

What Would an Empirical Counterexample to Motivational Internalism Require?

Voin Milevski

voinmilevski1@gmail.com

University of Belgrade

Research Article

Keywords: motivational internalism, amoralism, moral judgement, motivation, causal inference, underdetermination

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11047092/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

What Would an Empirical Counterexample to Motivational Internalism Require?

Abstract

An actual case in which sincere moral judgement occurs without corresponding motivation would refute an unrestricted psychological co-instantiation internalism, provided that the judgement can be identified without motivational accompaniment doing the classificatory work. The difficult question is evidential: what would justify describing an actual case as judgement without motivation rather than as judgement with weak, defeated, or masked motivation? This paper answers with an empirical counterexample criterion. The criterion requires independent warrant for the judgement and any restriction conditions; comparative support for absence over admissible masking models; sensitive, content-specific tests with prospective failure conditions; and temporal provenance linking any post-intervention motivation to the earlier state. Its first distinctive component is an admission rule under which masking must earn evidential standing rather than survive as a merely possible redescription. Its second is a temporal constraint distinguishing the revelation of a persisting motive from the activation of a prior disposition to form one or the generation of a new motive. Repeated sensitive null results can therefore favour absence even while masking remains logically possible. The criterion also yields an empirical-access boundary: psychologically substantive internalisms can answer to this evidence, whereas semantic, constitutive, and independently rationalist theses require different evidential routes. The result specifies both when an empirical amoralist has earned counterexample status and which internalisms that status could threaten.

Keywords: motivational internalism; amoralism; moral judgement; motivation; causal inference; underdetermination

1. Introduction

“Immoral—so what?” Foot’s amoral man can agree that some conduct is immoral and simply take no notice of that fact (1972, 310). A generation later, Svavarsdottir sharpened the challenge with the figure of a sophisticated moral cynic: an agent capable of sustained, serious moral reasoning who nevertheless remains unmoved by its conclusions (1999, 188–193). If either description is literally correct, motivational internalism appears to face exactly the kind of case it excludes. But there is a prior question: what would justify describing an actual agent in that way? Let J be the relevant moral judgement and M the corresponding motivation. For present purposes, represent the target co-instantiation claim schematically as $\Box(J \rightarrow M)$.¹ An actual counterexample must therefore warrant the description $J \wedge \neg M$. Writing down that conjunction is easy; earning both conjuncts from evidence is not. Suppose Clara sees an injured dog at the side of the road. She judges that she ought to stop, put the animal in her car, and take it to a clinic. She has time, money, and no competing obligation. Yet she walks past. Asked immediately afterwards, she sincerely reports that she felt no pull to help and had no motivation to stop.

Does Clara’s case refute motivational internalism? Only if two claims are true: that her avowal expresses the relevant moral judgement, and that no corresponding motivation is present. Motivational internalists deny that these claims can come apart in the required way, at least without further qualifications; externalists deny the necessity. Amoralists are meant to make the

¹ This notation does not decide between *de dicto* and *de re* readings, a distinction made explicit in Section 2

disagreement vivid. The empirical question is what would make the description true rather than merely stipulative.²

That question separates two uses of amoralist cases. The modal project stipulates, or asks us to conceive of, a genuine moral judgement without corresponding motivation and asks whether that combination is possible. The empirical project asks whether an actual or experimentally tractable agent instantiates the combination. The first turns primarily on conceivability and the individuation of moral judgement; the second turns on evidence. An empirical case must earn both parts of the description rather than stipulate them.

The modal logic fixes the stakes but not the paper's contribution. Many observations containing both *J* and *M* remain compatible with a merely contingent connection; one adequately identified actual bearer of *J* and not-*M*, by contrast, falsifies any unrestricted thesis that rules out that conjunction. The substantive problem is what evidence must establish before an actual case may responsibly receive that description. I answer with a single empirical counterexample criterion. Its paired-model component determines which masking hypotheses merit comparison; its temporal provenance component governs what interventions can show about an earlier state; and its scope conditions determine which internalisms the resulting evidence can threaten. Clara's profile is not irrelevant evidence. Inaction, absent felt pressure, and sincere denial all count. The difficulty is that this profile can arise from two different psychological organizations. On an absence model, there is no content-corresponding motive. On a masking model, such a motive exists, but its influence is blocked before it affects action, phenomenology, or self-

² I use 'moral judgement' as the paper's default term for the relevant first-person moral attitude. Nothing turns on choosing it over 'moral opinion' where the cited literature uses that expression. The important issue is how the target attitude is individuated, not the label attached to it.

ascription. The empirical problem is therefore not simply action versus inaction, but absence versus masked persistence.

Merely pointing out that masking is logically possible would be cheap. An internalist who can always posit an undetectable residual motive has made the thesis empirically invulnerable; an externalist who treats non-performance as absence assumes the crucial premise. The comparison must therefore be disciplined. I argue that a masking hypothesis earns evidential standing only when it has antecedent warrant independent of the verdict at issue, explains the dependencies in the observed profile, makes discriminating predictions, and accepts failure conditions. Evidence for a blocker, importantly, is not yet evidence for a motive.

Interventions create a second problem. Suppose Clara becomes responsive after a competing incentive is removed, the practical cost is lowered, attention is redirected, or an executive bottleneck is relieved. This is evidence of motivation after the manipulation, but not yet evidence that the motive was present beforehand. The intervention may reveal a persisting motive, activate a prior disposition to form one, generate a new motive, or change behaviour through a different pathway. I call the requirement that distinguishes these possibilities the provenance constraint.

The central claim is therefore not a collection of independent methodological recommendations. It is one criterion for empirical counterexample status. The paired-model test and temporal provenance constraint are its two evidential components; the grades introduced below mark increasing satisfaction of the criterion; and the empirical-access boundary states its scope. Applied to the evidence examined here, the framework yields limited but substantive verdicts: thin amoralist vignettes do not establish absence; the selected vmPFC findings do not

yet warrant counterexample status; and folk-attribution studies primarily measure classificatory practices rather than directly identifying the latent conjunction $J \wedge \neg M$.

The proposal extends an existing dialectic. Mele’s discussion of listlessness distinguishes loss of motivation from failures of motivational efficacy (1996), while Cholbi asks when actual depression, rather than a hypothetical description, could supply an empirical counterexample and stresses the need to compare moral with non-moral motivational deficits (2011, 28–45).

Bromwich shows how amoralist descriptions can import assumptions about genuine judgement and total motivational absence (2016, 456–463), and Suikkanen diagnoses a case-based standoff (2018). Prinz, approaching the issue from the other direction, offers empirical considerations in favour of an internal connection (2015). More recent work adds two complications: judgement–motivation covariance does not fix causal direction (Taylor 2026, 414), and the criteria for moral judgement may themselves be contested (Raskoff 2025, 12–14; Francén and Dernevik 2025).

What remains missing is a prospective rule for admitting and defeating masking explanations together with an account of when post-intervention motivation is evidence of motivational presence before the intervention. The framework developed here supplies those two elements and identifies the internalisms to which they apply.

The point can be put more sharply. Suikkanen observes that “internalists and externalists have different intuitions about the cases and they also describe them differently” (2018, 492). That is not merely sociological background. It identifies the evidential problem the paper addresses: a case cannot adjudicate between rival descriptions if the decisive psychological classification is simply built into the description at the outset.

I first specify what J and M must amount to for an empirical test to be informative. I then develop the absence–masking comparison and provenance constraint, apply them to vignettes,

vmPFC lesion cases, and folk-attribution studies, and formulate the criterion for counterexample status and its empirical-access boundary.

2. What an Empirical Counterexample Must Establish

The primary target is unrestricted, psychologically substantive co-instantiation internalism: the claim that sincere moral judgement necessarily co-occurs with at least some corresponding motivation, where the judgement can be identified without motivational accompaniment doing all the classificatory work. Internalist formulations differ over whether the necessity is conceptual or metaphysical, *de dicto* or *de re*, constitutional or non-constitutional, and whether rationality, normality, competence, agency, or other conditions restrict the antecedent. Those differences determine how directly empirical evidence can bear on the thesis (Asarnow and Taylor 2022, 139–157).

The framework has four principal domains of application. First, an unrestricted psychological co-instantiation thesis is directly vulnerable to a well-identified *J* without *M*. Second, a restricted thesis of the form $\Box((J \wedge R) \rightarrow M)$ is vulnerable only when *R*—relevant rationality, normality, competence, agency, or another condition—is independently warranted. Third, a *de dicto* or classificatory thesis may treat motivational absence as evidence that the state was not *J*, shifting the dispute to judgement individuation. Fourth, constitutive or independently rationalist theses require additional semantic, metaphysical, or rational evidence. The paired-model and provenance tests apply most directly to the first two domains; the latter two mark limits on what an empirical counterexample can establish.

This independence requirement matters for familiar restricted forms of internalism. Miller distinguishes unrestricted MI from versions restricted to normal, practically rational, or virtuous agents, and notes that a restriction such as “normal” cannot simply mean “an agent

whose moral judgements motivate” without making the thesis trivial (Miller 2008b, 233–235, 244). Thus *R* must have content and warrant that do not merely restate the motivational consequent.

The first difficulty arises before motivation is measured: what counts as *J* may itself be disputed. Francén and Dernevik distinguish constitutional from non-constitutional internalism; on the latter, they write that ‘an opinion does not count as a moral opinion unless the judge is motivated accordingly’ (2025, 5). But the evidential problem is older. Björnsson stresses that even sincere uses of ‘ought’ and ‘wrong’ do not by themselves identify a moral opinion and treats the issue explicitly as epistemological rather than ontological: identification must attend to the considerations given weight in deliberation, not merely to moral-sounding vocabulary (1998, 14–19). Svavarsdottir likewise argues that whether the sophisticated cynic makes genuine moral judgements cannot be settled by simply building suitable motivation into the mastery conditions for moral terms; the account of judgement must be defended on grounds that do not already decide the internalism dispute (1999, 188–193). The present framework therefore requires warrant for *J* that does not merely assume the disputed $J \rightarrow M$ connection. For de dicto views that make motivational accompaniment partly classificatory of *J*, evidence of motivational absence may therefore shift the dispute to the individuation of *J* rather than straightforwardly yielding a counterexample.

A recent transparency argument reinforces this caution from another direction. Raskoff argues that metaethicists often assume, without sufficient warrant, that the “psychological roles of moral terms or judgments” are transparent to competent speakers (2025, 1–2). If that is right, sincerity and conceptual competence are evidence about *J*, but they cannot by themselves settle *J*’s psychological role or its relation to *M*.

The second variable needs equal care. Let a motive corresponding to φ be a conative state with practical content favouring φ and a disposition to give that content some weight in deliberation, attention, planning, choice, effort, or action under independently specifiable conditions. This is weaker than successful guidance. A motive can be defeated, superfluous, phenomenally quiet, or merely regulative and still be motivationally real.³

This use of M is deliberately broader than “motivating reason” in Miller’s influential agential sense. Miller treats motivating reasons as considerations an agent takes to count in favour of action and distinguishes possessing such a reason from its being the reason for which the agent acts (2008a, 224–225). On that usage, first-person endorsement and a rationalizing role partly determine whether a consideration is a motivating reason. The present M instead tracks any content-corresponding conative organization with testable practical dispositions. Absence of an endorsed motivating reason therefore need not establish absence of M .

This breadth fixes the inferential target. The paper asks whether J and a content-corresponding M co-occur in the target token. It does not infer from co-occurrence that J caused M , that the agent acted for the moral consideration, or that motivation constitutes judgement. Content correspondence, causal dependence, and constitution must therefore remain distinct: a motive may correspond to a judgement without being caused by it, and may be caused by it without constituting it. Well-supported absence of broad M can threaten co-instantiation without vindicating a complete externalist theory or settling the metaphysics of moral judgement.

³ Björnsson and Francén distinguish guidance motivation from regulation motivation. Their motivating example stresses that ‘there is also a dispositional sense in which you are motivated’ even when the motivation does not causally affect intention formation (2026, sec. 1). I do not adopt their account wholesale. The point is narrower: failure to guide this action is not yet evidence that no corresponding motivational disposition existed.

This minimal characterization rules out an ‘epsilon motive’ introduced only to save a theory. A motive may be weak, defeated, unconscious, or causally idle in the target episode. It must nevertheless be disposed to make some content-sensitive difference under independently specifiable conditions. Selective attention, partial planning, willingness to incur a cost, latency, effort, and choice are possible manifestations. A state whose presence could make no difference under any such contrast has not been independently identified as a motive; it has merely been relabelled because the agent made a moral judgement.

Failure to act is compatible with motivation. Weak internalism requires only some desire-like state favouring the relevant action, not enough motivation to produce it (Lenman 1999, 441–443). Such motivation may instead shape dispositions, influence others, or guide future conduct (443). It may also remain ‘defeasible, though not cancellable’ (Corns and Cowan 2021, 76), or regulate reason-responsiveness without affecting the intention formed in the target episode (Björnsson and Francén 2026). These possibilities support plural, content-sensitive tests rather than identifying *M* with overt performance, while retaining testable counterfactual implications. The empirical burden is therefore double: warrant *J* without presupposing *M*, and warrant the absence of *M* through comparative evidence rather than surface appearance. The second task is harder because weak or defeated motivation is latent. The rest of the paper asks how evidence can discriminate absence from persistence without proving the motive in advance or allowing it to survive every failed test.

These distinctions delimit the conclusion. A case bears directly only on an internalism whose antecedent, classificatory conditions, and account of *M* it satisfies. Restricted theses require independent warrant for the relevant restriction; de dicto theses may contest whether the

state is *J*; idealized or virtuous-agent theses may exclude lesion pathology. The same evidence can therefore threaten one internalism while leaving another untouched.

‘Independent’ should not be confused with theory-free. No serious psychological attribution is theory-free. The weaker requirement is that the evidence doing the discriminating work have support beyond the verdict it is introduced to secure. This matters especially for *J*, because some forms of internalism make motivation part of the criterion for counting an attitude as a moral opinion, while pluralist accounts allow that boundary to vary. The evidence can test a connection only after its terms have been fixed without simply assuming the result.⁴

The empirical and modal projects are related but not interchangeable. An actual *J*-and-not-*M* case entails possibility only after both states have been warranted under defensible criteria.

Given an independent individuation of *J*, the present account specifies what would warrant not-*M* without building the contested connection into the classification.

Finally, falsification should not be confused with complete theory choice. One genuine exception can refute an unrestricted necessity thesis without vindicating every externalist theory. And one counterexample bearer is not one observation. Establishing that a single agent instantiates *J* and not-*M* may require repeated measurements, multiple tasks, interventions, controls, and longitudinal evidence. The token can be one; the evidence for the token need not be.

⁴ The qualification is not merely the generic point that evidence is theory-laden. Francén and Dernevik make the constitutive issue explicit: on one non-constitutional form of internalism, ‘an opinion does not count as a moral opinion unless the judge is motivated accordingly’ (2025, 5). Raskoff (2025, 12–14) supplies a different source of non-neutrality by challenging the transparency of psychological-role claims in metaethics.

3. Why the Familiar Negative Profile Does Not Identify Absence

Why not infer absence directly from Clara's three negative indicators? Because each measures something different. Non-performance bears on what won the practical conflict; felt pressure bears on phenomenology; denial bears on self-attribution. All are relevant, but none is an infallible inventory of the agent's motivational states. The point of the next three observations is not that these indicators are useless. It is that they need a causal interpretation before their conjunction can identify absence.

First, motives need not issue in action. An agent may judge that she ought to exercise and have some motivation to do so while fatigue, fear, habit, or a competing commitment wins the practical conflict. Behaviour reflects the resultant of a motivational organization, not a list of all motives contributing to it. Intentional action in favourable circumstances can therefore support a motive attribution, while inaction is generally less discriminating: a genuine motive may simply lose. This is one reason internalist theses have often been formulated in terms of some degree of motivation rather than guaranteed performance.

Second, phenomenal intensity does not transparently measure motivational strength. A vivid urge can be behaviourally weak; a standing commitment can be phenomenally quiet yet causally important. Motivational strength should not be identified with either felt intensity or behavioural outcome (Mele 1998, 23–36), and effort research likewise distinguishes potential motivation from the intensity actually mobilized in performance (Wright 2008, 682–701). A report of no felt pressure is therefore evidence, but it is not equivalent to the absence of every practical disposition corresponding to the judgement.

Third, motives need not be perfectly accessible to introspection. Sincere and attentive self-report is strong defeasible evidence, but agents can misclassify or overlook influences on

deliberation and action (see Nisbett and Wilson 1977, 231–259; Barnes 1997; Mele 2000). The relevant possibility need not involve a dramatic unconscious wish. It is enough that a weak or defeated disposition escapes first-person report while leaving content-sensitive traces elsewhere. Raskoff’s transparency argument supplies a useful metaethical caution: competence does not by itself guarantee accurate beliefs about the psychological roles of one’s moral judgements (2025, 12–14).⁵

The older literature already contains the two sides of the masking problem. Mele’s treatment of listlessness separates an agent’s moral judgement from the motivational organization that may be lost or rendered ineffective (1996, 727–753). Cholbi subsequently argues that actual depression would threaten internalism only if it selectively produced moral listlessness, and he appeals to contrasts with prudential motivation to assess that possibility (2011, 28–45). Listlessness may instead eliminate the normal inner and outer signs of motivation while a moral ‘optation’ remains present but functionally blocked, becoming manifest again when the wider motivational system recovers (Björnsson 1998, 120–121). Depression, exhaustion, and volitional obstacles can accordingly pose different challenges for internalism (Miller 2008b, 236–240). These accounts establish that surface non-response is ambiguous. They do not yet provide a general admission rule for masked-motive hypotheses or determine when later recovery evidences motivational continuity rather than reacquisition.

A longitudinal case makes the contrast especially clear. On Gert and Mele’s planet Alpha, first-person moral ought-beliefs persist after a catastrophic onset of “moral listlessness”

⁵ The possibility of unconscious or misdescribed motivation should not be treated as a licence to posit any hidden motive. It motivates the contrastive problem only. The masking hypothesis must still satisfy the admission and failure conditions in section 4.

removes moral motivation; continuity of propositional content before and after the episode makes continued judgement attribution especially intuitive (2005, 275–278). This contrasts with Björnsson’s use of later recovery as evidence that M may have persisted while functionally disabled. The dispute already has the structure of the paired models developed below: loss of motivation versus persistence under blockage.

Nothing here implies that a negative profile is worthless. Several fallible measures can jointly discriminate extremely well. What matters is whether their joint pattern distinguishes the rival organizations. Let E be the conjunction of non-performance, no reported pressure, and sincere denial. The absence hypothesis predicts E because no corresponding motive exists. A masking hypothesis can also predict E if an independently supported blocker suppresses the motive before it affects choice, phenomenal salience, or retrospective self-classification. The evidential question is how much more expected E is under absence than under an admissible masking hypothesis.

This is not merely an entailment question. A non-zero likelihood of E under masking does not show that E fails to favour absence, still less that absence and masking are equally probable. Priors, model complexity, background history, and the credibility of auxiliary structure all matter. The correct question is comparative: how much more expected is the observed profile under absence than under an admissible masking hypothesis? Compatibility defeats a claim of logical establishment; evidential parity requires much more.⁶

⁶ The likelihood notation is schematic. Nothing in the argument requires assigning numerical priors or likelihoods to mental-state models. Its purpose is to mark the difference between logical compatibility and comparative support: two models may both permit E while differing substantially in how expected E is given their independently supported structure.

The negative profile is especially limited because none of its components directly probes the counterfactual role on which the rival explanations disagree. If a content-corresponding motive exists but is blocked, it should affect some aspect of practical processing when the blocker is suitably varied while judgement and relevant incentives remain stable. If no such motive exists, there should be no effect attributable to a pre-existing content-corresponding state. This counterfactual difference provides the basis for discrimination.

4. The Paired-Model Test: When Masking Earns Evidential Standing

Return to Clara. Let $E = \{\text{non-performance; no reported pressure; sincere denial}\}$. The opening description fixes E but leaves open what produced it. We can now make the competing explanations explicit.

On the absence model, Clara has the moral judgement J but no content-corresponding motive m ; background preferences, habits, and situational attention explain why she walks past. On the masking model, she has J and a weak welfare-directed m , but a blocker b prevents that motive from influencing practical processing. To be more than a label for failure, b must have an independently specified psychological role.

Suppose Clara has a stable avoidance routine documented across non-moral helping contexts. When a task threatens open-ended delay, she disengages before forming an implementation plan and later classifies the episode as one in which she ‘never wanted to get involved.’ On the masking model, that routine blocks m before it becomes phenomenally salient and shapes retrospective self-ascription as well as action. The model thereby explains a dependency among the three elements of E rather than attaching a different excuse to each observation.

The models agree on the observed profile but disagree about a counterfactual. Hold J , its grounds, and relevant non-moral incentives as stable as the design permits, and vary b . If the motive exists, removing the blocker should reveal a content-specific difference in at least one intermediate practical variable—attention, planning, choice, latency, effort, or willingness to incur costs. If the motive is absent, the same manipulation should not produce a change attributable to a pre-existing content-corresponding state. This comparison is intentionally qualitative. Quantitative identification would require explicit measurement models and effect sizes; the philosophical point is the structure of a discriminating contrast.

This contrast uses familiar methodological resources for a specifically metaethical identification problem. Likelihood comparisons can favour one hypothesis over a specified rival without requiring a complete probability distribution over every possible alternative (Sober 2008, 7–10). Likewise, intervention is informative when changing a putative cause reveals a stable dependency rather than merely correlating with an outcome (Woodward 2003, 94–98). The novelty claimed here is not a new general theory of evidence or causation. It is the specification of what these comparative and interventionist constraints require when the disputed state is a minimal motive whose presence is invoked to protect a necessary judgement–motivation connection.

A masking model earns admission only if four conditions are met. First, independent evidence must support blocker b and its capacity to gate the relevant process. Second, there must be antecedent reason to predict content-specific m even when the disputed $J \rightarrow M$ connection is bracketed. Third, the model must specify in advance which observable or intermediate variable should change when b varies, in what direction, and why; not every element of E need change.

Fourth, it must specify outcomes that would reduce its support. A new masker introduced after every null is not an empirical rival.

Blocker admissibility and motive warrant must remain separate. A documented avoidance routine is evidence that Clara has a potential blocker, not that she has a dog-welfare motive. Conversely, a history of content-matched responsiveness can warrant predicting such a disposition without establishing that it was active in this episode. Requiring proof of m before admitting masking would trivialize the comparison; requiring no antecedent warrant would make masking costless to invent.

Suppose Clara has previously cared for injured animals and, before the target encounter, displayed content-matched signs of selective concern. Suppose further that the planning blocker can be manipulated independently—for example, because transport and waiting logistics have already been settled for unrelated reasons. In an otherwise matched case, the masking model predicts that animal-welfare content should now affect at least one practical variable. The absence model predicts no such effect from a pre-existing welfare-directed motive. Repeated failure under conditions in which the blocker is genuinely neutralized would therefore reduce the credibility of masking.

Model complexity cuts both ways. Any model that adds latent structure or auxiliary assumptions must repay that cost through explanatory or predictive success. Masking may often be more complex, but absence can also require a selective-loss mechanism. The comparison can therefore favour absence, preserve underdetermination, or favour masking.

This is where the framework improves on permissive redescription. Bromwich is right that an internalist need not accept every amoralist description at face value (2016). But

dialectical permission is not evidential standing. The paired-model test says when masking belongs in the serious comparison—and when further appeals to it should stop.

The burden is not merely defensive. Ratoff and Roskies make the point clearly in discussing psychopathy: an internalist who denies that an apparent amoralist makes genuine moral judgements “had better have some principled reason” independent of preserving the theory (2025, 64). The admission conditions extend that demand to blockers and residual motives.

5. Interventions and the Temporal Provenance of Motivation

The natural next step is intervention. If absence and masking differ over how the agent would respond when a blocker changes, manipulating the blocker appears to provide a discriminating test. Yet intervention creates a temporal non-entailment that the existing amoralism debate has not isolated clearly: evidence of motivation at t_2 does not by itself establish motivation at t_1 . The intervention may reveal a persisting motive, activate a prior content-specific disposition to form one, or generate a new motive. Because masking requires the first history, or a separately specified dispositional variant, the later response discriminates only when it carries evidence about the earlier motivational organization. This is the temporal provenance problem.

Three kinds of intervention should be distinguished. A revealing intervention targets an independently specified blocker while leaving the content and grounds of the moral judgement and candidate motive comparatively stable. A motive-producing intervention supplies a consideration, affective response, salience structure, or deliberative route capable of generating the motive. A non-specific behavioural intervention changes behaviour through convenience, reward, demand characteristics, coercion, habit, or general performance enhancement without

establishing a causal role for moral content. In Woodward's terms, only the first directly tests masked persistence; the others require controls that identify the pathway (2003, 94–98).⁷

The temporal contrast can be represented schematically. Under generation, not- $m(t1)$; intervention I at $t2$ causes $m(t2)$. Under revelation, $m(t1)$; I at $t2$ removes or weakens $b(t2)$, allowing the persisting motive to influence observable processing. Behaviour resembling m after I is compatible with both histories. The provenance constraint distinguishes them: a post-intervention response supports earlier masked motivation only if independent evidence links it to motivational organization already present before the intervention, rather than to motivation produced by the intervention itself.

Three possibilities must be distinguished. The motive may already be present at $t1$ but masked; the agent may instead have only a content-specific disposition to form it under a specified trigger; or the intervention may generate it through a pathway not antecedently present in content-specific form. The second is not the first. If an internalist makes the higher-order formation disposition the target, that disposition must itself have antecedent evidence and be distinguishable from a generic capacity to acquire new motives. Later motivation cannot by itself establish whichever earlier state the theory requires.

No single observation guarantees provenance. Persistence gains support when the judgement and its grounds remain stable, the manipulation independently targets the proposed blocker, content-matched traces precede intervention where feasible, non-moral controls estimate

⁷ 'Leaving the candidate motive intact' should be read comparatively, not as a demand for perfect psychological isolation. In realistic experiments the issue is whether the manipulation is better supported as changing the proposed blocker than as producing a new content-corresponding motive. Measuring J before and after the manipulation and using matched controls helps to constrain, rather than eliminate, alternative pathways.

generic compliance or effort, and distinct blocker-removing manipulations reproduce the response without sharing a motive-producing pathway. A formation disposition likewise needs an earlier content-specific signature—for example, selective attention to welfare-relevant information or incipient planning before the trigger. The aim is comparative control, not perfect psychological isolation.

Evidence strengthens through convergence. Stable judgement and matched controls improve on compliant action alone; a pre-intervention content-matched trace followed by the predicted response to *b* is stronger still; convergence across distinct blocker-removing interventions is stronger again. Each step reduces the ability of a generation model to explain the total pattern.

The same logic generates principled failure conditions. The masking model loses support when the putative blocker varies but preregistered content-sensitive outcomes do not appear. It also loses support when effects occur equally in matched non-moral tasks; when responsiveness emerges only after the intervention supplies a new moral reason or affective cue; or when no pre-intervention trace appears across heterogeneous measures and contexts in which *m* was predicted to matter. No context-free number of null results settles the issue. Measurement reliability, predicted effect size, and the heterogeneity of revealing conditions determine how diagnostic a null is.

The consequence is philosophically important: masking is not a permanent refuge. Repeated null results under interventions satisfying the provenance constraint can make absence predictively superior even while masking remains logically possible. In that situation, the externalist has not merely failed to find motivation. The practical role predicted by an admissible

masked-motive account has been tested and systematically failed to appear. That is the kind of evidence an actual counterexample requires.

Positive intervention results require equal discipline. If directing attention to suffering creates concern, reducing cost makes helping attractive for self-interested reasons, or social demand induces compliance, later action does not show that a corresponding motive persisted through the earlier episode. Evidence satisfying the provenance constraint supports no more than it identifies: a content-corresponding state at $t1$, not that J caused it or that the connection is constitutive or necessary.

The temporal provenance constraint should not be confused with causal-source internalism. It asks whether motivation evidenced at $t2$ was already present, or present in a specifically dispositional form, at $t1$. A different question asks whether the moral judgement produced or sustained that motivation. Miller's distinction between weak and strong MI makes the latter issue explicit: weak MI requires the necessary presence of some motivation, whereas strong MI adds the claim that the motivation has its source in the moral judgement itself (2008b, 234–235). Establishing temporal continuity does not establish this stronger causal provenance.

Measurement and manipulation should also be separated. A measure is useful when independently connected to the disputed practical role; an intervention is useful when it changes a proposed cause while leaving alternatives comparatively stable. Neural signals are especially risky because motivated choice decomposes into valuation, effort cost, cognitive control, and action selection. Without an established link to the content-sensitive role, either neural or behavioural correspondence may reflect a different realization, a generic deficit, a consequence rather than a constituent of motivation, or a desire to appear moral (Björnsson and Francén 2013, 6–10; Ratoff and Roskies 2025, 63–69).

The provenance constraint is diagnostic and prospective. Björnsson argues that restored functioning after listlessness can support persistence rather than creation anew, while distinguishing a strengthened optation, changed deliberative weight, and motivation added from another source (1998, 121–126). Together with Gert and Mele’s pre-listlessness evidence for continued *J*, this exposes the missing inference: history may make continued *M* tempting, but only evidence linking the later response to motivational organization present during the interval warrants continuity (Gert and Mele 2005, 277–281).

Preregistration has a philosophical rather than merely procedural role here. A masking hypothesis is vulnerable to a Duhem-style proliferation of auxiliaries: after a predicted signature fails, the theorist can redescribe the manipulation as insufficient, add another blocker, or relocate the motive to an unmeasured process. Fixing the blocker, manifestation measures, and failure conditions before the critical result does not eliminate auxiliary assumptions. It does, however, distinguish a model that risked error in advance from one reconstructed after the fact.⁸

Finally, discovery should be separated from confirmation. Exploratory searches can identify candidate traces or blockers, but stronger tests use those discoveries to predict held-out contexts, later time points, different tasks, or new cases. Counterexample status becomes more credible when the classification survives that separation.

6. Three Evidential Bases, Three Different Verdicts

The framework’s value depends on whether it changes how evidence is assessed. Vignettes, lesion cases, and folk-attribution experiments provide three tests of its discriminating value.

⁸ Preregistration does not solve the Duhem problem. A failed prediction can always motivate legitimate revision. Its value here is evidential bookkeeping: it records which blocker, manifestation, and failure criteria belonged to *M* before the critical outcome was known, making post hoc repair visible rather than impossible.

6.1. Amoralist vignettes

A thin vignette typically describes a competent and sincere agent who judges that she ought to φ , does not φ , reports indifference, and denies motivation. Such cases can be used in the modal project by stipulating total absence and asking whether the combination is coherent. Nothing in the present framework decides that conceivability question. Its target is the evidential use of the vignette: the inference from the narrative features to the claim that total motivational absence actually obtains.

On that evidential reading, a thin vignette does not yet contain a discriminating test. Nor does the bare internalist reply that ‘perhaps some motivation remains’ establish masking. To earn standing, the masking proposal needs an independently supported blocker, antecedent evidence for a content-specific disposition, a revealing condition, and a failure rule. If the vignette supplies only E , absence may be more parsimonious even though masking remains logically possible. If it adds a documented blocker but no evidence for m , the blocker does not supply the missing motive. If it supplies both and predicts a signature that later fails, the case can move towards absence. The framework ranks competing descriptions by evidential support rather than placing them on a par.

This clarifies the relation to Bromwich. Her critique undermines the idea that a surface description simply forces the internalist to concede an amoralist (2016, 456–463). The paired-model test accepts that diagnostic point but adds a termination condition. Once the internalist proposes a hidden motivational organization, the proposal must earn and retain evidential standing; failed manifestations cannot always be redescribed as further masking.

Svavarsdottir’s Patrick case provides a useful illustration. Patrick agrees that helping a politically persecuted stranger is his moral obligation, yet says that he “cannot be bothered to

help” and later shows no regret (Svavarsdottir 1999, 176).⁹ That makes the case vivid, but vividness does not decide whether the correct explanation is absence, defeated motivation, or a failure of genuine moral judgement. The paired-model test asks what additional evidence would discriminate those readings rather than treating one of them as part of the vignette’s surface description.

Tresan’s treatment of Patrick brings out a further distinction. Traditional amoralist intuitions often concern whether an unmotivated state still counts as a moral belief at all; Tresan argues that such intuitions support, at most, *de dicto* internalism (2006, 148–149). That classificatory question is distinct from the paired-model question addressed here: once *J* is warranted, does the evidence favour absence of *M* over masked persistence?

Miller’s volitional-impossibility strategy is a harder test because it aims to preserve moral judgement while establishing total motivational absence. In his adoption case, the mother continues to judge that giving up her child would be morally best but endorses the contrary aversion and treats the moral consideration as practically powerless (Miller 2008b, 238–240). The case can support local absence of an endorsed motivating reason. What it cannot establish from the vignette alone is that independently admissible persistence models fail under conditions designed to reveal the putative motive; that further achievement requires the prospective evidence characteristic of Grade 3 and, together with the remaining conditions, counterexample status.

⁹ Bromwich discusses the case in detail (2016, 461–462).

6.2. vmPFC lesion cases

The vmPFC literature is best treated as a stress test for the framework, not as an already realized exemplar of a decisive empirical counterexample. Lesion cases can provide temporal, mechanistic, and convergent evidence unavailable in a vignette, but the central variables remain contested: whether moral judgement is preserved, whether motivational impairment is content-specific, and whether broader changes in valuation, affect, effort, or executive integration explain the profile. Roskies' cases initiated precisely this dispute (2003, 51–66; 2006, 617–627), and competing readings remain live (see Cholbi 2006, 607–616; Smith 2007, 207–215; Leary 2017, 878–895; Ratoff and Roskies 2025, 65–66).¹⁰

Stronger vmPFC evidence would first support *J* without relying on motivational behaviour as its sole criterion. Repeated endorsement, inferential integration, generalization, reason-giving, comprehension, confidence, and stability before and after injury could triangulate the judgement attribution. Motivational measures could then span attention, planning, costly choice, effort, and action, with matched non-moral tasks controlling for generic deficits. No measure is theory-neutral or decisive; the point is convergence across indicators that do not themselves presuppose *M*.

This proposal does not settle the Roskies-Cholbi-Smith-Leary dispute by stipulation. It makes the disagreement empirically sharper. If patients retain stable endorsement, inferential role, comprehension, and generalization while motivational signatures disappear selectively, the

¹⁰ The vmPFC literature has generated exactly the kind of classification dispute at issue here. Roskies (2003, 51–66; 2006, 617–627) treats the cases as evidence for preserved moral belief with motivational disruption; Cholbi (2006, 607–616), Smith (2007, 207–215), and Leary (2017, 878–895) offer importantly different interpretations. Ratoff and Roskies (2025, 65–66) summarize the continuing disagreement.

judgement attribution gains support independently of overt motivation. If those judgement-sensitive capacities degrade in parallel, the case no longer supports the required *J*-and-not-*M* description. The same battery can therefore count against either reading rather than being designed to rescue one side. The battery is a triangulation strategy, not a sufficient operational definition of moral judgement: its point is to reduce dependence on any single disputed indicator while leaving room for substantive theories of *J* to challenge the package.

Accordingly, the vmPFC findings considered here should not be presented as already satisfying the counterexample standard. Garr's value-centric analysis emphasizes the possibility that task complexity, attention, and integration failures can mimic a specifically moral deficit (2024, 974–982). Lockwood and colleagues provide stronger lesion evidence for alterations in prosocial motivation, including effort-sensitive effects, but their results are component- and subregion-sensitive rather than a direct measure of motivation corresponding to a particular moral judgement (2024, 1403–1416). These findings show how mechanistic evidence improves on a thin profile while also exposing how demanding identification remains; they do not support a verdict about every strand of the wider vmPFC literature.

A future lesion or longitudinal case could nevertheless favour absence. If *J* remained well supported under independently defensible criteria, multiple sensitive content-matched measures repeatedly returned null effects, plausible generic deficits were controlled, and antecedently warranted masking models failed under heterogeneous revealing conditions, absence could become substantially better supported than masking. The inference would remain fallible, as psychological state attribution generally is. But fallibility does not reduce it to logical compatibility: a merely possible residual motive need not remain an evidentially competitive explanation.

A pattern meeting these conditions would be philosophically significant. Suppose a patient repeatedly expresses and reasons with a specific moral judgement under independently defended criteria, while pre-lesion behaviour and cognitive measures show a content-corresponding practical disposition. Suppose focal damage then selectively eliminates that disposition across multiple content-matched measures while leaving the judgement intact, and explanations involving general effort, comprehension, confidence, or valuation perform comparatively worse. That pattern would make the absence model predictively superior. It would not prove that all internalisms are false, since restricted or constitutive theses may individuate their scope differently. But it could constitute an actual counterexample to an unrestricted necessity claim. The framework therefore raises the standard while specifying how it could be met.

6.3. Folk-attribution experiments

Experimental philosophy asks whether ordinary speakers attribute moral belief or understanding to agents described as rational but unmotivated. Nichols found 85 percent willingness to attribute moral understanding to a psychopathic agent (2004, 74–75), whereas a word-for-word replication reported 48 percent and left rationality and residual motivation incompletely controlled (Björnsson et al. 2015; Ratoff and Roskies 2025, 67–69). Such instability shows that attribution is sensitive to variables a discriminating design must separate. These studies illuminate concepts and classificatory practices; they do not directly identify the described agent’s latent motivational organization.¹¹

¹¹ This is one reason folk-attribution results should not be read straightforwardly as measurements of the metaphysics of moral judgement. They can instead reveal which cues ordinary attributors treat as constitutive,

Some variation may reflect conceptual pluralism rather than noise: competent speakers can differ over whether unmotivated agents count as having moral opinions (Francén 2010, 120–125, 129, 145–148). The framework need not endorse that interpretation. It does require experiments to distinguish classificatory policy from evidence about the agent’s latent psychological organization.

A useful design would therefore vary evidence for genuine judgement, corresponding motivation, and a blocker independently. If motivation is constitutive of folk judgement attribution, removing motivational evidence should depress *J*-attribution even when sincerity and competence are secured. If motivation is merely a diagnostic cue, strong independent evidence for *J* should restore attribution; if blocker information matters, its effect should depend on antecedent evidence for a content-specific motivational trace. Such a design would distinguish attribution policies, not directly settle the metaphysical relation between *J* and *M*.

7. The Empirical Counterexample Criterion

The preceding analysis can now be stated as one criterion. An actual bearer warrants counterexample status against an unrestricted psychological co-instantiation internalism only when the following conditions are jointly satisfied:

- (1) Judgement warrant: *J* is supported under criteria that do not simply infer judgement from the motivational connection under test.
- (2) Model admission: absence is compared with masking models only when those models have antecedent warrant for both the proposed blocker and a content-corresponding motive.

sincerity-diagnostic, or defeasible in the presence of a blocker. Raskoff’s (2025, 12–14) transparency argument adds a further warning against inferring psychological structure directly from competent speakers’ classifications.

- (3) Discriminating test: admitted masking models face sensitive, content-specific tests with prospectively stated predictions and failure conditions, and absence outperforms them on the total evidence.
- (4) Temporal provenance: motivation observed after an intervention supports earlier masked presence only when the response is linked to motivational organization already present at the target time, rather than to the activation of a prior disposition to form a motive or the generation of a new one.
- (5) Scope warrant: any restriction condition *R* in the targeted internalism is independently supported, and the interpretation of *J* and *M* matches the version of internalism claimed to be threatened.

Satisfying these conditions warrants epistemic counterexample status in the intended sense: the evidence supports treating the actual bearer as *J* and not-*M* for the purpose of evaluating the target thesis. This is a comparative threshold, not certainty or the exhaustive defeat of every imaginable masking model. It does not provide deductive proof of psychological absence, settle every theory of moral judgement, or vindicate externalism as a complete metaethical position.

Clara's case illustrates how the criterion changes the verdict as evidence accumulates. Her initial inaction, absent felt pressure, and sincere denial make absence compatible with the profile, but masking remains possible; this is Grade 1. If a content-matched history supports a weak welfare-directed motive and an independently documented avoidance routine could block its influence, masking becomes an admitted rival and the evidence may yield only local support for one model; this is Grade 2. If the blocker is successfully varied and the preregistered content-sensitive traces repeatedly fail to appear while matched controls establish sensitivity, absence can outperform the serious persistence model; this is Grade 3. If *J* remains independently well

supported across those tests, alternative blockers lose support under heterogeneous revealing conditions, and any relevant restriction R is also warranted, the bearer may reach Grade 4: counterexample status.

The grades therefore record different evidential achievements rather than degrees of rhetorical confidence. Grade 1 establishes compatibility; Grade 2 gives local comparative support over specified rivals; Grade 3 requires prospective tests on which admitted masking models risk failure; and Grade 4 adds independently warranted J , relevant scope conditions, and an overall comparison strong enough to treat the bearer as J and not- M . The evidence is graded even though the modal consequence for the targeted necessity thesis is discrete.

The criterion does not mechanically privilege the simpler verbal description. Background evidence for a blocker or standing disposition can make masking credible despite a negative profile; repeated failures of the predictions that warranted masking can make continued persistence costly. A masking model that accommodates every outcome retrospectively may fit the data while earning little credibility: antecedent plausibility, model complexity, measurement reliability, and independent auxiliary evidence remain relevant to the comparison. In likelihood terms, both models may permit the evidence while differing in how expected it is given their independently supported structure (Sober 2008, 7–10).

A prospective study should accordingly specify before the critical outcome the target judgement, the relevant form of corresponding motivation, independently supported blockers, predicted manifestations when each blocker is reduced, and outcomes that would count against persistence. It should measure J across intervention, include content-matched and non-moral controls, and distinguish behavioural success from intermediate processing. A null is evidentially strong only when the test had a good chance to detect the effect predicted by the masking model;

this parallels the general idea that negative results acquire force through the severity or sensitivity of the test they survive (Mayo 2018, 83–86). Strength comes from convergent constraints on rival causal organizations, not from any single measure.

A possible amoralist, probable motivational absence, and an empirical counterexample are thus different achievements. The first is modal, the second comparative and psychological, and the third satisfies the full criterion and permits the modal consequence for the targeted thesis. Keeping them apart prevents both an easy externalist victory and an unfalsifiable internalism.

8. Objections

8.1. The framework makes counterexamples impossibly demanding

One objection is that motivational absence can never be observed directly, so demanding causal contrasts and provenance rigs the debate for internalism. The standard would indeed be unfair if an internalist could preserve *M* simply by insisting that some unmanifested disposition remains logically possible. That is not the proposal. Empirical attribution of latent psychological states is ordinarily fallible and comparative. Absence can be the best-supported explanation even when an unobserved state cannot be deductively excluded. The relevant question is whether serious alternatives were antecedently warranted, tested where they predicted a content-sensitive difference, and weakened when those predictions failed.

Counterexample status is epistemic rather than deductive: it requires evidence strong enough to warrant treating *J* and not-*M* as the best-supported description of an actual bearer. Although the counterexample bearer can be single, earning that description may require a substantial body of repeated, controlled, and longitudinal evidence. A masking hypothesis that remains only logically possible after losing the predictions that earned its admission does not block this inference. Nor does any null suffice merely because it follows an intervention. It is

informative only if the design could detect the predicted effect and if the intervention genuinely targeted the proposed blocker.

Psychological absence hypotheses are therefore not permanently inferentially disadvantaged. We routinely infer that a disposition is absent when sensitive opportunities for its manifestation fail under conditions in which serious rivals predict it. The distinctive requirements here are that the opportunities be content-sensitive and that temporal provenance be controlled, because generic action or prosociality can arise through pathways that do not instantiate the motive at issue.

8.2. A theory of motivation must be chosen first

A second objection grants the evidential analysis but argues that the real dispute is theoretical. A cognitivist externalist may individuate moral judgement independently of motivation; a constitutive internalist may build a motivational role into the relevant judgement. If one framework is independently correct, the empirical comparison seems secondary. This is partly right. No observational method settles every semantic or constitutive dispute.

The result is nevertheless nontrivial. The empirical counterexample project is often invoked precisely because actual patients, psychopaths, or experiments are supposed to pressure a modal thesis from outside the disputed theory. For that pressure to be independent, the case must not simply inherit the externalist's individuation of judgement and motive. But no empirical procedure can, by itself, settle a constitutive dispute over what counts as genuine moral judgement. Francén and Dernevik make the point vivid by distinguishing constitutional from non-constitutional internalism: on the latter, motivation can function as a conceptual restriction on what counts as a moral opinion even when the underlying mental-state kind can exist without motivation (2025, 4–6). The paired-model test therefore states what a theory-comparative

evidential argument must do once independently defensible criteria for *J* are in place; it does not promise a theory-neutral identification of *J*.¹²

Tresan's *de dicto/de re* distinction sharpens this boundary (2006, 145–149). A *de dicto* internalist may treat absent conation as evidence that a content-equivalent state is not a moral belief; on a *de re* reading, the moral-belief state itself necessarily has conation. Ridge (2015) prefers the latter. The present argument need not adjudicate between them: absence may shift a *de dicto* dispute to the individuation of *J*, whereas it bears directly on *de re* accompaniment only once the relevant state is independently identified.

The same boundary recurs in hybrid and pluralist accounts. Historical or evolutionary function may help individuate moral judgement despite absent token motivation (Bedke 2009, 193–202), and competent speakers may apply the concept with different motivational restrictions (Francén 2010, 129–145; Francén 2020, 360–361). Token behaviour and case intuitions cannot by themselves settle these questions. Attribution experiments can investigate classificatory policies; identifying latent psychological states requires the comparative causal evidence developed here.

Rationalist internalism marks a different limit. One response to the cynical hypothesis derives motivational requirements from enkratic constraints on deliberating agents (Ridge 2019, 845–850); the opposing reply is that those principles themselves depend on a posteriori assumptions about the psychological function of reason judgements (Björnsson and Francén

¹² This is also why the present framework should not be presented as a theory-neutral test of every internalism. If a version of internalism makes motivation part of the criterion for counting an attitude as a moral opinion, the identification of *J* cannot be settled by behavioural measurement alone. The empirical project becomes conditional on a defensible individuation of the target judgement.

2026, 170–193). The present framework does not decide that dispute. It prevents the negative profile from silently supplying either the rationalist principle or the externalist individuation that remains independently contested.

8.3. The minimal motive is too weak

A more serious objection charges circularity. The paper characterizes motivation through a disposition to make some content-sensitive practical difference under independently specifiable conditions, and then treats repeated failures of such differences as evidence of absence. Does the framework simply define away any motive that cannot be detected? The answer requires distinguishing two questions: what motivation metaphysically is, and what could justify attributing its presence or absence in an empirical counterexample. The framework offers an evidential criterion for the second question, not a reductive analysis of the first. It does not claim that every possible motive must manifest on the measures used here, or that sufficiently many nulls deductively entail not-*M*. It claims that a motive invoked as an empirical explanation must carry some independently testable counterfactual commitments if evidence is to favour its presence rather than merely leave it possible.

This yields an empirical-access boundary, but the boundary is methodological rather than definitional. An independently specifiable condition is one described and motivated without using the disputed verdict that *m* is present: for example, a documented reduction in implementation cost, removal of a measured competitor, change in attentional load, or manipulation of an executive bottleneck whose role is supported outside the target case. The condition should be specified before the critical outcome and should connect to the proposed manifestation mechanism independently of whether the motive is later detected. A theorist is free to posit motivation that has no such discriminating consequences. The point is not that this posit

fails to count as motivation; it is that behavioural, introspective, lesion, and intervention evidence of the present kind cannot confirm or disconfirm it.

Different internalisms answer to different evidence because they incur different commitments. Psychological versions that posit weak, defeasible, or regulative practical effects remain testable when those effects have discriminating manifestation conditions (Corns and Cowan 2021, 76; Björnsson and Francén 2026). Semantic or classificatory theses that make motivation partly criterial for moral judgement shift the dispute to the individuation of *J*; independently rationalist theses require arguments about rationality rather than a search for a psychological trace. Mixed views inherit mixed burdens. The boundary tracks what could count for or against each thesis, not a metaphysical definition of motivation.

The formation-disposition case illustrates the point. If moral judgement is said only to dispose an agent to acquire a motive under a content-specific trigger, the disposition is empirically accessible only if something distinguishes it from a generic capacity to acquire new motives. Without such a signature, later motivation cannot verify that the disposition was present beforehand. The access boundary constrains inference rather than legislating ontology.

8.4. The result is just ordinary underdetermination

The final objection is that the proposal merely imports ordinary causal methodology into metaethics. The framework openly relies on familiar ideas: intervention evidence concerns stable dependencies (Woodward 2003), likelihoods compare specified rivals (Sober 2008), and failed predictions are informative only when the test was capable of detecting the predicted effect (Mayo 2018). If the paper said only that competing explanations should be tested, the objection would succeed. The framework therefore claims no novelty for these ideas considered separately.

The contribution lies in showing that admission and provenance must be satisfied together in empirical amoralist cases. Masking must earn entry through antecedent warrant and predictive commitments, while motivation elicited by the ensuing test must be linked to the earlier state rather than produced by the test itself. Without admission, masking is costless; without provenance, the evidential significance of a successful intervention is temporally ambiguous. Their conjunction determines how negative and positive results can differentially support absence or persistence in the target episode.

Earlier work supplies important parts of this picture. Mele and Cholbi clarify what listlessness and depression would need to show. Björnsson rejects ad hoc internalist accommodation and allows that sufficiently clear actual cases could count against internalism (1998, 118–120); Suikkanen diagnoses a case-based stalemate (2018, 492); and Raskoff warns against assuming that competent speakers transparently know the psychological role of moral judgement (2025, 1–2). None of these points alone determines which masking hypotheses merit evidential standing or whether a post-intervention response reveals persistence rather than formation. The present proposal connects those questions and makes their failure conditions explicit.

The resulting burdens are symmetrical in the relevant sense. Surface non-response cannot simply become not-*M*; serious persistence hypotheses must be compared. But mere logical possibility cannot preserve masking indefinitely; it must retain the antecedent warrant and predictive success that admitted it. Temporal provenance adds that later motivation supports earlier motivation only when the continuity is earned.

9. Conclusion

An empirical counterexample has unusual force against an unrestricted psychological co-instantiation thesis: once the target judgement is independently warranted as J , one actual bearer with J and no corresponding M refutes a thesis that rules out that conjunction. That modal point is familiar; the neglected problem is evidential. Inaction, weak phenomenology, and sincere denial may support absence, but none constitutes absence or settles whether the state qualifies as J on every internalist account.

The evidence must therefore be comparative. J must be identified without simply presupposing the connection under test. Masking must have antecedent warrant, predict content-sensitive dependencies, and risk losing support when those predictions fail. If intervention elicits motivation, later responsiveness supports earlier presence only when temporal provenance links it to motivational organization already present at $t/$. A disposition merely to form a motive is a different target and matters only if the internalist thesis independently requires it.

Masking is neither a permanent refuge nor something null results deductively eliminate. Repeated sensitive nulls under revealing conditions can favour absence while residual motivation remains logically possible. Conversely, a motivational posit with no independently testable counterfactual commitments is not refuted by such nulls, but neither is it confirmed by failed searches. Applied to the evidence examined here, the standard explains why thin vignettes do not establish absence, why the selected vmPFC findings do not yet warrant counterexample status, and why folk-attribution studies do not directly identify $J \wedge \neg M$. The resulting counterexample, if achieved, would bear directly on psychologically substantive co-instantiation internalism; other internalisms remain answerable to the distinct evidential routes identified above.

The aim is not to show that amoralist cases can always be redescribed. It is to specify when a redescription has earned evidential standing, when it has lost it, and when the resulting absence verdict reaches counterexample status. That is what an empirical counterexample to unrestricted psychological co-instantiation internalism would have to earn.

References

- Asarnow, S., and D. E. Taylor. 2022. Judgment internalism. In *The Oxford Handbook of Moral Psychology*, ed. M. Vargas and J. M. Doris, 139–157. Oxford: Oxford University Press.
<https://doi.org/10.1093/oxfordhb/9780198871712.013.8>.
- Barnes, A. 1997. *Seeing through self-deception*. Cambridge: Cambridge University Press.
- Bedke, M. S. 2009. Moral judgement purposivism: saving internalism from amorality.
Philosophical Studies 144: 189–209. <https://doi.org/10.1007/s11098-008-9205-5>.
- Björnsson, G. 1998. *Moral Internalism: An Essay in Moral Psychology*. Doctoral dissertation, Stockholm University.
- Björnsson, G. 2002. How emotivism survives immoralists, irrationality, and depression.
Southern Journal of Philosophy 40: 327–344. <https://doi.org/10.1111/j.2041-6962.2002.tb01905.x>.
- Björnsson, G., and R. Francén. 2013. Internalists beware—we might all be amorality!
Australasian Journal of Philosophy 91 (1): 1–14.
<https://doi.org/10.1080/00048402.2012.665373>.
- Björnsson, G., and R. Francén. 2026. Reason judgments—political or motivational? In *Oxford Studies in Metaethics*, vol. 21, ed. R. Shafer-Landau, 170–193. Oxford: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780197903612.003.0008>.

- Björnsson, G., J. Eriksson, C. Strandberg, R. Francén, and F. Björklund. 2015. Motivational internalism and folk intuitions. *Philosophical Psychology* 28 (5): 715–734.
<https://doi.org/10.1080/09515089.2014.894431>.
- Björklund, F., G. Björnsson, J. Eriksson, R. Francén, and C. Strandberg. 2012. Recent work on motivational internalism. *Analysis* 72: 124–137. <https://doi.org/10.1093/analys/anr118>.
- Bromwich, D. 2010. Clearing conceptual space for cognitivist motivational internalism. *Philosophical Studies* 148: 343–367. <https://doi.org/10.1007/s11098-008-9331-0>.
- Bromwich, D. 2016. Motivational internalism and the challenge of amorality. *European Journal of Philosophy* 24: 452–471. <https://doi.org/10.1111/ejop.12053>.
- Cholbi, M. 2006. Belief attribution and the falsification of motive internalism. *Philosophical Psychology* 19 (5): 607–616. <https://doi.org/10.1080/09515080600901939>.
- Cholbi, M. 2011. Depression, listlessness, and moral motivation. *Ratio* 24 (1): 28–45.
<https://doi.org/10.1111/j.1467-9329.2010.00481.x>.
- Corns, J., and R. Cowan. 2021. Moral motivation and the affective appeal. *Philosophical Studies* 178: 71–94. <https://doi.org/10.1007/s11098-020-01421-2>.
- Foot, P. 1972. Morality as a system of hypothetical imperatives. *Philosophical Review* 81 (3): 305–316. <https://doi.org/10.2307/2184328>.
- Francén, R. 2010. Moral motivation pluralism. *Journal of Ethics* 14: 117–148.
<https://doi.org/10.1007/s10892-010-9074-y>.
- Francén, R. 2020. Reconsidering the meta-ethical implications of motivational internalism and externalism. *Theoria* 86: 359–388. <https://doi.org/10.1111/theo.12245>.
- Francén, R. 2023. Finding wrong. *Mind* 132: 493–523. <https://doi.org/10.1093/mind/fzac041>.

- Francén, R., and J. Dernevik. 2025. Disagreement for pluralists. *Inquiry*. Published online 28 April 2025. <https://doi.org/10.1080/0020174X.2025.2497344>.
- Gert, J., and A. Mele. 2005. Lenman on externalism and amorality: An interplanetary exploration. *Philosophia* 32 (1–4): 275–283. <https://doi.org/10.1007/BF02641624>.
- Garr, A. K. 2024. The role of the ventromedial prefrontal cortex in moral cognition: a value-centric hypothesis. *Philosophical Psychology* 37 (4): 970–987. <https://doi.org/10.1080/09515089.2023.2166820>.
- King, A. 2018. The amoralist and the anaesthetic. *Pacific Philosophical Quarterly* 99 (4): 632–663. <https://doi.org/10.1111/papq.12225>.
- Leary, S. 2017. Defending internalists from acquired sociopaths. *Philosophical Psychology* 30 (7): 878–895. <https://doi.org/10.1080/09515089.2017.1336529>.
- Lenman, J. 1999. The externalist and the amoralist. *Philosophia* 27 (3–4): 441–457.
- Lockwood, P. L., J. Cutler, D. Drew, A. Abdurahman, D. S. Jeyaretna, M. A. J. Apps, M. Husain, and S. G. Manohar. 2024. Human ventromedial prefrontal cortex is necessary for prosocial motivation. *Nature Human Behaviour* 8 (7): 1403–1416. <https://doi.org/10.1038/s41562-024-01899-4>.
- Mayo, D. G. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.
- Mele, A. 1996. Internalist moral cognitivism and listlessness. *Ethics* 106 (4): 727–753. <https://doi.org/10.1086/233670>.
- Mele, A. 1998. Motivational strength. *Noûs* 32: 23–36. <https://doi.org/10.1111/0029-4624.00085>.
- Mele, A. 2000. *Self-Deception Unmasked*. Princeton: Princeton University Press.

- Miller, C. B. 2008a. Motivation in agents. *Noûs* 42 (2): 222–266. <https://doi.org/10.1111/j.1468-0068.2008.00679.x>.
- Miller, C. B. 2008b. Motivational internalism. *Philosophical Studies* 139: 233–255. <https://doi.org/10.1007/s11098-007-9115-y>.
- Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgement*. Oxford: Oxford University Press.
- Prinz, J. 2015. An empirical case for motivational internalism. In *Motivational Internalism*, ed. G. Björnsson, C. Strandberg, R. Francén Olinder, J. Eriksson, and F. Björklund, 61–84. Oxford: Oxford University Press.
- Nisbett, R. E., and T. D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84 (3): 231–259. <https://doi.org/10.1037/0033-295X.84.3.231>.
- Raskoff, S. Z. 2025. The transparency fallacy in metaethics. *Synthese* 206: article 69. <https://doi.org/10.1007/s11229-025-05162-9>.
- Ratoff, W., and A. L. Roskies. 2025. Moral motivation. In *The Cambridge Handbook of Moral Psychology*, ed. B. F. Malle and P. Robbins, 55–77. Cambridge: Cambridge University Press.
- Ridge, M. 2015. Internalism: Cui bono? In *Motivational Internalism*, ed. G. Björnsson, C. Strandberg, R. Francén, J. Eriksson, and F. Björklund, 135–149. Oxford: Oxford University Press.
- Ridge, M. 2019. Internalists relax: we can’t all be amorlists! *Philosophia* 47: 845–850. <https://doi.org/10.1007/s11406-018-0010-x>.

- Roskies, A. 2003. Are ethical judgements intrinsically motivational? Lessons from ‘acquired sociopathy’. *Philosophical Psychology* 16: 51–66.
<https://doi.org/10.1080/0951508032000067743>.
- Roskies, A. 2006. Patients with ventromedial frontal damage have moral beliefs. *Philosophical Psychology* 19 (5): 617–627. <https://doi.org/10.1080/09515080600901947>.
- Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.
- Smith, M. 2007. The truth about internalism. In *Moral Psychology*, vol. 3: The Neuroscience of Morality, ed. W. Sinnott-Armstrong, 207–215. Cambridge, MA: MIT Press.
- Sober, E. 2008. *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.
- Strandberg, C. 2013. An internalist dilemma, and an externalist solution. *Journal of Moral Philosophy* 10: 25–51. <https://doi.org/10.1163/174552412X625754>.
- Suikkanen, J. 2018. Judgment internalism: an argument from self-knowledge. *Ethical Theory and Moral Practice* 21: 489–503. <https://doi.org/10.1007/s10677-018-9923-5>.
- Svavarsdóttir, S. 1999. Moral cognitivism and motivation. *Philosophical Review* 108: 161–219.
- Taylor, L. 2026. On the correlation between moral belief and motivation. *Journal of Value Inquiry* 60: 413–427. <https://doi.org/10.1007/s10790-024-09982-9>.
- Tresan, J. 2006. De dicto internalist cognitivism. *Noûs* 40 (1): 143–165.
<https://doi.org/10.1111/j.0029-4624.2006.00604.x>.
- Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Wright, R. A. 2008. Refining the prediction of effort: Brehm's distinction between potential motivation and motivation intensity. *Social and Personality Psychology Compass* 2: 682–701. <https://doi.org/10.1111/j.1751-9004.2008.00093.x>.