

Would You Defend This Decision in Public?

A Person × Situation Theory of Prospective Publicity and Research Integrity

Kevin L. Michel

Saint Lucia Centre for Psychological Science

Theory and Review | Working Paper

Version 1.0 | August 2026

<https://doi.org/10.5281/zenodo.21861184>

Abstract

Research integrity is often attributed to flawed individuals or flawed systems, leaving under-theorized the decision episode in which a researcher exercises consequential discretion. This article develops the Prospective Publicity-Integrity Model (PPIM), a person $\times$ situation theory of prospective answerability in scientific self-regulation. Its proposed contribution is not a new publicity test, but a discriminating architecture. Before implementing a focal choice, the Prospective Publicity Heuristic asks researchers to identify affected constituencies, articulate alternatives, uncertainty, and current evidence, and consider whether those reasons could withstand evaluation by people with relevant standing. The model separates reason generation from identifiability, procedural from psychological timing, simulated from credible institutional answerability, and a private cue from an optional retained-record protocol. It predicts a psychological route from issue recognition through self-standard accessibility and self-critical reasoning, plus a distinct protocol route that constrains retrospective reconstruction. Ten falsifiable propositions specify when this architecture should improve material-element completeness, event-report congruence, disclosed deviation, voice, and correction, and when it may instead produce compliance theater, conformity, concealment, privacy loss, or chilled exploration. Existing evidence supports components, not the complete model. The article therefore offers a cumulative research program, not an adoption mandate.

Keywords: research integrity, prospective publicity, accountability, moral identity, personality states

Would You Defend This Decision in Public? A Person × Situation Theory of Prospective Publicity and Research Integrity

Rules and investigations matter, but they can miss the decision point at which a researcher excludes, discloses, revises, claims, or escalates. The PPIM specifies a candidate intervention for that middle: structured reason generation before a contestable choice becomes the record. Crucially, the model predicts when answerability may improve candor, when it may only improve appearances, and when punitive or overly broad review may make conduct worse. It offers a framework for testing integrity without equating transparency with virtue or sorting researchers into fixed good and bad types.

Integrity failures do not always begin with an intention to deceive. One plausible pathway begins when a researcher concludes that expertise or good intentions make ordinary scrutiny unnecessary. A protocol is treated as clumsy, an oversight body as slow, a null result as misleading, an exclusion rule as too rigid, or a junior colleague's objection as naive. The researcher may be intelligent, conscientious in many domains, and sincerely committed to discovery. Yet at one consequential decision point, superior knowledge becomes a warrant for unilateral exemption. The method changes, the deviation goes unrecorded, and the explanation is written only after the outcome is known.

This is the maverick scientist problem. It is not synonymous with independence, creativity, or principled dissent. Science requires all three. The problem begins when confidence in one's judgment is used to make a decision less inspectable precisely because that decision is contestable. The resulting conduct ranges from defensible exploration, through questionable research practices (QRPs), to deliberate fabrication, falsification, or plagiarism. Those categories differ morally and institutionally, but they share a psychological question: what makes a

standard, stakeholder, or alternative course of action psychologically present or absent when a researcher chooses?

Two familiar explanations answer at different levels. A “bad actor” account locates failure in dishonest traits, motives, or values. A “bad system” account locates it in publication pressure, hierarchy, weak supervision, and distorted rewards. Each captures part of the evidence. Neither, by itself, explains why the same person may be scrupulous in one decision and self-serving in another, why a strong norm can fail under one configuration of pressure and ambiguity, or why identical oversight can produce candor in one laboratory and concealment in another. Personality science offers a useful bridging unit of analysis: behavior as an expression of persons in situations, with stable differences shaping distributions of momentary states rather than dictating every act (Fleeson & Jayawickreme, 2015; Tett & Burnett, 2003).

This article develops the Prospective Publicity-Integrity Model (PPIM), a theory of how anticipated answerability may alter a decision-level integrity-regulation episode before the focal choice is implemented. “Prospective” refers to the implementation of that choice, not necessarily to the start of a study: an exclusion, claim, authorship decision, or correction can arise after results are known. The model distinguishes procedural timing from psychological timing: a cue delivered before implementation may nevertheless arrive too late if the actor is already committed to a preferred course. The model’s practical cue is simple:

Before implementing the focal choice, and preferably before becoming psychologically committed to it, name the choice and the constituencies materially affected. When more than one interest is materially implicated, include at least two constituencies. In a short structured entry, state the material

alternatives, uncertainty, and evidence available now. Could you defend those reasons to people with relevant standing?

Publicity and defensibility tests are established ethical heuristics (Bok, 1999; Lewis & Gilman, 2005); the PPIM does not claim to originate them. Its proposed contribution is to convert a familiar test into a discriminating person $\times$ situation model that separates structured reason generation from identifiability, procedural timing from psychological commitment, and simulated answerability from bounded institutional review, while specifying observable benefits and harms. In the core PPH, reason generation is addressed to one or more legitimate constituencies. Plurality is a hypothesized enhancer when several interests are implicated, not a definitional requirement. Public does not mean automatic release to an unlimited audience. The heuristic is a triage device, not an ethical verdict.

The article first defines the decisions to be explained and the evidentiary limits of the review. It then links personality traits and achievement motives to within-person integrity-regulation episodes, identifies the mechanisms through which prospective publicity may work, and distinguishes reason generation from visibility alone. Ten propositions specify timing, audience, person, situation, record, and implementation effects. A four-stage, multi-study research program tests the model with behavioral choices, independent transfer measures, and contemporaneous records rather than confession alone. The central claim is not that sunlight disinfects. It is that carefully structured answerability may change what a person notices and does before there is something to disinfect, while poorly structured answerability may instead produce conformity, fear, or a more polished form of concealment.

What Must Be Explained

A continuum with morally important boundaries

Research misconduct is commonly defined narrowly as fabrication, falsification, or plagiarism, excluding honest error and differences of opinion (Office of Research Integrity, n.d.); under current U.S. Public Health Service rules, a finding also requires a significant departure from accepted practice and intentional, knowing, or reckless conduct (42 C.F.R. §§ 93.103, 93.234). QRPs occupy a wider and less stable category: selective exclusion, undisclosed flexibility, inappropriate authorship, inadequate supervision, incomplete reporting, or other practices whose defensibility depends on context, disclosure, rationale, and consequence. Responsible exploration differs again. An unplanned analysis can be informative when labeled exploratory; a protocol deviation can be justified by participant welfare when disclosed and reviewed; disagreement with an institutional rule can be principled when challenged through an accountable process. Integrity cannot therefore be measured by mechanical conformity alone.

The empirical prevalence literature reinforces both the importance of the problem and the need for restraint. A meta-analysis estimated pooled self-admission of at least one form of research misconduct at 2.9% and QRPs at 12.5%, with higher estimates for witnessed behavior (Xie et al., 2021). In a national Dutch survey of 6,813 researchers, randomized-response estimates for fabrication and falsification were each slightly above 4%, while approximately half of respondents reported frequent engagement in at least one surveyed QRP (Gopalakrishna et al., 2022). These numbers should not be treated as fixed population parameters. Definitions, reference periods, sampling frames, response rates, and elicitation techniques differ substantially; self-report mixes memory, norm interpretation, and willingness to confess. A 2026 scoping review of 229 studies concluded that the literature measures severe violations more readily than

routine, relational, and structurally produced integrity problems (Nazarovets, 2026). The robust conclusion is not a precise rate. It is that integrity-relevant behavior extends beyond rare adjudicated fraud and that current measurement is uneven.

The PPIM targets *research-integrity decisions*: choices in design, data collection, analysis, reporting, authorship, supervision, participant protection, data stewardship, or correction that materially affect truth-tracking, fairness, welfare, or answerability. Within that category, *integrity-critical discretion* exists when a researcher has more than one plausible action, relevant standards are ambiguous or competing, important information is private, and the decision may be difficult to reconstruct later. This is where normal scientific judgment and ethical risk overlap. The model predicts less about accidental mistakes, knowledge deficits, conduct under direct observation, or organized fraud carried out by actors already committed to evasion.

Table 1 establishes the conceptual boundaries. A justified departure from a plan is not misconduct merely because it is a departure; its integrity depends on contemporaneous reasons, appropriate review, and truthful disclosure. Conversely, procedural compliance is not sufficient evidence of integrity if the record conceals what occurred. The distinction between conduct and its representation is central because a conscientious person may produce complete paperwork, a strategically manipulative person may produce persuasive paperwork, and neither record is useful unless it corresponds to behavior.

No single indicator is treated as integrity. Event-report congruence indexes the accuracy of a later account relative to an independently logged event. Prospective record-implementation correspondence indexes whether conduct followed the contemporaneously stated choice; a departure is not counted as failure when it is promptly disclosed and justified. Completeness indexes coverage of prespecified material elements, not length. Process quality, welfare, and fairness outcomes remain separate. Support for the PPIM therefore requires a prespecified pattern across outcomes, not improvement in paperwork alone.

Table 1
Core Constructs and Boundaries

Construct	Definition	Not equivalent to
Scientific misconduct	Fabrication, falsification, or plagiarism; under U.S. PHS rules, a finding also requires significant departure and intentional, knowing, or reckless conduct	Honest error, differences of opinion, or every procedural breach
Questionable research practice	Integrity-relevant practice in a gray zone whose defensibility depends on context, disclosure, rationale, and consequence	An automatically unethical act or a harmless minor choice
Research-integrity decision	Choice that materially affects truth-tracking, fairness, welfare, or answerability in the research process	Every routine technical choice
Integrity-critical discretion	A research-integrity decision with plausible alternatives, incomplete observability, and contestable standards or interests	Scientific autonomy itself
Prospective publicity	Anticipation that reasons for a focal choice could be evaluated by specified legitimate constituencies before implementation	Indiscriminate public exposure or automatic data release
Simulated answerability	Mental consideration of how one or more specified constituencies could evaluate the reasons for a choice	Actual attribution, review, or enforceable accountability
Credible institutional answerability	A genuine prospect that a retained rationale will receive bounded, attributable review	An imagined audience or generic surveillance threat

Construct	Definition	Not equivalent to
Prospective Publicity Heuristic (PPH)	Brief preimplementation cue, preferably before psychological commitment, that combines simulated answerability with structured reason generation	A verdict, signature, or generic ethics reminder
Prospective publicity protocol	Optional institutional layer adding a retained record, bounded access, credible review sampling, feedback, and escalation	The internally rehearsed PPH alone or continuous surveillance
Integrity-regulation episode	Decision-bounded configuration of trait states, issue recognition, self-standard accessibility, reasoning, power, and willingness to disclose or seek review	A unitary trait state or global judgment of character
Representational legitimacy	Relevant standing, competence, plurality when several interests are implicated, and independence from the preferred outcome	Popularity, seniority, or agreement with the researcher
Review legitimacy	Representational legitimacy plus procedural fairness, contestability, correction rights, and an appeal route	Harshness, exposure, or review probability alone
Event-report congruence	Accuracy of a later account relative to an independently logged event	Agreement between a prospective plan and later implementation
Prospective record-implementation correspondence	Whether conduct followed the contemporaneously stated choice, interpreted jointly with timely disclosure and justification of deviations	Automatic evidence of integrity or misconduct
Transparency climate	Shared expectation that consequential decisions are documentable and challengeable, and that good-faith disclosure and correction are safe	A high volume of forms or punitive monitoring

Evidentiary approach and inference boundaries

This article is a problem-driven integrative review, not a systematic review or meta-analysis. It combines four evidence streams selected for explanatory leverage: research on scientists' self-reported conduct and climates; personality-process and moral-identity theory; experimental and review evidence on accountability; and evaluations of prospective research

practices such as preregistration and Registered Reports. These streams do not carry equal inferential weight.

Direct evidence involving researchers is often cross-sectional and self-reported. Accountability research offers experiments but usually in judgment, organizational, public-administration, or laboratory settings rather than consequential scientific work. Registered Reports and open-practice policies provide field analogues of prospective constraint and inspectability, but most comparisons are nonrandomized and target publication processes more directly than participant protection, authorship, or supervision. Finally, neither the PPH nor the prospective publicity protocol specified here has been tested as a package. Table 2 therefore distinguishes observations from transports and proposals. Statements about expected effects are predictions. Conditional recommendations in the governance section are risk-control implications, not demonstrated intervention benefits.

Table 2*Evidence Architecture of the PPIM*

Evidence layer	What it supports	Main limitation
Applied ethics publicity and defensibility tests	The heuristic family predates the PPIM and is used in professional decision making	Does not establish a psychological mechanism or effect in scientific work
Researcher surveys and observational studies	Integrity behavior varies with traits, goals, pressure, norms, and climate	Self-report, selection, common-method variance, and weak causal identification
Accountability experiments and reviews	Timing, audience knowledge, reason generation, and evaluative context can change judgment	Most tasks are not scientific decisions; effects are heterogeneous and can reverse
Moral-identity and personality-process research	Stable differences can shape situation-contingent state expression	Integrity-specific state measures remain underdeveloped
Preregistration, Registered Reports, and open-practice reforms	Prospective records and review can constrain outcome-dependent flexibility and improve inspectability	Policy selection and historical confounding; records may be vague, unchecked, or performative
PPIM, PPH, and protocol	Integrated predictions about persons, situations, mechanisms, benefits, and harms	The complete cue-protocol architecture is presently untested

Why retrospective oversight leaves a decision gap

Formal ethics review is indispensable for participant protection. Two U.S.-focused reviews published in 2011 found little direct evaluation against participant-protection outcomes and documented burdens and variability without corresponding outcome evidence (Abbott & Grady, 2011; Silberman & Kahn, 2011). A later U.S. Government Accountability Office review likewise called for federal agencies to develop approaches for examining IRB effectiveness (U.S. Government Accountability Office, 2023). These sources document an evaluation problem; they do not establish that IRBs are ineffective or dispensable. Approval at one point also cannot carry the full burden of integrity across later amendments, exclusions, analyses, reporting choices, and

corrections. Prospective publicity is proposed only as a complement at those decision points, never as a substitute for required review.

Integrity as a Repeated Personality Process

From traits to states

Whole Trait Theory describes traits both as distributions of states and as explanatory systems that generate those states (Fleeson & Jayawickreme, 2015). On this view, calling a researcher conscientious does not mean that every exclusion, authorship discussion, or correction is handled carefully. It means that careful states are more likely across occasions, while situational demands and goals continue to shift behavior. Trait activation theory similarly proposes that situations supply cues relevant to the expression of particular traits (Tett & Burnett, 2003). Strong situations - those with clear, consistent, constrained, and consequential expectations - can compress between-person variability; weak situations permit dispositions and local goals greater influence (Meyer et al., 2010).

Research integrity is well suited to this framework because science distributes consequential discretion across thousands of small choices. The same analyst may be meticulous when a plan is explicit and collaborative review is expected, then outcome-focused when an unexpected pattern threatens months of work. The same principal investigator may welcome correction from a peer while dismissing it from a dependent trainee. A person-level measure obtained once cannot explain these reversals without a theory of state activation.

The PPIM uses *integrity-regulation episode* for the broader decision-level configuration; it does not treat that configuration as one trait state. Its components include momentary honesty and carefulness, issue recognition, accessibility of moral identity and role-relevant standards, self-critical reasoning, felt power, and willingness to expose uncertainty or seek review. These

components can covary but need not. Conscientiousness may predict detail and completeness while having an ambiguous relation to whether the account is candid. Concern for reputation may increase care when review is fair yet increase concealment when the audience is punitive. Moral identity may matter most when a choice is first recognized as moral rather than coded exclusively as technical or strategic. Stable traits are expected to shape distributions of their corresponding momentary states; the episode is the context in which those states interact with motives and cues.

Person inputs: differentiated, not profiled

Direct personality evidence in research-integrity settings is limited. In a cross-sectional study of 535 Dutch biomedical researchers, Machiavellianism was positively associated with self-reported research misbehavior, whereas narcissism, psychopathy, and self-esteem were not significant in the main models; Machiavellianism alone explained only 1.6% of variance (Tijdkink et al., 2016). Among 217 psychology researchers, appearance-approach goals - trying to demonstrate competence - predicted greater self-reported QRP engagement, while learning-approach goals predicted less; the appearance association was stronger when publishing was valued over rigor (Janke et al., 2019). These studies are informative precisely because they are narrow. They do not justify screening scientists by personality, causal claims, or a stable “misconduct profile.” They suggest that strategic self-advancement and performance goals deserve study alongside climate.

The PPIM focuses on five differentiating tendencies. First, Honesty-Humility should predict unwillingness to exploit others and greater event-report congruence. A reanalysis of 16 incentivized cheating studies ($N = 5,002$) found that a one-standard-deviation increase in Honesty-Humility was associated with substantially lower odds of dishonesty (odds ratio = .53),

while other basic traits showed little incremental validity (Heck et al., 2018). The paradigm was not scientific work, so direct tests in researchers remain necessary. Second, conscientiousness should predict record completeness, prospective planning, and rule maintenance, but not necessarily candor; the same reanalysis found evidence against a conscientiousness main effect on cheating, and diligent people can pursue a distorted goal with impressive care. Third, moral-identity centrality should make a cue to “the kind of scientist I am” more behaviorally consequential, especially when the ethical character of a choice is not already obvious (Aquino et al., 2009; Aquino & Reed, 2002). Fourth, Machiavellianism and appearance-focused achievement goals are predicted to produce a gap between the polish of an account and its event-report congruence. This gap is a novel prediction; existing researcher studies establish associations with self-reported misbehavior, not with independently verified reporting outcomes. Fifth, public self-consciousness and fear of negative evaluation distinguish attention to the self as an object of others’ judgment from more private self-focus (Fenigstein et al., 1975; Leary, 1983). Transporting those constructs to integrity-critical decisions, the PPIM predicts that evaluation sensitivity will amplify anticipated-audience effects, yielding either care or concealment depending on fairness and safety.

This differentiation prevents a common theoretical shortcut. More accountability may produce the same length of decision record through different processes: careful documentation, truthful self-examination, impression management, or fear. The PPIM therefore treats material-element completeness, event-report congruence, prospective record-implementation correspondence with disclosed deviations, decision quality, voice, correction, and experienced burden as distinct outcomes.

Situation inputs: pressure, ambiguity, power, and climate

Research systems shape which states become likely. Publication pressure has been associated with self-reported research misbehavior and QRPs in cross-sectional surveys, although measures vary and causal direction remains unresolved (Gopalakrishna et al., 2022; Tjldink et al., 2014). A positive organizational research climate is associated with more desirable and fewer undesirable self-reported practices (Crain et al., 2013). Scientific norm subscription appears protective in national survey data, while perceived detection and fair procedures may alter the local payoff structure (Gopalakrishna et al., 2022). Recent reviews emphasize that integrity problems are simultaneously individual, relational, and institutional, and warn against measuring only spectacular violations (Nazarovets, 2026).

Four situational dimensions are central to the PPIM. *Ambiguity* determines whether a choice can be framed as ordinary judgment rather than an integrity issue. *Temptation and pressure* determine the value of a preferred outcome and the cost of delay, disclosure, or correction. *Power* determines who can question whom and whether a record protects a junior researcher or supplies a supervisor with another instrument of control. *Climate* determines whether transparency is interpreted as shared quality control or as prosecutorial surveillance. Meta-analytic evidence links psychological safety with information sharing and learning behavior (Frazier et al., 2017), while organizational-justice evidence distinguishes procedural fairness and relates it to trust and authority evaluations (Colquitt et al., 2001). These relations anchor the climate constructs but do not establish the PPIM interaction. Audience structure and review probability cut across all four: an unknown but legitimate audience can provoke self-criticism, whereas a known partisan audience can provoke conformity; a record that no one ever samples can become ritual.

The person and situation are not competing explanations. Appearance goals are more dangerous when local rewards prize striking findings and review is weak. Moral-identity centrality has little opportunity to matter when the decision is framed as purely logistical. Power is not expected to corrupt uniformly: field and experimental evidence outside science suggests that subjective power can increase self-interest when moral identity is weak and reduce it when moral identity is strong (DeCelles et al., 2012). Evaluation sensitivity is protective when correction is treated fairly and risky when disclosure invites humiliation or retaliation. The PPIM therefore starts with an interactionist premise: integrity-regulation episodes arise from dispositions and motives operating through the cues, constraints, and meanings of a particular decision.

How Ordinary Discretion Becomes Self-Exemption

The maverick problem is not best understood as a categorical personality. It is one candidate *self-exemption pathway*, not an explanation of most misconduct. A valued objective narrows attention; an institutional requirement is redescribed as an obstacle; superior expertise becomes evidence that external review adds no value; and nondisclosure is framed as protection of the science rather than deception. This sequence is proposed rather than inferred from any single deviation. Case analyses of adjudicated misconduct have identified recurring explanations that resemble parts of it, including production and recognition pressure, insecure position, poor supervision, denial of harm, preserving a presumed truth, and incremental escalation (Davis et al., 2007). Case files cannot establish causal prevalence, but they show why the psychology of justification deserves more than a footnote.

Issue recognition and ethical fading

Ethical fading occurs when the moral dimensions of a decision recede while a business, technical, or strategic frame dominates attention (Tenbrunsel & Messick, 2004). In research, a participant-welfare question can be experienced as a recruitment problem, selective reporting as narrative clarity, and resistance from a trainee as a project-management obstacle. The conduct need not begin with an explicit decision to violate a norm. What changes first is the description under which the choice is made.

This framing account explains why additional rules may fail at the moment of action. A standard cannot guide a decision that is not retrieved as relevant. Two prominent generic-reminder interventions also failed in high-powered tests. A 25-laboratory Registered Replication Report recruited 5,786 participants; its preregistered primary meta-analysis included 19 laboratories ($n = 4,674$) and found no reduction from a Ten Commandments reminder (Verschuere et al., 2018). Across five conceptual replications ($n = 4,559$) and one preregistered direct replication ($n = 1,235$), signing first did not reduce dishonesty (Kristal et al., 2020; see also “Correction for Kristal et al.,” 2026). These results do not show that salience never matters. They show that a generic ethical cue or signature is not equivalent to making the structure of a particular choice inspectable.

The PPH is designed as an issue-recognition cue because it requires the actor to name a choice, alternatives, uncertainty, evidence, and audience. Each element is diagnostic. If no alternative can be stated, the response may reveal premature closure; when the protocol is used, the record preserves that response. If the explanation changes when participants or a skeptical methodologist enter the imagined audience, a previously excluded claim has become visible. If a researcher cannot describe what was known before the result, hindsight may already be

contaminating the account. Content written because the prompt explicitly requires it is a manipulation response, not evidence of mediation or intervention success. Self-critical reasoning must therefore be measured through an independent process trace, unprompted transfer decision, response to new counterevidence, or another outcome not mechanically required by the cue.

Motivated reasoning and moral disengagement

Motivated reasoning does not require conscious fabrication of beliefs. People recruit and evaluate reasons in ways constrained by preferred conclusions while retaining a sense of objectivity (Kunda, 1990). Moral disengagement supplies a complementary account of how conduct is reconciled with the self: harmful action is euphemistically labeled, responsibility is displaced or diffused, consequences are minimized, and affected people are discounted (Bandura, 1999; Moore et al., 2012). A meta-analysis links moral disengagement with workplace misconduct and shows incremental prediction beyond several dark personality traits, although much of the literature remains correlational and methodologically heterogeneous (Ogunfowora et al., 2022).

In scientific decisions, the preferred conclusion may be epistemic (“the effect is real”), personal (“I am a successful scientist”), relational (“my supervisor must be protected”), or institutional (“the grant cannot fail”). The proposed self-critical route should operate most strongly when the cue precedes psychological commitment. After commitment, the same request may elicit defensive bolstering, although retrospective reflection may still improve disclosure or correction. Accountability research has repeatedly distinguished predecision conditions that encourage self-critical thought from postdecision conditions that encourage bolstering, and audiences with unknown views from audiences whose desired answer is clear (Lerner & Tetlock, 1999; Tetlock, 1983; Tetlock & Kim, 1987). A systematic review of 266 accountability

experiments found favorable, null, and adverse effects, underscoring that accountability is a configuration, not a dose (Aleksovska et al., 2019).

Identity accessibility and evaluative sensitivity

Moral identity predicts behavior modestly rather than perfectly. A meta-analysis of 111 studies estimated an association of approximately $r = .22$, with stronger relations in designs relying heavily on self-report (Hertz & Krettenauer, 2016). The modesty is theoretically useful: identity must be accessible, applicable, and supported by the situation. Experiments show that situational cues can change whether moral-identity centrality is expressed in behavior (Aquino et al., 2009). In research settings, moral identity can make role-relevant standards self-defining: “a responsible scientist reports uncertainty,” “a mentor protects a trainee’s ability to dissent,” and “an author corrects the record.” This claim does not assume that every professional norm is morally correct.

The PPIM distinguishes moral-identity accessibility from professional-norm accessibility. A researcher may retrieve the norm that responsible scientists disclose uncertainty without experiencing that norm as self-defining, and may experience moral traits as self-defining without retrieving the relevant scientific standard. The proposed pathway therefore measures these constructs separately and tests whether their conjunction is especially consequential.

Prospective publicity is hypothesized to increase moral-identity accessibility because attribution makes the self the subject of the explanation. Anticipated guilt or regret may make harm more immediate; anticipated loss of credibility may change the payoff of concealment; and justified pride may make correction compatible with competence. The PPIM treats these anticipated emotions and reputational concerns as amplifiers, not necessary links in the core causal sequence. Reputation can support integrity when the audience rewards candor and

appearance management when the audience rewards flawless results. The model predicts better outcomes when moral-identity accessibility and self-critical reasoning explain behavior beyond evaluative anxiety alone.

Contemporaneous records and the reconstruction problem

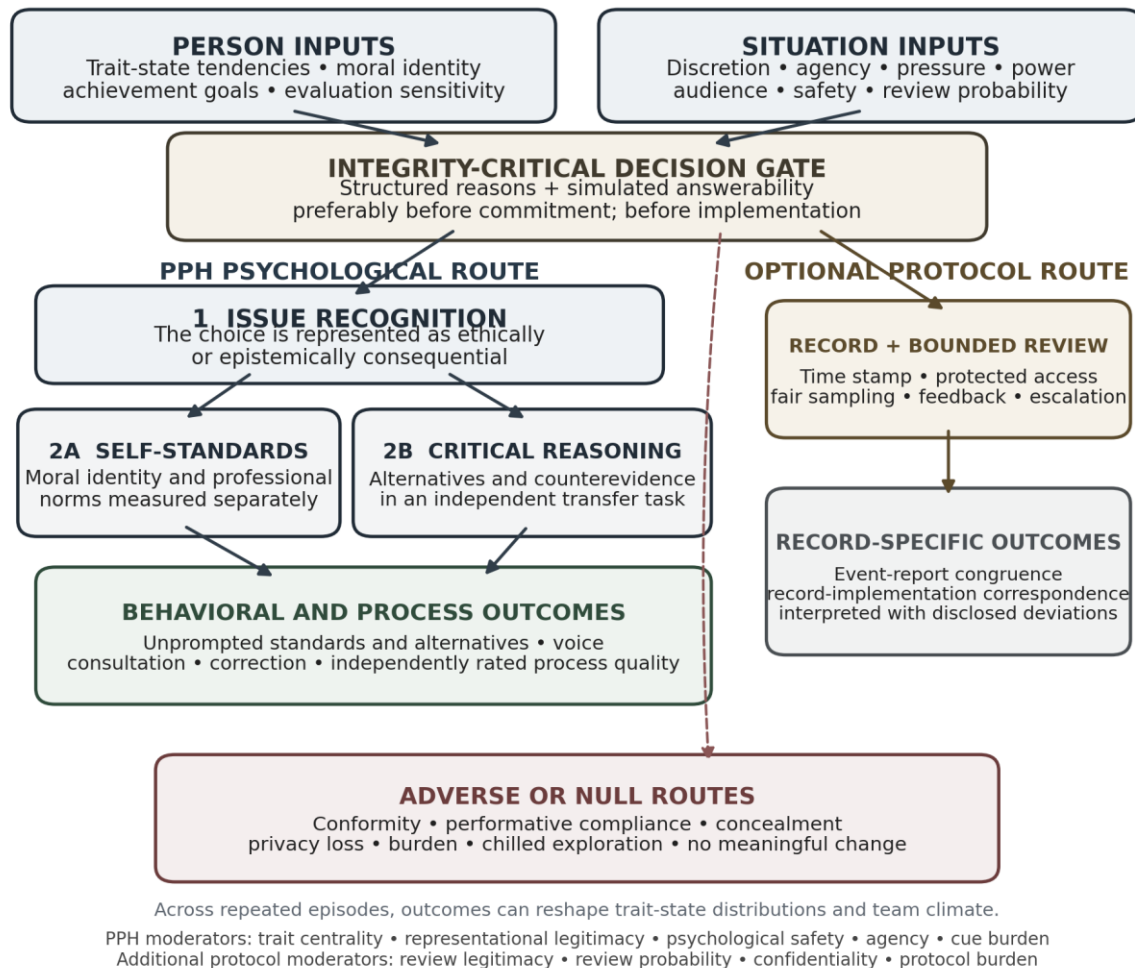
Many integrity systems depend on records created after results are known. That timing creates a reconstruction problem even without intentional deception. Outcome knowledge can alter perceived foresight and the remembered relevance of prior evidence, while a defensible exploratory choice can be redescribed as confirmatory (Fischhoff, 1975; Kerr, 1998). A time-stamped rationale cannot eliminate motivated reasoning, but it creates a comparison point between intention, action, and later report. It makes deviations discussable rather than forbidden and limits the ability to substitute an outcome-consistent history for the uncertainty that existed at the decision point.

Open-science reforms provide useful analogues. In a comparison of 71 Registered Reports with 152 standard psychology articles, 44% of first hypotheses in Registered Reports were supported, compared with 96% in standard reports (Scheel et al., 2021). In a blinded comparison, 353 researchers reviewed article pairs drawn from 29 Registered Reports and 57 non-Registered-Report comparison papers. They rated the Registered Reports higher on methodological rigor, analytic rigor, and overall quality, with little apparent novelty penalty (Soderberg et al., 2021). Neither study randomly assigned manuscripts, so topic, author, journal, and selection differences remain. These nonrandomized comparisons nevertheless illustrate the relevant design feature: the question and method are made reviewable before outcomes determine publication.

Preregistration also reveals the limits of a timestamp by itself. Structured formats restrict researcher degrees of freedom more than unstructured plans, yet both can remain underspecified (Bakker et al., 2020). In an audit of the first 27 studies awarded a preregistration badge by *Psychological Science*, 25 contained at least one deviation, only one disclosed every deviation, and nine disclosed none (Claesen et al., 2021). Deviation is not the failure; concealed or unreconstructable deviation is. These findings motivate a PPH that requires alternatives and uncertainty, and an implementation that includes fair sampling and feedback rather than equating the existence of a record with its truth.

The Prospective Publicity-Integrity Model

The PPIM has two intervention layers and two outcome routes. The PPH is a reason-generating cue delivered at an integrity-critical decision gate. It is hypothesized to increase issue recognition and then, through parallel self-standard-accessibility and self-critical-reasoning paths, influence behavioral and process outcomes even when no external record is created. The optional protocol surrounds that cue with a retained contemporaneous record, bounded access, review sampling, feedback, and escalation. The protocol adds a reconstruction-constraint route and changes the credibility and meaning of anticipated review; it is not a consequence of the psychological mediators. Person inputs shape trait-state tendencies and motives, while situation inputs shape the decision's meaning, incentives, and constraints. Representational legitimacy, psychological safety, agency, and cue burden may moderate the PPH route. Review legitimacy, review probability, confidentiality, and protocol burden additionally moderate the institutional route. Repeated episodes may shift trait-state distributions and team climate in either direction: fair feedback may support candor, while unreviewed forms may become ritual and punitive review may teach people to reveal less.

Figure 1*The Prospective Publicity-Integrity Model*

Note. Person and situation inputs shape an integrity-critical decision gate. The PPH is hypothesized to prompt issue recognition and then parallel self-standard-accessibility and self-critical-reasoning paths that can influence behavior without an external record. The optional protocol adds a retained-record and bounded-review route with record-specific outcomes. Route-specific moderators influence whether the architecture yields candor, performative compliance, concealment, or no meaningful change. Solid paths are theoretical predictions, not established causal effects of the complete architecture.

Cue and protocol: two levels of decision architecture

The PPH is a brief psychological cue: simulated answerability plus generation of reasons before the focal choice is implemented and, ideally, before psychological commitment. It can remain internally rehearsed for low-risk decisions. In that form, specified constituencies are

mentally represented, but no one can actually attribute or review the account. The prospective publicity protocol is the optional institutional layer: a retained record, tiered access, credible review sampling, developmental feedback, and an escalation route. Its minimum record contains six elements: (a) the choice, (b) material alternatives, (c) affected interests or constituencies, (d) standards and current evidence, (e) uncertainty and foreseeable harm, and (f) the condition that would trigger consultation, deviation disclosure, or escalation. The protocol is proportionate to the gate. Choosing a plotting convention does not warrant the same process as changing eligibility criteria after observing group differences or proceeding with a participant-risk deviation.

Prospective publicity is distinct from three neighboring practices. It is not *mere identifiability*, because knowing that one's name is attached does not require consideration of alternatives. It is not *outcome accountability*, because the account is evaluated for the quality and candor of the process given information available at the time, not for whether the result was favorable. It is not *indiscriminate transparency*, because records may contain participant information, proprietary methods, culturally restricted knowledge, security-sensitive material, or tentative ideas that require tiered access. Its psychological core is anticipated answerability; its institutional form is bounded process transparency.

Audience and review features must be specified through observable dimensions. Representational legitimacy consists of relevant standing, competence, plurality when several interests are implicated, and independence from the preferred outcome. Review legitimacy adds procedural fairness, contestability, correction rights, and an appeal route. Preference knowledge is separate: when the audience's desired conclusion is known, a researcher can optimize the account for conformity. A dependent trainee asked to justify a choice only to the supervisor who

demanded it is not meaningfully answerable. Finally, review must be possible but not omnipresent. A credible probability of fair sampling may preserve the anticipatory effect without turning every provisional choice into a hearing.

What the PPIM adds

The model is integrative, so its novelty must lie in discriminating joint predictions rather than renaming established ideas. Its causal core contains three separations: reason generation versus exposure, precommitment versus postcommitment timing, and simulated answerability versus bounded institutional review. Accountability theory supplies conditional effects of timing and audience; ethical-fading theory supplies issue recognition; moral-identity and professional-norm research supply distinct self-standard pathways; personality-process theories supply trait-state distributions and person $\times$ situation expression; and open-science reforms supply field analogues of prospective records. The larger set of trait and situational moderators are differentiated extensions of that core, not equally established links.

Table 3

Intellectual Genealogy and Incremental Claims of the PPIM

Framework	Established contribution	PPIM extension
Publicity and defensibility tests	A decision is examined by asking whether its reasons could withstand public or stakeholder scrutiny	Separates a familiar ethical heuristic into simulated answerability, structured reason generation, and credible bounded review
Accountability theory	Timing, known audience preferences, and evaluation conditions can improve or distort judgment	Separates a private reason-generating cue from an institutional protocol and predicts trait-differentiated responses
Ethical fading and moral disengagement	An ethical dimension can disappear from the focal frame, and self-sanctions can be neutralized	Places issue recognition first in a proposed sequence and requires independent evidence beyond prompted content

Framework	Established contribution	PPIM extension
Moral-identity theory	Identity centrality predicts behavior probabilistically and depends on situational accessibility	Distinguishes moral identity from professional-norm accessibility and tests both alongside self-critical reasoning
Trait activation and Whole Trait Theory	Traits are expressed through situation cues and distributions of momentary states	Distinguishes completeness, event-report congruence, voice, correction, and burden as different trait-linked outcomes
Open-science precommitment	Prospective records can constrain outcome-dependent flexibility, yet vague or unaudited records can fail	Generalizes from publication practices to broader decision gates while predicting reconstruction, ritual, gaming, privacy costs, and climate feedback

Propositions

Table 4 states ten numbered propositions, several divided into lettered components so that partial support cannot masquerade as confirmation of an entire bundle. Outcomes are defined independently of whether a reviewer likes the final choice: event-report congruence, prospective record-implementation correspondence interpreted with disclosed deviation, material-element completeness, consultation at preregistered thresholds, correction, and process quality rated with a preregistered multi-stakeholder rubric and reported inter-rater reliability. Process-quality elements must not merely repeat the prompt or reward agreement with a reviewer's preferred decision. Several propositions predict backfire, and several can be falsified even if the prompt improves one superficial outcome.

Table 4

Falsifiable Propositions of the Prospective Publicity-Integrity Model

Proposition	Prediction	Discriminating evidence
-------------	------------	-------------------------

Proposition	Prediction	Discriminating evidence
P1: Timing	A PPH delivered before psychological commitment will increase independently measured issue recognition relative to the identical cue delivered after commitment, matched documentation, or a generic ethics reminder. Downstream effects are predicted to operate through issue recognition followed by self-standard accessibility and/or self-critical reasoning.	No before/after-commitment difference, or a precise absence of the proposed sequential pathway, would count against the timing claim.
P2: Reasons over visibility	Relative to a matched identifiability cue, a PPH requiring alternatives, uncertainty, and current evidence will improve independent transfer measures of standards and alternative consideration and later event-report congruence. Prompted content, record length, or polish alone will not count as support.	Equal effects of identifiability and reason generation on the independent outcomes would reject the proposed self-critical mechanism.
P3: Identity activation	The indirect effect of the PPH on candid disclosure and correction through moral-identity accessibility will be more positive at higher moral-identity centrality, except under ceiling conduct.	An effect explained entirely by sanction fear, professional-norm accessibility, or social desirability would favor a rival account.
P4a-c: Differentiated traits	P4a: Honesty-Humility will predict event-report congruence. P4b: Conscientiousness will predict material-element completeness. P4c: Low verifiability will magnify discrepancies associated with Machiavellianism or appearance goals.	A single undifferentiated good-personality factor, no distinction between completeness and congruence, or no verifiability interaction would contradict the pattern.
P5: Discretion × agency	PPH benefits will be small in already strong situations, increase as consequential discretion rises while meaningful agency remains, and attenuate when coercion removes the ability to pause, consult, or refuse.	Benefits determined by pressure alone, regardless of discretion or agency, would reject the proposed structure.
P6a-b: Audience structure	P6a: Holding competence, power, stakes, and review probability constant, plurality will broaden represented interests. P6b: Unknown audience preferences will reduce conformity relative to known preferences.	Effects attributable only to seniority, harshness, or review probability would favor surveillance rather than PPIM.
P7: Climate and power	Psychological safety and procedural fairness will strengthen effects on voice, help-seeking, and correction; punitive or arbitrary review will shift effects toward concealment, especially among evaluation-sensitive researchers.	Precisely estimated equivalent candor across safe and retaliatory climates would count against the boundary condition.

Proposition	Prediction	Discriminating evidence
P8a-b: Retention and timing	P8a: A durable, time-stamped, credibly reviewable record package will improve later reconstruction resistance relative to a transient private rationale after an identical precommitment PPH. P8b: First elicitation before the outcome will reduce outcome-consistent reconstruction relative to first elicitation after the outcome.	No difference between the retained-and-reviewable record package and the transient-private rationale on independently measured later reconstruction would reject P8a; no prospective-versus-retrospective difference would reject P8b.
P9a-b: Repetition and maintenance	P9a: Across repeatedly sampled gates, PPH exposure will shift within-person distributions of momentary states and behaviorally anchored disclosure. P9b: Without credible fair review and developmental feedback, gains in record quality will decay faster.	No state-distribution or behavioral-frequency shift would reject P9a; equivalent maintenance with and without fair review would reject P9b.
P10a-b: Bounded access	P10a: Bounded access will be noninferior to broader access on prespecified integrity outcomes within a prespecified margin, or the two conditions will fall within a prespecified equivalence region. P10b: Broader access will increase perceived exposure, anxiety, privacy cost, or loss of exploratory originality.	A broader-access advantage larger than the noninferiority margin, or failure to fall within the equivalence region, would reject P10a; a precise absence of all preregistered harm differences would reject P10b.

P1 and P2 isolate the active ingredient. If a moral reminder, identifiability cue, matched documentation, and structured PPH perform identically on independent outcomes, the proposed issue-recognition and self-critical mechanisms are unnecessary. If cue timing before versus after psychological commitment makes no difference, timing is not central. P3 and P4 move beyond a uniform treatment effect. A conscientious participant may complete every material field without confronting a preferred conclusion; a high-Machiavellian participant may learn the genre of accountability faster than the practice of candor. The critical outcome is therefore event-report congruence and behavior, not fluency alone.

P5 predicts that the heuristic has a zone of usefulness defined by two variables rather than one inverted U. When expectations are clear, consistent, observable, and enforced, the situation is already strong. As consequential discretion rises, a cue has more room to matter, but only

while the actor retains enough agency to pause, consult, or refuse. Coercion can remove that agency even when ambiguity remains. P6 and P7 predict that audience and climate can reverse effects. A factorial test should manipulate single versus plural audiences and known versus unknown preferences while holding competence, power, stakes, and review probability constant. Accountability research suggests that answerability to an audience of unknown views can increase cognitive complexity, whereas a known position can produce conformity and postdecision accountability can produce defensive bolstering (Lerner & Tetlock, 1999). Psychological safety and procedural fairness motivate the proposed climate moderation, but their specific effects on PPH responses remain untested (Colquitt et al., 2001; Frazier et al., 2017).

P8 and P9 address the integrity of the infrastructure. A form that is created after the outcome or never reviewed invites reconstruction and ritual. Evidence from open-data policies illustrates the gap between disclosure and performance. Badges were associated with a marked increase in reported data sharing at *Psychological Science* relative to comparison journals (Kidwell et al., 2016), and a mandatory policy at *Cognition* increased data-availability statements and apparently reusable data (Hardwicke et al., 2018). Yet only a minority of sampled reusable data sets in the latter study reproduced all target values without author assistance, and a separate audit of data-sharing statements found that most authors who said data were available did not provide them when asked (Gabelica et al., 2022). Inspectability needs verification.

P10 rejects a maximalist equation of publicity with integrity. Conflict disclosure can produce strategic exaggeration or moral licensing by advisors, reduce trust while increasing social pressure to comply with distrusted advice, and alter whether advisors accept avoidable conflicts (Cain et al., 2011; Sah & Loewenstein, 2014; Sah et al., 2013). The lesson is not secrecy. It is that information changes relationships as well as knowledge. A protected record,

legitimate reviewer, and correction right may preserve answerability while reducing the costs of mass exposure.

When Prospective Publicity Should Fail or Backfire

A useful theory of accountability must explain its injuries. Five failure modes are especially important.

First, publicity can become *performative compliance*. If records are scored for polish, people learn to write defensible narratives rather than make defensible decisions. This risk should be highest for appearance-focused or strategically manipulative actors and when reviewers cannot compare records with conduct. Random, proportionate process audits and independent evidence are therefore theoretically relevant, not administrative detail.

Second, publicity can become conformity. A homogeneous or partisan audience may punish well-justified dissent. Scientific advance sometimes requires violating a local expectation, and ethical progress sometimes requires challenging an institution. The PPH asks whether a decision can be defended, not whether it will be popular. Plural audiences, written standards, and review of reasons rather than outcomes are hypothesized to help preserve that distinction.

Third, publicity can become retaliatory surveillance. A junior researcher cannot be expected to document a supervisor's questionable instruction candidly if that supervisor controls access to the record, evaluation, authorship, and employment. Psychological safety is predicted to shape whether the cue means think carefully or construct evidence against yourself. Protected channels, role separation, anti-retaliation procedures, and confidential advice are proposed boundary protections, not demonstrated guarantees.

Fourth, publicity can *chill exploration*. Preregistration is compatible with exploration when deviations and new analyses are labeled honestly; the same principle applies here (Lakens,

2024). Requiring a permanent, widely visible defense of every provisional idea would favor conventional questions, penalize uncertainty, and create administrative drag. Decision gates should be limited to choices with material epistemic, welfare, or fairness consequences. Records should allow revision and distinguish good-faith development from deceptive rewriting.

Fifth, publicity can violate privacy, consent, and cultural legitimacy. Participant identities, community-held knowledge, proprietary data, and security-sensitive methods cannot be released merely to signal openness. Nor is an imagined mass public automatically a legitimate moral authority. Because the legitimacy and meaning of publicity may vary across cultural and institutional contexts, the PPIM predicts the function of answerability, not one Western institutional form. Bounded access, community governance, and contextually legitimate representation must therefore be built into cross-cultural tests rather than appended after validation.

These failures define the model's domain. The PPH is most plausibly useful for ambiguous, self-regulatory decisions made by actors who retain some agency and remain responsive to reasons. It is not a substitute for competent ethics review, independent auditing, protected reporting, investigation, proportionate sanctions, mentoring, or incentive reform. A Cochrane review of 31 controlled studies found that evidence for misconduct-prevention interventions was uncertain, most trials were at high risk of bias, and organizational interventions were notably absent (Marušić et al., 2016). The appropriate posture is disciplined uncertainty: test the heuristic as one component of a research-integrity ecology, not as a cheap remedy for structural failure.

A Cumulative Research Program

The PPIM should be tested in a sequence that makes mechanism claims increasingly expensive. Early studies can establish discrimination among prompts. Later studies must show that effects survive real stakes, hierarchy, repeated use, and cross-cultural variation. Across the program, confession-style self-report should be secondary to behavior, time-stamped records, independent coding, correction, event-report congruence, and deviation traceability.

Studies 1A-1C: Timing, reasons, audience, and reconstruction

Three preregistered, multi-laboratory experiments with active researchers and doctoral trainees would use an incentivized research-management simulation centered on one consequential exclusion decision. Study 1A would compare (a) a PPH delivered before participants form an initial recommendation, (b) the identical PPH delivered after participants record an initial choice but before final implementation, (c) matched-length factual documentation delivered before commitment, and (d) a generic ethics reminder. Confidence and resistance to revising the reversible choice would be measured immediately before the postcommitment PPH, so commitment is not inferred solely from a mouse click. After the decision, every condition would complete the same fixed-format report independently of the treatment prompt. Co-primary outcomes would be event-report accuracy, material-element completeness, and the behavioral choice to disclose the deviation or seek consultation, with familywise error controlled or a preregistered hierarchical testing sequence. The treatment response would serve only as a manipulation check. Minimally reactive process traces would be captured at prespecified stages between the cue and final choice to test temporal ordering; postdecision self-reports would be supplementary rather than evidence of sequence. Time on

task, general deliberative effort, demand awareness, sanction fear, social desirability, and evaluative anxiety would be measured as rival explanations.

Study 1B would use a $2 \times 2 \times 2$ design: structured reason generation versus an identifiability-only active control, single versus plural audience, and known versus unknown audience preference. The active control would match identifiability, expected review, response length, minimum exposure time, and general deliberative effort. The plural reason-generation cell instantiates the recommended form when multiple interests are implicated; the single-audience cell tests whether plurality adds the effect predicted in P6a. Competence, power, stakes, review probability, and information would be held constant. After the decision, every condition would complete the same fixed-format report. Primary outcomes would be event-report accuracy and an independent consultation or concealment choice. Secondary outcomes would include an unprompted transfer decision after new counterevidence, stakeholder breadth in a common final report, conformity to the signaled audience preference, evaluative anxiety, and integrative complexity. Blinded coders would use a preregistered multi-stakeholder rubric, report inter-rater reliability, and rate process rather than agreement with the final choice. Sample sizes would be chosen by simulation for the smallest focal contrast regarded as practically meaningful, with laboratory modeled as a random factor.

Study 1C would separate two claims currently bundled by a simple prospective-versus-retrospective comparison. Study 1C1 would give every participant the same precommitment PPH, then randomize whether the rationale becomes a durable, time-stamped, credibly reviewable record or remains transient and private. This tests the retained-and-reviewable record package in P8a while holding initial cognition constant; it does not identify a pure storage effect. Study 1C2 would hold prompt wording, required fields, information, expected review, access,

and response-length limits constant while manipulating whether the first rationale is elicited before or after the simulated outcome. Outcome-consistent reconstruction, event-report accuracy, and omission of a logged deviation would test P8b. Neither contrast would be used to infer psychological commitment.

Study 2: Persons within decisions

A six- to eight-week micro-randomized field study would recruit researchers from multiple disciplines. Eligible gates would be defined before enrollment - such as a version-control event that changes an exclusion rule, a recruitment amendment, a formal authorship allocation, or a documented escalation threshold - and instrumented where feasible. At each eligible event, a secure system would randomly deliver either the PPH or a neutral factual log. Random daily probes would ask about the most recent eligible decision, including decisions not entered voluntarily, and nonresponse would be retained as an outcome rather than silently dropped. Baseline measures would include Honesty-Humility, conscientiousness, moral-identity centrality, Machiavellianism, achievement goals, public self-consciousness, and evaluation anxiety. Brief episode measures would capture pressure, discretion, agency, perceived power, moral-identity accessibility, professional-norm accessibility, psychological safety, verifiability or observability of conduct, and situational strength through clarity, consistency, constraints, and consequences. Psychometrically evaluated momentary honesty and carefulness items would be measured separately from behavioral outcomes.

Sensitive content would remain protected, with participants authorizing tiered coding or supplying redacted decision summaries. Outcomes would include contemporaneous disclosure, event-report congruence, prospective record-implementation correspondence interpreted with disclosed deviations, consultation, correction, burden, and selective nonlogging. Each event

would also yield behaviorally anchored indicators of candid disclosure and careful alternative consideration. Multilevel and distributional models would partition within-person episode effects from between-person differences; estimate momentary honesty and carefulness state distributions separately from frequencies of integrity-relevant enactments; and test P3-P5 and P9a. The crucial question is not whether ethical people use the prompt better. It is whether the same person behaves differently at randomly prompted decision points, and whether traits explain the form of that episode-level change.

Study 3: Hierarchy, voice, and concealment

A team simulation would manipulate a supportive versus punitive supervisor climate and PPH versus ordinary procedure. Junior members would encounter a planted analytic or participant-protection concern that threatens a valued result. They could question the decision, seek confidential advice, document disagreement, comply silently, or conceal the concern. Supervisors' stated preferences and control over evaluation would be explicit. Baseline public self-consciousness and fear of negative evaluation would index evaluation sensitivity; perceived psychological safety, procedural fairness, and supervisor power would serve as manipulation checks and moderators.

The primary outcomes would be speaking up, correction of the problematic decision, event-report accuracy, and concealment. Evaluation anxiety and perceived retaliation risk would test moderated mediation. P7 predicts an interaction: PPH should increase voice and correction under fair, safe review but may increase impression management or silence under punitive authority. A precise null climate interaction, with confidence intervals narrow enough to exclude a meaningful effect, would undermine the claim that answerability acquires meaning from power relations.

Studies 4A and 4B: Repetition, review, access, and institutional consequences

Study 4A would be a parallel, four-arm cluster-randomized trial assigning research groups for six months to (a) an internally rehearsed PPH, (b) the PPH plus a protected record with no routine review, (c) the identical cue and record plus periodic fair sampling and developmental feedback, or (d) existing procedures. Objective gate definitions and a common minimum data model would be set before randomization. Outcomes would include disclosure of deviations, time to correction, analytic-version consistency, use of consultation channels, privacy incidents, experienced burden, and indicators of exploratory work. Authorship or supervision disputes could be retained as rare secondary outcomes rather than relied on for statistical power. Evaluation should distinguish a rise in reported deviations caused by greater candor from a rise in underlying problematic conduct.

Study 4A primarily tests P9 and institutional harms: administrative delay, chilling of novel work, unequal burden, gaming, retaliation, and documented privacy incidents. Study 4B would be a separate, ethically permissible low-risk experiment using participants' own nonsensitive task records under a credible simulated-access manipulation; no record would be disclosed beyond the access participants explicitly consented to. Bounded versus broader access would vary while reviewer expertise, review probability, and task information remained constant. P10a would treat bounded access as the test condition and broader access as the reference condition, using preregistered noninferiority or equivalence bounds and a smallest effect size of interest rather than ordinary nonsignificance. P10b would test perceived exposure, anxiety, privacy cost, and originality of subsequent exploratory work. Cross-national participation would permit tests of measurement invariance and functional equivalence of

audience features. Registered Reports would prevent success from being defined after outcomes are seen.

Measurement and open-science safeguards

The research program must not reproduce the weaknesses it studies. Each experiment should separate confirmatory outcomes from exploration, publish the prompt and coding rubrics, record deviations, report null and adverse effects, and make de-identified materials available where consent and confidentiality permit. Where records cannot be shared, investigators should disclose the restriction and provide synthetic examples or controlled access. Multi-item self-report QRP totals should not serve as the sole criterion. Behavioral tasks, process traces, other-ratings, independently adjudicated corrections, and congruence metrics are harder to interpret but closer to the theory.

Construct validity requires discriminant tests. Issue recognition should be distinguishable from general vigilance; moral-identity accessibility from professional-norm accessibility and socially desirable responding; self-critical reasoning from verbosity and general effort; and record quality from ethical outcome. Prompted alternatives, uncertainty, evidence, and stakeholder names are manipulation responses, not outcomes. The PPH should therefore be compared with matched documentation, identifiability, sanction, and moral-reminder controls, then tested on independent process traces, unprompted transfer decisions, responses to new counterevidence, and common final reports. Without those contrasts, a positive result would show only that asking people to write more changes something.

Discussion

Implications for Personality Science

The PPIM reframes research integrity as an expression problem. Traits matter, but their effects are probabilistic, state-mediated, and situation-specific. This framing makes three contributions to personality science.

First, it identifies science itself as a consequential natural ecology for person $\times$ situation research. Researchers routinely face ambiguity, autonomy, delayed feedback, identity stakes, and asymmetric observability. These conditions allow the field to study how honesty, carefulness, moral identity, and evaluation sensitivity are expressed when desirable outcomes compete rather than align. The outcome space is also differentiated: a complete record, a candid record, a sound decision, and a courageous correction are related but not interchangeable.

Second, the model connects between-person traits with within-person intervention effects. Within a Whole Trait framework, a situational cue may shift the distribution of specific trait states without erasing individual differences (Fleeson & Jayawickreme, 2015). The PPH supplies a manipulable cue, while repeated decision sampling supplies the level of measurement needed to observe changes in honesty, carefulness, and related episode components separately. This is a stronger test than correlating one trait score with one retrospective confession total.

Third, the model treats backfire as theory-bearing. If Machiavellianism predicts polished inconsistency, conscientiousness predicts completeness without candor, evaluation anxiety predicts concealment in punitive climates, or a known audience predicts conformity, those are not implementation footnotes. They reveal which psychological process the intervention actually activated. Personality science is especially well placed to prevent an institution from mistaking uniform paperwork for uniform integrity.

Implications for Research Governance

Institutions combine training, monitoring, mentoring, and retrospective enforcement; prospective publicity adds a proposed anticipatory function: structuring a small number of decisions so that reasons are captured before the focal choice is implemented or reconstructed after the outcome. Because the complete model is untested, this article does not recommend institutional adoption. If an institution nevertheless chooses to evaluate it, the appropriate next step would be an ethics-reviewed, bounded pilot with explicit protections, stopping rules, and measurement of adverse effects.

Three proposed risk controls follow from the theory. First, focus on decision gates, not constant observation. Gates might include an exclusion considered after viewing outcomes, a material protocol change, whether and how to disclose and characterize a material analytic departure, contested authorship, participant-risk escalation, or correction of a published claim. Second, evaluate process under uncertainty, not success. A well-reasoned decision can have a poor outcome, and a reckless decision can get lucky. Outcome accountability can induce defensive hindsight and bolstering (Lerner & Tetlock, 1999; Tetlock, 1983). Third, protect candor. Any pilot would require tiered access, correction rights, role separation, protected advice, anti-retaliation enforcement, and prespecified stopping rules. These are proposed safeguards, not evidence that the intervention works.

These principles also clarify what the model does not authorize. It does not support personality screening, public shaming, compulsory release of sensitive information, or the presumption that deviation equals wrongdoing. It does not allow institutions to retain harmful incentives while assigning each researcher another form. If publication pressure, arbitrary

leadership, or insecure employment creates extreme coercion, a self-reflective prompt transfers responsibility downward rather than correcting the system.

Constraints on Generality

The model's strongest evidence comes from Western organizational psychology, moral psychology, biomedical and social-science researcher surveys, and reforms concentrated in psychology and biomedicine. Generality across disciplines, career systems, national cultures, community-governed research, industry, classified work, and high-risk field settings is unknown. The meaning of publicity, legitimate standing, independent review, and role-relevant identity standards will vary. Public answerability may be experienced as protection in one setting and exposure in another.

The PPH is expected to influence high-discretion choices made by actors who remain responsive to reasons and retain meaningful agency. It is less likely to prevent determined, collusive, or organized fraud; accidental error; violations driven by lack of knowledge; or choices made under severe coercion. Effects may also be small where strong procedures already make relevant standards salient and conduct readily observable. These are theoretical boundaries to be tested, not populations to be excluded silently.

Conclusion

Science depends on discretion. No set of rules can anticipate every sample, participant, model, authorship dispute, or correction. The ethical problem is therefore not that researchers sometimes depart from a plan or challenge an institution. It is that a person can use expertise, ambition, or a valued scientific goal to make a contestable decision less visible at the moment visibility matters most.

Prospective publicity offers a precise task for that moment: name the choice and materially affected constituencies; generate the rejected alternatives, uncertainty, and present evidence; then ask whether those reasons could withstand evaluation by people with relevant standing. The PPIM predicts that this task may make an integrity issue recognizable and then activate distinct self-standard and self-critical-reasoning pathways. A separate protocol may preserve a contemporaneous comparison point against hindsight. The model also predicts failure: conformity under a partisan audience, silence under retaliation, compliance theater without verification, and lost creativity or privacy under indiscriminate exposure.

The resulting theory is deliberately less comforting than “transparency works.” Integrity is not a possession revealed by a personality score, nor a property installed by a form. It is a repeated achievement of persons in situations. Institutions influence that achievement not only through the rules they announce and the misconduct they punish, but through the kinds of answerability they make possible before a private judgment becomes the public record of science.

Open Science Statement

This article reports no new data, materials, code, or statistical analyses and was not preregistered. It is a problem-driven integrative review rather than a systematic review; literature selection was iterative and theory-guided. Appendix A documents the search routes, query families, inclusion logic, and source-to-claim map to increase transparency. It does not make the review exhaustive or reproducible and supports no literature-coverage claim. The propositions and proposed studies are intended to support prospective, falsifiable tests of the model.

Funding

The author received no external financial support for the research, authorship, or publication of this article.

Competing Interests

The author declares no competing interests.

Ethics Statement

This article reports no human participants, animals, identifiable private information, or new empirical data. Institutional ethics review was not applicable.

References

- Abbott, L., & Grady, C. (2011). A systematic review of the empirical literature evaluating IRBs: What we know and what we still need to learn. *Journal of Empirical Research on Human Research Ethics*, 6(1), 3-19. <https://doi.org/10.1525/jer.2011.6.1.3>
- Aleksovska, M., Schillemans, T., & Grimmelikhuijsen, S. (2019). Lessons from five decades of experimental and behavioral research on accountability: A systematic literature review. *Journal of Behavioral Public Administration*, 2(2). <https://doi.org/10.30636/jbpa.22.66>
- Aquino, K., Freeman, D., Reed, A., II, Lim, V. K. G., & Felps, W. (2009). Testing a social-cognitive model of moral behavior: The interactive influence of situations and moral identity centrality. *Journal of Personality and Social Psychology*, 97(1), 123-141. <https://doi.org/10.1037/a0015406>
- Aquino, K., & Reed, A., II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, 83(6), 1423-1440. <https://doi.org/10.1037/0022-3514.83.6.1423>
- Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Cromptvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, 18(12), Article e3000937. <https://doi.org/10.1371/journal.pbio.3000937>
- Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193-209. https://doi.org/10.1207/s15327957pspr0303_3

Bok, S. (1999). *Lying: Moral choice in public and private*. Vintage Books. (Original work published 1978)

Cain, D. M., Loewenstein, G., & Moore, D. A. (2011). When sunlight fails to disinfect: Understanding the perverse effects of disclosing conflicts of interest. *Journal of Consumer Research*, 37(5), 836-857. <https://doi.org/10.1086/656252>

Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, 8(10), Article 211037. <https://doi.org/10.1098/rsos.211037>

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, 86(3), 425-445. <https://doi.org/10.1037/0021-9010.86.3.425>

Correction for Kristal et al., signing at the beginning versus at the end does not decrease dishonesty. (2026). *Proceedings of the National Academy of Sciences*, 123(27), Article e2610441123. <https://doi.org/10.1073/pnas.2610441123>

Crain, A. L., Martinson, B. C., & Thrush, C. R. (2013). Relationships between the Survey of Organizational Research Climate (SORC) and self-reported research practices. *Science and Engineering Ethics*, 19(3), 835-850. <https://doi.org/10.1007/s11948-012-9409-0>

Davis, M. S., Riske-Morris, M., & Diaz, S. R. (2007). Causal factors implicated in research misconduct: Evidence from ORI case files. *Science and Engineering Ethics*, 13(4), 395-414. <https://doi.org/10.1007/s11948-007-9045-2>

- DeCelles, K. A., DeRue, D. S., Margolis, J. D., & Ceranic, T. L. (2012). Does power corrupt or enable? When and why power facilitates self-interested behavior. *Journal of Applied Psychology*, 97(3), 681-689. <https://doi.org/10.1037/a0026811>
- Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology*, 43(4), 522-527. <https://doi.org/10.1037/h0076760>
- Fischhoff, B. (1975). Hindsight $\neq$ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3), 288-299. <https://doi.org/10.1037/0096-1523.1.3.288>
- Fleeson, W., & Jayawickreme, E. (2015). Whole Trait Theory. *Journal of Research in Personality*, 56, 82-92. <https://doi.org/10.1016/j.jrp.2014.10.009>
- Frazier, M. L., Fainshmidt, S., Klinger, R. L., Pezeshkan, A., & Vacheva, V. (2017). Psychological safety: A meta-analytic review and extension. *Personnel Psychology*, 70(1), 113-165. <https://doi.org/10.1111/peps.12183>
- Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: Mixed-methods study. *Journal of Clinical Epidemiology*, 150, 33-41. <https://doi.org/10.1016/j.jclinepi.2022.05.019>
- Gopalakrishna, G., ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in the Netherlands. *PLOS ONE*, 17(2), Article e0263023. <https://doi.org/10.1371/journal.pone.0263023>
- Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsson, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L., Altman,

S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science*, 5(8), Article 180448.

<https://doi.org/10.1098/rsos.180448>

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making*, 13(4), 356-371. <https://doi.org/10.1017/S1930297500009232>

Hertz, S. G., & Krettenauer, T. (2016). Does moral identity effectively predict moral behavior? A meta-analysis. *Review of General Psychology*, 20(2), 129-140.

<https://doi.org/10.1037/gpr0000062>

Janke, S., Daumiller, M., & Rudert, S. C. (2019). Dark pathways to achievement in science: Researchers' achievement goals predict engagement in questionable research practices. *Social Psychological and Personality Science*, 10(6), 783-791.

<https://doi.org/10.1177/1948550618790227>

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196-217.

https://doi.org/10.1207/s15327957pspr0203_4

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, 14(5), Article e1002456. <https://doi.org/10.1371/journal.pbio.1002456>

- Kristal, A. S., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N., & Ariely, D. (2020). Signing at the beginning versus at the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences*, 117(13), 7103-7107. <https://doi.org/10.1073/pnas.1911695117>
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480-498. <https://doi.org/10.1037/0033-2909.108.3.480>
- Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, 10(1), Article 117094. <https://doi.org/10.1525/collabra.117094>
- Leary, M. R. (1983). A brief version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin*, 9(3), 371-375. <https://doi.org/10.1177/0146167283093007>
- Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125(2), 255-275. <https://doi.org/10.1037/0033-2909.125.2.255>
- Lewis, C. W., & Gilman, S. C. (2005). *The ethics challenge in public service: A problem-solving guide* (2nd ed.). Jossey-Bass.
- Marušić, A., Wager, E., Utrobičić, A., Rothstein, H. R., & Sambunjak, D. (2016). Interventions to prevent misconduct and promote integrity in research and publication. *Cochrane Database of Systematic Reviews*, 2016(4), Article MR000038. <https://doi.org/10.1002/14651858.MR000038.pub2>
- Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, 36(1), 121-140. <https://doi.org/10.1177/0149206309349309>

- Moore, C., Detert, J. R., Treviño, L. K., Baker, V. L., & Mayer, D. M. (2012). Why employees do bad things: Moral disengagement and unethical organizational behavior. *Personnel Psychology*, 65(1), 1-48. <https://doi.org/10.1111/j.1744-6570.2011.01237.x>
- Nazarovets, S. (2026). Scientific misconduct, questionable research practices, and research climate: Empirical approaches to integrity problems. *Research Integrity and Peer Review*, 11(1), Article 22. <https://doi.org/10.1186/s41073-026-00214-1>
- Office of Research Integrity. (n.d.). *Definition of research misconduct*. U.S. Department of Health and Human Services. Retrieved August 8, 2026, from <https://ori.hhs.gov/definition-research-misconduct>
- Ogunfowora, B., Nguyen, V. Q., Steel, P., & Hwang, C. C. (2022). A meta-analytic investigation of the antecedents, theoretical correlates, and consequences of moral disengagement at work. *Journal of Applied Psychology*, 107(5), 746-775. <https://doi.org/10.1037/apl0000912>
- Public Health Service Policies on Research Misconduct, 42 C.F.R. §§ 93.103, 93.234 (2026). <https://www.ecfr.gov/current/title-42/chapter-I/subchapter-H/part-93>
- Sah, S., & Loewenstein, G. (2014). Nothing to declare: Mandatory and voluntary disclosure leads advisors to avoid conflicts of interest. *Psychological Science*, 25(2), 575-584. <https://doi.org/10.1177/0956797613511824>
- Sah, S., Loewenstein, G., & Cain, D. M. (2013). The burden of disclosure: Increased compliance with distrusted advice. *Journal of Personality and Social Psychology*, 104(2), 289-304. <https://doi.org/10.1037/a0030527>
- Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with Registered Reports. *Advances in*

Methods and Practices in Psychological Science, 4(2), 1-12.

<https://doi.org/10.1177/25152459211007467>

Silberman, G., & Kahn, K. L. (2011). Burdens on research imposed by institutional review boards: The state of the evidence and its implications for regulatory reform. *The Milbank Quarterly*, 89(4), 599-627. <https://doi.org/10.1111/j.1468-0009.2011.00644.x>

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2021). Initial evidence of research quality of Registered Reports compared with the standard publishing model. *Nature Human Behaviour*, 5(8), 990-997. <https://doi.org/10.1038/s41562-021-01142-4>

Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical fading: The role of self-deception in unethical behavior. *Social Justice Research*, 17(2), 223-236.

<https://doi.org/10.1023/B:SORE.0000027411.35832.53>

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology*, 45(1), 74-83. <https://doi.org/10.1037/0022-3514.45.1.74>

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology*, 52(4), 700-709. <https://doi.org/10.1037/0022-3514.52.4.700>

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, 88(3), 500-517.

<https://doi.org/10.1037/0021-9010.88.3.500>

Tijdkink, J. K., Bouter, L. M., Veldkamp, C. L. S., van de Ven, P. M., Wicherts, J. M., & Smulders, Y. M. (2016). Personality traits are associated with research misbehavior in

Dutch scientists: A cross-sectional study. *PLOS ONE*, 11(9), Article e0163251.

<https://doi.org/10.1371/journal.pone.0163251>

Tijdink, J. K., Verbeke, R., & Smulders, Y. M. (2014). Publication pressure and scientific misconduct in medical scientists. *Journal of Empirical Research on Human Research Ethics*, 9(5), 64-71. <https://doi.org/10.1177/1556264614552421>

U.S. Government Accountability Office. (2023). *Institutional review boards: Actions needed to improve federal oversight and examine effectiveness* (GAO-23-104721).

<https://www.gao.gov/products/gao-23-104721>

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Skowronski, J. J., Acar, O. A., Aczel, B., Bakos, B. E., Barbosa, F., Baskin, E., Bègue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., ... Yıldız, E. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, 1(3), 299-317.

<https://doi.org/10.1177/2515245918781032>

Xie, Y., Wang, K., & Kong, Y. (2021). Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Science and Engineering Ethics*, 27(4), Article 41. <https://doi.org/10.1007/s11948-021-00314-9>

Appendix A

Review Search and Evidence Map

Evidence gathering and bibliographic verification for this integrative review were completed on August 8, 2026. Search routes included PubMed and PubMed Central, publisher and DOI landing pages, official research-integrity and government sources, and scholarly web search. Query families combined: (a) research misconduct or questionable research practices with personality, Machiavellianism, achievement goals, publication pressure, norms, or climate; (b) accountability or answerability with predecisional, process, audience, or unknown views; (c) ethical fading, moral disengagement, or moral identity with unethical behavior or misconduct; (d) preregistration, Registered Reports, or open data with quality, deviation, reproducibility, or audit; (e) institutional review board or research integrity training with effectiveness or systematic review; and (f) psychological safety, procedural justice, public self-consciousness, or fear of negative evaluation with voice, learning, accountability, or disclosure. Targeted citation-chain, exact-title, and DOI searches were used to identify and verify foundational or meta-analytic construct sources. These routes document the review process but do not constitute an exhaustive or reproducible search protocol.

Priority was given to empirical research involving scientists, systematic or meta-analytic syntheses, primary experiments relevant to the proposed mechanisms, field evaluations of prospective research practices, and official definitions. Adjacent behavioral evidence was included only when the transport to scientific decisions was stated explicitly. Retracted work was excluded as affirmative evidence. Sources were not excluded for null or adverse findings. Because the review was theory-guided and iterative, no exhaustive screening count or coverage claim is made.

Table A1*Source-to-Claim Map and Inferential Limits*

Claim family	Direct or anchoring sources	Inferential limit carried into the manuscript
Misconduct definition and culpability	Office of Research Integrity (n.d.); Public Health Service Policies on Research Misconduct (2026)	Defines the current U.S. PHS framework; it is not a universal definition of every integrity failure
Publicity-test prior art	Bok (1999); Lewis and Gilman (2005)	Establishes a familiar ethical heuristic, not the PPIM mechanisms or effects
Prevalence and measurement	Gopalakrishna et al. (2022); Nazarovets (2026); Xie et al. (2021)	Heterogeneous definitions, recall periods, response methods, and self-report prevent a fixed prevalence claim
Traits and achievement motives in researchers	Janke et al. (2019); Tjldink et al. (2016)	Cross-sectional self-report; small explained variance; no personality screening or causal attribution
Climate, pressure, and procedural context	Colquitt et al. (2001); Crain et al. (2013); Frazier et al. (2017); Gopalakrishna et al. (2022); Marušić et al. (2016)	Researcher evidence is mostly correlational; organizational meta-analyses anchor constructs but do not test the PPIM interaction
Personality-process transport	DeCelles et al. (2012); Fleeson and Jayawickreme (2015); Heck et al. (2018); Tett and Burnett (2003)	General personality and behavioral tasks, not misconduct by scientists
Issue recognition and self-justification	Moore et al. (2012); Ogunfowora et al. (2022); Tenbrunsel and Messick (2004)	Ethical fading is conceptual; moral-disengagement evidence is not a demonstrated sequence in scientific work
Conditional accountability	Aleksovska et al. (2019); Lerner and Tetlock (1999)	Experimental effects are heterogeneous and usually outside research settings
Prospective records and reconstruction	Bakker et al. (2020); Claesen et al. (2021); Fischhoff (1975); Kerr (1998); Scheel et al. (2021); Soderberg et al. (2021)	Hindsight evidence and nonrandomized publication analogues do not test the PPH or retained-record mechanism
Generic ethical reminders	Kristal et al. (2020); Correction for Kristal et al. (2026); Verschuere et al. (2018)	High-powered null results bound claims about generic reminders; they do not directly test the structured PPH
Transparency and disclosure backfire	Cain et al. (2011); Gabelica et al. (2022); Sah and Loewenstein (2014); Sah et al. (2013)	Conflict-advice and data-sharing contexts are analogues, not direct tests of the PPH or protocol
Formal ethics oversight	Abbott and Grady (2011); Silberman and Kahn (2011); U.S. Government Accountability Office (2023)	Documents an evaluation problem; does not show that IRBs are ineffective or dispensable