Utilitas (2025), **37**, 123–140
doi:10.1017/S0953820825000044

CAMBRIDGE UNIVERSITY PRESS

ARTICLES

The Function of Hypocrisy Norms

Matthew Jeffers[1] and Alexander Schaefer[2]

[1]Independent Scholar, Hartford, CT, USA and and [2]Department of Philosophy, SUNY at Buffalo, Buffalo, NY, USA
Corresponding author: Alexander Schaefer; Email: Schaef@buffalo.edu

Abstract
Moral condemnation of hypocrisy is both ubiquitous and peculiar. Its incessant focus on word–action consistency gives rise to two properties that distinguish it from other types of moral judgment: non-additivity and content independence. Non-additivity refers to the fact that, in judgments of hypocrisy, good words do not offset bad actions, nor do good actions offset bad words. Content independence refers to the fact that we condemn hypocrisy regardless of whether we would condemn the words or actions in isolation from one another. To make sense of these peculiar properties, we present a costly signaling model of social cooperation, in which hypocrisy norms allow a separating equilibrium to emerge, thus facilitating reliable communication and higher levels of social trust. We compare our *functionalist* account of hypocrisy to other philosophical accounts, arguing that a functionalist analysis better illuminates our moral practices and public discourse.

Keywords: hypocrisy; signaling; moral theory; game theory; functionalism

1. Introduction

Politicians and other public figures are often censured for failing to live up to their avowed moral commitments. Al Gore, for instance, continues to receive criticism for the alleged inconsistency between his environmental activism and his luxurious lifestyle.[1] Similarly, Joszef Szar, a prominent figure in Hungary's anti-LGBTQ Fidesz party, faced accusations of hypocrisy after being caught participating in a 25-man orgy. Szar was ultimately forced to resign as a member of the European Parliament.

The political ramifications of these episodes, together with the sensitive nature of the topics involved, tend to obscure the peculiarity of the discourse surrounding them. In the case of Al Gore, criticisms originated largely from right-wing commentators who, in normal circumstances, see no problem with owning large homes or traveling on jets. In the case of Joszef Szar, the most searing criticism came from leftists who generally believe that sexual acts between consenting adults, including gay orgies, are morally permissible. Apparently, when it comes to hypocrisy, it is not the actions themselves that bother us, but rather the inconsistency between our words and our actions. Few other moral judgments seem to exhibit this obsession with formal consistency.

[1]In 2017, an article was published with the title "Al Gore's Climate Change Hypocrisy Is As Big As His Energy-Sucking Mansion."

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

As these examples demonstrate, judgments of hypocrisy are ubiquitous in moral and political life. Moreover, as we explain in section 2, judgments of hypocrisy exhibit peculiar features. For instance, in contrast to typical moral judgments, we often condemn hypocritical behavior regardless of whether it causes any real or potential harm. The pattern of judgments that surrounds hypocrisy, we call this pattern *the hypocrisy norm*, merits special attention due to its ubiquity and peculiarity.

Notable contributions to our understanding of hypocrisy come from both philosophy and psychology. Philosophical investigations into hypocrisy have tended to focus specifically on hypocritical blame, that is, cases in which an agent blames another for violating some principle or standard that the blaming agent herself fails to act in accordance with (Cohen 2006; Fritz and Miller 2018; Isserow 2022; Piovarchy 2023a; Rossi 2018; Wallace 2010). Much of the psychological literature on hypocrisy also focuses on our blaming behavior.[2]

Our approach differs from most philosophical accounts by offering a *functionalist* theory, which, we argue, better explains hypocrisy's peculiar features. More specifically, our account aligns with recent attempts to understand objections to hypocrisy in terms of a societal need to maintain reliable moral signaling (Piovarchy forthcoming; Shoemaker and Vargas 2021). There is very little psychological research that explains the functional role that hypocrisy might play or why the development of the hypocrisy norm is a valuable social mechanism. A notable exception is a paper by Jordan et al. (2017), where the researchers investigate the signaling function of moral discourse and propose that hypocrisy norms are a reaction to false signaling. The functionalist account we develop here takes inspiration from this latter approach, and seeks to deepen our understanding of the relationship between signaling and hypocrisy norms.

To advance a functionalist understanding of the hypocrisy norm, we develop a formal model that helps to elucidate the strategic logic and causal mechanism involved in our moral practices surrounding hypocrisy. By identifying a crucial function of the hypocrisy norm, this model provides an explanation of certain intriguing features of our moral reactions to hypocrisy. In addition, we think the proposed explanation better supports the moral judgments revealed in the empirical studies undertaken by Jordan et al. (2017) than do most other philosophical accounts.

Before discussing the model, we describe what we call *the two puzzles of hypocrisy*. Following this clarifying section, we lay out a game-theoretic model that explains the puzzling features of the hypocrisy norm by casting it as a response to the problem of generating trust amid the possibility of deceitful and opportunistic communication (section 3). In essence, our claim is that hypocrisy norms can be explained as a strategy for raising the cost of sending false signals about one's commitment to a social norm. After explaining how this model offers a plausible account of the function of hypocrisy norms, we show, in section 4, that this account resolves the two puzzles of hypocrisy identified earlier. Finally, in section 5, we compare and contrast our approach with other philosophers' attempts to analyze hypocrisy, highlighting the benefits of functionalist moral theory for understanding the hypocrisy norm. We conclude by considering how our analysis bears on commons uses and abuses of moral discourse surrounding hypocrisy.

[2]For example, Batson et al. (1999), Batson et al. (2002), and Valdesolo and DeSteno (2007) focus on explaining why people fail to judge their own immoral actions as seriously as the immorality of others. Kreps et al. (2017) and Effron et al. (2018) focus on determining when inconsistency between words and actions is interpreted as hypocrisy rather than a legitimate change of opinion.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

2. The two puzzles of hypocrisy

Hypocrisy refers to a type of inconsistency between one's avowed moral commitments and one's own behavior. The revelation of this inconsistency generates strong moralized responses, even when the behavior is harmless. Although somewhat peculiar, this feature of hypocrisy is not its most puzzling. In this section, we describe two puzzles of hypocrisy. That is, two ways in which moral evaluations of hypocritical behavior deviate from what we call *normal moral situations*.

In a normal moral situation, holding all else equal, doing multiple bad things is worse than doing one bad thing. Similarly, multiple good actions are better than a single good action. This basic principle applies to physical actions, such as helping one's neighbor or shoplifting, as well as speech acts, such as praising human rights or catcalling strangers.[3] Moreover, good words and bad actions, or bad words and good actions, are preferable, from a moral point of view, to bad words combined with bad actions (O'Brien and Whelan 2022: 1706). To praise human rights is good, even if yesterday I shoplifted from the corner store. One could easily say, "well, yes, he did shoplift but at least he's a strong defender of human rights." Call this feature of normal moral situations the *additive property*.

A second feature that characterizes normal moral situations is that they are *content sensitive*, that is, their moral status depends on the actual actions or words themselves, or the sentiment behind them, rather than any formal relations that hold between a set of various words or actions. One could *not* plausibly say, "well, yes, she did shoplift, but as a credit to her moral consistency she also defrauded an elderly couple, which seems fitting for someone that would go around stealing from stores." Nor could one plausibly say, "while it might seem good that she donated money to charity, that action actually counts against her, because it proves she's inconsistent – after all, you know how often she defrauds the elderly." The behavioral inconsistency does not seem relevant to the rightness or wrongness of the actions. What typically matters in evaluating actions or words is their content, not their alleged consistency or inconsistency with other actions or words.

To express these features more precisely, consider a functional form that represents the moral evaluation of a certain *word–action pair*, that is, a certain speech act performed in conjunction with a certain physical action.[4] Let M be the overall moral evaluation of the pair, let w represent the word component, a the action component, and $m_w(w)$ and $m_a(a)$ the moral evaluation of the word and action components respectively. A normal moral situation would present us with a moral evaluation function of the following form:

$$M(w, a) = m_w(w) + m_a(a)$$

[3]Although we will frequently refer to "words," "pronouncements," and "speech acts," the kind of moral signaling we are concerned with extends beyond verbal communication. Moral signaling can be broadly interpreted and could plausibly include other forms of communication aside from speech, such as public writings or even making facial expressions in certain contexts. We thank an anonymous referee for asking us to clarify this.

[4]Turner (1990: 265) calls these "disparity pairs." Cf. Rossi (2018: 58–60). Note that, given our interest in social trust, our account focuses on the consistency between words and actions, not words and (non-manifested) beliefs. Some accounts explicitly acknowledge a kind of *doxastic hypocrisy*, in which one's words and actions are compatible, but both violate one's private beliefs (Kittay 1982).

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

The function *M* is additive, since it results from simply adding up the moral evaluations of the word and the action. This function is also content sensitive, because it depends upon the actual word, *w*, and action, *a*, that comprise the components of the word–action pair (*w*, *a*).

Moral judgments pertaining to hypocrisy defy this functional form. When we criticize others for hypocrisy it is often the case that we have no problem with their words or actions, taken separately. When combined, however, they trigger a negative moral judgment. One might, for example, think that eating meat is a morally acceptable practice, and also think that veganism is a praiseworthy lifestyle. Nonetheless, it is difficult to resist condemning someone who endorses a vegan diet, yet continues to eat meat. In this case, espousing vegan beliefs, far from making up for eating meat, seems to add an aspect of wrongness or objectionableness to this behavior.[5] This same example shows that judgments of hypocrisy are insensitive to the content. The moral agent in question has no issue with either veganism or meat-eating. Yet, the formal inconsistency between championing a vegan diet while eating meat strikes her as morally suspect. In other words, the hypocrisy norm exhibits the following two puzzling features:

Non-additivity: The wrongfulness or objectionableness of a hypocritical word–action pair is not a sum of the wrongfulness or objectionableness of the elements of that pair.
Content independence: To condemn a hypocritical word–action pair is not to condemn the words or actions themselves, nor the sentiment behind them; rather, it is to condemn the formal inconsistency between the words and the actions.

The task at hand, then, is to find an account of hypocrisy that will explain this strange divergence from normal patterns of moral evaluation. An adequate explanation will illuminate, or at least be consistent with, both non-additivity and content independence. The model we present below sketches out such an explanation.

3. A costly signaling model of hypocrisy

At least since Hobbes, political theorists have recognized the crucial importance of trust in facilitating social cooperation.[6] Indeed, according to some accounts, many of our norms are best understood as adaptive responses to the problem of *assurance* – that is, the problem of knowing that others will reciprocate trusting, cooperative behavior, rather than acting in an opportunistic and self-serving fashion (Thrasher and Handfield 2018; Ullmann-Margalit 2015). The program of studying norms as instances of adaptive strategies has illuminated several otherwise puzzling features of human behavior (Kameda et al. 2005).

We propose that, like many other norms, the hypocrisy norm can be analyzed as an adaptive strategy. That is, a strategic analysis highlighting issues of trust and communication illuminates various features of the hypocrisy norm.

To undertake this analysis, we start with three basic observations. First of all, hypocrisy has to do with discrepancies between what one communicates and what one is actually committed to and will likely follow through on. This is, in effect, a problem of trusting another person's moral sincerity. Second, this problem of trust depends

[5]Hypocrisy can be objectionable without being wrong, such as when one athlete criticizes another for dropping the ball, even if they are guilty of the same thing (Todd 2023).

[6]For a more recent treatment of trust, see Vallier (2020).

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

upon *asymmetric information*: I am familiar with my own moral commitments, but unsure about yours, and vice versa. Finally, the problem of asymmetric information generates a problem of credibility. It is easy for non-cooperators to deceitfully claim to be morally upstanding agents. Hence, we need to establish some way of credibly signaling our true intentions.

In the game theory literature on credible communication under conditions of asymmetric information, the most elucidating model is one that involves *costly signaling*. This section explains the intuitive structure of a costly signaling game, before turning to a more rigorous, mathematical exposition of the same model. Finally, it addresses some objections and limitations of this model for explaining features of the hypocrisy norm.

3.1 Informal explanation of the model

When interacting with other people, we often do not know their preferences or what we will call their "type." In other words, we do not know exactly which game we are playing. Is our co-player honestly committed to their espoused morality or is she a false signaler seeking to take advantage of us? If players can make an educated guess about the nature of their opponents, then we can model this uncertainty as a game of incomplete information. With some probability p we are dealing with someone who flouts our norms, while with some probability $1-p$ we are dealing with someone who shares our norms.

For ease of analysis, we will model a particular example, but one which is analogous to many scenarios in which the hypocrisy norm comes to bear. Our simple model involves two players. Imagine we are part of a vegan organization looking to hire a new employee. We would prefer to hire a morally committed vegan, but it is a nice cushy job, and we are worried that some applicants might attempt to deceive us about their moral and dietary inclinations. In the model developed here, we interpret player 1 as a prospective employee. She comes in two possible types: τ_V, a true moral vegan, and τ_M, a meat-loving omnivore. To add in communication, the "word" portion of the word–action pair, we allow player 1 to send a signal that she is an authentic vegan. This may occur before the interview, perhaps in her social media posts, or during the interview, where player 2, the vegan interviewer, will ask player 1 about her moral views. Regardless of her actual commitments, player 1 can tell the interviewer that she believes in meat eating (i.e., send a signal M) or veganism (i.e., send a signal V). Both types of player 1 therefore have two available actions, $\{V, M\}$, where V stands for *tell the interviewer I am a moral vegan* and M stands for *tell the interviewer I believe that eating meat is morally acceptable*.

Player 2, who receives this signal from player 1, must decide whether or not to cooperate with her, that is, to hire her. Accordingly, his pure strategy set is $\{C, N\}$, for "Cooperate" and "Not Cooperate." Player 2 would benefit from interacting with a vegan type, but would rather not hire the meat-eating type. Hence, he will choose to cooperate only if his belief that player 1 is a vegan, μ_2, is sufficiently strong.

But how could a signal from player 1 provide any information to player 2 about her intention to communicate? After all, both types have the option to send the vegan signal V, which means that such a signal will provide no information whatsoever about player 1's type. Player 2 is left in the dark, since both honest and deceitful types have *pooled* into the same strategy choice. In game-theoretical terms, this is referred to as a *pooling equilibrium*. What player 2 needs is some way of *separating* the two types, so that one type chooses to send one signal, while the other type chooses the other signal.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

A primary mechanism for increasing the integrity of communication systems – a mechanism present in phenomena as diverse as job searches (Spence 1978) and prey–predator interactions (Zahavi 1977) – is known as *costly signaling*. If, somehow, the vegan type can send a vegan signal that is costlier for the meat-eating type, then perhaps only the vegan type can rationally send that signal. In this case, the disparities in the costliness of sending V allow for what game theorists call a *separating equilibrium*, rather than a pooling equilibrium. This separating equilibrium will be further stabilized if the vegan type faces higher costs than the omnivore when sending a meat-eating signal, M.

As an explanation of the hypocrisy norm, we propose that it functions so as to raise the costs of sending a false signal. In other words, it adds a cost that asymmetrically falls upon deceitful signalers.[7] Intuitively, the hypocrisy norm seems to provide exactly this sort of mechanism. The strong moral revulsion and the desire to punish hypocrites creates high expected costs for would-be deceivers. Those disposed to violate the principles they extol know that, if caught, they will face social censure. They must therefore choose to risk such censure or to outwardly comply with moral principles that they secretly loathe. Of course, not all hypocrites are detected, but if the punishment is severe enough, it can deter hypocrisy even if there is a low probability of being detected.[8] This prospect of punishment adds an additional cost to the disingenuous hypocrite that is not borne by the honest cooperator. The honest signaler has no intention of violating the principle that she extols, so the probability of being punished for doing so is basically 0.[9] If these asymmetric costs are sufficiently high, the hypocrisy norm can transform a low-trust pooling equilibrium with dysfunctional communication into a high-trust separating equilibrium with reliable communication.[10]

While this intuitive account of costly signaling lays bare its mechanism and how it operates, a more rigorous model can provide greater precision.

3.2 Formal explanation of the model

Before examining a game tree and deriving the equilibrium solutions, let us precisely define our notation. The signaling game consists of:

- Three players: $i \in P = \{1, 2, \text{Nature}\}$
  - Player 1 is called the "sender," player 2 the "receiver," and "Nature" randomly determines the sender's type.
- A set of types: $T = \{\tau_V, \tau_M\}$
  - Each of these types is distinguished by its distinct utility function. The vegan type, τ_V, sincerely abhors the consumption of animal products, while the meat-eating type, τ_M, loves eating animal products.

[7]Separation is most likely when the cost is applied to *both* types of false signaler, since this helps to ensure reliable communication. This does not mean, however, that the receiver does not have a preference as to which type he would like to interact with. The vegan employer would prefer a hypocritical vegan employee over a non-hypocritical omnivore, but the vegan employer will still apply the judgment of hypocrisy to the hypocritical vegan, but not to the non-hypocritical omnivore. The judgment *of hypocrisy* is therefore content independent, even when other moral judgments are not.

[8]For those in the public spotlight, e.g., politicians, the probability of being caught is much higher.

[9]It would be exactly 0, except that sometimes our intentions change or we make mistakes. And sometimes we are falsely accused of violating rules.

[10]For a review of the importance of social trust, see Vallier(2018).

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

  - Both types seek a cooperative response from player 2, whether or not they are committed to the vegan norm. Both receive a baseline payoff of 0 when they do not manage to elicit cooperative behavior from the other player.
  - In some treatments, each type is considered a different player, so that Nature decides who the receiver is playing against. For ease of expression, we treat them as the same player with different utility functions at each information set.
- A set of actions A_i for each player $i \in P$.
- A utility function for each player/type (excluding Nature): u_1, u_2, mapping from $A_N \times A_1 \times A_2$ to the real numbers $\mathbb{R}$.

Nature, player 1, and player 2 interact in a game that unfolds as follows:

(1) Nature chooses player 1's type according to some commonly known probability distribution; player 1 knows this type, but player 2 does not.
(2) Player 1 chooses an action (her signal), following which player 2 updates his (probabilistic) beliefs about his opponent's type, that is, the probability that player 1 is type τ_V or τ_M.
(3) With these new beliefs, player 2 chooses his best response. (Player 1, knowing that player 2 is a rational agent, foresaw how her actions would affect his choice and chose her own actions accordingly, viz., so as to maximize her own utility.)

To understand the problem for which the hypocrisy norm provides a solution, first consider a game in which there is no communication cost to sending a signal *V* or *M* for either type. The worst a player can do is 0, while the maximum payoff is 1. Player 1 achieves a payoff of 1 in two scenarios: (1) she is a vegan type τ_V and succeeds in cooperating with player 2, or (2) she is a meat-eating type τ_M and succeeds in cooperating with (or, perhaps we should call it "exploiting") player 2. Player 1 receives 0 whenever player 2 withholds cooperation. Player 2, on the other hand, receives 1 when he succeeds in cooperating with a vegan-type player 1, 0 when he attempts to cooperate with a meat-eating-type player 1, and some intermediate payoff*a*, where $0 < a < 1$, whenever he withholds cooperation by deciding not to hire the prospective employee. The resulting game tree is presented in Figure 1.

For the game presented in Figure 1, there are no separating equilibria.[11] Semi-separating equilibria are possible only when the situation is so hopeless that player 2 will always choose to not cooperate (*N*).[12] In all other cases, only pooling equilibria will exist.[13] Because there is no cost to communication, dishonest players will exploit any cooperative behavior on the part of player 2, making cooperation (*C*) a risky strategy for player 2. Cooperation could, nevertheless, become rational, even without trust. This will happen when *p* is quite high (i.e., there are very few individuals who are not true vegan types) or when *a* is quite low (i.e., when failing to cooperate presents a major loss for player 2). In the case of high *p* and low *a*, player 2 will cooperate but his cooperation will be exploited. When *p* is low and *a* is high, cooperation will not occur at all. In both cases, player 2 could get a higher payoff if only he could separate the true vegans from the omnivores and cooperate with the former.

[11]See Appendix, Proposition A1 for a formal proof of this claim. The appendix is available at https://www.amschaefer.com/s/Appendix_Hypocrisy.pdf.

[12]See Appendix, Proposition A2 for a formal proof of this claim.

[13]See Appendix, Proposition A3 for a formal proof of this claim.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

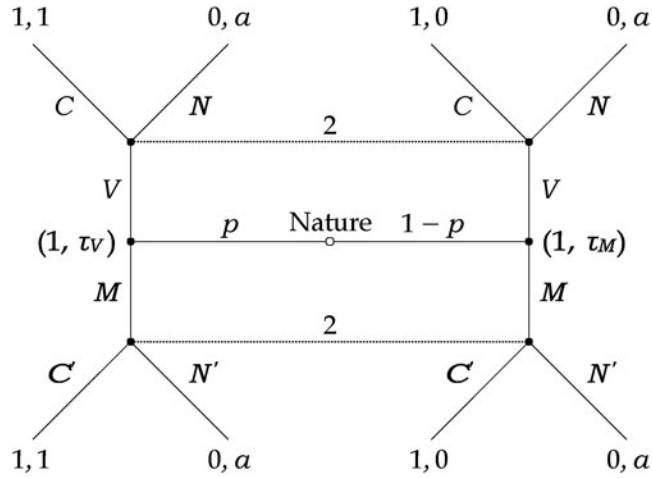

Figure 1. Signaling game with costless communication, $0 < a < 1$.

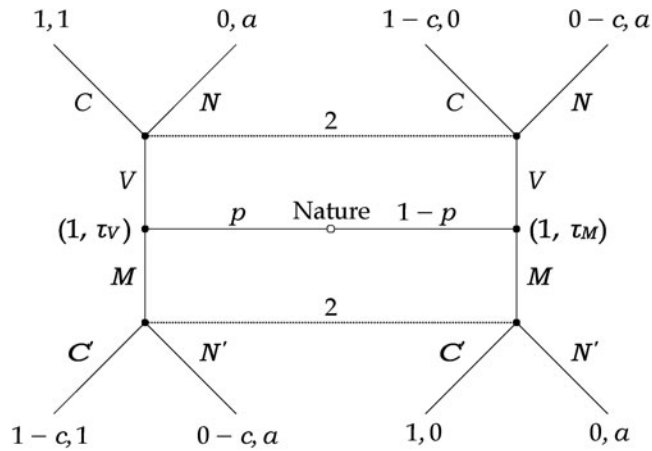

Figure 2. Signaling game with asymmetric communication cost, $0 < a < 1$, $c > 0$.

In other words, the problem of false signaling undermines the trust that might otherwise exist between interacting individuals. Yet, we observe that such trust does sometimes exist. We must, therefore, ask how society or biology might evolve so as to achieve a solution. The most straightforward way is to introduce a cost, c, to deceitful signaling (Figure 2).

By introducing a cost for inconsistency between one's avowed principles and one's behavior, can this new game mitigate the problem of hypocrisy? The answer to this question, it turns out, depends critically upon the value taken by the parameter c. If c is too low, specifically if $c < 1$, then any cooperative equilibrium will be a pooling one, replete with false signaling and hypocrisy.[14] Essentially, this means that imposing

[14]See Appendix, Proposition B1.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

a cost on false signaling fails to resolve the problem of social trust, even when player 2 is willing to cooperate with player 1. Moreover, this willingness to cooperate is contingent on one or more of the following assumptions being satisfied: (1) the probability of encountering a truly committed type, p, is high, meaning that player 2 is unlikely to be interacting with an M-type, and (2) non-cooperative outcomes are very unattractive in that the payoff to player 2 of choosing not to cooperate, a, is very low. Taken together, (1) and (2) mean that player 2 is willing to take the risk of cooperating, even if exploitation is possible, because either this risk is minimal or the gains to cooperation are large. When neither (1) nor (2) are satisfied, player 2 will not take the risk and cooperation will not occur.

In more promising scenarios, player 2 need not take a risk to cooperate; for a range of parameter c values, there exists a separating equilibrium. In fact, for $c > 1$, a separating equilibrium is the *only* possible equilibrium outcome.[15] The introduction of an asymmetric signaling cost, if sufficiently high, forces both types to identify themselves and thereby allows player 2 to distinguish between norm-adherents and non-adherents. As described above, this cost applies *asymmetrically* to hypocrites because they are the ones who risk being discovered and punished for their insincerity. At the same time, it applies *symmetrically* to both types of hypocrite, a type τ_M who sends a V signal and a type τ_V who sends an M signal. By generating a sufficient threat of punishment for disingenuous communication, a society can thereby establish a separating equilibrium, producing a much higher level of trust than in the costless communication model.

Unlike the costless model of Figure 1, this game also allows for the possibility of cooperative semi-separating equilibria.[16] A semi-separating equilibrium is one in which the signals sent are *partially*, though not fully, informative, because players of one type are more likely to send a particular signal than players of the other type. For example, if all vegan types sent a V signal and non-vegan types sent a mix of V and M signals, then the corresponding equilibrium would be a semi-separating one. These various results are summarized in Table 1.

Given the crucial importance of the value of c, we face the following question: how do we know whether c is high enough to deter hypocrisy in the real world?[17]

The phrasing of this question may be a bit misleading. Even assuming the same objective punishment, the subjective cost c will vary from individual to individual. Some individuals care more about their reputations and social ties, hence they are highly sensitive to any form of blame or outrage. Others care much less about such things and therefore face a lower c value. Consequently, for any reasonable punishment imposed, there will be a subset of the population for whom the subjective cost of this punishment suffices to deter them – for this subset, the subjective cost is $c > 1$ – while for others, it does not suffice – their subjective cost is $c < 1$. So, for a given level of punishment, some individuals will fall into the high-cost range, which marks a major improvement over the no-cost game of Figure 1. It is worth noting that empirical studies have found that our moral reactions to hypocrisy tend to be more severe than our moral reactions to other forms of deceit (Jordan et al. 2017). This suggests that the hypocrisy norm imposes a non-trivial reputational cost on hypocrites and may, therefore,

[15]See Appendix, Proposition B2.

[16]In fact, at the threshold value $c = 1$, pooling, separating, and semi-separating are all possible outcomes. See Appendix, Proposition B3.

[17]We thank an anonymous referee for pressing us to address this question.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

Table 1. Summary of cooperative equilibria

Value of c	Types of possible cooperative equilibria	Reliable signals?
$c < 1$	Pooling (cooperation only if $p > a$)	No
$c = 1$	Pooling, separating, or semi-separating	Determined by equilibrium type
$c > 1$	Separating	Yes

deter a sizeable subset of would-be hypocrites by ensuring that they face a sufficiently high value of c.

This analysis raises the further question: why not raise c to extremely high levels so as to deter *all* hypocrisy? The most likely answer is that punishment and blame can be costly (Shoemaker and Vargas 2021), and the cost increases with the severity of the blame or punishment. The socially optimal value of c, therefore, would take these costs into account and deter hypocrisy only up to the efficient level.[18] The ideal hypocrisy norm would optimize c so as to deter as many would-be hypocrites as efficiency allows, and no more.

For those that find themselves at the threshold value, $c = 1$, there remains the question of which equilibrium they will select.[19] One way to address this question would be through a dynamical analysis that would allow us to determine the basin of attraction for each potential equilibrium outcome.[20] Although potentially illuminating, such an analysis lies beyond the scope of this paper and is unnecessary for drawing the core lesson of the model analyzed here. That lesson is comparative in nature: a moral community with a hypocrisy norm is more likely to support trustworthy signaling and social trust among some subset of its members, and it will foster more trust and cooperation than one that does not have such a norm. The most realistic scenario is that some would-be hypocrites are reliably deterred, while others persist in their hypocritical ways. This scenario is, at least, the one we observe in the real world.

3.3 Clarifications and objections

The costly signaling model laid out above aims to clarify a potential social *function* of the hypocrisy norm. This norm provides a mechanism for increasing trust within a social network. In future work this conjecture should be subjected to empirical tests, but the current task is merely to establish its theoretical plausibility. To this end, there are two important objections that must be addressed.

First, if the hypocrisy norm functions to increase trust within one's group of interaction, why does it frequently appear to be directed at out-group members? Recall the two examples that opened this paper: Al Gore is criticized as a hypocrite, not by leftists, but by those on the right. Similarly, Jozef Szar faces harsh attacks primarily from leftists, not from his right-wing compatriots. The hypocrisy norm seems to be primarily directed at discrediting out-group members, rather than ensuring trust within the in-group.

[18]The background model here is similar to those deployed in the field of law and economics to analyze efficient deterrence, the basic idea being that the marginal benefit of deterrence should be set equal to its marginal cost. See Cooter and Ulen (2012: Ch. 12) for an overview.

[19]We thank an anonymous referee for pressing us to address this question.

[20]For an example of this approach applied to other kinds of signaling games, see De Silva and Sigmund (2024). See also Appendix, Proposition B3, where we show that the existence of the semi-separating equilibrium depends upon the ratio p/a.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

The impression that the hypocrisy norm is primarily directed at out-group members may be an illusion of our current cultural context. Politicians (along with celebrities) are some of the most scrutinized individuals in society. Consequently, their hypocritical acts are the most likely to be detected. In a cut-throat political environment, there are strong incentives to seize upon any opportunity to discredit one's political opponents or minimize the transgressions of one's political allies. To know whether the hypocrisy norm really is directed primarily at out-group members, we would need to examine its non-public, non-political applications. It is compatible with the costly signaling model that, in the context of political discourse, the hypocrisy norm is hijacked as a tool for political gain.

Yet, suppose that when examining non-public, non-political cases, we still find that the hypocrisy norm is directed at out-group members. Would this discredit the trust-establishing hypothesis? We think not. Nothing in our theory supposes that the hypocrisy norm applies equally to in-group and out-group members. In fact, for in-group members, the problem of trust is less pronounced. They can rely on mechanisms such as conformity bias and reciprocity, which lighten the pressure placed on disciplinary measures (Henrich and Boyd 2001; Trivers 1971). It may be that the hypocrisy norm arose primarily as a means of disciplining out-group members with whom one's in-group wishes to participate. In effect, the hypocrisy norm would raise the cost of deceitful signaling for out-group members by tarnishing the reputation of would-be deceitful exploiters along with their associates. Detectable hypocrisy will lead to bridge-burning and hostility, effectively undermining any mutually beneficial relationship that would otherwise have existed. Hence, whether the hypocrisy norm applies to in-group members, out-group members, or both, the explanatory model can shed light their function as trust-establishing mechanisms.

Another objection could ask how our account deals with "cancellation," that is, a hypocrite who admits that she does not live up to her avowed moral commitments.[21] Even though both kinds of hypocrite send a false signal, we tend to judge "honest hypocrites" much more gently than those who do not admit to their hypocrisy. This is not mere philosophical intuition. In one study, Jordan et al. (2017) asked respondents to evaluate traditional hypocrites, honest hypocrites, and control transgressors, that is, those who violate the moral principle without expressing any commitment to it. "Honest hypocrites," defined as those who avow a certain moral standard but also disclose that they sometimes fail to uphold that moral standard, fared much better than traditional hypocrites. In fact, subjects evaluated their moral goodness to be no worse than control transgressors. The authors conclude that "disclosure negates the false signal" of hypocrisy (Jordan et al. 2017). According to this objection, these findings raise an issue for our theory: if judgments of hypocrisy are about false signaling, and both the honest and the traditional hypocrite falsely signal, then our judgments of each should be somewhat similar.

Although our model does not explicitly consider the case of an honest hypocrite who partially or fully cancels her moral signal, it is fully consistent with this phenomenon. In fact, our model captures the fundamental reason why we let honest hypocrites off the hook: such signalers cancel out their signal in such a way that it poses no risk to the integrity of the system of moral signaling. Those disclosing their hypocrisy upfront are not attempting to send a false signal, so there is no reason to apply a harsh judgment to them. Far from generating a problem, the findings of Jordan et al. (2017) actually

[21]We thank an anonymous referee for pressing us to address this objection.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

corroborate our model's explanation of the hypocrisy norm, which is that it functions to create a separating equilibrium and thereby support honest communication.

4. The puzzles revisited

Recall the two puzzles that the hypocrisy norm poses: the additivity puzzle and the content independence puzzle. In the case of hypocrisy, we do not evaluate the morality of each separate activity and add them up to determine the degree of rightness or wrongness, instead we evaluate the level of inconsistency between the two activities. Moreover, we judge this inconsistency to be objectionable quite apart from the content of the words or actions that make up the word–action pair.

These two puzzles make perfect sense in light of the model developed in section 3. First, suppose Mario is an employer who calls out and condemns racist attitudes in society at large, but also exhibits racist bigotry in his hiring choices. The moral appropriateness of his words does not counterbalance or even offset the immorality of his actions; instead, his words, however laudable on their own, compound the wrongness of his actions. If we accepted additivity, that is, allowing Mario's words to offset his actions, we would fail to support a separating equilibrium that allows us to distinguish between norm violators and honest norm adherents. Social trust would deteriorate, because moral speech would become empty and uninformative.

Second, imagine Avital frequently extols the benefits of healthy living and the harms of smoking, including the harm posed to others. At the same time, Avital is an insatiable chain-smoker. Smokers and non-smokers alike will object to Avital's hypocrisy, regardless of their attitudes toward smoking. The outrage over Avital's hypocrisy stems largely from the fact that Avital falsely signals her adherence to non-smoking norms. She thus becomes less trustworthy in light of her failure to uphold her pronounced moral commitments. If content were all that mattered, then many, perhaps most, would not condemn Avital, since they have no problem with her anti-smoking words or with her smoking behavior. Consequently, they would have no way of deterring "false anti-smokers," that is, smokers who want credit for their purported capacity to resist smoking. The content insensitivity required to punish Avital's hypocritical inconsistency is thus crucial to obtaining the separating equilibrium described in section 3.

The signaling model shows that when the hypocrisy norm is in play, false signalers face a threat of social censure if their words and actions are inconsistent. The hypocrisy norm functions to punish inconsistencies between avowed beliefs and actions. Consequently, this decreases the probability that people will endorse values that they do not uphold and transforms opaque and unreliable platitudes into transparent signals about a person's true commitments. The hypocrisy norm flaunts additivity and exhibits content independence because this pattern of judgment functions to ensure reliable moral communication. The hypocrisy punishes false signalers, undermining the possibility of cheap talk and thereby establishing accurate expectations regarding the behavior of others.[22]

[22]This understanding of hypocrisy has the added benefit that it clarifies why harsh judgments of hypocrisy can occur independently of any identifiable, tangible harms. For instance, individuals who are entirely non-credible will be condemned for hypocrisy, even though there was never a real risk of them being believed. The reason is that such judgments are required to maintain the overall integrity of moral signaling, not to prevent any particular instance of hypocrisy.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

5. Advantages of the functionalist account

In this section, we connect our theory to existing studies of hypocrisy and highlight its strengths in comparison to these other accounts. We focus, in particular, on two competing types of account – moral equality accounts and moral commitment accounts (Piovarchy forthcoming) – neither of which offers a fully adequate theory of hypocrisy.[23] By focusing on the *function* of the hypocrisy norm, our theory offers a more general account of its characteristic features.

According to moral equality accounts, hypocritical blame is objectionable because the hypocrite unfairly blames others while exempting himself from blame for the same behaviors (Fritz and Miller 2019; Wallace 2010). By unfairly applying moral standards, the hypocrite implicitly denies the moral equality of persons (Fritz and Miller 2018). When we fail to acknowledge the moral equality of persons, we lose our moral standing to blame others for failing to live up to moral standards, even when the moral standards that we apply are, in themselves, correct or unobjectionable.

Some have pointed out that moral equality accounts have counter-intuitive implications and contradict our moral practices. For instance, Todd (2019: 372) observes that our moral reaction to merely *inconsistent* blamers is quite different from our moral reaction to *hypocritical* blamers, even though both apply moral standards unfairly. Another issue with the moral equality approach concerns its emphasis on the relationship between the hypocrite and the person whom the hypocrite is blaming. Wallace (2010) attributes hypocrisy's wrongfulness to the hypocrite's unequal treatment of himself compared to those he blames. Since "we all have an interest in being protected from the kind of social disapproval and opprobrium that are involved in blame" (Wallace 2010: 328), the hypocrite wrongfully heaps blame on others but not himself, thus treating his own interests as more important than those of others. If this way of cashing out moral equality accounts is correct, then the wrongness of hypocrisy lies in the fact that the *blamee* is treated unequally by the *blamer*.[24] However, in real-world moral reactions to hypocrisy, remarkably little attention is given to the hypocrite's "victim," that is, those whom the hypocrite criticizes. There does not seem to be any general practice of having the hypocrite apologize to those they blame. When hypocrites apologize, they typically apologize for failing to live up to their avowed moral standard; they do not apologize for blaming others who also did not live up to the standard. Finally, if the wrongness of hypocritical blame is rooted in a commitment to equality, why does the presence of a hypocrisy norm transcend cultures across time and place, including cultures where equality is not a prevailing moral value? The ubiquity of the hypocrisy norm suggests the presence of some alternative principle or explanation. In sum, the moral equality account fails to accord with the judgments and social practices surrounding hypocrisy.

Accounts of the second type, viz. "commitment accounts," assert that hypocrisy is objectionable because it indicates a lack of commitment to the moral standards that the hypocrite applies or claims to adhere to (Crisp and Cowton 1994; Friedman 2013; Piovarchy 2023a, 2023b; Rossi 2018; Todd 2019). Such accounts seem to better describe our judgments and practices than do moral equality accounts. Not only do

[23]There are other accounts of hypocrisy that fit into neither category, notably Isserow and Klein (2017), who develop an account based on the notion of moral authority.

[24]Fritz and Miller (2019) try to evade this problem by arguing that the wrongfulness of hypocritical blame does not lie in its failure to give others equal treatment but instead lies in the hypocrite having an unfair differential blaming *disposition* (Fritz and Miller 2019: 551).

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

such accounts satisfy content-independence and non-additivity, they also deal well with "subjunctive hypocrisy" (Isserow 2022), or cases in which the hypocrite blames someone for doing something wrong, even though the hypocrite *would have* behaved similarly if placed in the same circumstances. We judge such instances to be objectionable, it seems, because we judge the hypocrite, not for actually behaving in ways that contradict his moral pronouncements, but for lacking a commitment to the values and standards that underlie these pronouncements.

Nevertheless, as Todd (2019: 372) noted, commitment accounts do not offer a complete account of hypocrisy because they fail to clarify *why* we object to those who make moral pronouncements while lacking a true commitment to the values and standards underlying those moral pronouncements. In many cases, these pronouncements embody correct or unobjectionable values and standards, and in many cases these pronouncements seem to be harmless. *What reason can we have for objecting to the mere inconsistency between our commitments and our pronouncements?*

Our approach, which has much in common with Piovarchy (forthcoming) and Shoemaker and Vargas (2021), focuses on the *function* of the hypocrisy norm. In doing so, it offers an answer the question posed above. We object to mere inconsistency because trust and cooperation are imperative to social life.[25] The functionalist account explains key features of the hypocrisy norm by highlighting the norm's usefulness; even though some acts of hypocrisy are perceived to be harmless, the fact that people *act* like hypocrisy is harmful serves important interests shared by members of society. A society with a well-established hypocrisy norm is one in which we can trust that others will follow through on their avowed moral commitments. The hypocrisy norm is socially useful because it enables social trust, which, in turn, allows people to engage in fruitful cooperation. Rather than trying to discover if and how hypocrisy is morally wrong in some fundamental sense, our account focuses on its social function. And this, we contend, offers deeper insight into its special features.

Moreover, the formal model we developed in section 3 highlights how generally signaling considerations apply. These considerations are not restricted to *blame* per se. Any kind of hypocritical moral pronouncement – blame, praise, moral assertions or moralized claims about one's identity, and so on – must be subject to the threat of social punishment if we are to secure a separating equilibrium. Because the model is developed in abstract, mathematical terms, it elucidates the pure logic of signaling, rendering its generality more manifest.[26] Hypocritical blame is simply a special case of the kind of scenario that our costly signaling model describes.[27]

This claim about the generality of signaling finds corroboration in the empirical literature on hypocrisy. Jordan et al. (2017) and Alicke et al. (2013) show that people view hypocrisy as different from, and worse than, mere lying. So, there must be some feature above and beyond deception which allows us to distinguish between liars and hypocrites.[28] Our functionalist signaling account identifies this feature quite clearly: it is

[25]A series of models identify the unraveling of trust as a key component leading to a Hobbesian state of war. See Vanderschraaf (2006), Chung (2015), and Schaefer and Sohn (2021).

[26]Philosophers have applied costly signaling models to analyze a variety of other moral phenomena, suggesting that our account of hypocrisy may, ultimately, be one component of an even more general theory of moral signaling. See, e.g., Kogelmann and Wallace (2018).

[27]Although McKinnon (2002) does not explicitly discuss signaling, she does accurately describe how hypocrisy degrades moral communication and social trust, and she does so without unduly focusing on blame.

[28]See also Szabados and Soifer (2004).

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

the fact that certain pronouncements are meant to signal moral commitments. This suggests that the key consideration involved in our condemnation of hypocrites is not whether they *blame* or not. Instead, what we care about is whether they falsely signal their moral commitments, regardless of how they choose to do so. Our model demonstrates that a similar concern arises whether or not the signal in question comes in the form of blame. The essential role of the hypocrisy norm transcends any particular kind of moral expression and lies in its ability to foster social trust and cooperation.

Before concluding, it is worth noting that a functionalist interpretation of hypocrisy has another advantage: it may help to explain the near universality of the hypocrisy norm. If we are correct that the hypocrisy norm supports trust and cooperation, and if trust and cooperation enhance productivity and general welfare, then societies that adopt the hypocrisy norm are more apt to be successful. While noting a norm's benefit is not sufficient to account for its emergence or spread, there are various models of cultural evolution in which a norm's individual or social benefit helps to account for its proliferation (Boyd and Richerson 1985; Mesoudi 2011; Soltis et al. 1995; Ullmann-Margalit 2015). The fact that the functionalist account can pinpoint an extremely useful role for the hypocrisy norm thus allows it to underwrite a promising hypothesis about how such norms have become so ubiquitous.

6. Conclusion

The hypocrisy norm imposes a cost on inconsistency, thereby enhancing the reliability of moral signaling. Given this primary focus, it flouts the features that characterize normal moral judgments, most notably content dependence and additivity. The peculiarities of the hypocrisy norm, we have argued, are best explained by considering the function of hypocrisy, which is to avoid deception by separating reliable signalers from false signalers. Societies that fail to develop the hypocrisy norm risk falling prey to crises of trust, in which potential cooperators cannot reliably identify those who share their norms.

Our argument has taken the form of an inference to the best explanation. We have sought to develop a coherent and plausible account that makes sense of hypocrisy's most distinctive features. While this marks an important step in theory development, it is not the final step. Future work must design and run experiments that allow competing theories of the hypocrisy norm to be weighed against one another. Designing such experiments is beyond the purview of this paper. Instead of extending its empirical contribution, we will conclude by considering one of the paper's normative implications.

The theory we propose here is that the hypocrisy norm has the function of increasing the social trust by enhancing the reliability of moral signaling. Adopting this functionalist perspective implicates a standard of evaluation. Analyzed functionally, the hypocrisy norm is like a tool or a technology used for certain tasks. We can therefore evaluate it, or its use, in terms of its ability to accomplish said tasks (Van Schoelandt and Gaus 2018).

From this perspective, there may be cause for concern about the popular application of the hypocrisy norm. We opened this paper by citing two examples of politically motivated accusations of hypocrisy. The motivation for calling out Al Gore or Jozef Szar is probably not to promote social trust, but rather to discredit the opposing political party. In a climate of cut-throat political competition, bad motives are probably the

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

rule, not the exception. Rather than establishing trust, such accusations often look like opportunistic forays into moral discourse and egregious misapplications of the otherwise venerable hypocrisy norm. This kind of discourse may lead to increased polarization and the breakdown of social trust. In politics, the antidote becomes the poison.

On the other hand, calling out Al Gore for excess greenhouse gas emissions or Joszef Szar for his sordid sexual liaisons may still serve to deter other would-be hypocrites from falsely signaling their moral commitments – *even if that is not the goal of the accusers*.[29] Our theory is about the function of the hypocrisy norm, not the separate issue of its underlying psychology. Norms often serve functions without that function being the psychological explanation of why people follow them. While we readily admit that those who call out politicians for hypocritical behavior are often acting in bad faith, this need not prevent their accusations from offering a *bona fide* deterrent to disingenuous moral signaling. Insofar as political actors need to court centrists and independents, they should worry about accusations of hypocrisy, even if these come from members of an opposing political party.

Whether or not political discourse offers a compelling example, there are probably scenarios in which the hypocrisy norm is, in fact, misapplied and where this misapplication has harmful effects. In particular, accusing an out-group member of hypocrisy might benefit one's own standing or harm that person's credibility, even if the accusation is specious. Once recognized, however, the problem suggests its own potential solution. Rather than rewarding harmful, opportunistic accusations of hypocrisy, we should recognize that such accusations are often an abuse of moral language that undermines the social function of the hypocrisy norm. In addition, we should heap additional praise on those who identify hypocritical behavior at some personal cost to themselves or their in-group. In this way, by bearing in mind the social function of the hypocrisy norm, we can *re-code* partisan accusations as self-serving acts generating social harm, while calling out hypocrisy amid one's own ranks can be coded as a socially beneficial act worthy of moral praise. This offers a partial illustration of how the functionalist perspective allows us to hone our moral reactions, rendering them more precise and thus more useful. The functionalist perspective does more than merely *explain*, it can also *guide* moral judgment and conduct.

Acknowledgments. We are indebted to Mario Ivan Juárez García, as well as the attendees of our talk at the 2022 PPE Society Annual Meeting. Elizabeth Anderson offered several helpful examples as well as encouragement. And two anonymous referees helped us to hone the paper into its current form.

References

Alicke, M., Gordon, E., and Rose, D. (2013). Hypocrisy: What counts? *Philosophical Psychology*, **26**(5):673–701.

Batson, D., Thompson, E., Seuferling, G., Whitney, H., and Strongman, J. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, **77**(3):525–37.

Batson, D., Thompson, E., and Chen, H. (2002). Moral hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology*, **83**(2):330–39.

Boyd, R. and Richerson, P. J. (1985). *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.

Chung, H. (2015). Hobbes's state of nature: A modern Bayesian game-theoretic analysis. *Journal of the American Philosophical Association*, **1**(3):485–508.

[29]We thank Mario Juárez García for this observation.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

Cohen, G. A. (2006). Casting the first stone: Who can, and who can't, condemn the terrorists?. *Royal Institute of Philosophy Supplements*, **58**:113–36.
Cooter, R. and Ulen, T. (2012). *Law and Economics, Sixth Edition*. White Plains, NY: Addison-Wesley.
Crisp, R. and Cowton, C. (1994). Hypocrisy and moral seriousness. *American Philosophical Quarterly*, **31**(4):343–49.
De Silva, H. and Sigmund, K. (2024). Dynamics of signaling games. *SIAM Review*, **66**(2):368–87.
Effron, D. A., O'Connor, K., Leroy, H., and 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.
Friedman, M. (2013). How to blame people responsibly. *The Journal of Value Inquiry*, **47**:271–84.
Fritz, K. and Miller, D. (2018). Hypocrisy and the standing to blame. *Pacific Philosophical Quarterly*, **99**(1):118–39.
Fritz, K. and Miller, D. (2019). The unique badness of hypocritical blame. *Ergo: An Open Access Journal of Philosophy*, **6**(19):545–69.
Henrich, J. and Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, **208**(1):79–89.
Isserow, J. (2022). Subjunctive hypocrisy. *Ergo*, **9**(7):172–99.
Isserow, J. and Klein, C. (2017). Hypocrisy and moral authority. *Journal of Ethics and Social Philosophy*, **12**(2):191–222.
Jordan, J., Sommers, R., Bloom, P., and Rand, D. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science*, **28**(3):356–68.
Kameda, T., Takezawa, M., and Hastie, R. (2005). Where do social norms come from? The example of communal sharing. *Current Directions in Psychological Science*, **14**(6):331–34.
Kittay, E. F. (1982). On hypocrisy. *Metaphilosophy*, **13**(3–4):277–89.
Kogelmann, B. and Wallace, R. H. (2018). Moral diversity and moral responsibility. *Journal of the American Philosophical Association*, **4**(3):371–89.
Kreps, T. A., Laurin, K., and Merritt, A. C. (2017). Hypocritical flip-flop, or courageous evolution? When leaders change their moral minds. *Journal of Personality and Social Psychology*, **113**(5):730.
McKinnon, C. (2002). Hypocrisy and the good of character possession. *Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie*, **41**(4):715–39.
Mesoudi, A. (2011). *Cultural Evolution*. Chicago, IL: University of Chicago Press.
O'Brien, M. and Whelan, A. (2022). Hypocrisy in politics. *Ergo: An Open Access Journal of Philosophy*, **9**(63):1692–714.
Piovarchy, A. (2023a). Does being a "bad feminist" make me a hypocrite? Politics, commitments and moral consistency. *Philosophical Studies*, **180**(12):3467–88.
Piovarchy, A. (2023b). Situationism, subjunctive hypocrisy and standing to blame. *Inquiry*, **66**(4):514–38.
Piovarchy, A. (forthcoming). Hypocritical blame as dishonest signalling. *Australasian Journal of Philosophy*.
Rossi, B. (2018). The commitment account of hypocrisy. *Ethical Theory and Moral Practice*, **21**(3):553–67.
Schaefer, A. and Sohn, J.-Y. (2021). Unravelling into war: Trust and social preferences in Hobbes's state of nature. *Economics & Philosophy*, **38**(2):171–205.
Shoemaker, D. and Vargas, M. (2021). Moral torch fishing: A signaling theory of blame. *Noûs*, **55**(3):581–602.
Soltis, J., Boyd, R., and Richerson, P. J. (1995). Can group-functional behaviors evolve by cultural group selection?: An empirical test. *Current Anthropology*, **36**(3):473–94.
Spence, M. (1978). Uncertainty in economics. *Uncertainty in Economics*, **83**(3):355–74.
Szabados, B. and Soifer, E. (2004). *Hypocrisy: Ethical Investigations*. Peterborough, ON: Broadview Press.
Thrasher, J. and Handfield, T. (2018). Honor and violence. *Human Nature*, **29**(4):371–89.
Todd, P. (2019). A unified account of the moral standing to blame. *Noûs*, **53**(2):347–74.
Todd, P. (2023). Let's see you do better: An essay on the standing to criticize. *Ergo: An Open Access Journal of Philosophy*, **10**.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, **46**(1):35–57.
Turner, D. (1990). Hypocrisy. *Metaphilosophy*, **21**(3):262–69.
Ullmann-Margalit, E. (2015). *The Emergence of Norms*. Oxford: Oxford University Press.

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press

Valdesolo, P. and DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science*, **18**(8):689–90.

Vallier, K. (2018). Social and political trust: Concepts, causes, and consequences. *Knight Foundation Policy Paper*: 1–26.

Vallier, K. (2020). *Trust in a Polarized Age*. Oxford: Oxford University Press.

Vanderschraaf, P. (2006). War or peace?: A dynamical analysis of anarchy. *Economics & Philosophy*, **22**(2):243–79.

Van Schoelandt, C. and Gaus, G. (2018). Constructing public distributive justice: On the method of functionalist moral theory. In *New Perspectives on Distributive Justice*, ed. M. Knoll, S. Snyder and N. Şimsek. Boston: De Gruyter, 403–22.

Wallace, J. (2010). Hypocrisy, moral address, and the equal standing of persons. *Philosophy and Public Affairs*, **38**(4):307–341.

Zahavi, A. (1977). Reliability in communication systems and the evolution of altruism. In *Evolutionary Ecology*, ed. B. Stonehouse and C. Perrins. London: Palgrave, 253–59.

Cite this article: Jeffers, Matthew, Schaefer, Alexander 2025. The Function of Hypocrisy Norms. *Utilitas* **37**, 123–140. https://doi.org/10.1017/S0953820825000044

https://doi.org/10.1017/S0953820825000044 Published online by Cambridge University Press