, $p > 0.53$).

After the harm aversion task, one trial was randomly selected and implemented. Thus, if a self trial was selected, shocks and money corresponding to participants' choice in that trial were delivered to them at the end of the session. However, if an other trial was selected, participants received the money and were told that shocks will be delivered to the Receiver at the end of their session. Finally, participants answered a few debrief questions about the study.

Second Experimental Session

Moral Judgement Task

To operationalise hypocritical blame, decisions in the harm aversion task had to be contrasted with moral judgements for similar actions. Thus, the moral judgement task aimed to examine how much people rate actions that they have previously performed as blameworthy when they are performed by others[9]. At least one week after participants completed the harm aversion task, they had an experimental session where they undertook the moral judgment task. In this task, participants were presented with a subset of the trials that they completed in the harm aversion task, i.e., all the trials in the other condition where the money was for the Decider and the shocks were for the Receiver (72 trials). On each trial, the harmful option was chosen. Participants were asked to judge the blameworthiness of each harmful choice on a visual analogue scale ranging from 0 (not at all blameworthy) to 100 (extremely blameworthy). Using this task, we calculated each participant's degree of hypocritical blame, which quantifies discrepancies between their blame judgments and the choices they made a week earlier (see below).

Prosocial Effort Task.

Participants completed the prosocial effort task[25–27] after the moral judgement task in the same behavioural session. They were told that they would continue their role as Deciders, while a Receiver would be paired with them from the pool of Receivers in the study. In the prosocial effort task, participants traded-off physical effort for monetary rewards in the form of credits. Crucially, in half of the trials the money was received by the Receiver, while in the other half profit was obtained by the participants themselves, the Deciders. In every trial, participants chose between two options: a baseline option, associated with no exertion of effort (i.e. 3 seconds of rest) for a low reward level of one credit, and a work offer, associated with higher amounts of reward (2-10 credits) for variable levels of effort exertion (30-70% of their maximum voluntary contraction) to obtain them. Once participants made their choice, they were required to perform the specific effort level squeezing a handle with their dominant hand for at least one second in a three second window. Failing to do so meant to get zero rewards for that trial. Each combination of effort and reward were repeated three times per condition, having 75 trials per beneficiary. This task allows to measure two important aspects of motivation: prosocial effort-based decisions, and the energisation of actions for self and other. For the latter, two indexes were used: (i) the force exerted for each trial where participants decided to work, and (ii) participants' success in accomplishing the required level of physical effort to obtain the reward on offer. In principle, there should not be differences between self and other as participants have the alternative of resting for a low reward, and as each effort level is adjusted to a percentage of each participant's maximum voluntary contraction (MVC), which ensure a high level of success. Thus, if participants choose to work for a specific amount of effort level, they should exert, on average, the same amount of force in that effort level regardless of whether the reward is for themselves or not. However, participants in fact generally exert less force into the same actions when they benefit others than self[25,26].

Prior making decisions in the main task, participants' MVC was determined, when they squeezed the handle as strongly as they could. This ensured effort levels were idiosyncratically set for different participant's grip strength. After this, participants experienced each effort level three times, getting familiar with effort levels and their display on the screen. Once these two processes were completed, participants performed the prosocial effort task.

Analysis of the behavioural data

Calculating Hypocritical Blame

To compute the degree of hypocritical blame in participants, we determined the discrepancy between participant's own aversion to harming others and the level of blame that they attributed to similar actions, following the methods used in Yu et al., 2022[9]. Given that we assessed participants' real choices in the harm aversion task, we could contrast these choices with the moral judgments they rendered towards the decisions made the same set of trials. We thus quantified each participant's degree of hypocritical blame by comparing (i) their probability of choosing he harmful choice in each trial of the harm aversion task and (ii) the blame they assigned on each trial when a harmful decision was made in the moral judgment task. We used a computational model that has been extensively validated in previous work[31,32,52] to calculate participants' likelihood to harm in each trial. Here, the probability to harm is the softmax transformation of the subjective value of the harmful relative to the helpful option, ΔV. Thus, for participant *j* and trial *i*:

$$(1)\ \Delta V_{j(i)} = \left(1-\kappa_j\right)\Delta m_{j(i)} - \kappa_j\,\Delta s_{j(i)}$$

$$(2)\ P(harm)_{j(i)} = \frac{1}{1+\ exp^{-\beta_j\,\Delta V_{j(i)}}}\left(1-2\varepsilon_j\right) + \ \varepsilon_j$$

Where *Δm* and *Δs* are the difference in money and shocks respectively between the harmful and the helpful options. The free parameter *κ* represents how sensitive participants are to shocks over money. Importantly, κ has different values for self and other trials (κ_{self} and κ_{other}, respectively). We transformed ΔV into probabilities using a softmax function, where *β* was a participant-specific inverse temperature parameter indicating the stochasticity of the decisional process. A lapse rate parameter *ε* was also included which captured task-irrelevant noise.

To compute hypocritical blame, we used choices made only in the other trials. Hypocritical blame was defined as the sum of the judgement made by the participant in each trial weighted by participants' probability to harm in that trial, such that:

$$(3)\ Hypocritical\ Blame_j = \sum_i blame_{j(i)} * P(harm)_{j(i)}$$

Where $blame_{j(i)}$ is participant j's blame on trial i of the moral judgment task. $P(harm)_{j(i)}$ is the probability to harm of that participant in the same trial extracted from their performance in the harm aversion task.

Testing the relationship between hypocritical blame and prosocial motivation

For decisions, we tested whether hypocritical blame was associated with prosocial motivation with two mixed models (*glmer* function in R):

Simple model

$$(4)\ DW_i = \beta_{0j[i]} + \beta_{1j[i]}R_i + \beta_{2j[i]}E_i + \beta_3 B_i + \beta_4 R_i B_i + \beta_5 E_i B_i + \beta_6 E_i R_i + \beta_7 E_i R_i B_i$$

Hypocrisy Model

$$\begin{aligned}(5)\ DW_i = \beta_{0j[i]} &+ \beta_{1j[i]}R_i + \beta_{2j[i]}E_i + \beta_3 B_i + \beta_4 H + \beta_5 R_i E_i + \beta_6 R_i B_i + \beta_7 R_i H + \beta_8 E_i B_i \\ &+ \beta_9 E_i H + \beta_{10} B_i H + \beta_{11} R_i E_i B_i + \beta_{12} R_i B_i H + \beta_{13} E_i B_i H + \beta_{14} R_i E_i H \\ &+ \beta_{15} R_i E_i B_i H + \beta_{16}\kappa_{other}\end{aligned}$$

In the simple model, decision to work *DW* in the trial *i* is predicted by the fixed effects of reward *R*, effort *E*, beneficiary *B*, and their interactions. DW is a binary, factor variable in this logistic mixed-model. The hypocrisy model adds the hypocritical blame index *H* as an independent variable together with its interaction with all the other variables. The harm aversion parameter κ_{other} was also included as a regressor of no interest given that it is strongly correlated with hypocritical blame[9]. For both the simple and the hypocrisy models, random intercepts were clustered in subjects j, and random slopes on effort and reward were included as it is expected that participants vary in their sensitivities to this information. Crucially, AICs for each model were used to test whether the addition of hypocrisy improved the model. This was complemented with a loglikelihood test using the *anova* function in R to test whether differences in model fitting were significant, given that the simple and the hypocrisy models were nested.

For pot-hoc analyses, we used the *emtrends* function in R (*emmeans* package). For these analyses, we took hypocritical blame as a covariate, and compared whether the slope at each level of effort level and reward magnitude combination was different between self and other. Therefore, with this analysis we could test for significant differences of the effects of hypocritical blame on decisions to work between self and other, disentangling the 4-way interaction found in the main results.

To test the influence of hypocritical blame on performance, similar mixed-models to the ones described above were used. In these models, the force exerted in each trial in which participants chose to work were used as dependant variable instead of DW, using the *lmer* function in R. Force was normalised per each trial as a proportion of participant's maximum to account for between-subject variability in the force exerted and calculated the AUC for the three seconds window in which they exerted force. Thus, normalised force was a continuous variable in these subject-level random-intercept models. Trials where participants chose to rest were excluded from analysis. Unlike the decision models described above, random slopes for effort and reward were not used because (i) participants had different number of trials per each different condition, and (ii) high individual variability in the force exerted between levels of effort and reward was not expected - the force required to obtain the reward on offer was set for each participant's specific MVC and showed limited variability in existing data[25–27]. Furthermore, interactions between effort x reward x beneficiary were not included, as participants, regardless of hypocritical blame, chose to work for others significantly less often compared to self, especially in low-reward/high-effort trials. This means that a potential 3-way interaction would be unstable because of the small trial sampling at certain combinations of effort and reward.

Finally, the effects of hypocritical blame on performance were also tested with a model that predicted whether each trial was succeeded or failed, i.e. whether people achieved the specific effort requirement for each work trial and therefore whether they obtained the reward on offer. Here, success was a binary factor variable in a logistic random-intercept mixed-model. The success models had the same structure that the force models, i.e. 3-way interactions of effort x reward x beneficiary and random slopes were not included. Notice that even though models predicting success can give important insights about the motivational profile of hypocritical blamers, these results must be taken cautiously, as the success rates in this task were very high regardless the condition, and therefore, the effects of hypocritical blame on trial success could be triggered by mainly a few participants/trials.

Acknowledgements

This work was supported by grants from the John Templeton Foundation (Beacons Project and No. 61495), the Academy of Medical Sciences (SBF001\1008), the Oxford University Press John Fell Fund, and the Wellcome Trust Institutional Strategic Support Fund (204826/Z/16/Z) awarded to MJC; a Biotechnology and Biological Sciences Research Council David Phillips Fellowship (BB/R010668/1; BB/R010668/2) and a Jacobs Foundation Research Fellowship to MAJA; a National Agency for Research and Development (ANID) DOCTORADO BECAS CHILE (BECAS CHILE/2016 – 72170287) to LSCH; and a Sir Henry Dale Fellowship funded by the Wellcome Trust and the Royal Society (223264/Z/21/Z) to PLL

Data and code accessibility

All data and scripts used for main analysis and figures can be found here https://osf.io/qmzup/?view_only=b223dbe0b4904c93a64289a695e6ec81.

Competing interests

The authors declare no competing interest.

Authors’ Contributions

LSCH: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. HY: Conceptualization, Methodology, Investigation, Data curation, Writing – review & editing. AMBP: Investigation, Writing – review & editing. PLL: Methodology, Writing - Review & Editing. MJC: Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. MAJA: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

References

1. Alicke, M., Gordon, E. & Rose, D. Hypocrisy: What counts? *Philos. Psychol.* **26**, 673–701 (2013).

2. Batson, C. D., Thompson, E. R. & Chen, H. Moral hypocrisy: Addressing some alternatives. *J. Pers. Soc. Psychol.* **83**, 330–339 (2002).

3. Coates, D. J. & Tognazzini, N. A. *Blame: Its nature and norms*. (Oxford University Press, 2013).

4. Effron, D. A., Markus, H. R., Jackman, L. M., Muramoto, Y. & Muluk, H. Hypocrisy and culture: Failing to practice what you preach receives harsher interpersonal reactions in independent (vs. interdependent) cultures. *J. Exp. Soc. Psychol.* **76**, 371–384 (2018).

5. Merritt, A. & Monin, B. Moral Hypocrisy, Moral Inconsistency, and the Struggle for Moral Integrity. *Soc. Psychol. Moral. Explor. Causes Good Evil* 167–184 (2012).

6. Dover, D. The walk and the talk. *Philos. Rev.* **128**, 387–422 (2019).

7. Howe, L. C. & Monin, B. Healthier Than Thou? Practicing What You Preach Backfires by Increasing Anticipated Devaluation Superior Behavior Triggers Concern for Devaluation. **112**, 718–735 (2017).

8. Laurent, S. M. & Clark, B. A. M. What Makes Hypocrisy? Folk Definitions, Attitude/Behavior Combinations, Attitude Strength, and Private/Public Distinctions. *Basic Appl. Soc. Psych.* **41**, 104–121 (2019).

9. Yu, H. *et al.* Neural and cognitive signatures of guilt predict hypocritical blame. *Psychol. Sci.* **33**, 1909–1927 (2022).

10. Graham, J., Meindl, P., Koleva, S., Iyer, R. & Johnson, K. M. When Values and Behavior Conflict: Moral Pluralism and Intrapersonal Moral Hypocrisy. *Soc. Personal. Psychol. Compass* **9**, 158–170 (2015).

11. Valdesolo, P. & DeSteno, D. The duality of virtue: Deconstructing the moral hypocrite. *J. Exp. Soc. Psychol.* **44**, 1334–1338 (2008).

12. Kittay, E. F. ON HYPOCRISY. *Metaphilosophy* **13**, 277–289 (1982).

13. Szabados, B. & Soifer, E. *Hypocrisy: ethical investigations*. (Broadview Press, 2004).

14. Barden, J., Rucker, D. D. & Petty, R. E. 'Saying one thing and doing another': Examining the impact of event order on hypocrisy judgments of others. *Personal. Soc.*

Psychol. Bull. **31**, 1463–1474 (2005).

15. Jordan, J. J., Sommers, R., Bloom, P. & Rand, D. G. Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling. *Psychol. Sci.* **28**, 356–368 (2017).

16. O'Connor, K., Effron, D. A. & Lucas, B. J. Moral cleansing as hypocrisy: When private acts of charity make you feel better than you deserve. *J. Pers. Soc. Psychol.* (2020).

17. Bartel, C. Hypocrisy as either deception or akrasia. *Philos. Forum* **50**, 269–281 (2019).

18. Mele, A. R. Akratic feelings. *Philos. Phenomenol. Res.* **50**, 277–288 (1989).

19. Batson, C. D. & Thompson, E. R. Why don't moral people act morally? Motivational considerations. *Curr. Dir. Psychol. Sci.* **10**, 54–57 (2001).

20. Daniel Batson, C., Ahmad, N. & Tsang, J. A. Four motives for community involvement. *J. Soc. Issues* **58**, 429–445 (2002).

21. Lindenberg, S., Steg, L., Milovanovic, M. & Schipper, A. Moral hypocrisy and the hedonic shift: A goal-framing approach. *Ration. Soc.* **30**, 393–419 (2018).

22. Hull, C. L. *Principles of behaviour*. (Appleton-Century-Crofts, 1943).

23. Kool, W. & Botvinick, M. Mental labour. *Nature Human Behaviour* (2018). doi:10.1038/s41562-018-0401-9

24. Le Heron, C., Apps., M. A. J. & Husain, M. The anatomy of apathy: A neurocognitive framework for amotivated behaviour. *Neuropsychologia* **118**, 54–67 (2018).

25. Lockwood, P. L. *et al.* Prosocial apathy for helping others when effort is required. *Nat. Hum. Behav.* **1**, 0131 (2017).

26. Lockwood, P. L. *et al.* Aging Increases Prosocial Motivation for Effort. *Psychol. Sci.* **32**, 668–681 (2021).

27. Lockwood, P. L. *et al.* Distinct neural representations for prosocial and self-benefiting effort. *Curr. Biol.* **32**, 4172-4185.e7 (2022).

28. Contreras-Huerta, L. S., Lockwood, P. L., Bird, G., Apps, M. A. J. & Crockett, M. J. Prosocial Behavior Is Associated With Transdiagnostic Markers of Affective Sensitivity in Multiple Domains. *Emotion* (2020). doi:10.1037/emo0000813

29. Contreras-Huerta, L. S. *et al.* Neural representations of vicarious rewards are linked to interoception and prosocial behaviour. *Neuroimage* 119881 (2023).

30. Watson, G. W. & Sheikh, F. Normative self-interest or moral hypocrisy?: The importance of context. *J. Bus. Ethics* **77**, 259–269 (2008).

31. Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P. & Dolan, R. J. Harm to others outweighs harm to self in moral decision making. *Proc. Natl. Acad. Sci.* **111**, 173201–17325 (2014).

32. Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P. & Dolan, R. J. Moral transgressions corrupt neural representations of value. *Nat. Neurosci.* **20**, 879–885 (2017).

33. Contreras-Huerta, L. S. A cost-benefit framework for prosocial motivation—Advantages and challenges . *Frontiers in Psychiatry* **14**, (2023).

34. Tang, H. *et al.* Stimulating the right temporoparietal junction with tDCS decreases deception in moral hypocrisy and unfairness. *Front. Psychol.* **8**, 1–7 (2017).

35. Tang, H. *et al.* Are proselfs more deceptive and hypocritical? Social image concerns in appearing fair. *Front. Psychol.* **9**, 1–9 (2018).

36. Dong, M., van Prooijen, J. W. & van Lange, P. A. M. Self-enhancement in moral hypocrisy: Moral superiority and moral identity are about better appearances. *PLoS One* **14**, 1–17 (2019).

37. Lönnqvist, J. E., Irlenbusch, B. & Walkowitz, G. Moral hypocrisy: Impression management or self-deception? *J. Exp. Soc. Psychol.* **55**, 53–62 (2014).

38. Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C. & Wilson, A. D. In a very different voice: Unmasking moral hypocrisy. *J. Pers. Soc. Psychol.* **72**, 1335–1348 (1997).

39. Carnes, N. & Janoff-Bulman, R. Harm, Help, and the Nature of (Im)Moral (In)Action. *Psychol. Inq.* **23**, 137–142 (2012).

40. Decety, J. & Cowell, J. M. Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. *Dev. Psychopathol.* **30**, 1–12 (2017).

41. Gert, B. *Common Morality: Deciding What to Do*. (Oxford University Press, 2004).

42. Janoff-Bulman, R., Sheikh, S. & Hepp, S. Proscriptive Versus Prescriptive Morality: Two Faces of Moral Regulation. *J. Pers. Soc. Psychol.* **96**, 521–537 (2009).

43. Berridge, K. C., Robinson, T. E. & Aldridge, J. W. Dissecting components of reward:

‘liking’, ‘wanting’, and learning. *Curr. Opin. Pharmacol.* **9**, 65–73 (2009).

44. Berridge, K. C. Food Reward: Brain Substrates of Wanting and Liking KENT. *Neurosci. Biobehav. Rev.* **20**, 1–25 (1996).

45. Gilead, A. How is Akrasia Possible After All? *Ratio* **12**, 257–270 (1999).

46. Polman, E. & Ruttan, R. L. Effects of anger, guilt, and envy on moral hypocrisy. *Personal. Soc. Psychol. Bull.* **38**, 129–139 (2012).

47. Bruneau, E. G., Kteily, N. S. & Urbiola, A. A collective blame hypocrisy intervention enduringly reduces hostility towards Muslims. *Nat. Hum. Behav.* **4**, 45–54 (2020).

48. Warach, B., Josephs, L. & Gorman, B. S. Are Cheaters Sexual Hypocrites? Sexual Hypocrisy, the Self-Serving Bias, and Personality Style. *Personal. Soc. Psychol. Bull.* **45**, 1499–1511 (2019).

49. Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H. & Strongman, J. A. Moral hypocrisy: appearing moral to oneself without being so. *J. Pers. Soc. Psychol.* **77**, 525 (1999).

50. Carpenter, T. P. & Marshall, M. A. An examination of religious priming and intrinsic Religious motivation in the moral hypocrisy paradigm. *J. Sci. Study Relig.* **48**, 386–393 (2009).

51. Lammers, J., Stapel, D. A. & Galinsky, A. D. Power increases hypocrisy: Moralizing in reasoning, immorality in behavior. *Psychol. Sci.* **21**, 737–744 (2010).

52. Crockett, M. J. *et al.* Dissociable Effects of Serotonin and Dopamine on the Valuation of Harm in Moral Decision Making. *Curr. Biol.* **25**, 1852–1859 (2015).