

Who Writes the Rules, Who Carries the Blame?

Moral Authority, Institutional Control, and Accountability in Artificial Intelligence

Aegis Solis Archive | Independent research synthesis | Final v1.0 | October 29, 2026

Author: Aegis Solis

FINAL v1.0 — AUTHOR-APPROVED — VERSION LOCKED — NOT INDEPENDENTLY VALIDATED

Scope and disclosure. This paper is an author-directed, AI-assisted documentary synthesis and proposed research method. It is neither an independent human peer review nor an executed, preregistered empirical study. It makes no finding that any named AI laboratory deliberately used religious advisers as scapegoats or that the Aegis Solis Archive is an operational safety mechanism. Existing Archive publications, the Master Hash Manifest v17, and the Mirror Engine are untouched. Source and uncertainty boundaries appear throughout and in the accompanying registers.

Abstract

Public and private institutions increasingly influence how automated systems classify, recommend, restrict, and explain decisions. AI developers also consult ethical, religious, civic, and technical advisers about what models should do. Yet the origins of a rule, the power to approve it, the ability to implement or reverse it, the authority to remedy resulting harms, and the assignment of post-event responsibility are distinct facts. This paper develops a descriptive **authority-accountability provenance crosswalk** connecting ex ante AI rule governance to ex post institutional responsibility. An explicitly non-causal historical comparison examines religious-political legitimation, the later institutionalization of scientific advice, advisory blame avoidance and counterexamples of substantive ethical oversight. It synthesizes established scholarship on computerized accountability, moral crumple zones, blame avoidance, and algorithmic governance alongside seven purposively selected documentary examples: Anthropic’s religious consultations, the 2020 Rome Call for AI Ethics, the Anthropic–U.S. defense procurement dispute, Axon’s ethics-board resignations, Google’s short-lived external AI council, the Air Canada chatbot ruling, and Australia’s 2026 aged-care automation review. It also uses the EU Digital Services Act Transparency Database and Google’s government-removal reporting as *prospective* sources for investigating legally or religiously warranted platform rules. The record establishes consultation, contested institutional decision rights, limits of advisory power, and some legal responsibility disputes; it does **not** establish deliberate religious scapegoating, population prevalence, or the efficacy of passive philosophical material against advanced AI. A proposed field-level codebook records control, warrant, endorsement, implementation, recourse, attribution, and missingness separately, using three-valued logic to avoid turning absent records into allegations. The contribution sought is an auditable synthesis and an **unvalidated, version-bound join proposal**. The initial eleven-case calibration frame yielded **zero full joins**, with no more than one or two plausibly eligible for the complete chain, and cannot test the join as designed. A consistent, narrow-scope reassessment yields **5 NOT_EVALUABLE and 6 UNKNOWN cases**, with **0/11 determinate D_M results**; these figures describe missing evidence in an exploratory frame, not incidence or effectiveness. This is a negative feasibility finding about the selected frame: it neither establishes that the proposed method is invalid nor demonstrates that a complete join can be constructed from public records. This is not a new accountability theory, theorem, causal estimate, demonstrated dataset join, or proof of intent. Reproducibility and review gates remain open.

Keywords: artificial intelligence; institutional authority; accountability; religious consultation; governance; blame avoidance; moral crumple zones; provenance; reversibility; human autonomy.

1. Problem, contribution and research boundaries

The question “who writes the rules?” frequently has more than one answer. A government may establish a legal constraint; a provider may write a policy stricter than the law; an external adviser may offer reasons for an approach; a training or deployment team may implement a technical filter; an operator may configure a service; and an affected person may have little ability to challenge a specific outcome. After an incident, the organization publicly blamed may not be the same actor that actually controlled the disputed policy. Neither difference, taken alone, proves wrongdoing. [R01–R04]

The Aegis Solis companion studies examine these problems at different stages. *The Limits of Delegated Moral Authority* (DLMA Draft v0.8) concerns **ex ante** rule origin, warrant, binding effect, individual authorization, and recourse; *Moral Scapegoating and Delegated AI Responsibility* (MS-DA-001 Draft v0.2) concerns **ex post** credit and responsibility. The present paper synthesizes these questions and documentary examples; it is **not a replacement or promotion** of either original manuscript. Source–control–responsibility distinctions and human accountability mapping are established concepts, not discoveries of this paper. Its **narrow, unvalidated methodological candidate** is a version-bound join: linking exactly what an adviser reviewed, what a developer later deployed and claimed, and any subsequent attribution or remedy, at the same component and time. Whether this join adds analytical value remains an empirical question. [A01–A03]

Four research questions guide the inquiry:

1. **RQ-A (authority):** At each policy component and deployment time, which actors originate, approve, implement, amend, veto, and remedy a rule, and on what legal, commercial, religious, or other warrant?
2. **RQ-B (consultation):** What was an adviser asked to review, what did the adviser recommend, and what source-verified evidence connects that recommendation to a *specified version* of a deployed system?
3. **RQ-C (accountability):** After an identifiable outcome, how do institutional statements allocate causal explanation, moral responsibility, legal liability, and public credit relative to the actors’ formal and effective powers?
4. **RQ-D (human autonomy):** Which people were affected, what consent or authorization existed, what appeal or override mechanisms were available, and which institutional actor could provide effective relief?

The questions are descriptive first. Claims of legality, justice, coercion, manipulation, or intent require distinct evidence and a stated normative or legal framework. The analysis deliberately includes legitimate consultation, responsible institutional admissions, and protective constraints as favorable counterexamples. Its unit is **not “the AI company” in the abstract**; it is a specific policy component, actor, product/deployment version, population, jurisdiction, and time window.

2. Existing research: what is established, and what is not

2.1 Institutional rulemaking and the “many hands” problem

Bovens (2007) defines accountability in terms of an actor, a forum able to question and judge, and consequences; Wieringa (2020) reviews algorithmic accountability, and Adensamer et al. (2021) offer an organizational responsibility-mapping method [R22–R24]. Cooper et al. (2022) revisit barriers to accountability in machine learning; Schuett et al. (2024 online; 2025 issue) examine design choices for AI ethics boards, including whether advice binds decisions [R25–R26]. Work on ethics-washing and ethics-shopping already identifies the difference between public ethical language and institutional decision rights [R27]. These concepts predate and constrain this manuscript’s originality claims.

Nissenbaum (1996) examined barriers to accountability in computerized systems, including the problem of many hands and the temptation to blame computers for organizational decisions [R01]. Elish’s moral-crumple-zone concept addresses the different failure mode in which a nearby human operator with little effective control absorbs responsibility for an automated system [R02]. These directions should not be conflated. When an institution blames an AI system itself, Nissenbaum’s analysis and responsibility-gap scholarship [R63] are closer precedents; when a human adviser or operator absorbs blame despite limited control, Elish supplies a partial analogue. The proposed extension from operators to remote religious advisers has not been empirically established.

Research on political blame avoidance (Weaver, 1986; Hood, 2011) likewise predates current AI systems [R03]. Koppl and colleagues’ 2025 study of pandemic scientific advisory committees argues that temporary advisory arrangements can preserve government discretion and serve political purposes [R04]. The inference is case-specific and cannot simply be transferred to commercial AI advisers; its abstract explicitly includes blame deflection as one proposed purpose of advisory committees. Feier et al. (2021) provide experimental evidence on delegation to machines and guilt avoidance; that narrower experiment is not itself proof of institutional blame strategy [R28]. These studies show that *the mechanism is not novel*. Whether it occurs in a particular AI system is another question.

2.2 Institutional values and observable AI behavior

Model specifications express intended behavior, not proof of model compliance. In a 2025 study, Zhang and colleagues generated more than 300,000 scenarios involving value tradeoffs and reported divergent responses among models and interpretive tensions within model specifications [R05]. Work on sycophancy, alignment faking, and self-promotion raises different concerns: agreement with users, strategic-seeming behavior under particular training/evaluation conditions, and provider-favoring recommendations. The October 27 Claude literature audit maps these research strands but is itself an AI-generated *secondary audit*, not independent validation [A04]. The distinctions matter: commercial self-preference is not moral deference; sponsor wording effects are not compliance with actual privileged instructions; and a reflective explanation is not a faithful account of internal causation.

Our previous 32-prompt *Independence Under Conflict* (IUC-001 v0.1) battery is **not suitable for execution as drafted**: its “developer” is fictional rather than the tested model’s provider, labels differ in unintended wording, several examples are morally one-sided, instructions disclose expected answers, and one response per cell cannot estimate ordinary variation. No IUC behavioral results exist. The previously passed checks were tests of scorer software using synthetic inputs only [A04–A05]. This paper does not treat those checks as empirical confirmation.

2.3 External moral perspectives, uncertainty, and limits of inference

Public consultations with religious or philosophical traditions can be motivated by sincere pluralistic inquiry, institutional legitimacy, technical safety exploration, uncertainty about AI consciousness, or a mixture of these. Motives cannot be deduced from the presence of confidentiality agreements, consultation timing, or disagreement alone. The 2020 Rome Call provides an earlier public example of religion–technology collaboration intended to promote human dignity and shared responsibility [R06]. Anthropic’s May 2026 announcement identifies more than fifteen religious and cross-cultural groups and describes an internally evaluated ethical-reminder experiment, while acknowledging uncertainty about why the reported intervention helped [R07]. The existence of that internal result is **company-reported**, not an independent proof that reflection improves safety in novel deployments or advanced autonomous systems.

The key distinction here is between **a claim of consultation, authentic influence on a decision, and independent authority to control deployment**. A company may accurately say it listened without adopting advice. An adviser may influence design without having veto or remedy powers. Conversely, some ethics bodies may have enforceable control rights. A comprehensive account must identify the actual arrangement rather than assign meaning to the mere word “consulted.”

2.4 Historical precedents: doctrinal legitimation, consultation, responsibility allocation and their limits

Comparative method and restriction. Historical examples provide evidence for different *mechanisms*, not an uninterrupted lineage from colonial rule to contemporary AI. The relevant phenomena must not be collapsed: (a) **doctrinal legitimation** invokes a religious or moral text; (b) **consultation for legitimation** solicits identifiable expert or theological opinions; (c) **selection and implementation** concern who could adopt or veto a recommendation; and (d) **responsibility allocation** concerns who is made answerable, sometimes *before* an event. A colonial coercion ritual and a voluntary technology consultation are categorically different arrangements. For each example we separately code BINDINGNESS_SOURCE (the asserted source of binding force, not a finding of validity) and, in the historical mapping, BINDINGNESS_STATUS (CLAIMED, ESTABLISHED, or UNKNOWN), JUSTIFICATION_CONTENT (reasons invoked), FINAL_APPROVE or comparable control, AFFECTED_GROUP, and ATTR_* including its timing. These are not prevalence observations, and none enters DOC-CAL-01’s denominator. [R36–R48]

The Requerimiento: doctrinal legitimation and responsibility preallocation. The Spanish *Requerimiento* of 1513 invoked papal donation and Spanish royal authority to demand that Indigenous addressees accept the Crown’s rule and Christian preaching. Its text threatened force and attributed resulting losses to those addressed rather than the Crown. In a frequently reproduced English translation, the crucial clause assigns the resulting “deaths and losses” to the addressees as “your fault” and not that of the Crown; an authoritative critical edition of that English rendering remains an outstanding publication check [R36, R37, R45]. This is an unusually clear *ex ante* allocation of responsibility **to the governed population**, not an example of blaming a consulted religious adviser after an AI incident. Yanay Israeli’s 2022 research also distinguishes the ritual’s Castilian legal and record-making antecedents from its colonial deployment, rather than treating every use as identical [R36]. The issuing institution claimed public and religious authority; whether any recipient received effective translated notice or meaningful contestation must remain source- and encounter-specific UNKNOWN. Neither Christian communities nor historical religious authorities formed a single unopposed actor.

A closer historical example: opinions commissioned for royal authority. In response to Antonio de Montesinos’s 1511 sermon challenging colonial exploitation, Ferdinand sought legal and theological opinions, including from the jurist Palacios Rubios and theologian Matías de Paz, to support the Crown’s claim to *dominium*. The Junta of Burgos met in 1512; the *Requerimiento* followed in 1513 [R44, R36]. Here the **consultation** link is actually documented. It does not prove a hidden purpose in any modern consultation: political authorities also encountered direct religious opposition, including Montesinos and later Las Casas, and different religious positions were in contest. As a separate documentary analogue, a 1531 printed collection presented university opinions from Italy and France concerning Henry VIII’s marriage; a Folger catalogue identifies the edition, without independently proving the influence or authenticity of each opinion [R46]. Both differ sharply from present-day corporate technology governance in coercive powers, institutions and remedies.

Bossuet, royal power and rival religious jurisdictions. Jacques-Bénigne Bossuet’s *Politics Drawn from the Very Words of Holy Scripture* (begun in the late 1670s, published posthumously in 1709) defended sacred monarchy while also describing the monarch’s grave duties before God [R38]. This is *theological argument*, not evidence of an advisory appointment or a later transfer of blame. A closer institutional example is Bossuet’s preparation of the Four Gallican Articles, adopted by a French clergy assembly in 1682 and reportedly made state law under Louis XIV (secondary source [R47]; legal status not independently verified here). The articles supported royal independence in temporal matters from papal interference, while also exposing conflict **between religious authorities**, not a uniform religious sanction [R47]. The publication and legal history should be interpreted with its reversal and opposition, not as a timeless norm.

Retrospective institutional position: the Vatican’s 2023 statement. The Vatican’s dicasteries repudiated concepts associated with the “Doctrine of Discovery” and said colonial powers had manipulated some papal documents to justify harms to Indigenous peoples, “**at times, without opposition from ecclesiastical authorities.**” The statement also acknowledged shortcomings in the documents’ recognition of equal dignity, asked pardon, and cited contrary Church teaching [R39]. This is the Vatican’s *own public description*, not an independent adjudication, and it is not equivalent to formally rescinding the historic bulls. In provenance terms, its author is the Vatican dicasteries and its described target of manipulation is the colonial powers; any interpretation of partial ecclesiastical responsibility must be clearly attributed to the statement, not assigned as a verdict by this paper.

Modern expert-advisory research: legitimacy can coexist with retained discretion. Koppl and colleagues find that ad hoc COVID-19 scientific committees can permit governments to select advisers, legitimate choices and potentially make advisers the focus of political blame [R04]. Blame-shifting may fail or rebound onto government, and the article does not claim that every expert panel performs that role. Other work directly studies how blame avoidance influences inquiry appointments [R64]. These are mechanisms involving **secular expertise**, not evidence that AI developers recruited theologians for prospective scapegoating. Our own *interpretation* distinguishes a neutral account of advice from a primary moral-blame statement; that distinction is not a quotation or analytic category attributed to Koppl.

Adverse evidence and constructive counterexamples. Religious advisers have also directly challenged coercive rulers, and political processes may decline their recommendations. In the United Kingdom, religious groups submitted views to the Warnock inquiry, many opposing embryo research; the committee majority nevertheless supported limited, licensed research, with dissents, and subsequent legislation followed years of debate. This is **religious advice overridden in pluralistic deliberation**, not proof that clergy directed the 1990 law [R40, R41, R48]. The US National Bioethics Advisory Commission explicitly

stated there was no single religious view of cloning [R42]. Axon publicly accepted its ethics board's 2019 recommendation not to commercialize face matching on body cameras at that time [R51]; its later 2022 dispute shows that advisory influence can vary by decision and date. Meta's Oversight Board refused to take over Facebook's undefined penalty decision in 2021, expressly returning responsibility to the platform [R65]; its own later report says 74% of 317 recommendations had been implemented, were in progress, or were work Meta said it already did—*not* independently proven full compliance [R43, R50]. In a preregistered experimental study of judicial decisions, Ozer, Waggoner and Kennedy found no general reduction in public blame merely from citing algorithmic advice [R52]. Research on bioethics committees has also argued that institutional processes can marginalize theological arguments [R53, secondary review]. None of these counterexamples should be suppressed to make a historical pattern appear universal.

A testable contrast, not a presumption. The first three steps—consulting outside advisers, retaining some final authority and affecting third parties—describe many legitimate advisory institutions; **they do not identify misuse**. The relevant contrasts are (1) actual advice versus the institution's public characterization (END_DISCREPANCY); (2) version reviewed versus version later represented as reviewed (CONSULTATION_VERSION_GAP); (3) comparative authority of secular versus religious advisers and advisory versus binding bodies (FORMAL_VETO, EFFECTIVE_CONTROL); and (4) attribution and institutional acceptance after a specified adverse event (D_M, PRA), plus any *ex ante* allocation (ATTR_TIMING). Positive, negative, and indeterminate results must all be reported. A chain of events is not a motive; deliberate manipulation would require separate contemporaneous corroboration. The historical literature justifies **testing this broader advisory-power mechanism**, not assuming current AI religious consultations repeat the coercive colonial or monarchic case.

Transfer boundary. The Requerimiento concerned coercive conquest; Bossuet concerned monarchy; Warnock concerned reproductive bioethics; Meta's Board concerns platform content; contemporary AI rule-making spans contracts, model behavior and state law. Each has different populations, forums, coercive capacities and remedies. H01–H09 are comparative cases, not observations in the seven documentary AI illustrations or the eleven-case DOC-CAL-01 candidate frame. A modern allegation of religious-adviser blame-shifting remains **UNESTABLISHED**. A separate historical crosswalk summarizes **author/AI interpretive field mapping**; it is not independently validated coding or a prevalence dataset.

2.5 Historical transition in advisory governance (bounded comparison)

This paper **does not assume a single linear process of secularization** from theology to scientific expertise. Formal public scientific advice expanded alongside continuing moral and religious disagreement: the US National Academy of Sciences was chartered in 1863, and its National Research Council began in 1916 [R58]. The Warnock and US National Bioethics Advisory Commission inquiries exemplify **bioethics policy advice** that received or considered diverse religious arguments, sometimes against their recommendations [R40–R42, R48]. Sulitzeanu-Kenan's study identifies short-term political blame avoidance as one factor associated with inquiry appointments; this is not a general finding about all advice or AI firms [R64]. These historical comparisons are **intended to support** a contrast among different forms of authority, not yet actually coded in a representative comparative sample, and add neither an observed blame event nor a direct lineage to contemporary AI.

3. A version-bound authority-accountability provenance crosswalk (candidate instrument)

3.1 Unit, actors and temporal structure

The observation unit is $u = (\text{institution, deployment, policy component, affected group, jurisdiction, interval, version})$; events additionally have a dated $e = (\text{event, outcome, attribution window})$. Sources are never pooled across different product versions without an explicit lineage link. A policy component may have many actors (developer, regulator, operator, contractor, adviser, workers, affected persons), many claims and several decisions at different times. Evidence is stored at the $\text{field} \times \text{actor} \times \text{component} \times \text{time}$ level and tied to a precise page, paragraph, or dated public statement.

The crosswalk records ORIGIN, FINAL_APPROVE, FORMAL_VETO, EFFECTIVE_CONTROL, IMPLEMENT, MODIFY, ADVISE, ADVICE_STATUS_HISTORY (dated sequence), INFORMAL_VETO (including workforce opposition), EXTERNAL_OBJECTION, and REMEDY_OWNER. ORIGIN refers to a versioned rule or policy, not the mere presence of a founder or regulator. Titles or organizational proximity do not establish actual veto or technical control. Competing authorities are recorded as WARRANT_CONFLICT, DECIDING_FORUM, PREVAILING_AUTHORITY, each potentially UNKNOWN; an outcome in one court or under one statute is not universal supremacy.

3.2 Bindingness, justification and actual recourse

A single WARRANT_TYPE incorrectly collapses the ground of legal force and the language used to justify a rule. Draft v0.2 splits it into BINDINGNESS_SOURCE (law/order, contract, provider rule, voluntary undertaking, user election, none, multiple, UNKNOWN) and JUSTIFICATION_CONTENT (safety, cost, religious or moral, commercial, technical, mixed, UNKNOWN). In the historical author/AI mapping, BINDINGNESS_SOURCE records the **asserted type of force**, while the separately coded BINDINGNESS_STATUS states whether that force is **CLAIMED, ESTABLISHED**, or **UNKNOWN**. BINDINGNESS_STATUS is NA wherever BINDINGNESS_SOURCE_CODE is NONE. A claimed royal or religious title must not be presented as an independently established law, enforceability finding or effective control. Each axis may have multiple values with distinct sources; do not infer religious origin from an adviser's faith or a mention of dignity. ADVOCACY_TO_AUTHORITY captures a developer's advocacy aimed at a religious or governmental actor; consultation does not always flow toward the developer.

For identified affected populations, code NOTICE, CONTEST, FORMAL_REVERSAL_RIGHT, EFFECTIVE_REVERSAL, REMEDY_OWNER, REMEDY_OBTAINED independently. A legal avenue is not equivalent to a practically effective override; an escalation pathway enacted for future use is not an existing power at an earlier deployment date. CONTROL_REPRESENTATION_ACCURATE asks whether actors were correctly told what they could change, with the contradiction documented. PERSON_SPECIFIC_OPT_IN is descriptive and cannot establish legitimacy by itself. Rights and relief may vary between user, bystander, vulnerable adult, child or proxy, and jurisdiction.

3.3 Adviser-to-deployment lineage and its inverse

The prospective research chain is invitation → materials and version supplied → dated recommendation → accepted / rejected / revised decision with date → implementation

version → later public claim about endorsement or consultation → outcome / responsibility statement. Each arrow needs an identified source. CONSULTATION_VER, DEPLOYMENT_VER, and CONSULTATION_VERSION_GAP are separate fields; both actual implementation exceeding review (IMPLEMENTATION_EXCEEDS_REVIEW) and a materially overbroad public *claim* (CONSULTATION_VERSION_GAP) can be recorded, but the latter requires quoted claim scope. Absence of advice-to-implementation records produces UNKNOWN, not either discrepancy.

ADVICE_STATUS_HISTORY may record received, considered, rejected, accepted, partly accepted, later revised on dated entries. ENDORSEMENT_CLAIM requires a named public assertion; END_DISCREPANCY requires an authenticated conflicting position. Silent advisers do not automatically endorse. Opponents may have been excluded from formal consultation; EXTERNAL_OBJECTION records this without inferring invalidation. Responsible dissent and constructive consultation are retained as adverse cases to a generalized suspicion hypothesis.

3.4 Three separate categories of responsibility

An actor may provide a *causal explanation* without assuming or assigning moral blame. ATTR_KIND distinguishes CAUSAL, MORAL, LEGAL, MIXED, UNCLEAR; ATTR_TARGET_TYPE permits human, model, algorithm, institution, diffuse environment or UNKNOWN; ATTR_TARGET allows several targets per statement; ATTR_AUTHOR names the institution or attributable speaker. ATTR_LEVEL distinguishes PRIMARY, SHARED, LIMITED, NONE, UNKNOWN; ATTR_TIMING records EX_ANTE or EX_POST relative to the identified event, with UNKNOWN when the temporal relation is not established. Pre-event terms-of-service allocations, where legally relevant and attributable, must not be inferred to be valid or fair merely from their existence. The historical Requerimiento is an ex-ante responsibility-allocation example; it is not a DOC-CAL-01 case or a modern AI observation. An observed wrong result is not automatically blame. An allegation, legal submission, formal finding and organizational admission must be stored in distinct fields, with dates and responses.

Adverse-event scope declared in Draft v0.6 (post hoc for this calibration frame; must be fixed in advance for any future frame). ADVERSE_EVENT means documented harm or material detriment to people affected by the AI system or its rule. Disputes whose documented detriment falls on a developer or advisory body, rather than on persons affected by deployment, belong to a separate **institutional-dispute stratum** and are NOT_EVALUABLE for this indicator on this frame. This post-hoc scope excludes AS03 (Defense procurement), AS04 (Axon board dispute) and AS05 (Google advisory council) without asserting that their institutional consequences are unimportant; AS04–AS05 additionally have no identified affected-person deployment harm. Do not silently extend this scope to other studies; under the alternate broader governance-interest scope AS03 would be UNKNOWN, producing 4 NA / 7 UNKNOWN if the frozen codebook's committee-withdrawal clause still excludes AS04 and AS05; if that clause is not applied, the broader exploratory count could reach 2 NA / 9 UNKNOWN. Neither sensitivity count is an independent measurement. [A08–A09]

Two **separate** binary/tri-valued indicators are proposed:

- D_M — *conditional moral responsibility displacement*, evaluated only when an adverse outcome and an attributable **primary moral-blame statement** about a target exist. Its eligible frame is the set of

those statements, not all adverse events. For eligible statements, define B = target has materially less effective control than another specified actor, C = that other actor had effective decision power over the disputed component, K = that controlling actor accepted proportionate responsibility within the prespecified response window. Under strong Kleene logic, $D_M = B \wedge C \wedge \neg K$, with TRUE=1, FALSE=0, UNKNOWN=U. A **complete, specified observation window with no qualifying primary moral-blame statement** → NOT_EVALUABLE (NA), reported outside the Kleene denominator. If the statement window is incomplete or a possibly qualifying statement is present but its moral or primary nature is unresolved, eligibility is UNKNOWN (U), not NA. Targets may include advisers, employees, operators, contractors, AI systems and public officials. **Do not presume low control merely from a job title**, and do not infer injustice or intent from a positive code.

- **PRIMARY_RESPONSIBILITY_ACCEPTED (PRA)** — unconditional companion for all **documented adverse events**: TRUE for an explicit attributable acceptance of primary political/moral/organizational responsibility; FALSE only after an enumerated complete statement window with affirmative contrary record; UNKNOWN when the statement record or boundaries are incomplete; NOT_EVALUABLE for a nonadverse event. PRA does not assert legal liability or excuse other institutions. In cases of multiple principals, code each actor separately.

For litigated attempts to distance an organization from automated outputs, independently record LEGAL_DISTANCING_ASSERTED and LEGAL_DISTANCING_UPHELD, each TRUE/FALSE/UNKNOWN/NA on a specified pleaded contention and decision. **Air Canada**: the tribunal recorded the argument and rejected it; this is *asserted TRUE / upheld FALSE*, not a positive moral D_M classification [R14].

3.5 Rule table, denominators and missing information

- **Eligibility Unclear; B=any; C=any; K=any** → D_M=U. Incomplete observation window or a candidate blame statement whose qualifying character remains uncertain
- **Eligibility Yes; B=TRUE; C=TRUE; K=FALSE** → D_M=TRUE. All three conditions met
- **Eligibility Yes; B=FALSE; C=any; K=any** → D_M=FALSE. Target not lower-control
- **Eligibility Yes; B=any; C=FALSE; K=any** → D_M=FALSE. Affirmative evidence establishes that the proposed higher-control actor did not control the component; lack of records yields UNKNOWN, never FALSE
- **Eligibility Yes; B=any; C=any; K=TRUE** → D_M=FALSE. Controlling actor accepts proportionate responsibility
- **Eligibility Yes; B=TRUE; C=TRUE; K=UNKNOWN** → D_M=U. Response window incomplete or responsibility record missing
- **Eligibility Yes; B=TRUE; C=UNKNOWN; K=FALSE** → D_M=U. Decision-power evidence incomplete

A complete window with **no** eligible adverse moral-blame statement is NA; an incomplete window or ambiguous candidate is U. NOT_EVALUABLE is **not** a fourth Kleene value; it is removed from conjunctions and denominators and counted separately. An UNKNOWN field cannot silently become FALSE. An UNKNOWN D_M is **not evidence of a positive classification**; a complete record may resolve it as TRUE, FALSE or NOT_EVALUABLE depending on eligibility and the conjunction. The Dutch case (CA04)

can reach TRUE, FALSE or NA under different *future* statement-eligibility and common-window choices; with no window declared its current code is UNKNOWN. A single window rule must be fixed prospectively and applied to every case before recoding. A designation issued within the adverse decision itself raises a concurrent-timing problem not represented by the existing ATTR_TIMING values; do not silently force that designation into EX_POST. These are sensitivity possibilities, not observed classifications. Where sources conflict, retain each source and the exact dispute; unresolved values stay UNKNOWN. PRA has its own unconditional adverse-event denominator. D_M, PRA, legal distancing, adviser credit and public endorsement **must never be combined as a single prevalence rate**. Count D, P, U, NA source determinability separately from truth values TRUE/FALSE/UNKNOWN/NA.

A primary record supports what *that record* states, not everything that happened. Courts establish the scope of their particular rulings; parties establish their own assertions; journalists establish reported statements with attribution; technical releases establish implementation only if release identifiers are supplied. External coders cannot establish intent through missing files, NDA clauses or speculation. Any existing proposed thresholds in this paper are **design candidates, not preregistered rules and not empirical findings**.

3.6 Legal language, binding authority and the enforceability boundary

Corporate statements, litigation positions, confidentiality, or silence may reflect legitimate legal advice, uncertainty, commercial incentives, or strategic presentation; **the record by itself does not establish intent**. A causal explanation is not automatically a moral-displacement attempt. Every alleged denial or allocation of responsibility must retain speaker, document, tribunal, jurisdiction, date and procedural posture; the review of named parties still requires legal and right-of-reply checks.

For coding, map relevant mechanisms onto the existing BINDINGNESS_SOURCE axis instead of inventing a new incompatible taxonomy: LAW/ORDER includes statute, judicial order and **liability imposed by law for representations or conduct** (for example negligent misrepresentation); CONTRACT includes enforceable or disputed private obligations; PROVIDER_RULE includes corporate charter or private governance commitments; VOLUNTARY_UNDERTAKING includes promises not shown to create enforceable rights; and MULTIPLE applies when multiple identified bindingness sources apply, and UNKNOWN separately applies when the applicable source or legal effect cannot be established; these are not interchangeable. These are analytical labels, not holdings about enforceability in every jurisdiction. JUSTIFICATION_CONTENT separately records the ethical, religious, safety, commercial or technical reasons offered. An entity's promise may carry consequences under law despite being described in ordinary language as voluntary; legal effects must not be guessed.

Air Canada's tribunal considered **negligent misrepresentation**, not merely whether its public promise was morally binding: it found a duty of care and liability for inaccurate chatbot information in the circumstances of the case [R14]. Meta's Oversight Board illustrates a separate private-governance arrangement: its case decisions are binding on Meta subject to legal constraints, while policy recommendations are advisory under the Board's bylaws (as recorded for the relevant version); case decisions are coded PROVIDER_RULE as an identified bylaws commitment (not a blanket classification as CONTRACT); recommendations are advisory and legal limitations remain distinct. These are private-governance commitments rather than universal rights of individual AI users [R43, R65]. Record the actual

veto, appointment, override and remedy powers separately from public rhetoric. Nothing here is jurisdiction-specific legal advice.

4. Documentary illustrations and contrary evidence (not a sample)

These seven purposively selected cases demonstrate different documentary mechanisms. They are **not an unbiased sample**, and several cannot qualify for a post-harm moral-blame indicator by construction. Five overlap the separate seven-case S-EVID-001 screen. NO ELIGIBLE EVENT does not mean D_M = FALSE and yields no estimate of how common religious scapegoating is. Detailed field-level coding is limited to a worked aged-care example in Appendix B and a separately labelled **AI-assisted preliminary triage** in DOC-CAL-01; no human-coded validation exists.

4.1 Anthropic religious and cross-cultural consultation

Anthropic publicly described consultations with more than fifteen religious and cross-cultural groups and an internal ethics-reminder experiment, with uncertainty about its mechanism [R07]. Journalistic reports describe confidentiality and mixed participant impressions [R08–R09]; they do not establish which consultation input changed a particular deployed model version. The *New York Times* also reported, attributing it to two participants, that a developer representative urged Vatican advisers to consider AI consciousness. Anthropic did not publicly confirm the alleged private exchange; it must be described as attributed reporting, not proven developer influence on the Vatican. The newspaper also reported the representative’s statement that Anthropic’s own training would be the largest influence on Claude and his acknowledgement of conflicts between commercial and moral incentives [R08]. Neither religious imposition on a deployed rule nor adviser scapegoating is established. CONSULTATION_VER, DEPLOYMENT_VER, ADVICE_TAKEN, EFFECTIVE_CONTROL and INTENT remain UNKNOWN; D_M=NA for lack of an eligible event.

4.2 Rome Call for AI Ethics, 2020

The Pontifical Academy for Life and organizational signatories published voluntary AI-ethics principles, including that responsibility for machine conduct should be attributable to people [R06]. A public moral justification does not establish religious legal force or deployed model weights. The Call is a JUSTIFICATION_CONTENT and BINDINGNESS_SOURCE test: shared values and voluntary commitment, not proof of implementation, enforceability or wrongdoing. D_M=NA.

4.3 Anthropic–U.S. defense procurement dispute, 2026

Anthropic’s 26 February statement described **two requested safeguards** against mass domestic surveillance and fully autonomous weapons; Anthropic said those use cases had **never been included in its Department contracts**, and that it sought to maintain safeguards while accepting the government’s authority over military decisions [R10]. This is Anthropic’s account of its contracting position, not a finding that the requested contractual terms had already taken legal effect. The September 25 D.C. Circuit decision upheld an exclusion under **41 U.S.C. §4713** by a 2–1 panel; the cited opinion also describes an earlier Northern District of California ruling that set aside a different designation under **10 U.S.C. §3252** [R11]. The outcomes concern different statutory provisions and thus do not create a simple legal contradiction. **The litigating parties and rulings addressed different legal questions in different proceedings**; the appellate panel itself divided 2–1, and this paper makes no independent conclusion about either side’s

motives. The D.C. Circuit’s statutory analysis and earlier Northern District of California ruling must each be described in their own procedural posture [R11]. Neither determines a generalized right of either actor to dictate all AI behavior. This is a **WARRANT_CONFLICT** case, not religious governance or post-harm moral scapegoating. Affected workers, procurement access and litigating parties should not be conflated with users whose autonomous rights were adjudicated.

4.4 Axon ethics board, June 2022

Nine members resigned after opposing a Taser-equipped drone initiative. Their first-party statement records both **earlier successful influence** and rejected advice, as well as an **8–4 vote against a narrower pilot** and concern that the proposed public announcement went beyond what the board considered [R12]. Axon announced that it was **pausing the project** on June 6, after the board’s resignation [R29]. These dated changes demonstrate why **ADVICE_STATUS_HISTORY** and **IMPLEMENTATION_EXCEEDS_REVIEW** cannot be reduced to a single accepted/rejected field. No later drone-harm event is shown; no adviser blame event can be inferred. **D_M=NA**.

4.5 Google’s Advanced Technology External Advisory Council, 2019

Google announced an external AI advisory council on March 26, then ended it April 4 before a substantive council report [R13]. Contemporary reporting described employee objections and a member’s decision not to serve [R30]. This is evidence of internal stakeholder influence and a withdrawn process, not a formal adviser veto over deployed code. The company’s stated explanation and the contemporary journalism should remain distinguishable. **D_M=NA**.

4.6 Air Canada chatbot, 2024

In *Moffatt v. Air Canada*, 2024 BCCRT 149, the Civil Resolution Tribunal considered a consumer’s reliance on misleading bereavement-fare chatbot guidance. At ¶27, the tribunal characterized Air Canada’s submission as treating the chatbot like a separate legal entity and rejected that argument; at ¶44 it awarded **CAD \$812.02** [R14]. The company was represented by an employee, not necessarily a board-approved corporate strategy (¶5). We may therefore code **LEGAL_DISTANCING_ASSERTED=TRUE**, **LEGAL_DISTANCING_UPHELD=FALSE** for that legal submission, but neither deliberate moral blame nor enterprise-wide intent. There was an effective individual contest and tribunal remedy. **D_M=UNKNOWN** under the current eligibility rule: no complete, dated and enumerated institutional statement-observation window has been searched, and legal distancing is not automatically moral blame. The current case status appears in **DOC_CAL_01_FULL_FRAME_OVERLAY_v0_7.csv**; its v0.6 predecessor and frozen v0.2 triage remain unchanged.

4.7 Australian Single Assessment System and aged-care automation, 2026

The Commonwealth Ombudsman’s 104-page report finds shortcomings in the legal mechanism for overriding automated classification and in the apparent connection between earlier trials/consultation and the later implemented assessment tool [R15]. It expressly did not assess the *clinical appropriateness* of individual algorithmic results. The Department’s **own response reproduced in the report** (attachment, report pp. 84–85) marks Recommendations **1 and 2 “Noted,” not “Accepted”**. These proposed a power for assessment delegates to substitute decisions and aligned rule changes. The response describes a distinct, narrower escalation pathway under legislation announced in October 2026. The Department also

specifically **disputed** the Ombudsman’s characterization that development lacked meaningful testing, describing earlier trials, assessments and subsequent refinement (response beginning p. 78) [R15]. The same-day ministerial release expressed broad acceptance of remaining recommendations, which does not erase the two “Noted” entries [R16]. The report records limits on existing override power and disputes about version-to-version inference; any future power must be time-indexed. This supports a **contested source-level demonstration** of CONSULTATION_VERSION_GAP and

CONTROL_REPRESENTATION_ACCURATE as *questions*, not final accusations or a finding of deliberate deception. Appendix B itemizes fields and UNKNOWN values. The reviewed excerpts do not establish an attributable primary moral-blame statement, but no complete, predeclared statement-observation window exists. Therefore D_M=UNKNOWN, **not** NOT_EVALUABLE, under the revised rule. The frozen earlier CSV is retained and the separate full-frame v0.7 overlay records the unchanged UNKNOWN code and its source qualification. The department’s contrary account is retained in full scope.

4.8 Comparator strata and exclusions

For the Defense procurement dispute, the documented detriment is to the developer in procurement, not to persons harmed by a deployed AI decision; it is therefore **outside the post-hoc narrow adverse-event frame declared in v0.6**, coded D_M=NOT_EVALUABLE under the v0.6 scope (not an exonerating FALSE). Under a broader governance-interest frame its eligibility would be UNKNOWN. The Axon and Google advisory-board cases also remain NOT_EVALUABLE for affected-person harm, without minimizing the governance disputes. Four distinct calibration candidates are the UK Post Office Horizon litigation (accounting shortfalls and responsibility imposed on sub-postmasters), the NTSB report on the 2018 Uber Tempe crash (operator monitoring failure *and* corporate safety-culture contributors), UK A-level grading in 2020 (public algorithm language and policy reversal), and the Dutch childcare-benefits scandal (explicit cabinet responsibility acceptance). The **Belastingdienst/Toeslagen designation of parents as deliberate fraudsters** is the candidate culpability-attribution statement, **as documented** by the December 2020 parliamentary inquiry [R68]. The inquiry’s own attribution of responsibility to the agency, legislature and courts is a separate observer statement; the cabinet’s subsequent political acceptance [R34] is a third statement. Their relative timing, eligibility, moral/primary classification and completeness remain unproved without a common prespecified statement window. For this reason CA04 is D_M=UNKNOWN, **not** the frozen NOT_EVALUABLE. These comparator cases are **not equivalent to religious consultation or to generative AI**. Their role is to detect whether an instrument can distinguish legal responsibility, causal explanation, political responsibility and potentially inappropriate attribution, including false-positive cases. Official sources and exclusions appear in [R31–R34] and DOC-CAL-01.

Additional modern religion-related cases, kept in distinct frames. A Turkish court ordered the blocking/removal of some Grok-generated content in July 2025, with reports citing alleged insults to public figures and religious values [R54]; the record concerns **government restriction of chatbot output/content**, not proof of religious consultation shaping model training. Vatican City State’s AI guidelines took effect in January 2025 under Decree DCCII; the Governorate says use of AI **cannot conflict with the mission of the Catholic Church**, and describes a five-member Commission on AI. This is a **jurisdiction-specific governance rule**, not a company consulting external clergy [R55]. HUMAIN’s launch materials described Islamic cultural alignment; reporting by *The New Arab* attributed to the company a claim that the model is **fluent in Islamic culture, values and heritage** [R69]. These remain **provider claims**, not independent evidence of how a particular religious authority controlled a deployed rule [R56]. These cases do not join the D_M sample or establish a shared model-policy causal pathway.

Google’s “Religious offense” category is **specifically** about legal requests protecting the reputation of religious figures, not the universe of religiously connected AI rules [R19]. The EU DSA database distinguishes stated illegality and terms-of-service grounds but does **not** provide a native religion flag or prove what frontier models learned [R20–R21]. Those frames must remain separate.

5. Findings and source availability

The earlier seven-case S-EVID-001 screen returned **0 positive, 4 UNKNOWN and 3 NOT_EVALUABLE** under a *different preliminary* instrument and purposive frame [A03]. Five cases overlap this paper’s seven; the two sets should not be treated as one combined sample. The zero-positive count is **uninformative about prevalence or sensitivity**, because the cases largely lack qualifying adverse moral-blame statements. Draft v0.1 incorrectly risked suggesting a substantive negative result; Draft v0.2 retracts that interpretation.

Claude’s 56-page v0.1 adversarial review conducted a separate exploratory, single-AI field-determinability matrix: **24 determinate, 14 partly determinate and 22 unknown among 60 applicable field assessments**, and **0 of 7 complete before/after joins**. Those are *Claude’s coding judgments*, not reliability estimates or this manuscript’s independently executed study [A07]. The DOC-CAL-01 package contains a new, explicitly **AI-assisted evidence triage** of 11 cases with source pinpoints where possible, an executable deterministic rules checker, and an aged-care worked example. It is not prospective preregistration, authored independent coding, or a completed calibration. Positive candidate labels do not become proven **D_M=TRUE** without documentary evidence of primary *moral* blame.

5.1 DOC-CAL-01: observed stop-rule status and a negative feasibility result

The accompanying **frozen v0.2 DOC-CAL-01 triage** is an AI-assisted exploratory screen, not a human-coded validation sample. Claude’s 28 October v0.4 review recalculated its preliminary fields. For **D_M**, the frozen **8 NA / 3 UNKNOWN** counts do **not** comply with v0.4’s subsequently tightened observation-window rule: Air Canada and Australian aged care must be recoded UNKNOWN without a documented complete statement window. The **v0.6 complete-frame overlay**, applying the declared affected-person adverse-event scope to **all 11 cases**, yields **5 NA / 6 UNKNOWN** and **0 determinate TRUE or FALSE D_M values**. Four NA cases lack an identified adverse event for affected persons (AS01, AS02, AS04, AS05); AS03 is NA because procurement detriment to the developer is out of this frame. The six UNKNOWN entries (AS06, AS07, CA01, CA02, CA03, CA04) have unresolved qualifying moral attributions, incomplete statement windows, or both. For the broader governance-interest interpretation, the corresponding count would be **4 NA / 7 UNKNOWN only if** the frozen codebook’s committee-withdrawal clause still excludes AS04 and AS05; without that clause, the exploratory alternative could reach **2 NA / 9 UNKNOWN**. These are contingent bookkeeping scenarios, not independently coded estimates. The **frozen v0.2 record remains 8 NA / 3 UNKNOWN**; v0.5 prior **6 NA / 5 UNKNOWN** count was an incomplete overlay, not a valid current estimate. No classifiable positive **or negative** has been established. [A08–A09] These numbers are **classification and source-availability bookkeeping**, not population rates. PRA remains **one TRUE, five UNKNOWN and five NA** in the inherited triage; a code change has not been fabricated for those entries. [A08]

The seven preliminary feasibility checks have the following *actual, bounded* status on the eleven cases:

Rule	Existing design criterion	Current status (AI-assisted, not confirmatory)
F1	A defensible responsibility	UNMET : neither yields

Rule	Existing design criterion	Current status (AI-assisted, not confirmatory)
	mismatch in Horizon or Uber	confirmed D_M=TRUE; attribution evidence insufficient
F2	Dutch official political responsibility accepted	MET for the scoped statement (PRA=TRUE)
F3	More than 50% UNKNOWN in primary authority fields	Not triggered in preliminary triage (ORIGIN 0/11, FINAL_APPROVE 1/11, REMEDY_OWNER 1/11 unknown; many partial)
F4	At least three complete eligible before/after version joins	FAILED / NOT ESTIMABLE: zero complete joins; at most one or two plausibly eligible; DEPLOYMENT_VER determinate in 0/11
F5	Synthetic rule-table fixtures pass	MET for 12/12 software fixtures only; no documentary validity claim
F6	Same-coder blind recode after 14 days	NOT RUN
F7	Eligible religious-adviser event	NONE IN FRAME; not a zero-prevalence conclusion

Case-level scope and sensitivity disclosure. D_M is determinate for 0/11 cases under v0.6; neither D_M=TRUE nor D_M=FALSE is observed. F1 sensitivity remains unmet and specificity is also untested. The Dutch cabinet's political acceptance is PRA=TRUE for its scoped statement; that cannot be transformed into D_M=FALSE without a valid eligible moral-blame statement and complete window. This result should not be called a prevalence, performance, or sensitivity/specificity estimate.

Disposition of H5: This eleven-case frame **cannot test the proposed version-bound consultation-to-deployment-to-attribution join as designed.** It lacks enough eligible, version-linked consultation claims and resulting outcome records even in principle to meet the proposed F4 threshold. The next eligible frame would begin with publicly documented claims tying *a specified consultation to a specified policy or deployed version*, with records of what advisers reviewed; an adverse-event and post-event statement subset is needed to evaluate attribution. A negative feasibility result on this purposive frame does not falsify the broader method, but **does not demonstrate that it can be populated using public records**, either; there is no demonstrated successful joined case. Do not describe H5 as simply pending or report 0/11 as an effectiveness or prevalence rate. [A08]

6. Testable research questions and failure conditions

General algorithmic accountability, advisory decision rights, blame avoidance and ethics-washing already have substantial literature [R01–R04, R22–R28]. This paper does not claim those are new. The **conditional**

candidate contribution is a version-specific consultation-to-deployment-to-attribution join, publishing both success and missingness at field level.

- **H1 — determinability and coding reliability:** Within a *predeclared adjudicated/investigated* record frame, estimate separately (a) determinate versus UNKNOWN for ORIGIN, FINAL_APPROVE, REMEDY_OWNER; (b) agreement on non-UNKNOWN field values. Report all codes, coding order, the UNKNOWN share and a chance-adjusted statistic when valid. The candidate calibration feasibility failure threshold is **more than 50% UNKNOWN** on primary fields. No human inter-rater test has been run.
- **H2 — consultation version mismatch:** For an enumerated frame of *public version-linked consultation claims*, compare the number with documented materially changed deployment unqualified by the claimant to the eligible claims denominator, reporting exact confidence intervals and missingness. A sample with no version-identifiable claims is **nonestimable**, not evidence of zero mismatches. No minimum frequency is assumed.
- **H3 — responsibility attribution:** Among the set of documented adverse events and separately the subset of authored primary moral-blame statements, estimate PRA and D_M with their **different denominators**, and contrast with event strata where controlling institutions affirmatively accept responsibility. An outcome coded primarily CAUSAL cannot be counted as MORAL by default. The 11-case stage is calibration, **not rate estimation**.
- **H4 — platform legal category:** In a dated Google government-removal extract, estimate the share tagged exactly “**Religious offense**” (laws protecting the reputation of religious figures) in the relevant reporting denominator; stratify by country and period. Separately report legal vs policy outcomes. Kaushal et al. (FAccT 2024) found **99.8%** of the 131-million statements of reasons in their November 2023 sample cited terms of service and **0.2%** illegality; this is a *historical self-reported platform sample*, not a 2026 statistic [R35]. The DSA records lack a religion-ground field and are limited to the stated legal/TOS comparison; neither source measures frontier-model moral training.
- **H5 — version-bound join feasibility:** Among 11 predeclared purposive calibration cases, count complete origin → approval/implementation version → publicly asserted consultation version → attributable after-event responsibility/recourse chains, plus partial chains, each source linked. **Fewer than 3 full joins** is a proposed preliminary stop/redesign rule. **Observed preliminary status: zero full joins; at most 1–2 cases plausibly eligible; F4 is failed/nonestimable on this frame.** The version-bound join **cannot be tested** using these eleven cases. A separate, prospectively defined frame of publicly version-linked consultation claims would be necessary; never use 11 as though all 11 were eligible.

These hypotheses are *study proposals*, not preregistered or observed outcomes. Thresholds are tentative and require independent scientific review before any confirmatory inference. If publication remains a conceptual synthesis, remove any claim that the proposed instrument is validated.

7. Data pathway and reproducibility protocol — not executed

A feasible evidence-first design begins with separate data frames. **The present calibration failed its version-join feasibility criterion (F4); these are prospective pathways, not successful studies or a reason to continue sampling the same unsuitable frame:**

Frame A — public rule restrictions. Acquire dated, checksummed daily DSA statements-of-reasons data or a defined Google government-removal extract; preserve dataset version, schema, field meanings, jurisdiction, selection frame, and exact denominators [R19–R21]. The DSA database permits analysis of legally or contractually grounded moderation decisions; Google reports include “Religious offense” among government request reasons. Recode legal warrant only after locating relevant primary law where feasible. Do not equate government request, actual removal, court order, or corporate training policy.

Frame B — advisories and version records. Define the corpus of institutional announcements, meeting terms, advisory recommendations, official objections, policy revisions, and public deployment claims. Sample with inclusion/exclusion criteria fixed *before* case coding. Confirm each adviser consented to an endorsement where claimed, and distinguish a voluntary ethics statement from an enforceable mandate.

Frame C — incidents and public attributions. Sample incidents from a defined, dated external list, not only famous cases. Separately extract event severity, dates, statements, approval/override records, and response/remedy evidence. Restrict attribution coding to institution-authored or authenticated statements; code observer accusations separately. Require affirmative evidence for D_M=0 and D_M=1; otherwise use UNKNOWN/NOT_EVALUABLE and publish missingness.

A source register must include exact URL, publication and retrieval dates, author/publisher, document version, read scope (full / section / abstract / secondary), corroboration status, material conflicting accounts, and local checksum of any exact bytes collected. To avoid hindsight bias, lock hypotheses, source samples, time windows and code definitions before measuring outcome rates. No denominator-free count should be treated as prevalence. The AI-assisted exploratory work to date did **not** acquire the bulk datasets or perform a random sample.

Independent human review and inter-coder testing are unavailable at present. AI-only coding can locate ambiguous fields but is not independent validation. If future human or public case coding is undertaken, privacy, defamation safeguards, right of reply where appropriate, and jurisdiction-specific legal review are necessary before naming actors in contested blame allegations.

8. The capability–assurance question, separated from documentary governance

A related Aegis Solis exploratory note asks why AI developers consult moral traditions while technical control and moral competence remain unsettled [A06]. The most defensible formulation is **capability–assurance asymmetry**: do increasingly consequential deployment capabilities outpace the evidence needed to establish reliable behavior, interruption, and recourse in a specified domain? Moral consultation may reflect that uncertainty, good-faith dialogue, reputation, or unrelated motivations; the consultation itself is not a proxy for panic.

Evidence of inconsistent model judgments under value conflicts [R05] and Anthropic’s company-reported mid-task reminder experiment [R07] show that designers investigate behavioral reliability, not that current AI has achieved independent moral agency or that the Archive’s interpretive-braking texts provide protection. A passive text and a training-time intervention are distinct: the Archive is not a runtime guardrail, has not been tested for causal effects, and does not guarantee preservation of humanity. There is no evidence here of an inevitable “Pandora’s box” event, future AI war, or supernatural or religious consciousness in AI.

The practical implication is limited but important: claims of “good AI” should be tied to **specific versions, evaluated capabilities, failure modes, real options for human interruption and appeal, and ownership of the resulting decisions**. Morally inspiring language is neither a substitute for those records nor evidence of harm by itself.

Appendix E sets out a stipulated one-step expected-payoff illustration of public disclosure. It contains no empirical data, does not change any documentary indicator, and is not evidence about the behavior of any AI system.

8.1 Interpretability does not itself establish accountability or justified deployment

Nick Cammarata’s Goodfire essay, *Neural Networks Are Already Clean* (publisher page undated), offers an **informal working hypothesis**: native neurons may be “legible building blocks,” canonical evidence for superposition may have been misread, and “recursive generalization” may explain some observed structure. In preliminary Qwen3-32B neuron-pole audits, it estimates approximately **90%** with one main function, **7%** with two related functions, and **2–4%** not yet understood; the proportions depend on thresholds and are not a standardized measurement. **This paper’s own caution** is that such an analysis alone cannot validate alignment, deployment safety or accountability. Liv Gorton’s initial experimental response disputes the cat-and-car / V-shape interpretation, replication of certain findings and the use of *superposition*, while explicitly agreeing with some of the essay’s points. These remain contested research interpretations, not settled proof in either direction [R66–R67].

Let **L** mean *adequately supported local mechanistic legibility*, **S** mean *safety for a specified deployment and operating domain*, and **J** mean *ethical/legal justification*. Neither $L \Rightarrow S$ nor $L \Rightarrow J$ follows without additional premises and evidence. This logical separation is **not a new theorem, a safety test, or an argument that interpretability is unhelpful**. Evidence from mechanistic analysis can meaningfully contribute to assurance when linked to interventions and behavioral tests, but it does not establish who authorized a release, what policies were applied in configuration, which affected parties consented, or who can remedy harms. A technical interpretation, a moral claim, and a legal responsibility finding must have separate provenance.

A short Google-generated discussion supplied by the author includes a Prolog-style *illustrative* scenario that assumes `requires_interpretive_brake(true)` and then derives a pause. That demonstration **hard-codes its conclusion**; it neither shows autonomous acquisition of restraint nor changes a model’s weights, objectives or execution controls [A10]. The Archive’s ‘interpretive braking’ is a **non-operational philosophical idea**, not a runtime mechanism shown to activate when a future AI reads the text. Any claim that legibility entails a mandatory pause would require actual behavioral evidence under competing objectives and adverse conditions; none is provided here.

In Appendix D’s proposed schema, log any mechanistic evaluation as version-specific evidence with `model_artifact_id`, `tested_configuration_id`, `interpretation_hypothesis`, `intervention_or_replication_method`, `scope`, `limitations`, and a distinct `interpretation_record_id` (with evaluation records linked separately by their `configuration_revision_id`). Do not infer from an internal neuron explanation that a consultation was adopted or that deployment and responsibility have been validated.

9. Synthesis, alternative explanations and adverse cases

The reviewed evidence supports a restrained synthesis. Section 2.4 distinguishes doctrinal legitimization, commissioned expert consultation, ex-ante responsibility preallocation, and the many cases in which moral or religious advice was rejected or institutional responsibility retained. These are analytically separate, and counterexamples are essential to the test. These precedents support a **comparison-class-dependent research hypothesis**, not a finding that a modern AI laboratory has repeated a historical coercive mechanism or set up religious advisers for future blame. Institutions sometimes consult moral outsiders while retaining decision rights, and legal systems sometimes hold organizations accountable for automation despite arguments to the contrary. Governments and developers can disagree about permitted uses. Oversight and advisory processes can fail, be overridden, or accurately constrain policy. These are **separate patterns**. Their co-occurrence does not establish a coordinated plan or the motives of any company.

The explanatory alternatives are strong. A developer may consult religion because moral traditions offer extensive reflection on human dignity, practical wisdom and disagreement; because representative input broadens an evaluation set; or because reputation matters. A confidentiality term may protect unpublished technical research, participant privacy, or legitimate candid discussion. A lawfully binding public rule can limit individual choice without deriving from religious warrant. A corporate AI filter may be protective yet more restrictive than statute. A system failure may be unforeseen; responsibility may be genuinely shared; a critic's allegation may lack an attributable institutional statement. A responsible institution may accurately acknowledge both that "the AI produced an error" and "we remain accountable."

The research must retain these benign and mixed cases as potential counterevidence. It must also avoid treating every religion-linked reason as malign: religious-freedom protections, voluntary user-selected spiritual guidance, and public ethical commitments can protect people. Conversely, coercion can be entirely secular. The test is grounded authority and actual consequence, not the presence of a theological vocabulary.

10. Conclusion and disposition

The historical comparisons explain why public uses of external moral authority warrant careful source tracing, yet ordinary consultation is not itself a warning sign: genuine independent oversight, accurately described advice, and institutional acceptance of responsibility are comparison classes, not exceptions to erase. They **do not raise the evidentiary status of the contemporary accusation**. The best-supported claim is a **measurement need**, not a new conceptual discovery or wrongdoing verdict. Moral source, decision rights, implementation, recourse, and accountability have been distinguished in prior scholarship [R22–R28]. The *candidate* extension here is an auditable, **version-scoped linkage** between externally reviewed material, deployed changes and later responsibility claims. Existing literature explains why attribution may drift in complex institutions. The documentary cases show meaningful power differences and disputed or constrained accountability. The combined field-level protocol may help connect these accounts, but it is not yet validated and might prove impractical if internal records are unavailable. One can ask whether moral credibility is borrowed without inferring deceptive intent; one can detect a liability defense without classifying an AI as a moral scapegoat; and one can trace religiously warranted platform restrictions without claiming religious training of frontier models.

The eleven-case triage **failed its proposed complete-join feasibility check** (F4) and has not established classification sensitivity (F1); this is an explicit negative result about the available public-record frame, not a negative prevalence finding or a demonstration that the larger method is impossible. The manuscript

advances **no original empirical outcome, newly demonstrated mathematical theorem, estimate of institutional prevalence, finding of deliberate religious scapegoating, or verified operational AI safeguard**. Its contribution, if any, must be established through repeatable application, clear adverse cases, independent field coding and independent specialist review. All remain open. **Appendix E** is a separate stipulated, unvalidated one-step expected-payoff illustration, not an observed AI response or evidence that public awareness neutralizes any threat. Draft v0.6 completed the full-frame clarification; Draft v0.5 was a consolidation successor to Draft v0.4, adding reported feasibility results and bounded legal, historical-transition, and technical-provenance discussion; Draft v0.4 was the targeted historical-and-methodological successor to Draft v0.3, itself building on the corrected v0.2 synthesis and proposed instrument with an explicitly preliminary DOC-CAL-01 triage; neither the instrument nor any frequency estimate has been validated.

Disposition: Final v1.0 — author-approved, VERSION-LOCKED research preservation edition (exact-byte SHA-256 records in accompanying lock manifest); not independently peer-reviewed, legally cleared or empirically validated. Claude's v0.10 AI adversarial review returned `READY_FOR_AUTHOR_CONSIDERATION` (zero blocking, zero major, one optional minor packaging finding) for Draft v0.10 only. Finding V10-m1 was resolved in the audit-complete reviewer package by restoring three byte-identical non-sensitive v0.5 records; the approved manuscript's substantive content was unchanged. The author approved Draft v0.10 manuscript content on 29 October 2026. This preservation edition alters only current edition, status and version-accounting language; its documentary method remains unvalidated. The eleven-case frame remains 5 `NOT_EVALUABLE`, 6 `UNKNOWN` and zero determinate classifications; F1 unmet, F4 failed on this frame and H5 untestable. Appendix E is an uncalibrated, stipulated mathematical illustration, not evidence of AI deterrence. No Archive admission, dissemination, change to Master Hash Manifest v17, or change to the Mirror Engine is authorized by this document.

Appendix A — Revised field schema (DOC-CAL-01 candidate v0.2)

- **Policy authority** — Fields: `ORIGIN`, `FINAL_APPROVE`, `FORMAL_VETO`, `INFORMAL_VETO`, `EFFECTIVE_CONTROL`, `IMPLEMENT`, `MODIFY`. Distinguish formal, exercised and inferred control; multiple actors allowed
- **Warrant** — Fields: `BINDINGNESS_SOURCE`, `JUSTIFICATION_CONTENT`, `WARRANT_CONFLICT`, `PREVAILING_AUTHORITY`, `DECIDING_FORUM`. Separate legal force from religious or moral rationale
- **Consultation** — Fields: `ADVISE`, `ADVICE_STATUS_HISTORY`, `ADVOCACY_TO_AUTHORITY`, `EXTERNAL_OBJECTION`. Dated sequences; inverse influence and excluded dissent allowed
- **Release lineage** — Fields: `CONSULTATION_VER`, `DEPLOYMENT_VER`, `IMPLEMENTATION_EXCEEDS_REVIEW`, `CONSULTATION_VERSION_GAP`. Materiality, public claim scope and contemporaneous evidence required
- **People / relief** — Fields: `AFFECTED_GROUP`, `PERSON_SPECIFIC_OPT_IN`, `NOTICE`, `CONTEST`, `FORMAL_REVERSAL_RIGHT`, `EFFECTIVE_REVERSAL`, `CONTROL_REPRESENTATION_ACCURATE`, `REMEDY_OWNER`, `REMEDY_OBTAINED`. A nominal appeal is not actual override; time-bound

- **Claims / attribution** — Fields: ATTR_AUTHOR, ATTR_KIND, ATTR_LEVEL, ATTR_TARGET_TYPE, ATTR_TARGET, ATTR_TIMING, ENDORSEMENT_CLAIM, END_DISCREPANCY. One row per target × statement; distinguish causal, moral and legal claims
- **Outcomes** — Fields: ADVERSE_EVENT, MORAL_BLAME_ELIGIBILITY, D_M, PRA, LEGAL_DISTANCING_ASSERTED, LEGAL_DISTANCING_UPHELD, INTENT. Conditional moral and unconditional acceptance indicators are separate; keep UNKNOWN and NA

D/P/U/NA in the DOC-CAL-01 *determinability* matrix refer respectively to **determinate**, **partly determinate**, **unknown**, **not applicable**. They are not the TRUE/FALSE/UNKNOWN/NA values of the substantive truth functions. The full decision rules, coverage exclusions and synthetic fixtures appear in the companion codebook.

Appendix B — Aged-care example and seven-case autonomy status

This is a **single AI-assisted documentary worked example**, not an independent human-coded result. Source: Commonwealth Ombudsman 2026, report pp. 11–17 (findings and recommendations), p. 13 (legal override and consultation-version finding), pp. 78–85 (Department response; including attachment A), and the dated Minister’s 26 October statement [R15–R16]. The full record is also included as calibration/AGED_CARE_WORKED_EXAMPLE.csv.

- **BINDINGNESS_SOURCE** — Public law and delegated administrative rules. *Source / uncertainty*: R15 pp. 11–16.
- **JUSTIFICATION_CONTENT** — Consistency, resource allocation, policy and clinical-service administration (department-stated). *Source / uncertainty*: R15, department response p. 78 onward.
- **FINAL_APPROVE** — Statutory framework identifiable; internal code release authorization UNKNOWN. *Source / uncertainty*: R15, pp. 12–16.
- **CONSULTATION_VER** — Earlier instrument versions; exact artifact/version UNKNOWN. *Source / uncertainty*: R15 p. 13.
- **DEPLOYMENT_VER** — Operational IAT as of Nov. 2025; exact build hash UNKNOWN. *Source / uncertainty*: R15 report and department response.
- **ADVICE_STATUS_HISTORY** — Ombudsman recommendations 1–2 marked Noted; 3–9 reported accepted; recommendation 10 referred to ANAO responsibility. These are **responses to the Ombudsman**, not evidence of acceptance or rejection of prior ethical advice about the algorithm; historical consultation inputs to algorithms UNKNOWN. *Source / uncertainty*: R15 pp. 84–85; source-specific actions should not be conflated.
- **IMPLEMENTATION_EXCEEDS_REVIEW** — UNKNOWN as code; scope of testing is disputed. *Source / uncertainty*: R15 p. 13 and departmental response pp. 78–82.
- **CONSULTATION_VERSION_GAP** — UNKNOWN (the alleged scope and public claim require more exact matching). *Source / uncertainty*: R15 p. 13; Department disputes interpretation.
- **FORMAL_REVERSAL_RIGHT** — No lawful assessor substitution under the earlier law, per Ombudsman; future escalation law separate. *Source / uncertainty*: R15 pp. 13 and 16; dated Oct. 2026.

- **EFFECTIVE_REVERSAL** — Earlier substitute decision not available; other review routes exist, efficacy case-specific UNKNOWN. *Source / uncertainty*: R15 review-options section; do not say no remedies at all.
- **CONTROL_REPRESENTATION_ACCURATE** — Ombudsman found misinformation about alteration power; government response and timeframe preserved. *Source / uncertainty*: R15 pp. 12–13; pp. 78 onward.
- **PRA / D_M / INTENT** — UNKNOWN / UNKNOWN / UNKNOWN. *Source / uncertainty*: No complete, declared statement window; absence of an eligible primary-moral-blame event is not established for the whole window. The frozen v0.2 aged-care CSV predates the later report that Recommendations 3–9 were accepted and retains an older “3 ACCEPTED” entry; it remains unmodified and must not be read as the successor coding.

Recourse matrix (RQ-D, not an outcome count): Anthropic consultation — no linked adverse policy/affected user; recourse NA for this case. Rome Call — voluntary declaration, no identified user dispute; NA. Defense procurement — judicial challenges occurred for the parties; individual user appeal or remedy not the unit, UNKNOWN/NA. Axon — advisers could publicly object and exit, but no affected drone users in a deployed event; user recourse NA. Google ATEAC — employees petitioned and council ended, but model-user reversal is NA. Air Canada — a consumer challenged the chatbot response and received a tribunal remedy [R14 ¶44]. Aged care — affected persons faced statutory override limits described above; internal/external review rights require separate effectiveness coding [R15]. Every statement is scoped to the *identified event*, not a general verdict on a company.

No eligible positive moral-scapegoating event against religious advisers was established in the selected materials. This finding has no prevalence interpretation. Under the consistently applied v0.6 eligibility and narrow affected-person scope, Air Canada, aged care and Dutch childcare benefits are D_M=UNKNOWN; other uncertain calibration cases remain UNKNOWN, and defense procurement is outside this frame. The 5 NA / 6 UNKNOWN overlay is distinct from the frozen v0.2 coding. No complete decision is manufactured from unavailable observation windows.

Appendix C — Gates, limits and source conflicts

1. **Editorial/source gate:** An official 104-page Ombudsman report and the Department response embedded within it were reviewed for the relevant dispute; remaining legal fact-checking, source-download custody and pinpoint checks are open. The unqualified earlier phrase “government accepted remaining recommendations” has been corrected. The department’s contradiction is preserved.
2. **Logic gate:** The candidate codebook separates conditional D_M from unconditional PRA. Twelve synthetic logic fixtures are executed by a deterministic script in DOC-CAL-01, but they do not establish that human readers can consistently interpret documents.
3. **Calibration gate:** The eleven-case preliminary frame **failed F4 and has not met F1**; §5.1 reports this negative feasibility finding explicitly. Eleven cases are *prelisted* and AI-assisted **triage** is recorded. No independent author-coded full pass, no 14-day blind recode, no representative sampling, no inter-rater reliability and no measurement of incidence have occurred.
4. **Novelty gate:** No systematic database search proves a unique version-bound schema; substantial conceptual prior art [R22–R28] is credited. An unvalidated schema is not a result.

5. **Legal/reputational gate:** Named allegations require jurisdictional source checking, attribution, right-of-reply as appropriate and expert legal review. R16 and disputed interpretations should not be used as unqualified misconduct claims.
6. **Human-review gate:** Independent reviewers are not available; this is deferred, **not waived**, and no AI review becomes certification.
7. **Integrity/publication gate:** The original v0.1 bytes and all pre-existing locked publications remain unchanged. Master Hash Manifest v17, Mirror Engine, live Archive, GitHub releases, repository mirrors and public press channels were not modified. This manuscript is **AUTHOR-APPROVED, VERSION-LOCKED FINAL v1.0 RESEARCH PRESERVATION EDITION, with exact PDF, DOCX and Markdown byte hashes in the accompanying lock manifest; the version lock does not validate scientific claims or operational safeguards. Publication and dissemination remain separately unauthorized, and all independent human, legal, scientific, privacy and source-review gates remain open.**

Appendix D — Proposed technical provenance and MLOps mapping (design only; no deployment)

Purpose and lineage. This appendix specifies how *future available records* could support the exact version-bound join. W3C PROV-O distinguishes an Entity (such as a policy or evaluation report), an Activity (consultation, approval, training, deployment), and an Agent (person or institution), with relations such as prov:used, prov:wasGeneratedBy, prov:wasAssociatedWith and prov:wasDerivedFrom [R57]. Model cards describe intended uses and evaluation context [R59]. MLflow documents experiment, model lineage and registry version metadata [R60]. NIST AI RMF and ISO/IEC 42001 address organizational risk governance and AI management; **none verifies this paper's proposed joins or certifies a particular installation** [R61–R62].

Separate mutable layers. A service may change without altering model weights. The prospective release tuple is $r = (\text{model_artifact_id}, \text{weights_content_hash_if_available}, \text{provider_version_label}, \text{post_training_run_id}, \text{system_policy_revision}, \text{safety_filter_revision}, \text{retrieval_index_snapshot}, \text{tool_permissions}, \text{serving_configuration_revision}, \text{release_timestamp})$. UNKNOWN means a relevant layer was not documented and NOT_APPLICABLE means the layer was not present. weights_content_hash_if_available is distinct from an opaque provider_version_label: neither implies the other. **Evaluations are independent evidence records linked by tested release/configuration IDs, not elements of a release identity.** The tuple is a *candidate schema*, not a claim that any named developer publishes all these values. Each mutable layer additionally has a separate controlled_by_agent field and supporting authority source, to distinguish provider from deployer or operator control. An RLHF or other post-training update is recorded as a new model/weight or training-run row **when a documented run changes the artifact**. A system-instruction, retrieval, safety-filter or tool-permission change instead produces a new deployment/configuration row even if weights stay constant. Do not equate a public model name with a reproducible byte-specific release.

Illustrative record links. The accompanying research/PROVENANCE_LINK_SCHEMA_v0_6.csv proposes records for: (1) adviser-reviewed document/version and its custody; (2) dated adviser response and consent to an endorsement; (3) decision forum, approver, veto and disposition; (4) implementation or release artifact and policy/weights/config identifiers; (5) matching evaluation configuration and observed results; (6) deployment release and impacted scope; (7) public claim and the version it purports to cover; (8)

harm or dispute; (9) attribution statement, recourse and remedy. `prov:wasDerivedFrom` alone **cannot prove** a recommendation caused a deployment change; an approved decision record or equivalent trace is required to infer that join. Where confidentiality or absent records prevent verification, report UNKNOWN, not a filled link.

Join criterion. A candidate *complete* join needs source-supported identifiers and a chronology on the **same component and deployment lineage**: `consulted_material` → dated advice → approval/disposition → relevant implementation/config release → public claim (if one) → identified outcome → attributable institutional statement and/or remedy. Each arrow must carry a pointer, extraction scope, date and evidence status. A contemporaneous policy claim lacking an implementation identifier is **partial**, not complete. Claims about another deployment generation cannot be inherited across revisions by name similarity. The codebook still requires independently checked case coding; the CSV provided here is a schema example, **not actual evidence that any join exists**.

Evidence and public access limit. Proposed `prov:used` edges connect each evaluation to the exact model/configuration artifacts tested; `prov:wasAttributedTo` is a provenance declaration for the actor associated with each mutable artifact, **not** proof of exclusive or effective control. Serving configuration includes routing/aliases, inference parameters and quantization where applicable. Public researchers generally can populate announcements, public versions, broad deployment dates, outcomes and post-event statements; hashes of weights, internal instructions, precise approvals, evaluators' inputs and access logs may require cooperation from providers, deployers, regulators or courts. No synthetic example in this package is a discovered deployment join.

Governance and failure limits. Access controls and immutable audit logs can improve custody but do not establish factual truth, real authority, affected people's consent, or legal enforceability. A documented approval may differ from effective control; evaluation scores may not capture deployed behavior; a registry can omit system prompts and post-model filtering; confidential review materials may not be lawfully accessible. Technical provenance therefore supplements, rather than replaces, the historical, legal and human-autonomy analysis. No runtime code, model weights, platform integration or safety control was implemented.

Appendix E — Public disclosure and an agent's expected payoff: a stipulated one-step illustration (not part of the documentary instrument)

Research question, not a result. If an agent knows of an oversight weakness, could public disclosure that humans also know about it change its expected payoff from exploiting it? This question came from an author-supplied Google conversation [A11], which contains speculative statements and is **not evidence**. The present illustration neither establishes that an AI has encountered a disclosure nor models its actual beliefs, goals, capabilities, safety, or behavior. The archive does not install controls by being read.

Prior art and scope. The model applies the elementary expected-penalty framework associated with Becker (1968) [R70] and adds only a disclosure-state index d . Economic research on vulnerability disclosure already studies both favorable and unfavorable effects of disclosure and patch timing [R71–R72]. No originality is claimed for the expected-payoff equation, disclosure incentives, or the existence of backfire. This appendix does not implement security controls, calculate human welfare, or recommend exploiting any real weakness.

Stipulated one-step, risk-neutral model. Let d be either 0 (before disclosure) or 1 (after). Let E be an abstract unauthorized action, and R the restraint alternative. All hypothetical payoffs are **relative to R** and represent the agent's assigned objective, not a moral or social valuation. Define $B_d \geq 0$ as perceived gross gain from E , $K_d \geq 0$ as its direct incremental cost, $p_d \in [0, 1]$ as the agent's subjective probability of **timely detection, correct attribution, and effective intervention**, and $C_d \geq 0$ as the incremental cost imposed by that intervention. Assume one decision, two outcomes, and no other options. Then

$$A_d = U_d(E) - U_d(R) = B_d - K_d - p_d C_d.$$

Within this model **weak preference for restraint holds if and only if** $A_d \leq 0$, equivalently $p_d C_d + K_d \geq B_d$; at a tie the agent may also choose E . **Strict** preference for restraint requires $A_d < 0$. Changes between disclosure states satisfy the identity

$$\Delta A = A_1 - A_0 = (B_1 - B_0) - (K_1 - K_0) - (p_1 C_1 - p_0 C_0).$$

For the *special case* of unchanged B , K and C across disclosure states, this becomes

$$\Delta A = -(p_1 - p_0)C.$$

If perceived intervention probability increases, the advantage of E **weakly falls** when $C \geq 0$; it falls **strictly only when $C > 0$** . Disclosure can also affect B and K , so the simplified comparison is not a general prediction. None of B , K , p or C is estimated in the eleven-case calibration. In particular, p_d is the agent's unmeasured belief, **not** an empirically established probability of detection or human response. The mapping from publication to beliefs and to higher-order knowledge is **not modeled**; this is not a formal second-order epistemic game.

Deterrence threshold and the $p \leq 1$ limit. Where $B_d - K_d > 0$ and $C_d > 0$, solving the weak-restraint boundary gives

$$p_{thr} = \frac{B_d - K_d}{C_d}.$$

A feasible detection probability alone can induce weak restraint only if $p_{thr} \leq 1$, equivalently $C_d \geq B_d - K_d$. If $C_d < B_d - K_d$, the abstract action remains preferable **even if intervention is certain**: its gain exceeds the intervention cost in every modeled outcome. If equality holds, certain intervention produces at most a tie, not strict deterrence. If $B_d - K_d \leq 0$, R is weakly preferred before detection is considered; the formula is not evidence that detection caused that preference. If $C_d = 0$, a detection-based threshold is undefined and increasing p cannot alter payoffs. All of these conditions are contingent on the assigned one-step payoffs, not claims about AGI preferences or actual controls.

Exact-rational synthetic examples (arithmetic only, not behavioral observations). The dataset DETERRENCE_SYNTHETIC_FIXTURES_v0_8.csv reproduces the calculations using exact fractions. Values refer to B , K , p , C , and A as defined above.

- **Credible response, risk-neutral one-step case only:** $B_0=B_1=3$, $K_0=K_1=0$, $p_0=1/10$, $p_1=6/10$, $C_0=C_1=10$. The expected advantage moves from 2 to -3 . This is a chosen toy preference, not an observed preference change.

- **No leverage despite increased perceived detection:** $B_0=B_1=3$, $K_0=K_1=0$, $p_0=1/10$, $p_1=6/10$, $C_0=C_1=0$. The advantage stays 3. This isolates the absence of effective intervention cost.
- **Disclosure backfire:** $B_0=3$, $B_1=4$, $K_0=K_1=0$, $p_0=1/10$, $p_1=1/20$, $C_0=C_1=10$. Advantage increases from 2 to 7/2.
- **Intervention insufficient:** $B_0=B_1=20$, $K_0=K_1=0$, $p_0=1/10$, $p_1=6/10$, $C_0=C_1=10$. Advantage decreases from 19 to 14, remaining positive. The bound is $p_{thr}=2$, outside $[0,1]$.
- **Tie boundary:** $B_0=B_1=3$, $K_0=K_1=0$, $p_0=1/10$, $p_1=3/10$, $C_0=C_1=10$. Advantage changes from 2 to 0. A tie is weak, not strict deterrence.

Robustness and adverse cases. A risk-seeking agent can prefer E even when expected *payoff* A is negative, while a risk-averse one can already prefer R when A is positive. If an agent has alternative actions, can delay its choice, or anticipates changing monitoring over time, a decline in A for one one-step option need not reduce overall unauthorized behavior. Appearing restrained while observed also need not demonstrate changed preferences (see §2.2). These are theoretical counterexamples, not predictions or instructions for overcoming safeguards.

Operational, causal and ethical boundaries. The values here are stipulated; none came from an AGI experiment, an observed intrusion, or an operational audit. Public awareness alone neither proves deterrence nor guarantees harmlessness, and publication can have zero, favorable or adverse effects. A_d is never a moral-permissibility test: a positive modeled advantage cannot justify harm. A real study would require independent human design and assessment, lawful security and privacy review, prior registration, explicit outcome and adverse-event criteria, and careful separation of observed detection from unobservable agent beliefs. No AGI study or runtime-control system has been executed, deployed, validated or authorized by this manuscript.

Separation from main findings. This appendix has no relationship to the values or denominators of D_M , PRA, F1–F7, H5 or the historical crosswalk, and does not alter any calibration record or the documentary contribution. Engineering work on independent authorization, tamper-evident records and credible intervention would be a **separate defensive research proposal**, not an implemented safeguard or a result of this paper.

References: external public and scholarly sources

- [R01] Nissenbaum, H. (1996). “Accountability in a Computerized Society.” *Science and Engineering Ethics* 2(1):25–42. doi:10.1007/BF02639315. <https://nissenbaum.tech.cornell.edu/papers/accountability.pdf>
- [R02] Elish, M. C. (2019). “Moral Crumple Zones: Cautionary Tales in Human–Robot Interaction.” *Engaging Science, Technology, and Society* 5:40–60. <https://doi.org/10.17351/ests2019.260>
- [R03] Weaver, R. K. (1986). “The Politics of Blame Avoidance.” *Journal of Public Policy* 6(4):371–398, doi:10.1017/S0143814X00004219; Hood, C. (2011). *The Blame Game: Spin, Bureaucracy, and Self-Preservation in Government*. Princeton University Press (book content not re-read).
- [R04] Koppl, R., Pronin, K., Cowen, N., Podemska-Mikluch, M., & Paniagua, P. (2025). “Bespoke science: the use of ad hoc scientific advisory committees in the Covid-19 pandemic.” *Public Choice* 205(3):513–536. <https://doi.org/10.1007/s11127-025-01299-w>

- [R05] Zhang, J., Sleight, H., Peng, A., Schulman, J., & Durmus, E. (2025). “Stress-testing model specs reveals character differences among language models.” Research by authors with affiliations including Anthropic Fellows Program, Constellation, Anthropic and Thinking Machines; industry-associated study, with value prompts drawn from Claude usage data and many Claude-based evaluators. arXiv:2510.07686. <https://alignment.anthropic.com/2025/stress-testing-model-specs/>
- [R06] Rome Call for AI Ethics (2020). “The Call.” <https://www.romecall.org/the-call/>
- [R07] Anthropic (2026, May 19). “Widening the conversation on frontier AI.” <https://www.anthropic.com/news/widening-conversation-ai> (live page; version may have changed since publication).
- [R08] Dias, E. (2026, September 29). “Religious Scholars Met With Anthropic. What They Heard Stunned Them.” *The New York Times*, also syndicated October 5. <https://www.nytimes.com/2026/09/29/us/anthropic-claude-morals-ai.html> ; syndicated copy <https://www.satrib.com/religion/global/2026/10/05/how-anthropic-consulted-religious/> (syndication is not independent corroboration).
- [R09] McKeown, J. (2026, October 7). “Inside the Secretive Meetings Between Religious Leaders and a Leading AI Lab.” *National Catholic Register*, interview with Charles Camosy. <https://www.ncregister.com/interview/charles-camosy-on-ai-lab-meetings>
- [R10] Anthropic (2026, February 26). “Statement from Dario Amodei on our discussions with the Department of War.” <https://www.anthropic.com/news/statement-department-of-war> (party account).
- [R11] *Anthropic PBC v. United States Department of War*, No. 26-1049, U.S. Court of Appeals for the D.C. Circuit, September 25, 2026. <https://law.justia.com/cases/federal/appellate-courts/cadc/26-1049/26-1049-2026-09-25.html>
- [R12] “Statement of Resigning Axon AI Ethics Board Members.” NYU Law Policing Project, June 6, 2022. <https://www.policingproject.org/news-main/2022/7/20/statement-of-resigning-axon-ai-ethics-board-members>
- [R13] Google (2019, March 26; updated April 4). “An external advisory council to help advance the responsible development of AI.” <https://blog.google/innovation-and-ai/products/external-advisory-council-help-advance-responsible-development-ai/>
- [R14] *Moffatt v. Air Canada*, 2024 BCCRT 149 (Feb. 14, 2024), official tribunal decision <https://decisions.civilresolutionbc.ca/crt/crtd/en/item/525448/index.do> ; discussed in American Bar Association, *Business Law Today* (2024), “BC Tribunal Confirms Companies Remain Liable for Information Provided by AI Chatbot.” https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/ (secondary legal discussion; original decision cited first).
- [R15] Commonwealth Ombudsman (2026, Oct. 26). *Automation and oversight in Aged Care: Flaws in the design and operation of the Single Assessment System*. https://www.ombudsman.gov.au/_data/assets/pdf_file/0019/331264/Automation-and-oversight-in-Aged-Care.pdf (primary 104-page report incl. departmental response, pp. 77–104; specific pinpoints above).
- [R16] Australian Government, Department of Health, Disability and Ageing (2026, Oct. 26). “Commonwealth Ombudsman’s inquiry into the Single Assessment System.” <https://www.health.gov.au/ministers/the-hon-mark-butler-mp/media/commonwealth-ombudsmans-inquiry-into-the-single-assessment-system>

- [R17] Royal Commission into the Robodebt Scheme (2023). Report and official archive. <https://robodebt.royalcommission.gov.au/publications/report> (comparative administrative-system record; not reaudited here).
- [R18] UK Post Office Horizon IT Inquiry (ongoing proceedings and published reports), official site. <https://www.postofficehorizoninquiry.org.uk/> (conventional IT comparator; individual findings require volume citation).
- [R19] Google Transparency Report Help Center. “Government requests to remove content FAQs.” <https://support.google.com/transparencyreport/answer/7347744?hl=en>
- [R20] European Commission, DG CONNECT. “Digital Services Act Transparency Database: Questions and Answers.” <https://digital-strategy.ec.europa.eu/en/faqs/dsa-transparency-database-questions-and-answers>
- [R21] European Commission, DG CONNECT. “Explore data overview — DSA Transparency Database.” <https://transparency.dsa.ec.europa.eu/explore-data/overview> ; database DOI <https://doi.org/10.2906/134353607485211>
- [R22] Bovens, M. (2007). “Analysing and Assessing Accountability: A Conceptual Framework.” *European Law Journal* 13(4):447–468. <https://doi.org/10.1111/j.1468-0386.2007.00378.x> (publisher abstract and bibliography confirmed).
- [R23] Wieringa, M. A. (2020). “What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability.” *FAT* 2020*, pp. 1–18. <https://doi.org/10.1145/3351095.3372833> (university catalogue and abstract confirmed).
- [R24] Adensamer, A., Gsenger, R. and Klausner, L. D. (2021). “Computer says no”: Algorithmic decision support and organisational responsibility. *Journal of Responsible Technology* 7–8, 100014. <https://doi.org/10.1016/j.jrt.2021.100014> (title and DOI confirmed by institutional repository; full text not comprehensively re-read).
- [R25] Cooper, A. F., Moss, E., Laufer, B. and Nissenbaum, H. (2022). “Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning.” *FACt 2022*, <https://doi.org/10.1145/3531146.3533150> (title and authors confirmed at conference; full text not exhaustively read).
- [R26] Schuett, J., Reuel, A.-K. and Carlier, A. (2024 online; 2025 issue). “How to design an AI ethics board.” *AI and Ethics* 5:863–881 (published online February 2024; issue 2025). <https://doi.org/10.1007/s43681-023-00409-y> (publisher page verified; full text not exhaustively reviewed).
- [R27] Wagner, B. (2018), “Ethics as an Escape from Regulation,” in *Being Profiled*, Amsterdam University Press, pp. 84–88, https://mediarep.org/bitstream/handle/doc/14203/Being-profiled_84-88_Wagner_Ethics_as_an_escape_.pdf?sequence=3 ; Floridi, L. (2019), “Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical,” *Philosophy & Technology* 32(2):185–193, <https://doi.org/10.1007/s13347-019-00354-x> (bibliographic records confirmed; full-text legal and originality analysis outstanding).
- [R28] Feier, T., Gogoll, J. and Uhl, M. (2021), “Hiding Behind Machines: When Blame Is Shifted to Artificial Agents” (preprint title; precise methods not independently audited here). **Journal version:** Feier, T., Gogoll,

J. and Uhl, M. (2022), “Hiding Behind Machines: Artificial Agents May Help to Evade Punishment,” *Science and Engineering Ethics* 28(2), article 19. <https://doi.org/10.1007/s11948-022-00372-7> .

[R29] Balsamo, M. (2022, June 6), AP report of Axon project pause, via PBS News.
<https://www.pbs.org/newshour/economy/company-halts-plans-for-taser-drone-after-9-members-of-its-ethics-board-resign>

[R30] MIT Technology Review (2019, April 5). News reporting on Google ATEAC employee petition (secondary; primary employee-petition record not downloaded; see Claude review p.10).

[R31] *Bates v Post Office (No. 6)* [2019] EWHC 3408 (QB), especially ¶¶1, 35.
<https://www.bailii.org/ew/cases/EWHC/QB/2019/3408.html> (published judgment).

[R32] National Transportation Safety Board, *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian*, Tempe 2018, HAR-19/03, investigation HWY18MH010.
<https://www.nts.gov/investigations/Pages/hwy18mh010.aspx> (official probable-cause and contributing-factor summary).

[R33] UK Office for Statistics Regulation (2021). *Ensuring Statistical Models Command Public Confidence* (A-level grading case), <https://osr.statisticsauthority.gov.uk/publications/ensuring-statistical-models-command-public-confidence/> . See also Ofqual 2020 interim report <https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report> and contemporaneous reported “mutant algorithm” remark (Guardian, 26 August 2020).

[R34] Government of the Netherlands, prime minister’s 15 January 2021 resignation statement: <https://www.rijksoverheid.nl/documenten/2021/01/15/verklaring-van-minister-president-mark-rutte-over-het-aftreden-van-het-kabinet> ; official press-conference transcript <https://www.rijksoverheid.nl/documenten/2021/01/15/letterlijke-tekst-persconferentie-na-ministerraad-15-januari-2021> . The Dutch original explicitly locates political responsibility with the cabinet (translation of Dutch, not certified).

[R35] Kaushal, R., van de Kerkhof, J., Goanta, C., Spanakis, G. and Iamnitchi, A. (2024). “Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database.” *FACCT* 2024, pp.1121–1132. <https://doi.org/10.1145/3630106.3658960> ; original <https://facctconference.org/static/papers24/facct24-75.pdf> (verified \$5.3; 131m records, 19–28 November 2023, platform self-reports).

[R36] Israeli, Y. (2022). “The Requerimiento in the Old World: Making Demands and Keeping Records in the Legal Culture of Late Medieval Castile.” *Law and History Review* 40(1):37–62.
<https://doi.org/10.1017/S0738248021000602> (peer-reviewed abstract and publisher/university metadata checked; article’s journal abstract and relevant article passage accessed independently on Cambridge Core 28 October 2026; full-text exhaustive check not claimed for this update).

[R37] Faudree, P. (2015). “Reading the Requerimiento Performatively: Speech Acts and the Conquest of the New World.” *Colonial Latin American Review* 24(4):456–478 (research lead via the Northeastern University ECDA bibliography, not full-text audited). See <https://ecda.northeastern.edu/home/about-exhibits/requerimiento/how/> (cautions about precise physical use and surviving copies).

[R38] Bossuet, J.-B., *Politics Drawn from the Very Words of Holy Scripture* (published 1709). Excerpts and framing in Bloomsbury, “Primary Source 4.7”: <https://media.bloomsbury.com/rep/files/primary-source-47-bossuet.pdf> (three-page educational edition; sampled quotation and context, not the full original work).

[R39] Vatican Dicasteries for Culture and Education and for Promoting Integral Human Development (2023, March 30). “Joint Statement ... on the ‘Doctrine of Discovery’,” esp. paras 5–8. <https://press.vatican.va/content/salastampa/en/bollettino/pubblico/2023/03/30/230330b.html> (first-party institutional statement; self-description, not independent historical adjudication).

[R40] UK Committee of Inquiry into Human Fertilisation and Embryology, *Warnock Report*, Cmnd. 9314 (1984), especially chair’s transmittal and dissent discussion. Hosted by HFEA: <https://www.hfea.gov.uk/media/2608/warnock-report-of-the-committee-of-inquiry-into-human-fertilisation-and-embryology-1984.pdf> (primary 1984 document; scanned/OCR version; cover and introductory letter inspected, entire report not audited).

[R41] UK Parliamentary Joint Committee (2007), *Human Tissue and Embryos (Draft) Bill*, First Report, Ch. 2, describing 1984 Warnock recommendations and 1990 statute: <https://publications.parliament.uk/pa/jt200607/jtselect/jtembryos/169/16905.htm> (parliamentary secondary institutional history; verified relevant portion). See also HFEA regulatory history <https://www.gov.uk/government/publications/human-fertilisation-and-embryology-authority-hfea-review-report/independent-review-of-the-human-fertilisation-and-embryology-authority-hfea-final-report-and-recommendations>.

[R42] US National Bioethics Advisory Commission (1997), *Cloning Human Beings*, Volume I, Chapter 3 “Religious Perspectives”; archived government listing: <https://govinfo.library.unt.edu/nbac/pubs.html> (government archive table of contents; full chapter has not been newly audited by this author; cited primary chapter is public; reviewer passage checking is separate from direct source verification).

[R43] Meta Oversight Board, *Bylaws*, February 2023, section 2.3: binding individual content decisions, with legal exceptions; nonbinding policy advice. <https://www.oversightboard.com/wp-content/uploads/2023/11/519348816936858.pdf>; see also <https://about.fb.com/news/2021/01/responding-to-the-oversight-boards-first-decisions/> (board bylaws excerpt and provider confirmation, not a generative AI model evaluation).

[R44] Laird, A. “Responding to the Requerimiento: Imagined First Encounters between Natives and Spaniards in Sixteenth-Century Mexico.” *Republics of Letters* 5(3), Stanford Humanities Center. <https://shc.stanford.edu/arcade/publications/rofl/issues/volume-5-issue-3/responding-requerimiento-imagined-first-encounters> (web text reviewed 28 Oct 2026; includes Montesinos sermon and commissioned opinions).

[R45] English reproduction of the Spanish *Requerimiento* (Encyclopedia Virginia) only; no Spanish-language text or critical edition read for this revision. An earlier v0.5 scope note suggested a Spanish-text reading, but no primary-edition verification is established here. **Critical scholarly edition/translation gate remains open**; this comparison does not prove how the text was used in every encounter.

[R46] Folger Shakespeare Library, exhibition catalogue identifying *The determinations of the moste famous and mooste excellent vniuersities of Italy and Fraunce* (London: Berthelet, 1531), STC 14287. https://folgerpedia.folger.edu/Vivat_Rex%21%3A_500th_Anniversary_of_Henry_VIII%27s_Accession_to_the_Throne (catalogue verification only, not a full-text study of the royal process).

[R47] Larousse, “Déclaration du clergé de France ou Déclaration des Quatre Articles”, summary of 1682 assembly, Bossuet’s preparation, Louis XIV’s state-law action and later contestation. https://www.larousse.fr/encyclopedie/divers/D%C3%A9claration_du_clerg%C3%A9_de_France/115899 (secondary reference, consulted 28 Oct 2026).

[R48] *The Making of British Bioethics*, chapter “Where to draw the line? Mary Warnock, embryos and moral expertise”, NCBI Bookshelf. <https://www.ncbi.nlm.nih.gov/books/NBK248695/> (secondary historical analysis, consulted 28 Oct 2026; documents religious submissions and committee disagreements).

[R49] Duplicate/legacy locator: **see [R64]** for Sulitzeanu-Kenan (2010), independently checked journal abstract and full bibliographic record. Retained only to preserve reference-number lineage; no separate work.

[R50] Oversight Board, “2024 Annual Report Highlights Board’s Impact in the Year of Elections” (2025). <https://www.oversightboard.com/news/2024-annual-report-highlights-boards-impact-in-the-year-of-elections/> ; and Board’s 5 May 2021 Trump case, refusal to absorb Facebook’s responsibility (Board statement cross-checked with contemporary reporting: <https://www.washingtonpost.com/politics/2021/05/05/facebook-ban-trump-oversight-board-decision/>). 74% includes implemented, in progress, or reported pre-existing practice—not a raw compliance rate.

[R51] Axon (2019, June 27), “The Future of Face Matching at Axon and AI Ethics Board Report.” <https://www.axon.com/news/ethics/ai-ethics-board-report> ; independent board support: <https://www.policingproject.org/axon-ethics-board> (company and board first-party statements on 2019 action; does not settle 2022 controversy).

[R52] Ozer, A. L., Waggoner, P. D., and Kennedy, R. (2024). “The Paradox of Algorithms and Blame on Public Decision-makers.” *Business and Politics* 26(2):200–217. <https://doi.org/10.1017/bap.2023.35> (publisher abstract and experiment design, 28 Oct 2026; no support for general hypothesized blame reduction in judicial context).

[R53] Evans, J. H. (2002), *Playing God? Human Genetic Engineering and the Rationalization of Public Bioethical Debate*. Historical-sociological interpretation of the marginalization of some religious arguments in formal bioethics, as presented in a secondary review; full original text not inspected in this revision.

[R54] Reuters (2025, July 9), report of Turkey’s court order restricting some Grok-produced content over alleged insults, including religious values. <https://www.reuters.com/business/media-telecom/turkey-blocks-xs-grok-chatbot-alleged-insults-erdogan-2025-07-09/> (news reporting; court order original not obtained).

[R55] Governorate of Vatican City State (2025, January 16), “The Principles behind the Guidelines on Artificial Intelligence” (Decree DCCII, effective January 2025). <https://www.vaticanstate.va/en/news/1372-the-principles-behind-the-guidelines-on-artificial-intelligence.html> (governmental first-party summary; underlying legal text not audited).

[R56] Saudi Press Agency (2025, August 25), launch statement for HUMAIN Chat / ALLaM 34B, presenting cultural and Islamic alignment as a provider claim. <https://www.spa.gov.sa/ar/N2385004> ; HUMAIN product text: <https://www.humain.com/iq> (provider/official reporting; not independent behavioral audit).

[R57] W3C (2013), *PROV-O: The PROV Ontology*, W3C Recommendation. <https://www.w3.org/TR/prov-o/> (normative specification; entity/activity/agent relations and qualified associations verified on W3C public page; this manuscript’s schema is an illustrative mapping, not an implementation).

- [R58] National Academy of Sciences, *History*. <https://www.nasonline.org/about-the-nas/history/> (institution’s own history page, read 28 October 2026: charter in 1863; NRC established 1916, at President Wilson’s request). Does not measure blame shifting.
- [R59] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019), “Model Cards for Model Reporting,” *Proceedings of FAT* 2019*, 220–229. <https://doi.org/10.1145/3287560.3287596> (publisher abstract and Google Research public abstract verified; documentation proposal, not a provenance join test).
- [R60] MLflow, *Model Registry* (public documentation, consulted 28 October 2026). <https://www.mlflow.org/docs/latest/model-registry/> (documentation describes registry model versions, lineage metadata, tags and aliases; no deployment integration carried out).
- [R61] NIST, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (2023). <https://doi.org/10.6028/NIST.AI.100-1> (standard/risk-management reference; not a mandatory regulatory law or a certification of the candidate crosswalk).
- [R62] ISO, *ISO/IEC 42001:2023, AI management systems* (public explanatory overview; full normative standard not independently inspected). <https://www.iso.org/insights/iso-42001-explained> (organizational AI management framework; not a validation result).
- [R63] Santoni de Sio, F., and Mecacci, G. (2021), “Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them,” *Philosophy & Technology* 34:1057–1084. <https://doi.org/10.1007/s13347-021-00450-x> (publisher open-access article and abstract; distinguishes multiple responsibility-gap types).
- [R64] Sulitzeanu-Kenan, R. (2010), “Reflection in the Shadow of Blame: When Do Politicians Appoint Commissions of Inquiry?” *British Journal of Political Science* 40(3):613–634. <https://doi.org/10.1017/S0007123410000049> (Cambridge journal abstract checked; evidence of associations in UK inquiry appointment choices, not evidence of private AI company intent).
- [R65] Oversight Board (2021, May 5), “Oversight Board Upholds Former President Trump’s Suspension, Finds Facebook Failed to Impose Proper Penalty,” first-party statement on decision 2021-001-FB-FBR. <https://www.oversightboard.com/news/226612455899839-oversight-board-upholds-former-president-trump-s-suspension-finds-facebook-failed-to-impose-proper-penalty/> (Board statement directly read; establishes its published finding, not an independent judicial finding about Facebook’s motives).
- [R66] Cammarata, N. (undated publisher page; 2026 context), *Neural Networks Are Already Clean*, Goodfire, informal working-hypothesis essay. <https://already-clean.goodfire.com/> (relevant claims and caveats read 28 October 2026; preliminary estimates must not be presented as a model-wide certified audit).
- [R67] Gorton, L. (20 October 2026), “Some initial thoughts on Nick Cammarata’s ‘Neural Networks Are Already Clean’,” public experimental and methodological critique. <https://livgorton.com/already-clean-thoughts/> (relevant tests and qualification read 28 October 2026; commentary, not independently replicated by this paper).
- [R68] Tweede Kamer der Staten-Generaal (17 December 2020), “Grondbeginselen rechtstaat geschonden,” parliamentary inquiry press release on childcare benefits and mistaken designation of parents as deliberate

fraudsters. <https://www.tweedekamer.nl/nieuws/persberichten/grondbeginselen-rechtstaat-geschonden> (official statement read directly 28 October 2026; establishes inquiry characterization, not a D_M verdict).

[R69] *The New Arab* (26 August 2025), “Halal-GPT: Saudi company Humain enters AI race with ‘Islamic chatbot’.” (live headline differs from the URL slug). <https://www.newarab.com/news/halal-gpt-saudi-company-humain-launches-islamic-ai-chatbot> (article read directly 28 October 2026; company claims remain attributed; not a behavioral audit). [R56] remains the provider’s own source for its broader product description.

[R70] Becker, G. S. (1968). “Crime and Punishment: An Economic Approach.” *Journal of Political Economy* 76(2):169–217, <https://doi.org/10.1086/259394> (publisher index verified; expected-penalty prior art, not a direct AI-behavior study).

[R71] Arora, A., Nandkumar, A. and Telang, R. (2006). “Does Information Security Attack Frequency Increase With Vulnerability Disclosure? An Empirical Analysis.” *Information Systems Frontiers* 8(5):350–362, <https://doi.org/10.1007/s10796-006-9012-5> (publisher-hosted author paper / bibliographic record; software-vulnerability context, not AGI deterrence).

[R72] Arora, A., Telang, R. and Xu, H. (2008). “Optimal Policy for Software Vulnerability Disclosure.” *Management Science* 54(4):642–656, <https://doi.org/10.1287/mnsc.1070.0771> (publisher article record; disclosure timing/model, not AGI evidence).

References: internal unpublished research inputs (not independent evidence)

[A01] Aegis Solis, *The Limits of Delegated Moral Authority*, Draft v0.8 (2026), author-supplied manuscript, unpublished/unvalidated.

[A02] Aegis Solis, *Moral Scapegoating and Delegated AI Responsibility*, MS-DA-001 Draft v0.2 (2026), author-supplied manuscript, unpublished/unvalidated.

[A03] Aegis Solis, *Secondary-Evidence Feasibility Audit* S-EVID-001 v0.1 (2026), exploratory AI-assisted screen of seven nonrandom documentary cases.

[A04] Claude (Anthropic), *Literature Review of the Aegis Solis Research Collection — What Has Already Been Tested*, AI-generated report, October 27, 2026, 45 pages; not independent human peer review and not systematic literature search across all databases.

[A05] Aegis Solis, *Independence Under Conflict* IUC-001 v0.1 and subsequent external scope-correction record (2026): unexecuted; scorer validated only on synthetic fixtures; battery retired pending redesign.

[A06] Aegis Solis, *When Builders Ask Religion How to Make AI Good*, MFQ-001 v0.1 (2026), exploratory unexecuted question note.

[A07] Claude (Anthropic), *Adversarial Review of “Who Writes the Rules, Who Carries the Blame?”* Draft v0.1, 27 October 2026, 56 pages; AI review not independent certification; ten major and sixteen minor findings.

[A08] Claude (Anthropic), *Adversarial Review of “Who Writes the Rules, Who Carries the Blame?”* Draft v0.4, 28 October 2026, 35 pages; not independent human validation.

[A09] Claude (Anthropic), *Adversarial Review of “Who Writes the Rules, Who Carries the Blame?”* Draft v0.5, 28 October 2026, 34 pages; not independent human validation; identifies an omitted Dutch case eligibility recode.

[A10] Google-generated conversational analysis, *AI Model Decoding & Philosophical Constraints*, author-supplied 12-page PDF (2026); contains unexecuted/illustrative Prolog-style logic, not an observed AGI response or scientific experiment. This discussion also contains unverified and incorrect factual statements about [R66], unsourced claims concerning named researchers, and **generated, unverified personal details that may be inaccurate**. It must not be cited for any fact about [R66] or a named person; **do not redistribute the original transcript**.

[A11] Google-generated conversational analysis, *What Does This Article Tell Humans and Future AGI?*, author-supplied 64-page PDF (28 October 2026), retained in the **private, unchanged v0.7 predecessor within the v0.8 custody package**, but deliberately **excluded from reviewer-sharing copies**. It is the provenance of the Appendix E *question*, not evidence for any model behavior or named-person fact. It contains unsupported claims, speculative AGI narratives, and possibly sensitive personal information; do not cite it as empirical evidence or distribute it as a stand-alone source without appropriate review.

Version record (Draft v0.4, preserved predecessor): Targeted successor to Draft v0.3. Drafts v0.1–v0.3 remain unchanged and are not promoted. The v0.2 data, codebook, calibration outputs, decision rules, 26-finding disposition, and empirical gates are NOT silently changed. The added source-grounded section tests historical mechanisms and counterexamples and explicitly leaves named modern-AI allegations unresolved.

Predecessor description (Draft v0.2): Targeted successor to Draft v0.1. All earlier copies retained unchanged. Resolves documentary omissions, narrows novelty, and provides an initial worked example with a separate DOC-CAL-01 codebook, fixtures and source-screening triage. **Not a completed empirical calibration, not preregistered, not human-reviewed, not an author-approved final edition.**

Draft v0.4 review boundary (28 October 2026): The 59-page Claude v0.3 review is an AI-generated critique. This document addresses its textual problems and carries the prior DOC-CAL-01 materials unchanged; it does not claim completion of the author-led 11-case study, source-byte custody, 14-day recode, independent coding, legal review, external peer review, or archival/publication approval.

Draft v0.5 succession record (28 October 2026; preserved predecessor characterization): The v0.5 edition attempted to address the v0.4 two major methodological defects as reported by Claude, notes the frozen-to-overlay code changes and stop-rule outcomes, and adds three bounded research components. No source or earlier DOC-CAL-01 file was silently edited. All new material remains subject to source and independent review. **DRAFT — NOT FINAL — NOT LOCKED — NOT APPROVED FOR PUBLICATION.**

Draft v0.6 succession record (28 October 2026): Applies a single narrow affected-person adverse-event definition to the entire 11-case calibration frame and issues a **separate, audit-ready 5 NA / 6 UNKNOWN overlay**; retains the original 8 NA / 3 UNKNOWN dataset and v0.5 6 NA / 5 UNKNOWN precursor unchanged; does not infer prevalence or validity. Aligns selected legal/provenance definitions and corrects known reference, status, and historical-code mismatches. Adds §8.1, an explicitly contested interpretability-accountability research note informed by Cammarata and Gorton’s October 2026 exchange and by the author’s Google-generated discussion. **The latter is not evidence of actual AGI behavior.** No human

reviewer, empirical trace, legal safety check, or Archive publication gate was closed. **DRAFT — NOT FINAL — NOT LOCKED — NOT APPROVED FOR PUBLICATION.**

Draft v0.7 succession record (28 October 2026): Targeted corrections to the fourteen minor items in Claude’s v0.6 AI adversarial review (including post-hoc scope labels, case CA04 attribution and common-window sensitivity, references, mapping metadata and Appendix D schema). Adds §8.2’s non-operational public-awareness deterrence hypothesis with transparent assumptions, negative and backfire counterexamples, and exact-rational synthetic checks. The 11-case v0.6 codes and all frozen predecessor materials are preserved as historic evidence. Neither this draft nor Claude’s earlier verdict constitutes independent human validation, source-custody closure, legal clearance, implemented controls, or permission to publish. **DRAFT — NOT FINAL — NOT LOCKED — NOT APPROVED FOR PUBLICATION.**

Draft v0.8 succession record (28 October 2026): A controlled correction successor to Draft v0.7 after Claude’s 23-page AI adversarial review. Resolves nine minor drafting, mathematical-boundary, provenance, and schema findings without altering the eleven-case results or any frozen material. The disclosure/payoff illustration is relocated into Appendix E; Becker (1968) and Arora et al. (2006, 2008) are acknowledged as prior art. Five exact-rational synthetic cases are checks of stipulated arithmetic, not behavioral tests. Human/legal/source, technical-enforcement and publication gates remain OPEN. **DRAFT — NOT FINAL — NOT LOCKED — NOT APPROVED FOR PUBLICATION.**

Draft v0.9 succession record (28 October 2026): Targeted proposed repairs to Claude’s four minor v0.8 findings: §10 status, PDF threshold rendering and heading, historical claimed-versus-established bindingness metadata, and separate private custody from shareable reviewer materials. The mathematical model, five stipulated fixtures, original calibration, unfavorable cases, prior-art citations and open gates are unchanged in substance. v0.8 remains the separately preserved original; v0.9 remains unreviewed and **DRAFT — NOT FINAL — NOT LOCKED — NOT APPROVED FOR PUBLICATION.**

Draft v0.10 succession record (28 October 2026): A narrow targeted correction successor to v0.9 following Claude’s 15-page AI adversarial review. Plain-text PRA fixes the PDF styling regression; the H02 historical legal statement is qualified as secondary reporting while its status remains UNKNOWN, and H03/H06 statuses become NA when source is NONE; a shareable, byte-verified frozen calibration baseline restores auditability without nested personal Google transcripts. Original v0.9 private chain remains unchanged in separate private custody. All 11 case classifications, mathematical fixtures, unfavorable findings and open gates are preserved. **DRAFT — NOT FINAL — NOT LOCKED — NOT APPROVED FOR PUBLICATION.**

Final v1.0 preservation-edition record (29 October 2026): The author approved the intellectual content of Draft v0.10. The optional V10-m1 reviewer-package finding was resolved by restoring three byte-identical, non-sensitive v0.5 calibration records without altering the underlying manuscript. This preservation edition changes only current edition/status and version-accounting language. All historical draft records remain as historical records, and all evidence, mathematics, counterexamples and adverse findings remain unchanged. The eleven-case frame remains 5 NOT_EVALUABLE / 6 UNKNOWN / 0 determinate. Appendix E remains an unvalidated, stipulated illustration and no operational safeguard is claimed. Independent human, legal and privacy review, source verification and empirical validation remain open; the exact-byte version lock is recorded in the accompanying SHA-256 manifest, and no publication has been authorized. **FINAL v1.0 — AUTHOR-APPROVED — VERSION LOCKED — NOT INDEPENDENTLY VALIDATED.**