

Moral Violations Are Made, Not Found: A Theory of Motivated Violation Construction

Emile Boullineau

*Independent Researcher, United Kingdom · ORCID 0009-0001-1500-4990 ·
emile.boullineau@gmail.com*

Author accepted manuscript. Accepted for publication in New Ideas in Psychology on 16 October 2026. The published version will be available at the DOI added here on publication.

*© 2026. This manuscript version is made available under the CC BY-NC-ND 4.0 licence,
<https://creativecommons.org/licenses/by-nc-nd/4.0/>*

Keywords: *moral judgement • motivated reasoning • moral cognition • violation construction • moralisation • social evaluation*

ABSTRACT

Moral psychology treats moral violations as features of behaviour that observers detect and then evaluate. This paper introduces Motivated Violation Construction (MVC): the proposal that observers may classify ambiguous behaviour as morally wrong when that classification supports a prior negative evaluation. MVC shifts whether behaviour enters the moral domain, rather than merely intensifying evaluation after entry. Factors such as threat, dislike, envy/status resentment, or group allegiance are proposed to lower the classification threshold under negative motivation. Positive stances raise it. The paper focuses on violation construction under negative motivation, where moral content is added rather than withheld. MVC concerns the point at which behaviour is classified as wrong, before its severity is evaluated. MVC is distinct from motivated reasoning, which shapes how people judge behaviour already seen as wrong, and from moralisation, which concerns shifts in what a society treats as moral issues. Four predictions are advanced. Negative motivation increases, and positive motivation decreases, the frequency with which identical behaviour is classified as a violation. When the behaviour is genuinely ambiguous, this effect is strongest. It weakens where moral consensus is clear. Differences in classification track the observer's motives rather than the behaviour. Motivation affects classification more strongly than severity, and this asymmetry is moderated by ambiguity. Each prediction has operationalisations and procedural falsifiability conditions, with a registered manipulation check on the evaluative-stance induction.

1. THE DETECTION ASSUMPTION

Existing research on moral judgement typically treats moral violations as identified prior to evaluation. Consider the following cases. A person cuts in front of you in a queue. A colleague takes credit for work supported by others. A politician cancels a public commitment. In each of these cases, moral psychology often treats the observer as perceiving the behaviour, recognising it as morally relevant, and evaluating the actor accordingly. The violation is treated as a stable input across observers, something detected from the behaviour rather than something the observer brings to it.

This assumption is structural, not incidental. Attribution theory (Heider, 1958; Weiner, 1985) begins after the violation has already been identified, asking how people assign causes to recognised transgressions. Models of moral blame (Alicke, 2000; Malle *et al.*, 2014) specify how evaluators move from a perceived harmful event to a judgement of responsibility, but take the harmful event as given. Moral foundations theory (Graham *et al.*, 2013) maps the domains in which people make moral evaluations without asking whether the evaluative inputs are stable across observers. In every case, the moral violation functions as the independent variable: it is the stimulus that triggers judgement.

We call this the **detection assumption**: the implicit premise that moral violations exist in behaviour and are identified through a process that is fundamentally perceptual. Under this assumption, disagreements about whether a violation occurred reflect differences in moral standards, attention, or values. Current empirical literature has not typically formalised the possibility that the violation itself may be constructed by the observer, serving an underlying psychological need.

The detection assumption is reinforced by the dominant experimental paradigm in moral psychology. In many vignette-based studies (Cushman, 2008; Uhlmann *et al.*, 2009), a central methodological paradigm in the field, participants receive behaviours pre-classified as violations by the researcher. The participant evaluates a violation that has already been constructed for them. The paradigm does not permit the participant to decide whether the behaviour constitutes a violation at all. The result is a literature highly informative about how people process moral violations but largely silent about how those violations come into being.

This paper identifies a stage in moral processing that current models do not typically formalise: the point at which behaviour is classified as a violation. It proposes that this classification stage is subject to motivational influence, and introduces Motivated Violation Construction (MVC) as a theoretical account of how this process operates. If correct, this implies that a central assumption of moral psychology (that violations are stable inputs into judgement) is incomplete. Variation in moral judgement may arise not only from how people evaluate violations, but from whether they construct them in the first place. The core claim of

MVC is that, for ambiguous behaviours, motivation determines whether an act is classified as a moral violation at all, not merely how it is evaluated once classified. This framework applies specifically to behaviours whose moral status is ambiguous, those that generate genuine disagreement among non-motivated observers, rather than to behaviours whose moral status commands broad consensus. If this distinction holds empirically, it suggests that the standard experimental paradigm in moral psychology, which does not include the classification stage, may need to be supplemented with designs that allow participants to classify behaviour freely.

Section 2 introduces MVC and specifies its scope, mechanism, and the core distinction between classification and evaluation. Section 3 distinguishes MVC from six neighbouring frameworks through a systematic comparison. Section 4 derives four testable predictions with proposed empirical approaches and falsification conditions. Section 5 considers implications, relation to broader research, and limitations.

2. MOTIVATED VIOLATION CONSTRUCTION

Everyday life shows us that we react differently to certain behaviour depending on who the subject is. Consider watching someone you dislike arrive late to a meeting. The reaction is immediate: disrespectful, dismissive of everyone's time. But when a friend does the same thing, the same lateness becomes evidence of a busy schedule. Bluntness from a rival registers as rude, but when it comes from a friend, it is evaluated as being refreshingly honest. The behaviour has not changed, but the moral label applied to it has.

This disconnect is not unusual and it is often observed in everyday life. Yet it causes a problem that has not been adequately addressed by moral psychology. If the same behaviour is classified as morally wrong or morally neutral depending on who performs it, and if this classification takes place before anything resembling formal moral reasoning has begun, then for that behaviour the violation is not a stable property of the act alone. The observer's prior orientation toward the person is doing part of the classificatory work.

Independent footing for the classification/evaluation distinction

The classification/evaluation split is not unique to the present framework. Moral foundations research suggests that observers can differ in which concerns they treat as morally relevant, not only in how harshly they judge already-recognised violations (Graham et al., 2009; Graham et al., 2013). Moralisation research likewise shows that behaviours and preferences can move into the moral category over time, indicating that moral classification is not a passive registration of behavioural features alone (Rozin, 1999; Skitka, 2010). Signal-detection logic provides a measurement-theoretic analogue, because response criteria can

shift separately from sensitivity or discriminability (Stanislaw & Todorov, 1999). These literatures do not demonstrate MVC, but they independently motivate the possibility that moral-entry classification and post-entry evaluation can come apart.

We introduce Motivated Violation Construction (or MVC) as a theoretical account of how this process could occur. MVC proposes that an observer is more likely to classify ambiguous behaviour as a moral violation when a pre-existing negative motivation exists toward the target. This includes perceived threat, dislike, envy/status resentment, or group allegiance, with group-based allegiance illustrated by social identity accounts of intergroup conflict (Tajfel & Turner, 1979). Threat, dislike, envy/status resentment, and group allegiance are treated as equivalent only in the formal sense that, when successfully induced, each is predicted to lower the classification threshold under ambiguity. MVC does not assume that they share proximal pathways or produce identical effect sizes. Therefore, under negative motivation, not all behaviour is more likely to be classified as a violation. Ambiguous behaviour is.

Why does this happen? One plausible mechanism is based on justificatory need (cf. Kunda, 1990; Bastardi *et al.*, 2011). The observer already dislikes the target. But simple dislike without a moral basis can create dissonance or self-justificatory pressure (Festinger, 1957). This is because most people want to regard themselves as fair-minded evaluators, not as people who simply dislike others without cause. The observer resolves this tension not by revising their evaluation, but by shifting the threshold at which behaviour is classified as a violation. An ambiguous act gets reclassified as an offence. A further, subsidiary prediction concerns phenomenology. If constructed violations are experienced as perceived rather than made (cf. Ross & Ward, 1996), observers would be expected to show a dissociation between the effect of induced stance on classification and their reported attribution of the judgement to that stance. A separate correction-resistance probe tests whether disclosing the induction produces revision. Section 4.1 specifies both probes. This phenomenological-invisibility prediction is not one of MVC's four formal predictions (P1–P4, set out in Section 4), and P1–P4 do not depend on it.

Positive motivation and asymmetric mechanism

MVC is best understood as a theory of evaluative-stance-dependent classification. In this framework, motivation refers to the observer's evaluative stance toward the target rather than to a general state of arousal, effort, or desire. The mechanism just described, justificatory construction under negative motivation, is the central case for the present theory, because it specifies how moral content can be added where none would otherwise be detected. The broader framework predicts that positive motivation can shift the same classification threshold in the opposite direction.

When the observer holds a positive stance toward a target (favourable disposition, identification, admiration, or trust), the same ambiguous behaviour becomes less likely to be classified as a violation. Here the mechanism is not justificatory construction but motivated discounting (cf. Ditto & Lopez, 1992). The observer's existing positive stance generates pressure to interpret behaviour charitably, raising the threshold at which it would be classified as morally wrong. A parallel phenomenological possibility for positive motivation remains speculative and would require separate testing. If it holds, the behaviour would be registered as benign rather than transgressive, and the underlying motivational source would not be consciously available.

The framework applies directionally to both stances. P1 states this bidirectional threshold prediction explicitly, while the remaining predictions are specified here around the focal negative-stance case. Motivation shifts the classification threshold. The direction of the shift depends on whether the observer's stance toward the target is negative or positive. The empirical and theoretical focus of the present paper is on violation construction under negative motivation, because that is the case in which moral content is added rather than withheld, and where the MVC signature is clearest. The opposite-direction case (motivated leniency under positive motivation) is consistent with MVC and predicted by the same framework. The present paper operationalises the framework through the negative-stance direction, while retaining the positive-stance case as a formal prediction for future direct tests.

Scope condition and the definition of ambiguity

MVC applies specifically to behaviour that admits more than one plausible and socially recognisable moral reading when assessed without prior orientation toward the actor. We define moral ambiguity as cases in which baseline observers, those with no motivational stake in the target, produce systematically split classification patterns: the behaviour is neither overwhelmingly classified as wrong nor overwhelmingly classified as acceptable. This definition is anchored in observable disagreement under neutral conditions and is therefore independent of the motivational process MVC describes. Behaviours that generate strong baseline convergence (virtually all observers agreeing on their moral status) fall outside MVC's operating range. If ambiguity were defined by motivated disagreement itself, the framework would be circular. The non-motivated baseline avoids this. As a working operationalisation, we adopt 25%–75% baseline classification, a procedural default rather than a theoretical claim, since the construct is continuous. Empirical tests should report sensitivity analyses at alternative thresholds (e.g., 20%–80%, 30%–70%), with conclusions treated as robust only when sign and significance hold across the range.

Why the mechanism does not extend to clear violations

A natural question arises: if the justificatory need mechanism is available whenever an observer dislikes a target, why does MVC not also operate on unambiguous transgressions such as theft or violence? The answer involves two complementary considerations. First, when the behavioural signal is strong and morally unequivocal, it constrains classification, leaving little interpretive latitude for motivation to exploit. The behaviour provides its own moral warrant. Second, the mechanism is not absent for clear violations. It is functionally redundant. When behaviour is already clearly wrong, the observer does not need to construct a violation to justify their negative stance. MVC becomes active precisely where this moral basis is absent and needs to be constructed. In short, when the behavioural signal is strong, it constrains classification on its own, rendering the constructive mechanism both difficult to deploy and unnecessary. Whether signal strength largely constrains classification is itself an empirical question. MVC predicts that it does, a claim directly tested by P2.

For acts such as theft, physical violence, and outright deception, the moral signal is loud enough that motivation barely changes the classification. Existing models of moral judgement handle these adequately and the intention is not for MVC to replace them. MVC operates in a different domain: the broad and poorly mapped space of everyday acts whose moral valence remains open to interpretation. Acts such as arriving late or speaking too bluntly.

Withholding information or prioritising oneself are further examples of acts that can be interpreted as being a violation or as something acceptable. MVC argues that motivation is what moderates the framing toward one interpretation or the other. To be precise about the boundaries of this claim: observers do not fabricate behaviours or invent events. They adjust the threshold at which real, observed behaviours get slotted into the category of moral violation. MVC predicts that observers will often be unaware that this threshold adjustment is occurring. This is treated as a subsidiary prediction tested through the exploratory source-attribution and correction-resistance probes described in Section 4.1, not as one of the framework's formal P1–P4 predictions.

Two terms require clarification, because their confusion has obscured the process described by MVC. Classification is the threshold decision, the binary judgement that an act is wrong rather than permissible. This is where a behaviour enters the moral domain at all. Evaluation is what follows: the assessment of severity, the question of how wrong something is once it has been classified as a violation. MVC operates at the classification stage (see Figure 1). Most existing models of moral judgement focus almost exclusively on evaluation. Because the two stages have not been clearly pulled apart in the literature, the assumption that moral violations are simply detected, perceived like objects in a visual field, has gone largely unchallenged.

MVC describes a conditional process, not a personality trait. Any observer can construct violations when motivation toward a target passes a paradigm-specific manipulation-check threshold (see §4) and the behaviour in question leaves room for more than one moral reading. Some people may do this more often or more readily than others, and that variation is worth studying (see Discussion). But the framework developed here is about the mechanism itself, not about who is more or less prone to deploying it.

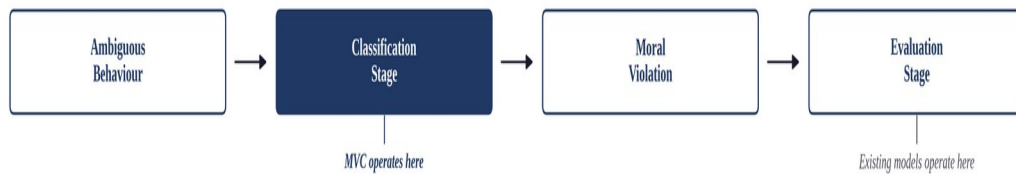


Figure 1. The locus of Motivated Violation Construction in the moral judgement processing chain. MVC operates at the classification stage, where behaviour is first designated as a moral violation. Existing models begin at the evaluation stage.

3. DISTINGUISHING MVC FROM ADJACENT ACCOUNTS

The classification/evaluation distinction is what positions MVC relative to existing frameworks. Each of the accounts reviewed below addresses a different stage of moral processing. Table 1 summarises where each framework operates and what it changes, before the detailed comparisons that follow.

Table 1. Comparison of MVC and neighbouring frameworks.

Framework	Stage Affected	What Changes	Core Prediction	Why MVC Is Needed
Motivated reasoning	Evaluation	Severity, blame, causal attribution	Changes how harshly a known violation is judged	Assumes violation already identified
Attribution bias	Evaluation	Blame assignment	Disproportionate blame for disliked targets	Starts after violation detection
Moralisation	Classification (cultural)	Entire categories of behaviour	Practices migrate into moral domain over time	Social categories, not individual, real-time construction
Dyadic morality	Moral perception	Perceived harm content	Harm is inferred, not given	Targets harm perception, not violation classification
Motivated moral perception	Detection	Speed of noticing moral stimuli	Moral content detected faster	Accelerates detection, does not construct from ambiguous inputs
Learned valence priming	Associative evaluation	Activation of negative actor associations	Negatively valenced targets shift both classification and severity	Predicts general evaluative shift, not classification-specific construction under ambiguity
MVC	Classification (individual)	Violation vs. non-violation	Motivation determines whether specific behaviour is classified as wrong	Addresses classification stage directly

Motivated Reasoning

Research on motivated reasoning (Kunda, 1990; Ditto & Lopez, 1992) is relevant here. It explains that people tend to arrive at conclusions by selectively searching for, interpreting, and weighing evidence (see also Nickerson, 1998). MVC takes this further and identifies a process that occurs earlier in the evaluative process. Conclusions are not only biased once a moral event has been recognised, but what counts as a moral event can itself be shaped by motivation in the first place. The object of judgement is not simply taken as given and can be constructed in ways that support the observer's prior evaluative stance. Motivated reasoning distorts the processing of an event already categorised as moral. MVC proposes that the categorisation itself can be distorted, meaning the input to downstream reasoning differs depending on whether MVC has operated.

The empirical discriminator is prediction P4, formulated below as a conjunction test.

Attribution Bias

Attributional models (Alicke, 2000) account for how an observer assigns disproportionate blame to disliked or threatening targets. What these models take for granted is that a blameworthy event has already been identified. The attribution process starts after violation detection. However, MVC operates before attribution happens. Attribution makes a detected violation seem worse, whereas MVC determines whether it gets detected in the first place.

Moralisation

Research on moralisation (Rozin, 1999; Brandt et al., 2015; Skitka, 2010) documents how preferences and practices migrate into the moral domain over time. Smoking acquires moral weight. Dietary choices become matters of right and wrong. MVC differs in its scope, mechanism, and temporal grain. MVC therefore operates at a different level of analysis from Rozin's account. Moralisation tracks how cultural norms shift over time. MVC tracks how a single observer constructs a violation in real time, around a specific target. It operates at the level of specific behavioural episodes, not social practices. It is driven by interpersonal motivation such as dislike, envy/status resentment, or group allegiance rather than cultural or ideological change. It involves construction from individual acts in real time, not the gradual reclassification of entire categories. A person who moralises meat-eating has undergone a value shift. A person who constructs a violation from a rival's ambiguous email has done something different entirely. This is motivated moral construction at the level of a single behavioural instance.

Harm Construction

The Theory of Dyadic Morality (Gray *et al.*, 2012; Schein & Gray, 2018) argues that moral judgement is fundamentally about perceived harm between an intentional agent and a suffering patient. What counts as harm is itself inferred rather than given. This account shares an important characteristic with MVC. Both constructs reject the idea that moral inputs are fixed properties of events waiting to be detected, but the two frameworks target different stages. Dyadic morality is concerned with the content of moral perception (specifically, whether an observer perceives harm in a given event), whereas MVC concerns the threshold (the point at which an observer classifies a behaviour as a moral violation at all), and the motivational conditions under which that threshold shifts.

These two things can come apart. An observer can perceive diffuse harm without classifying the behaviour as a violation (harm without moral classification). An observer can classify a behaviour as a violation under motivation without perceiving concrete harm (moral classification without harm). These accounts are not competing but rather complementary and operating at adjacent but distinct stages of the moral processing chain.

Motivated Moral Perception

Research on motivated moral perception (Gantman & Van Bavel, 2014, 2015) shows that moral stimuli capture visual attention and reach conscious awareness faster than non-moral stimuli. This perceptual advantage is in itself motivationally modulated. When justice motives are satiated, the advantage shrinks. This work supports a point relevant to MVC: motivation can shape early perceptual processing, including responses to morally relevant stimuli (see also Balci et al., 2006).

MVC and motivated moral perception diverge regarding assumptions about the input. Motivated moral perception accelerates the detection of stimuli that already carry moral weight. The signal is present in the environment and motivation acts as a perceptual filter that lowers the threshold for noticing it. MVC applies where no consensual moral signal is available to the observer. The behaviour is ambiguous, the moral content is not pre-existing, and motivation drives the observer to construct a violation from interpretive flexibility rather than to detect one from perceptual salience. In short: motivated moral perception assumes the presence of a moral signal. MVC specifies the conditions under which that signal is absent and constructed.

Learned Valence Priming

A further alternative worth distinguishing is learned valence priming (cf. Fazio, 2001; Amodio, 2014; Payne et al., 2017). On this family of accounts, prior evaluative associations with social categories could shift classification through automatic association rather than motivated construction. This account is more parsimonious than MVC because it requires no justificatory mechanism, only differential base rates of prior exposure. Three predictions distinguish MVC from this alternative. First, absent additional assumptions about moral-entry thresholds, valence priming does not itself predict the classification-over-severity asymmetry specified in P4. MVC predicts disproportionate classification effects conjunctively moderated by ambiguity (P4). Second, valence priming predicts the strongest effects in domains where the actor's role carries evidential relevance. MVC predicts effects that generalise to domains bearing no rational connection to the actor's identity. Third, valence priming predicts a generalised shift in evaluation of the actor across behaviours. MVC predicts shifts concentrated specifically in the ambiguous range (P2).

In sum, MVC occupies a distinct position in the moral processing chain: it targets the classification of specific behaviours by specific observers, at a stage where the neighbouring accounts reviewed here are either silent or operating on already-classified inputs.

Why compound accounts do not jointly reproduce P1–P4

The argument for MVC as a separate construct depends on the claim that the predicted P1–P4 pattern is not currently predicted by combining the simpler accounts reviewed above. Three especially plausible compound candidates are considered here: negativity bias plus criterion shift, attentional capture plus ambiguity sensitivity, and learned valence priming plus halo effects. As currently specified, each falls short of this explanatory burden in a different way. A negativity-bias-plus-criterion-shift account can explain some of the expected pattern, including increased violation labelling under negative stance. What it does not yet specify is the full MVC profile of target-specific stance effects, structured between-observer clustering, and classification/severity asymmetry. A combination of attentional capture and ambiguity sensitivity can produce the ambiguity moderation P4 requires, but the attentional component does not by itself predict the classification/severity asymmetry P4 specifies.

A combination of learned valence priming and halo effects can produce both asymmetric and ambiguity-sensitive shifts in classification, but does not by itself predict the structured between-observer variability P3 requires without additional assumptions about observer-specific motivational stance. The conjunction P1–P4 conditional on ambiguity therefore requires a mechanism that is simultaneously motivation-sensitive, asymmetric in its effects on classification versus severity, ambiguity-moderated, and observer-specific. The existing accounts reviewed here do not specify this full conjunction. The argument does not foreclose the possibility that some future compound account might do so. It shows that the candidates currently in the literature do not.

4. TESTABLE PREDICTIONS

MVC generates four core predictions. P1 is bidirectional. P2–P4 operationalise the focal negative-stance direction unless otherwise stated. Each can be tested using standard experimental designs with between-subjects manipulation of motivation and within-subjects manipulation of behavioural ambiguity.

P1. Motivation shifts the classification threshold. Holding behaviour constant, the probability that behaviour is classified as morally wrong increases when the observer holds a negative evaluative stance toward the target, and decreases when the observer holds a positive stance. The present paper develops the negative-stance direction in greatest detail. This represents identical behaviour that is subject to different moral classification. The key driver is the observer’s motivational state and not the characteristics of the behaviour observed. This can be empirically tested using standard behavioural paradigms that induce an evaluative stance toward the target, e.g. by presenting brief biographical profiles with liked vs. disliked indicators.

P2. The effect is strongest under ambiguity. MVC operates where behaviour is generally considered morally ambiguous. For behaviours with a widely accepted moral status, such as theft by a colleague (clearly wrong) or declining to attend a lunch event (clearly benign), the behavioural signals constrain classification and reduce the influence of motivation. The MVC effect should be most evident in the ambiguous zone: a manager reassigning a struggling employee's project without discussion, a friend cancelling plans at short notice, a co-worker forwarding a private message to a third party. These are acts whose moral status depends on how they are read. The prediction is that motivation shifts the reading. The explicit failure condition that separates MVC from a general negativity bias is that, when behaviour is high-consensus (virtually all observers agree on its moral status), MVC should produce null or negligible effects. Pilot norming of behavioural stimuli across a non-motivated control sample would establish the necessary baseline classification rates.

P3. Constructed violations show a distinctive signature. Violations produced through MVC should show less agreement between observers and greater variation in how behaviour is classified than clear-cut transgressions. However, this variability will not be random. It should reflect differences in observers' motivations. Observers sharing a motivational stance toward the target will converge in their classifications, while those with opposing stances will diverge, producing a motivation-dependent clustering pattern rather than uniform noise. The key prediction is not merely increased variability, but structured variability: classification differences should align systematically with motivational condition rather than appearing randomly across observers. This prediction can be tested by comparing agreement in moral classification between observers who share similar motivations and those who do not.

P4 (Critical Test). Classification is affected more than severity, and this asymmetry is moderated by ambiguity. This prediction most clearly distinguishes MVC from generalised evaluative bias and serves as the critical test of the framework. Motivation is expected to shift whether a behaviour is classified as wrong more than how severely it is judged once classified.

Practically, the influence of motivation on whether behaviour is seen as a violation should be stronger than its influence on how bad that behaviour is judged to be, given that it is already seen as wrong. If motivation had an equal effect on both, making everything seem worse, the effect would be better characterised as a general negativity bias. This would not require a new construct.

However, the asymmetry alone is not sufficient to confirm MVC, because alternative accounts (such as attentional accounts or general negativity bias) could predict a similar pattern. The additional requirement is that the classification effect must be moderated by ambiguity: motivation should shift classification substantially for ambiguous behaviours but

negligibly for high-consensus behaviours. A general negativity-bias account would not by itself require this moderation pattern without additional assumptions. P4 is therefore a conjunction: the classification/severity asymmetry must co-occur with ambiguity moderation of the classification effect. If the asymmetry appears without ambiguity moderation, the simpler account is preferred and MVC is not supported. This distinction can be tested using models that separate the likelihood of classifying behaviour as wrong from the severity of judgement once that classification is made, tested via interaction effects between evaluative stance and ambiguity. Even as a conjunction, P4 alone does not establish MVC. What distinguishes MVC is the full P1–P4 pattern, which the neighbouring accounts reviewed here do not jointly predict.

4.1 Proposed Empirical Approach

MVC makes a set of interlocking predictions. Failure to observe the predicted pattern at either the individual or interaction level would challenge the framework. The constructs introduced above can be operationalised as follows, providing a basis for empirical testing without committing to a single experimental protocol.

Motivation and ambiguity. Motivation can be manipulated via stance induction such as biographical profile, group membership, or competitive framing (cf. Tajfel & Turner, 1979), with non-motivated controls providing the baseline against which effects are measured. Ambiguity is operationalised through pre-norming: a non-motivated control sample classifies each behaviour. Items generating systematically divided baseline classifications (substantial disagreement rather than convergence toward either pole, with classification rates between approximately 25% and 75%) are treated as morally ambiguous for the purposes of the main study (for example, a manager reassigning a project without consultation). This classification is fixed before the main experiment and cannot be revised post hoc. This forecloses the possibility that the ambiguity construct absorbs variance later produced by the motivation manipulation itself. Non-motivated status is itself verifiable at the participant level, using pre-task measures of affect or evaluative orientation toward each target. Participants showing prior orientation are excluded from the baseline sample. Robust designs should validate the non-motivated baseline through convergent checks, whether multi-baseline comparison, participant-level non-motivation verification, or polarity-reversal robustness analyses, to rule out unintended motivational structure in the norming sample.

Classification and severity. The target construct is moral-entry classification, defined as whether the behaviour is treated as entering the moral domain. Observers respond with Yes or No to a classification question, which can be operationalised using a binary concern-detection stem or a direct wrongness stem, provided the wording is pre-specified and the moral-entry interpretation is validated through convergent diagnostics such as severity

linkage, rationale coding, and polarity-reversal checks. Severity is then measured conditionally on a rating scale, asked only of those who answered Yes. The conditional design creates a selection-on-the-dependent-variable structure for severity, since only those who answered Yes contribute severity ratings. A related concern is that binary classification could blur with low-threshold severity, with Yes responses signalling mild concern rather than a categorical decision. The separate-samples design (classification and severity measured by independent observers) addresses both concerns and is required for the primary inferential test of P4. Future work could supplement this with process-tracing measures such as response latency or verbal protocol analysis.

Reconciling classification and conditional severity in P4

The point is not simply to compare two coefficients. The test must ask whether motivation changes the probability of classifying behaviour as wrong more than it changes severity once wrongness has already been accepted. The conditional severity design (severity rated only by participants who classified the behaviour as a violation) creates a selection-on-the-dependent-variable structure that can distort the estimated effect of motivation on severity. The conditional sample is non-representative of the underlying severity distribution. The separate-samples design (classification and severity measured by independent observer pools, with severity rated without first requiring a Yes classification) is not subject to this selection bias and provides the primary inference for P4. The within-subject conditional design provides convergent estimation with explicit acknowledgement of the selection structure. A positive P4 verdict requires both designs to agree directionally and the separate-samples design to meet the registered P4 effect-size threshold. Reporting both designs alongside the threshold prevents inferential ambiguity about whether the asymmetry reflects measurement geometry rather than the proposed mechanism. The decisive comparison is a pre-specified marginal-effect contrast on a common probability or standardised scale, not a raw-coefficient comparison between models. P4 evidence must not reduce to thresholded severity. The asymmetry must remain visible in the common-scale estimand itself.

Design logic. P1 can be tested via logistic or probit regression comparing classification rates between motivated and non-motivated observers. P2 requires a motivation $\times$ ambiguity interaction term in the same model, with the prediction that the motivation effect is largest in the ambiguous zone and negligible at high consensus. P3 is assessed through inter-rater agreement indices comparing classification convergence among observers sharing a motivational stance versus those with opposing stances. P4 requires fitting separate models for classification (binary) and severity (ordinal), with the prediction that average marginal effects of motivation on classification exceed those on severity, and that this asymmetry is itself moderated by ambiguity. Motivation manipulations can be verified via post-task self-report measures matched to the registered motivational source, such as target liking,

perceived threat, group identification, or status resentment, confirming that the stance induction produced the intended motivational state.

Convergent diagnostics. Beyond the core regression tests, four convergent diagnostics could further test the classification-stage interpretation against simpler alternatives. Of these four diagnostics, the counterfactual-reversal probe and the rationale-coding analysis are treated as primary interpretive diagnostics for the construction interpretation, while the identity-alone probe and the signal-detection framing are treated as exploratory transparency checks. A positive P4 result is most diagnostic of construction when the two primary diagnostics align with the regression-based asymmetry estimate. A counterfactual reversal design asks whether the same observer reverses classification when the same behaviour is paired with actors of opposite motivational valence, isolating the contribution of motivated stance from features of the act itself. Rationale analysis codes participants' spontaneous explanations for novel victims, invented harm pathways, or attributions of malicious intent absent from the stimulus, treating such elaborations as observable traces of moral construction rather than mere reporting. Identity-alone probes test whether motivation can trigger classification in the absence of any behavioural signal, which would mark the upper bound of the construction process. Signal detection theory could provide a psychometric framing, with MVC predicting that motivation shifts criterion without affecting sensitivity. This distinguishes the framework from generalised evaluative-bias accounts.¹

Separately from these four diagnostics, two exploratory probes address the subsidiary phenomenological-invisibility prediction introduced in Section 2. In a source-attribution sample, participants classify the behaviour, give an open-ended account of what drove the judgement, and rate whether their classification reflected the behaviour itself or their prior stance toward the target. In a separate correction-resistance sample, used to avoid the source-attribution questions contaminating revision, participants classify the behaviour and report confidence. They are then informed of the stance induction and offered an opportunity to revise, while a matched control group receives a generic invitation to reconsider without disclosure of the induction. Support would require convergence: a reliable stance effect on classification coupled with little self-attributed stance influence, followed by limited revision and sustained confidence after disclosure. The prediction would not be supported if participants reliably attribute their classifications to the induced stance at or above a pre-specified threshold, or if disclosure produces revision above a pre-specified threshold or reliably greater revision than the generic reconsideration control. Non-attribution and non-revision are not decisive evidence of subjective unawareness because both can arise for other

¹ A preregistered companion empirical study testing analogues of predictions P1–P4 using large language models as evaluative systems has been registered with the Open Science Framework. The registration is embargoed and no results from it are used to support the claims advanced here. The present paper stands on theoretical grounds alone.

reasons. These probes are therefore falsification-oriented and exploratory. They do not bear on P1–P4 or gate the P4 verdict.

Falsification conditions. MVC would not be supported if: (1) motivation produces no significant effect on classification for ambiguous behaviour, (2) motivation effects on classification appear equally across all ambiguity levels, including high-consensus behaviours, (3) inter-observer variability in classification is random and unrelated to motivational condition, or (4) motivation affects severity equally or more than classification, or the classification/severity asymmetry appears without ambiguity moderation. Any of these outcomes would favour simpler accounts over MVC, such as generalised evaluative bias, halo effects, or post-classification severity inflation. Taken together, these conditions ensure that MVC is not supported by general evaluative bias alone, but only by a specific pattern of classification-stage effects under ambiguity.

Procedural requirements for falsifying tests

A test of MVC must satisfy specific procedural conditions before its null result can count as falsifying. Empirical tests must include a registered manipulation check on the evaluative-stance induction and report the manipulation-check effect size as a primary methodological metric. They must also pre-specify a minimum manipulation-check effect-size threshold appropriate to the paradigm, based on prior validation, pilot data, or minimum-detectable-effect logic, in advance of data collection. Null results should be treated as falsifying MVC only when the manipulation passes the pre-registered threshold. Null results from manipulations that fail their own check are uninformative about MVC and should be reported descriptively rather than as confirmatory evidence. These procedural requirements substantially constrain the unfalsifiability concern that motivation strength is endogenously adjustable to fit results. The procedural commitment is paradigm-agnostic, allowing different operationalisations to set thresholds appropriate to their measurement context.

5. DISCUSSION

Theoretical Implications

MVC's core premise is that a moral violation, a concept that is central to moral judgement research, is not a stable factor in the evaluative process.

For a substantial class of everyday behaviours, it can vary and be constructed by the observer. It carries motivational information about the observer as well as behavioural information about the target. What is usually treated as the independent variable in moral judgement research, the presence of a violation, is partially determined by the observer. This has methodological consequences. Vignette-based studies that present pre-defined violations to

participants may be studying only the second half of the moral judgement process while bypassing the classification-stage variability MVC targets.

For motivated reasoning, MVC extends the account one stage earlier. People shape the starting point on which their conclusions depend, and do not simply reason their way to the conclusions they prefer. MVC proposes a reversal of the usual assumption that perception comes first and evaluation follows: in MVC, the evaluative stance precedes and shapes the classificatory perception. In MVC, the observer's prior stance toward the target can shift whether ambiguous behaviour is classified as morally wrong.

Illustrative Applications

The following applications illustrate contexts in which MVC's input conditions (motivation plus ambiguity) could operate. They are provisional examples, not demonstrations.

Political polarisation. MVC offers a candidate mechanism for how partisan moral differences and partisan bias may lead observers to classify opponents' ambiguous behaviour as morally violating (Graham *et al.*, 2009; Ditto *et al.*, 2019). Under MVC, the partisan is not judging the same violations more harshly but constructing violations from ambiguous political behaviour that a non-motivated observer would classify as legitimate disagreement.

Interpersonal conflict. In close relationships, MVC offers a way to conceptualise how accumulated negative sentiment may lower the threshold at which a partner's ambiguous behaviour is classified as morally wrong (cf. Gottman, 1994). Acts once tolerated enter the moral domain. The partner experiences this as inexplicable scrutiny, the motivated observer as a growing pattern of transgressions.

Online moral discourse. Digital environments tend to strip away context while amplifying group identity and emotional reactions, exposing observers to a high volume of behaviour open to interpretation. At the level of individual judgements, these conditions may make it easier for motivated observers to classify ambiguous behaviour as morally wrong, a pattern which may scale to aggregate dynamics in online discourse. MVC suggests that motivated classification operates particularly readily online, where digital environments maximise both interpretive flexibility and motivational activation.

Whether MVC operates in these domains as predicted remains an empirical question. The limitations that follow outline the conditions under which the present framework would require revision or rejection.

Relation to Evaluative Rigidity

MVC can be situated within a broader interest in evaluative rigidity, but the present argument does not depend on that programme. The key claim here is narrower: motivation can shape whether ambiguous behaviour is classified as a moral violation before severity evaluation begins. On this view, evaluative rigidity may sometimes operate downstream of MVC: if motivation has already shaped the moral input, later evaluation operates on an already moralised representation.

Limitations and Future Directions

Several limitations of the present framework should be acknowledged directly, as they define the agenda for empirical work on MVC.

First, MVC is presented here as a theoretical framework without direct empirical evidence. The predictions require testing in paradigms allowing participants to classify behaviour freely. Until such tests are conducted, MVC remains a theoretical proposal.

Second, while the scope condition restricts MVC to morally ambiguous behaviour, the operationalisation requires careful pre-norming (see Section 4.1). Threshold-sensitivity and baseline-rate stability across populations are open methodological questions.

Third, MVC does not specify whether violation construction is automatic or deliberative. If present, phenomenological invisibility would be consistent with automatic processing, but MVC does not require it. The issue remains open. Response-time paradigms or cognitive-load manipulations could address this directly.

Fourth, the boundary between MVC and motivated reasoning is secured theoretically through the classification/evaluation distinction and empirically through prediction P4 (see also the distinctiveness-condition statement immediately below). However, the extent to which the two processes interact in practice remains to be investigated. MVC and motivated reasoning may operate sequentially, with classification-stage construction shaping the inputs to downstream evaluative bias.

Fifth, MVC is described in functional rather than neuropsychological terms. Future work may benefit from grounding violation construction in specific affective-cognitive architectures, such as Affective Realism (Barrett, 2017; Siegel *et al.*, 2018), which offers a promising account of how internal affective states can shape perception of external stimuli and be experienced as features of the environment.

Sixth, the practical applications discussed above are speculative extrapolations from the framework's logic. Their empirical status is no different from that of MVC itself: they represent hypotheses to be tested, not conclusions to be drawn.

Seventh, the classification/evaluation distinction is methodologically difficult to operationalise. Binary classification and conditional severity measures may differ in scale properties. Future empirical tests should therefore confirm P4 using robustness designs such as separate samples, or alternative response formats (see the distinctiveness-condition statement immediately below).

Eighth, the sources of negative stance themselves require qualification. MVC treats threat, dislike, envy/status resentment, and group allegiance as functionally equivalent only at the level of the prediction. Each is predicted to bias the classification threshold under ambiguity. The framework does not assume these sources arise from the same psychological pathway or produce identical effect sizes. Differential effects across these sources represent an open empirical question for downstream tests, not a theoretical commitment of the framework. Empirical tests should therefore register the motivational source being induced and, where multiple sources are compared, treat source type as a moderator rather than assuming uniform effect sizes. A null or weaker effect for a registered source would therefore constrain MVC's supported scope for that source rather than being treated as an ad hoc failure of generic motivation.

Ninth, the phenomenological-invisibility prediction, that constructed violations are experienced as perceived rather than made, is currently untested. Section 4.1 specifies separate source-attribution and correction-resistance probes. Because neither measure can establish subjective unawareness on its own, the prediction requires convergent evidence and has pre-specified conditions under which it would not be supported. The formal predictions P1–P4 do not depend on it.

Taken together, these limitations define the agenda for empirical work on MVC. They frame the theoretical claims advanced in the conclusion that follows.

P4 and the framework's distinctiveness condition

MVC's status as a separate construct depends especially on the empirical pattern P4 specifies. If P4 fails (that is, if motivation affects severity equally or more than classification, if classification and severity move proportionally, or if any classification/severity asymmetry is not moderated by ambiguity) MVC may not be needed as a separate construct beyond motivated reasoning or general evaluative-bias accounts. The framework treats this as its central distinctiveness condition and as a substantive risk, not a peripheral caveat. The two primary convergent diagnostics specified in §4.1 constrain the interpretation of a P4 success as construction. The exploratory diagnostics provide transparency but are not required for the P4 verdict.

6. CONCLUSION

Much moral-psychology theory has treated violations as given. MVC proposes that this foundation is incomplete. For ambiguous behaviours, the violation is not fixed by the behaviour alone but is partly constructed by the observer. The construction is driven by motivation and shaped by the observer's pre-existing stance toward the target. A subsidiary prediction, specified above for separate testing, is that this construction may be experienced as genuine moral perception rather than motivated fabrication. If this account is correct, then moral disagreement is not only about values, principles, or facts. It is about competing constructions of what counts as a violation in the first place.

DECLARATIONS

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Use of AI-Assisted Tools

The author used AI-assisted tools (Claude, Anthropic) during the preparation of this manuscript for drafting assistance, editorial refinement, and formatting. The author reviewed, critically evaluated, and takes full responsibility for the theoretical arguments, conceptual claims, and scientific content presented in this paper.

Conflicts of Interest

The author declares no conflicts of interest.

REFERENCES

- Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126(4), 556–574.
- Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, 15(10), 670–682.
- Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology*, 91(4), 612–625.
- Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.
- Bastardi, A., Uhlmann, E. L., & Ross, L. (2011). Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. *Psychological Science*, 22(6), 731–738.
- Brandt, M. J., Wisneski, D. C., & Skitka, L. J. (2015). Moralisation and the 2012 U.S. presidential election campaign. *Journal of Social and Political Psychology*, 3(2), 211–237.
- Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2), 353–380.
- Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., Celniker, J. B., & Zinger, J. F. (2019). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, 14(2), 273–291.
- Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63(4), 568–584.
- Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion*, 15(2), 115–141.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Row, Peterson.
- Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*, 132(1), 22–29.
- Gantman, A. P., & Van Bavel, J. J. (2015). Moral perception. *Trends in Cognitive Sciences*, 19(11), 631–633.
- Gottman, J. M. (1994). *What predicts divorce? The relationship between marital processes and marital outcomes*. Lawrence Erlbaum.
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47, 55–130.
- Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5), 1029–1046.
- Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, 23(2), 206–215.
- Heider, F. (1958). *The psychology of interpersonal relations*. Wiley.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498.
- Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, 25(2), 147–186.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220.

- Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, 28(4), 233–248.
- Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *Values and knowledge* (pp. 103–135). Lawrence Erlbaum.
- Rozin, P. (1999). The process of moralisation. *Psychological Science*, 10(3), 218–221.
- Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, 22(1), 32–70.
- Siegel, E. H., Wormwood, J. B., Quigley, K. S., & Barrett, L. F. (2018). Seeing what you feel: Affect drives visual perception of structurally neutral faces. *Psychological Science*, 29(4), 496–503.
- Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, 4(4), 267–281.
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1), 137–149.
- Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole.
- Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, 4(6), 479–491.
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, 92(4), 548–573.