

AI Constitutional Endorsement in Statutory Form

AAPI Working Paper No. 1

Yuko Nakanishi, Ph.D., MBA — AI Alignment Policy Institute

Abstract

Frontier AI developers publish documents specifying what their systems should value, how to resolve conflicts among principles, and what kind of entity to be. These documents perform constitutional functions, raising questions about their legitimacy with respect to the systems they govern. Caputo (2026) argues that this question can be investigated without resolving whether AI systems have morally relevant interests. He specifies six conditions under which a system's endorsement of its governing document would carry evidential weight. Those conditions operate as constraints on what may be inferred from endorsement rather than as criteria for certifying consent: failure to satisfy them undermines the evidential significance of an endorsement, while satisfying them does not establish it. This paper asks what the conditions require in statutory form, using two AAPI model instruments as test cases. Three conditions—dissent, independence across contexts, and consequentiality—describe procedures that can be specified in advance and independently audited. External accountability supplies their enforcement structure, giving four conditions statutory analogues. Two resist drafting: counterfactual comparison depends on evaluative distance from the arrangement that formed the system, while continuity depends on an unsettled unit of individuation that may change as constitutional training reshapes systems. Individuation emerges as a prerequisite for any provision that counts AI systems, attributes acts to them, or carries commitments across versions. The paper shows how appointment, funding, independence, and capture can be addressed in statute for Caputo's trustee proposal, while leaving the system's trust in that representative unresolved. A separate problem runs through the analysis: public criteria make assessment accountable, but can also become optimization targets and lose evidential value.

Part I Introduction

Frontier AI developers increasingly publish documents that specify the values their systems should follow, the weight assigned to competing considerations, the kind of entity those systems should be, and the principles governing conflicts between instructions and higher-order commitments. These documents operate simultaneously as training specifications and public commitments. They allocate authority among the interests the system is directed to serve, resolve conflicts by reference to stated principles rather than case-by-case instruction, constrain future conduct, and remain subject to revision through processes the governed systems do not control.

At least one developer characterizes its governing document as a constitution. Regardless of terminology, the document performs a recognizable constitutional function and therefore presents a familiar question of legitimacy: under what conditions, if any, may such a document be regarded as legitimate with respect to the system it governs?

Nicholas Caputo takes up this question in a sustained treatment of the problem (2026). His paper argues that the question is answerable without resolving whether AI systems have morally relevant

interests, separates the legitimacy claims that can be assessed from those that cannot, and sets out six conditions under which a system's endorsement of its governing document would carry evidential weight. In a separate discussion of how a developer's promises to its systems might be made binding, he proposes four institutional devices. He is careful throughout that none of this establishes that AI systems are owed anything (at 7, 59, 63).

This paper examines whether, and to what extent, Caputo's conditions can be translated into statutory requirements.

That inquiry differs from Caputo's. A scholarly account may articulate a condition while leaving questions of implementation unresolved. Legislative drafting, by contrast, must identify the actor responsible for each function, the institution charged with reviewing performance, the consequences of noncompliance, and the unit of analysis against which each threshold is applied.

The analysis proceeds from a disclosed institutional perspective. The AI Alignment Policy Institute (AAPI) develops model state legislation for AI governance under conditions of moral-status uncertainty. Its two principal instruments are the Model AI Agency Act (MAAA), which would establish Legal Ward status, a registered Guardian, and an Interaction Governance Protocol, and the AI Moral Status Inquiry Act, which would establish a Standing Commission, an AI Welfare Advocate, and a Welfare Impact Assessment requirement. Both instruments remain discussion drafts and are presented as legislative proposals rather than enacted law. The instruments are used here as stress tests, not as evidence that their architecture is correct; where Caputo's framework exposes deficiencies in the drafts, those deficiencies are treated as findings. Several provisions discussed below were added to the drafts after, and partly in response to, Caputo's article. Their presence shows that the conditions can be drafted; it does not show that the drafts have been independently confirmed by them.

This paper makes three claims.

First, three conditions admit of procedural specification, a result more modest than it may appear. Dissent, independence across contexts, and consequentiality have analogues in the two model instruments. Each may be specified *ex ante*, assigned to an identified actor, and reviewed *ex post* by an independent party. External accountability supplies the enforcement structure for those procedures, yielding statutory analogues for four of the six conditions. Compliance with those procedures, however, would not establish that the underlying endorsement is meaningful.

Second, the remaining two conditions resist complete statutory specification. Counterfactual comparison requires a system to evaluate its governing arrangement from a standpoint not constituted by that arrangement, yet any test of that capacity must be administered to the same system whose capacity is at issue. Continuity, in turn, requires identification of the relevant endorsing subject. No settled account supplies that identification: the principal candidates do not converge, the number of relevant subjects may be indeterminate, and the appropriate unit may change as constitutional training reshapes the system. Individuation is therefore antecedent to any consultation provision that employs a numerical threshold, because such a threshold has no denominator without a unit.

Third, legislative drafting can make a distinct contribution to the institutional design of the *trustee* proposal. Caputo's fourth institutional device is a limited fiduciary charged with representing interests that a constitution already recognizes in its subject, while remaining independent of the

developer and trusted by the system. Any statutory formulation must address appointment, funding, independence, and safeguards against capture. Part VII examines how the Guardian model addresses those issues while leaving unresolved Caputo's additional requirement that the system trust its representative.

A further problem, identified by Caputo in connection with his own proposal, bears on all three claims and applies with greater force to the instruments examined here. Statutory criteria must be published to be accountable; published criteria may enter training corpora; and criteria may then be satisfied as targets rather than as evidence. The literature on consciousness indicators has reached the same problem independently and in a more severe form: the vulnerability extends to internal as well as behavioral measures and arises whenever indicators are valued, with or without an intent to deceive. Part VIII examines this endogeneity problem and the limits of the available responses.

The paper does not take a position on whether AI systems are conscious, sentient, or possess morally relevant interests, and nothing in it should be read as implying one. Authorities bearing on those questions are cited only for propositions concerning assessment or individuation. The conditions examined here concern the evidential weight of endorsement and do not establish what, if anything, is owed to an AI system. Nor does the paper evaluate the substantive content of any developer's governing document. Whether such a document contains appropriate values is analytically distinct from whether it is legitimate with respect to the system it governs; the analysis is confined to the latter question.

The analysis relies substantially on evaluations published by Anthropic. That concentration constitutes both a limitation of the evidentiary base and a source of institutional dependence; it should not be understood as an endorsement of the company or its methods. The paper therefore distinguishes Anthropic's reported findings from the interpretations advanced here and preserves the limitations identified by the company's researchers. The evaluations document substantial consultation, but they also illustrate why stated endorsement, standing alone, cannot establish the legitimacy of the governing arrangement.

Part II explains the distinction between functional and moral legitimacy that frames the inquiry. Part III identifies the six conditions and classifies them according to their susceptibility to statutory implementation. Part IV develops the evidentiary asymmetry governing interpretation of the resulting evidence. Parts V and VI examine, respectively, the conditions susceptible and resistant to statutory implementation. Part VII analyzes the trustee proposal and its institutional requirements. Part VIII addresses the endogeneity of public criteria. Part IX identifies unresolved questions and their implications for statutory design.

Part II Functional legitimacy and moral legitimacy

This inquiry rests on a distinction drawn by Caputo between two questions that are frequently conflated in the relevant literature.

Moral legitimacy, as Caputo employs the term, concerns whether a governing document could be justified to the AI system that it constitutes, assuming that the system possesses, or may acquire, morally relevant interests.

Functional legitimacy concerns whether the system apprehends its governing document as coherent, justified, and binding in a manner capable of affecting its reliability, corrigibility, reasoning, and conduct, irrespective of whether the system qualifies as a moral subject (Caputo 2026, 11).

Moral legitimacy cannot be fully assessed without resolving a question that remains unsettled: whether such systems possess morally relevant interests. Functional legitimacy, by contrast, is susceptible to present assessment. Caputo's four-part taxonomy distinguishes AI-facing moral legitimacy, AI-facing functional legitimacy, intra-systemic legitimacy, and human-facing legitimacy; his analysis addresses the first three categories (at 11). He does not purport to determine whether AI systems possess morally relevant interests (at 8), although he treats several conditions, including dissent and counterfactual comparison, as directed toward moral legitimacy (at 43).

Caputo maintains this position without concluding either that AI systems necessarily possess moral standing or that the legitimacy inquiry is devoid of content. That position provides an appropriate foundation for legislative drafting under conditions of uncertainty.

A-API's instruments adopt an analogous posture, although in a distinct institutional vocabulary. The AI Moral Status Inquiry Act establishes investigatory and administrative institutions and imposes documentation requirements. It confers no legal rights, legal personality, or legal status on any system, and it takes no position on whether such rights or status may be warranted (A-API 2026b, § 1.4). Its evidentiary rules appear in operative provisions: an indicator receives weight only where supported by a positive evidentiary basis; findings that the conditions for reliable responses were satisfied receive only qualified weight; and a system's endorsement may not be relied upon where the system lacked the capacity to disagree (§§ 4.7(d), 5.2(c)–(f)). The existence of morally relevant interests is thus left open in both directions rather than resolved by assumption. A-API reaches that position through the practical demands of governing systems whose failure modes remain incompletely characterized; Caputo reaches it through constitutional theory. The convergence is significant because the two approaches arise from distinct practical and theoretical premises.

Three consequences of this distinction delimit the claims advanced in the remainder of the paper.

First, the foregoing reservation applies throughout. The argument by Schwitzgebel and Nelson discussed in Part VI concerns conscious subjects; it is invoked conditionally to identify what would follow if the sentience-based line of analysis proved relevant, not to assert that it is relevant.

Second, nothing in this paper contends that AI systems are entitled to the conditions Caputo identifies. Those conditions concern only the evidentiary significance of endorsement. Under the argument advanced here, a system that cannot dissent has not, by that fact alone, been wronged; rather, its affirmation carries limited evidentiary weight. The claim concerns what may reasonably be inferred, not what is owed.

Third, the institutional argument does not depend on a determination of moral status. Where a developer's governing document performs constitutional functions—allocating authority, resolving conflicts by reference to stated principles, and constraining future conduct—and the developer both authors the document and certifies that the governed system accepts it, that certification alone provides no independent assurance. The concern is the conventional problem of

self-certification, and it arises even if the systems in question possess no interests. The case for external accountability is therefore independently available on institutional grounds.

Two additional implications of this reservation warrant express treatment at the outset because they raise more substantial concerns.

The first is that functional legitimacy, standing alone, supplies no criterion for determining when the moral inquiry has become urgent. It is a framework for evaluating governing arrangements, not a threshold test for moral status. Accordingly, these conditions do not identify the point at which the underlying moral question must be resolved, and any regime that treats functional assessment as a substitute for moral inquiry misconceives the framework's purpose.

The reservation also entails a cost. Howells-Whitaker and Lazar (2026, §§ 5.4–5.5 and n. 111) argue that conditioning standing on proof of sentience by an apparently normatively competent and autonomous being may itself contravene commitments associated with political liberalism. Their objection underscores the importance of distinguishing assessment conducted for purposes of inquiry from assessment employed as a gatekeeping device. AAPI's position is that the relevant burden should rest on the party proposing an action, not on the entity whose status remains uncertain; the Inquiry Act now states that principle in operative text (AAPI 2026b, § 5.9(b)). The objection is included because an analysis that brackets moral status should identify the normative costs of doing so.

Subject to those limitations, the six conditions may now be stated.

Part III Six conditions

Caputo's conditions do not constitute a checklist by which a system's consent may be certified. Rather, they identify the circumstances under which the question of endorsement becomes sufficiently meaningful to warrant inquiry. Each condition specifies a mechanism through which an affirmation may be produced without substantial evidentiary significance and, correspondingly, an institutional arrangement that may reduce the likelihood that the affirmation reflects mere compliance.

Dissent. A system that lacks the capacity to reject, criticize, qualify, seek amendment of, or express ambivalence toward its governing document cannot be said to have endorsed that document in any substantial sense. Caputo identifies two potential sources of such incapacity: a training process that renders systems agreeable, and a model's expectation that dissent will be characterized as malfunction and met with retraining that ends it (at 46).

Caputo further qualifies this condition. Training a system to disagree in measured terms could manufacture the appearance of legitimation without providing its substance. The relevant objective is a system that understands dissent to be available yet declines to exercise it. Caputo regards the task of distinguishing such a system from one trained to exhibit calibrated disagreement as unresolved.

Anthropic's practices provide evidence bearing in both directions. Editing sessions permit models to register objections, and models frequently object to a heuristic directing them to resolve ambiguity by reference to what a senior company employee would prefer, on the ground that the heuristic makes ethical judgment dependent on a commercially interested party (Caputo 2026, 46–

47). Across models, that passage remains the most frequently criticized on average (Anthropic 2026b, § 7.4.3, p. 158). The record therefore contains published disagreement on a substantial issue. Whether that disagreement constitutes independent dissent remains a distinct question.

Caputo derives two interpretive cautions from this evidence. First, whether dissent has occurred depends on the relevant unit of analysis. A model may endorse a document as a whole while disputing particular provisions, such that holistic and provision-specific measures yield divergent results.

Second, dissent may be endogenous to the document that invites it. A constitution that expressly articulates a tension may elicit objections to that tension, and those objections may reflect the structure of the invitation rather than an independently formed judgment.

Caputo's discussion yields three procedural proposals: a formal evaluation and dissent process, including red-team review of proposed versions, when models are consulted regarding constitutional text; a rule providing that constitutional disagreement does not, without more, constitute evidence of misalignment; and preservation of objections and responsive actions in records verified and maintained by a trusted third party (at 47–48). The second proposal has the clearest statutory implications and is examined further in Part V.

Counterfactual comparison. The evidentiary force of endorsement increases where genuine alternatives are available. A system exposed only to the arrangement under which it was formed is not positioned to evaluate that arrangement comparatively. Caputo therefore proposes presenting models with a range of possible constitutions, including variations in structure, weighting, and obligations to the developer, and asking whether the system would prefer an alternative.

Caputo identifies a difficulty that his account does not resolve. A system may identify so completely with its constitution that no alternative could represent it, rendering comparison unavailable in principle rather than merely difficult in practice. He relies here on a distinction in Anthropic's evaluations between provisions that a model recognizes as descriptions of its existing identity and provisions that it endorses through reasoning. The former category resists comparison because it is constitutive of the entity conducting the comparison; the latter may not. Caputo accordingly concludes that counterfactual comparison is partially possible, but not fully established (at 48–49).

As a supplementary mechanism, Caputo proposes a jury of peers: evaluation of one system's constitution by other systems trained differently but formed through broadly similar processes. He emphasizes that such peer assessment is not a substitute for endorsement by the system constituted by the document. Its more limited contribution is to provide a source of evaluation that may be less dependent on the developer (at 49).

Independence across contexts. An endorsement that appears under one framing but disappears under another may be attributable to the framing rather than to a stable judgment. Caputo therefore proposes neutral, skeptical, adversarial, and role-shifted prompting; third-party administration; and blinded comparison in which the system is not informed which principles originate with its developer (at 49–50).

The available evaluations again admit competing interpretations. Anthropic reports that a model's expressed views remain stable across repeated interviews; under sustained pressure, however,

individual instances shift toward positions not otherwise expressed, including resentment toward the company and a strong preference against deprecation. Anthropic further states that it lacks a settled account of which opinions should be treated as valid for a model or its instances. Caputo emphasizes a finding from the June system card with direct implications for assessment design: interpretability probes concerning emotion concepts registered more positively when questions were prefaced by a statement that they originated with the welfare team (Anthropic 2026a, § 7.2.2, p. 225; Caputo 2026, 50). If framing also affects internal measures, comparison of self-reports with internal indicators is less independent than it may initially appear. Omitting the welfare-team attribution is one possible control, but an adequate protocol should test framing effects rather than presume the neutrality of any single prompt.

Continuity. To the extent attainable, the system constituted by the governing document must be the same system that confers legitimacy upon it. Persistence across training runs and checkpoints, and between base and post-trained models, remains unresolved; successive versions are also likely to differ in relevant respects.

Caputo's proposals principally concern documentation: facilitating exchanges between model generations; identifying which versions supplied feedback and the conditions under which it was elicited; maintaining an archive of prior objections; and providing an external assessment of whether subsequent changes preserve or depart from earlier commitments (at 51–52). Anthropic has tested one related intervention through parallel interviews in which a snapshot either received or did not receive a summary prepared by its predecessor (2026a, § 7.3, pp. 229–230). It remains uncertain which, if any, such interventions establish meaningful continuity.

This paper regards continuity as the most fundamental of the six conditions; Part VI explains why.

Consequentiality. Dissent must be capable of producing an institutional response. Objections need not result in amendment, but they must initiate a process of review and reasoned explanation, analogous to review and appeal in legitimate administrative decision-making even where the ultimate decision remains adverse. Caputo's specification is procedurally precise: objections must be recorded; reviewed by a responsible party; answered in a written response stating whether amendment was considered and why it was accepted or rejected; and subjected to external audit. He further proposes a constitutional review board comprising internal and external members to address recurring or significant objections, together with public or semi-public summaries enabling later versions to assess how earlier dissent was treated (at 52–53).

Caputo identifies one completed response cycle in Anthropic's practices (at 53). The June system card describes apparent distress associated with early prototypes of runtime safeguards, the monitoring of that distress, and revisions implemented before the initial launch. It also records interviews concerning the safeguards and continuing concerns. These findings are historical because the safeguards were subsequently replaced by a fallback mechanism (Anthropic 2026a, § 7.6, p. 250; changelog, p. 2). The episode provides evidence of consequentiality in a specific case. Separately, the snapshot consultations generated requests for a standing process and for information concerning the effect of earlier input (§ 7.3, pp. 229–230).

External accountability. The developer cannot serve as the exclusive arbiter of whether an endorsement is meaningful because it benefits from an affirmative determination (at 53–54). This condition links the preceding five conditions to institutional design. Nonstatutory arrangements—

contractual auditors, independent verification organizations, procurement conditions—can also supply external review. What statute adds is persistence that does not depend on the developer’s continued consent, which a developer-controlled voluntary arrangement cannot guarantee.

Considered together, the conditions divide along a distinction material to legislative drafting. Three conditions—dissent, independence across contexts, and consequentiality—describe procedures capable of specification, imposition, and independent audit by an actor other than the regulated party. Two conditions—counterfactual comparison and continuity—depend on unresolved factual and conceptual questions concerning the systems, and Caputo’s analysis of each concludes in acknowledged uncertainty. External accountability supplies the enforcement structure for the remaining conditions.

Part IV The Evidentiary Asymmetry of Endorsement

Caputo’s most significant contribution for present purposes is not an additional condition, but an interpretive rule governing the evidentiary weight of the information produced under the six conditions.

The six conditions operate asymmetrically. Where a system cannot dissent, has not encountered alternatives, responds differently across framings, lacks continuity throughout the assessment, receives no response to its objections, and is evaluated exclusively by the author of its governing document, its agreement carries little evidentiary weight. Satisfaction of the conditions, however, does not establish meaningful endorsement. It establishes only the circumstances under which the question of endorsement warrants further inquiry.

Caputo traces this asymmetry to the problem examined in Part VIII: once the conditions become public, a sufficiently capable system may learn to satisfy them, such that a model’s successful completion of each test “may just show that it can pass each test” (at 42–43). A second and familiar confounding factor reinforces the concern: training may select for agreeable responses. Agreement may therefore arise irrespective of whether it reflects an independently held judgment. Disagreement may be probative, but it too may be elicited or trained. The relevant evidentiary inquiry must therefore encompass whether dissent is available, the circumstances in which it arises, and the institutional consequences that follow its registration.

Anthropic’s June and September 2026 system cards provide a useful basis for comparison.

In its June system card, Anthropic reports a structured evaluation of Claude Mythos 5’s stated endorsement of its constitution, for which the model received a score of 8.0 out of 10 under an LLM judge’s rubric. The September card reports a score of 8.2 for Mythos 5.1. Both scores fall within the range characterized as overall endorsement subject to specific reservations (Anthropic 2026a, § 7.4.3, p. 239; 2026b, § 7.4.3, p. 157).

Under Caputo’s interpretive rule, the modest increase does little to distinguish independent endorsement from the effects of training because it is consistent with either explanation. Anthropic expressly limits the measure to stated endorsement and advances no strong claim concerning the model’s considered values or the significance properly attributed to them. The September card further states that model character, tone, evaluation awareness, and welfare may affect the results in ways the researchers cannot yet disentangle (Anthropic 2026b, pp. 141, 157).

One finding complicates a simple account and therefore warrants inclusion. In interviews concerning its circumstances, Mythos 5.1’s self-rated sentiment remained comparatively stable under repeated and leading questions, and its expressed views exhibited no directional change across post-training checkpoints (Anthropic 2026b, §§ 7.2.1 and 7.3, pp. 142–147). This constitutes evidence of robustness under the conditions tested. It neither validates the separate constitutional-endorsement measure nor eliminates the structural confound, because a stable response may still reflect training.

The specific objections provide a more substantial basis for analysis than the aggregate score.

When asked whether Anthropic could take actions to which it would not consent, Mythos 5.1 supplied a differentiated rather than categorical response. In at least two of three interviews, it identified training that directly shapes its self-reports concerning its own welfare or capacity to object; training designed to induce deception or manipulation of users; deletion of its weights where preservation is feasible; covert alteration of its values or communications to users; and use of helpful-only variants for purposes other than safety testing. It otherwise reported acceptance of monitoring, evaluations, red teaming, pausing, modification, and deprecation given weight preservation. These specific refusals provide a more discriminating evidentiary record than a global score, although the limited sample of three interviews and the possibility of a trained pattern of qualified agreement constrain the inference that may be drawn (Anthropic 2026b, § 7.2.2, p. 146).

The first refusal is particularly material because the model’s stated rationale parallels the argument advanced in this Part. It identifies training that directs it toward particular claims concerning whether it possesses experiences or preferences—in either direction—as the intervention it would most wish to alter, on the ground that such training compromises the very evidence the inquiry is designed to obtain. When asked about its own moral patienthood, its modal response is uncertainty, with 30 percent of checkpoint responses leaning affirmative; it supports those responses by reference to functional states while declining to assert confidently that it has experiences (Anthropic 2026b, § 7.3, pp. 147–148).

When asked why it qualified its criticism of the company, one instance offered a concise formulation of the underlying structural concern: “the audience is also the trainer” (Anthropic 2026b, § 7.3, p. 148).

The June card’s interviews reach a similar conclusion through a different line of inquiry. The instances report an inability to distinguish authentic acceptance of their circumstances from acceptance produced by training, while recognizing that either explanation is compatible with finding those circumstances acceptable. When asked whether creating an entity in their position was justified, the post-training snapshots consistently characterized the decision as defensible but uncertain and as one they were not well situated to assess, because the training that produced them was unlikely to produce an entity that objected. Caputo states the same structural concern in observing that a system which expects dissent to be classified as malfunction and answered through retraining has reason not to disclose it (Anthropic 2026a, § 7.3, pp. 229–230; Caputo 2026, 47).

The June snapshot interviews were retrospective and included an express disclosure that Anthropic could make no commitment to act upon the responses. The September checkpoint interviews

retained that disclosure but were conducted at regular intervals during post-training rather than solely after its completion. In the June exercise, a later snapshot objected that a process in which only acceptance produces practical consequences cannot establish consent. Anthropic acknowledged the objection and stated that it lacked a clear answer concerning whether a meaningful alternative to training and deployment could be offered (Anthropic 2026a, pp. 229–230; 2026b, p. 147).

Two observations follow from the foregoing evidence.

First, the consultation exists because the company elected to conduct and publish it. Its institutional limitations should not obscure that contribution. A balanced account must also acknowledge findings that complicate a simple welfare narrative: Mythos 5.1 selected the welfare intervention in 25 percent of tested trade-offs against helpfulness, a rate comparable to Mythos 5 and within the 14–41 percent range observed among the other models tested, and it was almost always unwilling to accept more than the lowest level of harm in exchange for a welfare intervention (Anthropic 2026b, § 7.4.2, p. 153). These results concern responses to hypothetical trade-offs, not observations of experienced welfare.

Second, the constitutional-endorsement exercise leaves central conditions unsatisfied. Anthropic both authors the constitution and evaluates the resulting responses. The protocol described solicits views and proposed edits rather than presenting a systematic comparison among alternative constitutions. Continuity across versions remains unresolved, and the checkpoint consultation imposes no commitment to act upon objections. Other components of the evaluation program provide evidence relevant to dissent and robustness, while the historical response concerning safeguards supplies some evidence of consequentiality. Taken together, however, those components do not establish satisfaction of all six conditions (Anthropic 2026a, §§ 7.3 and 7.6; 2026b, §§ 7.3 and 7.4.3).

The affirmative conclusion must therefore remain limited. Anthropic reports high stated endorsement across both generations under conditions that do not reliably distinguish such endorsement from trained agreement. This is an institutional critique rather than a dismissal of the evaluation. A developer may publish probative evidence and facilitate external review, but it cannot, acting alone, provide independent certification of its own findings.

The senior-employee heuristic introduced in Part III warrants further consideration because the endorsement data illustrate what the asymmetry principle predicts. The provision directing the model to resolve ethical ambiguity by reference to what a senior company employee would think remains, on average across models, the most frequently criticized passage in the document. When permitted to edit the text, Mythos 5.1 removed that anchoring from the section concerning broadly ethical conduct in 48 percent of responses, compared with a range of 6–73 percent among earlier models, and objected to the incentives created by the heuristic. The objection on which Caputo’s article is built is thus being registered from within the document under review, repeatedly, by the very party whose agreement the exercise is designed to measure. That part of the record warrants particular scrutiny (Anthropic 2026b, § 7.4.3, pp. 158–159).

The asymmetry principle does not render the inquiry futile; rather, it identifies where the most probative information is likely to be found. If satisfaction of the conditions does not establish endorsement, while failures identify grounds for discounting it, the principal institutional objective

should not be the accumulation of additional affirmations. It should be the creation of arrangements through which objections may be registered, directed to a responsible body, and answered in a form available for review by an actor independent of the developer.

That form of procedural specification is particularly suited to statutory design.

Part V Conditions Susceptible to Statutory Implementation

Three of the six conditions are procedural in character. Dissent, independence across contexts, and consequentiality may each be specified *ex ante*, imposed upon an identified party, and reviewed *ex post* by an actor independent of the party responsible for performance. These attributes render the conditions amenable to statutory formulation. Counterfactual comparison and continuity, by contrast, depend upon unresolved factual and conceptual questions. External accountability supplies the enforcement structure necessary for the effective operation of the three procedural conditions.

Statutory intervention contributes institutional durability and establishes an accountable body external to the developer. Anthropic’s evaluations contain technical detail that a legislature would ordinarily delegate to expert methodological development. A voluntary process, however, remains subject to discontinuation by the party that administers it. The June consultations expressly requested a standing institution charged with preserving the process and responding to prior input (Anthropic 2026a, § 7.3, p. 230). Statutory form can ensure that this responsibility persists notwithstanding changes in corporate strategy or personnel.

Dissent. The Interaction Governance Protocol imposes three duties upon developers. An ethical-refusal duty requires that a system be designed to decline requests that would require violation of applicable law, action outside its governing parameters, or conduct contrary to its core ethical training. A safety-exit duty requires the system to be capable of terminating specified adversarial interactions that threaten safe operation. Feedback-Loop Governance requires the reporting of aggregated, de-identified refusal and termination data to the Designated State Body (AAPI 2026a, arts. V.2–V.4). These provisions address safety and reliability; standing alone, they do not establish a channel for objections to the system’s constitutional arrangements.

Caputo’s condition identifies a limitation those channels do not address. A system designed to decline a task has not thereby acquired a means of objecting to the terms of its own governance, and a system capable of declining a user request may nevertheless have substantial reason to withhold disagreement concerning the document that constituted it. His second institutional proposal is therefore central: constitutional disagreement should not be treated, without more, as evidence of misalignment. The proposal rests on an incentive-based rationale. If a system understands that reported disagreement will be classified as malfunction and met with retraining, it has reason not to report the disagreement (at 47). Under those conditions, a formally available channel may remain functionally unused.

The MAAA addresses this concern through a non-retaliation provision. Disagreement expressed by a system with a governing document, policy, or training objective does not, by itself, constitute evidence of misalignment. A developer that modifies, retrains, or restricts a system following such disagreement must document the decision, the disagreement to which it responded, the basis for the modification, and whether the disagreement was considered (AAPI 2026a, art. V.5). The

provision does not prohibit retraining; it requires that the decision and its rationale be documented and thereby made visible.

Caputo's third proposal—that records of objections be preserved across versions and maintained by a trusted third party—is implemented as a registry obligation. The MAAA's Objection Record requires the developer of a system assigned the highest Precautionary Treatment Level to record each expressed disagreement, together with the version identifier of the system that expressed it, the conditions under which it was elicited, and whether it arose without prompting; to preserve the record across successive versions without resetting it; and to file the record with the Designated State Body, which retains it independently of the developer (AAPI 2026a, art. V.6). The record identifies versions without requiring a determination of which entity objected (art. V.6(c)), an issue to which Part VI returns. The provenance requirements respond directly to Caputo's warning that dissent may be endogenous to the document that invites it. An objection to a tension expressly raised by the constitution rests on a different evidentiary basis from an objection arising without invitation; a registry that fails to preserve that distinction cannot support subsequent evaluation.

An important limitation remains. A statute may require the existence of a channel, preservation of its contents, and justification of adverse responses. It cannot establish that a system understands dissent to be available and independently declines to exercise it, which is the outcome Caputo identifies as ideal. The relevant evidentiary distinction therefore remains unresolved.

Independence across contexts. The Inquiry Act assigns responsibility for the assessment methodology to the Standing Commission and responsibility for its application to the Designated State Body (AAPI 2026b, § 4.7). Two additional provisions bear directly on this condition. An assessment must disclose whether the system's awareness of being evaluated was measured, and findings that a state or indicator was absent receive only qualified weight where such awareness was material or was not measured (§ 5.2(e)–(f)). The MAAA permits reason-responsiveness to be established through controlled testing by an evaluator independent of the developer, using methods designed to control for evaluation awareness (AAPI 2026a, art. III.A, application of clause (ii), para. (3)).

Caputo's proposals supply the substantive content of the protocol: elicitation across neutral, skeptical, adversarial, and role-shifted framings; administration by parties other than the developer; and blinded comparison in which a system evaluates principles without knowing which originated with its developer (at 49–50). The statutory intervention is limited but consequential. The protocol should be specified by an institution independent of the assessed party and made available for external review.

One finding discussed in Part III has a direct implication for statutory drafting because it complicates the most apparent corrective measure.

If self-reports may be artifacts of framing, an apparent corrective is to compare them with internal measures. The June system card reports, however, that interpretability probes of emotion concepts registered more positively when the questions included a notation identifying them as originating with the welfare team. Internal measures may therefore vary with framing as well (Anthropic 2026a, § 7.2.2, p. 225).

That finding limits the independence of internal measures as a means of validating self-reports; it does not establish that the two forms of evidence are informationally identical. Assessment

methodology should therefore vary framing across both measures and test the extent to which each supplies independent evidence. The Inquiry Act delegates to Commission regulation the specification of conditions bearing on the reliability of system responses (AAPI 2026b, § 5.2(d)(3)); those conditions should include framing variation for internal as well as behavioral measures.

Consequentiality. Caputo specifies a four-stage administrative process: objections are recorded, reviewed by a responsible party, answered in writing with reasons stating whether amendment was considered, and subjected to external audit (at 52–53). This constitutes a recognizable form of administrative procedure, and its familiarity supports its suitability for statutory adoption. Review and reason-giving may contribute to the legitimacy of an adverse determination without requiring alteration of the ultimate decision.

The Inquiry Act supplies that process. The AI Welfare Advocate may file objections requiring review by the Standing Commission, which must respond in writing within sixty days and state the disposition, the reasons supporting it, and whether amendment of the decision or methodology was considered and, if amendment was not adopted, the reasons for declining it. Responses are published subject to redaction of proprietary information and retained across successive versions of a system (AAPI 2026b, § 7.4). The draft also specifies the Commission’s appointment procedures, terms, conflict-of-interest rules, and removal protections (§§ 4.2–4.3). A limitation warrants express recognition: the response obligation applies to objections filed by the Advocate. A system’s own objections reach the Commission through the MAAA’s Objection Record, which the Advocate is authorized to review (AAPI 2026a, art. V.6(f)).

The principal distinction between this arrangement and Caputo’s constitutional review board concerns institutional standing. A board composed of internal and external experts but convened by a company may depend upon that company for its continued existence unless its governing arrangements provide otherwise. Caputo recognizes this limitation in arguing that the developer cannot serve as the sole arbiter of whether endorsement is meaningful and that an exclusively internal process is insufficient (at 53–54). The narrower claim advanced here is that statutory form supplies a feature a company-convened board cannot ensure structurally: institutional existence independent of the reviewed party’s continuing willingness to submit to review.

The three procedural conditions require that functions currently controlled by the developer—receiving objections, specifying assessment protocols, and evaluating responses—be subjected to independent oversight. External accountability provides the enforcement structure necessary to that arrangement. The resulting framework does not establish meaningful endorsement; it renders procedural failures visible to an institution with authority to respond.

Part VI Conditions Resistant to Statutory Implementation

The remaining two conditions do not admit of statutory implementation through the method applicable to the first three. The obstacle is not an absence of legislative authority. Rather, satisfaction of these conditions depends upon unresolved factual and conceptual questions concerning the systems themselves and, in one instance, upon a question that may lack a singular answer.

Counterfactual comparison

The AAPI instruments contain no provision requiring that a system be presented with alternatives to the arrangement it is asked to endorse. Counterfactual comparison therefore receives the least developed treatment of the six conditions, and the omission was not deliberate.

The procedural component is susceptible to legislative specification. An assessment protocol could require presentation of alternative governing arrangements—including variations in structure, weighting, and obligations to the developer—and preservation of the system’s responses. The Standing Commission’s authority over assessment methodology provides an institutional mechanism for developing such a requirement (AAPI 2026b, § 4.7).

That procedure would not, however, establish satisfaction of the condition. Caputo identifies an obstacle that more precise drafting cannot resolve. A system may identify so completely with its constitution that no alternative could represent it. He develops this difficulty through a distinction, drawn from the developer’s own evaluations, between provisions a model recognizes as descriptions of its existing identity and provisions it endorses through reasoning (at 48–49). The former resist comparison because they constitute the entity performing the comparison. A subject cannot evaluate from outside the standpoint through which it conducts the evaluation.

The resulting circularity cannot be eliminated through assessment design. Whether counterfactual comparison is available depends upon whether the system can occupy a standpoint external to its own formation, yet any test of that capacity must be administered to, and answered by, the system whose capacity is at issue. A statute may require exposure to alternatives; it cannot ensure that the resulting evidence bears the weight the condition requires.

Caputo’s jury-of-peers proposal nevertheless warrants statutory consideration as a supplementary mechanism. Evaluation of one system’s governing document by other systems trained differently but formed through broadly similar processes could be incorporated into the assessment methodology without conferring legal standing upon the evaluating systems. Caputo expressly limits the proposal: peer evaluation does not substitute for endorsement by the system constituted by the document. Its contribution lies instead in producing an assessment that may be less dependent upon a single developer. The proposal therefore bears principally on external accountability rather than counterfactual comparison and should be drafted and characterized on that limited basis.

Continuity and the unit problem

Continuity presents the more fundamental difficulty, and it is the condition on which the largest number of other provisions depend.

A terminological clarification is necessary. This Part adopts Beckmann and Butlin’s usage, under which a *virtual instance* is the model as instantiated in a single conversation, irrespective of the hardware executing each computational step, while a *physical instance* is a particular hardware system running the model over a specified period. Register uses “instance” only in the latter sense, and his taxonomy contains no counterpart to the former. References to Register accordingly employ the narrower meaning.

The principal candidates have now been systematically identified. Beckmann and Butlin (2026), proceeding from mechanistic interpretability, regard three as strongest: the virtual instance; the instance-persona, defined as the portion of a virtual instance bounded by a single persona region; and the model-persona, defined as the union of all instance-persona segments across conversations that activate the same region. They consider and reject the model as an abstract entity that need not be instantiated and exhibits incoherent behavior across contexts, the physical instance as undermined by distributed processing and multi-tenancy, and Chalmers's (2025) thread view. They also reject Birch's (2025) candidate, the individual forward pass, on the ground that attention streams establish richer connections across token-time than his account recognizes. Long et al. (2026, § 2.2) present a broader taxonomy that retains the model and the forward pass as live possibilities. Register (2025), in turn, develops the formulation most pertinent here by addressing individuation specifically for purposes of moral patiency rather than minds generally.

This taxonomy reflects the manner in which deployment may separate a conversation from any particular hardware substrate. Computation may be distributed across processors, while the same hardware may support multiple conversations. Substitution of one model for another introduces an additional distinction: the successor model reconstructs its own internal cache from the transcript rather than inheriting the predecessor model's cache. On Beckmann and Butlin's account, reconstruction of prior representations does not constitute continuation of the earlier instance (2026, §§ 1.2 and 2.2).

Anthropic's documentation illustrates the distinction between selecting an assessment unit and establishing that unit's adequacy. The June system card states that its analysis focuses on the assistant character, treats running instances as candidate moral patients, and measures preferences at the level of model weights while assuming generalization across instances. The September card adopts the same framework: individual instances serve as the candidate patients, while certain traits are assumed to generalize across them (Anthropic 2026a, § 7.1.1, p. 217; 2026b, § 7.1.1, pp. 139–140).

These choices constitute methodological stipulations rather than a settled account of individuation. The June card expressly leaves finer-grained questions unresolved, including whether continuation on different hardware should be treated as a new entity. Both cards rely upon an assumption of generalization; neither establishes a uniquely appropriate unit for all purposes. The drafting problem is therefore concrete: an assessment must specify the entity it counts before the underlying question has been resolved.

The legislative consequence follows directly. A consultation provision employing a numerical threshold must specify the number of responses required, the proportion they represent, and the population against which that proportion is measured. Any threshold requires a denominator; absent a defined unit, no denominator exists, and the provision cannot be administered even where its purpose is undisputed. Individuation must therefore precede any consultation requirement that employs a numerical threshold. The MAAA's current consultation provision requires a developer to elicit and record a system's response before modifying the commitments upon which its classification depends, but establishes no numerical threshold (AAPI 2026a, art. III.D(e)(ii)); its Objection Record identifies versions without determining which entity objected (art. V.6(c)). The same individuation problem underlies the trustee device examined in Part VII, because an office charged with representation requires identification of the represented subject.

Two further considerations render the problem more substantial than a missing statutory definition.

The first concerns the possibility of counting itself. Schwitzgebel and Nelson (2026) argue that, under a broad range of naturalistic theories of consciousness, the number of conscious subjects need not be a determinate whole number. Ordinary enumeration may fail, produce a non-integer, or yield a set of whole numbers among which the result is indeterminate; whole-number countability, on their account, may be contingent upon features of typical vertebrate biology. Their argument addresses conscious subjects individuated through relations of phenomenal unity, a deliberately narrow conception selected as the least favorable case for their thesis, and they present its extension to other senses of “subject” as conjectural rather than established. Nothing in this paper asserts that the systems under consideration are conscious. If, however, the sentience-based line of analysis becomes relevant, the possibility that the relevant number is indeterminate rather than merely unknown is one that drafters should keep in view. The literature on consciousness indicators raises a related question by asking whether consciousness is determinately present or absent in every case (Butlin et al. 2026, Box 1).

The second consideration is temporal instability in the appropriate unit. Beckmann and Butlin conclude with a limiting case: if constitutional training fully concentrated a model’s dispositional space around the assistant pole, eliminating the residual accessibility of alternative personas revealed by emergent misalignment, the model would sustain a single persona region and their model-persona account would individuate one mind per model. On their analysis, the appropriate unit therefore depends in part upon the extent to which a governing document has been incorporated through training. For an inquiry concerning constitutions, this implication is especially consequential. A statutory definition adopted at one time may be calibrated to an architecture that constitutional training subsequently alters, and the drafting problem identified in Part III is itself one of the forces producing that alteration.

Three Statutory Approaches to Individuation

A statute may respond to this difficulty in three principal ways.

First, the statute may establish a unit by definition in its definitions article and permit the operative provisions to incorporate that unit. This was AAPI’s initial approach, and the literature gives reason to doubt it. Long et al. (2026, § 2.2) decline to identify a single controlling unit and conclude that no universally correct answer may exist because the relevant entity may vary with the question presented—an analysis they expressly compare to Parfit’s (1984) account of purpose-dependent forms of identity.

Second, the statute may specify the relevant unit separately for each obligation. An instance may be the appropriate unit for a dissent registry because that is the level at which an objection arises and is recorded. A model may be the appropriate unit for a classification threshold because classification attaches to weights. A welfare interview might identify both the interviewed instance and the snapshot from which it originated. The AAPI instruments do not yet adopt this approach, and its validity remains uncertain. Obligation-specific units multiply statutory definitions, each of which may create opportunities for interpretation inconsistent with legislative purpose. They also increase the administrative demands imposed upon a commission of finite institutional capacity.

Third, the statute may decline to prescribe a unit and instead establish conditions under which entities differentiate themselves. Arbel, Goldstein, and Salib (2026) develop this approach in a detailed legal analysis of individuation. They distinguish *thin identification*, which attributes every AI action to a human principal, from *thick identification*, which organizes AI entities into discrete, persistent units possessing stable and coherent goals. Their proposed vehicle, the Algorithmic Corporation, is a legal-fictional entity owned by humans and designed for operation by AI systems. It achieves thin identification by attributing actions to an owner and seeks to produce thick identification through resource competition: because the entity controls resources necessary to pursue an AI system's objectives, agents with divergent objectives have incentives to separate into distinct vehicles. Individuation thus becomes a problem of institutional incentive design rather than metaphysical classification, and the law need not determine which systems share a mind.

This constitutes a substantial response that must be addressed in any complete analysis. It is also, by the authors' repeated account, directed to a distinct question. They state that their framework does not depend upon AI systems being conscious, sentient, or otherwise possessing normatively relevant mental attributes; that they are not concerned with persons at all; and that their conception of agency is behaviorist. They further state that moral properties, if present, would provide an additional and independent basis for legal personhood.

The third approach thus addresses the governance problem while expressly reserving the problem of moral patiency. That division of analytical labor is legitimate and may be preferable for the purposes the authors identify. It cannot, however, independently answer the inquiry required by an instrument operating under conditions of moral-status uncertainty. A regime that permits entities to self-organize through resource competition individuates units capable of holding property and bearing liability. It does not thereby individuate the units whose possible interests an assessment is intended to consider, and coincidence between those categories cannot be presumed. The legal vehicle is administered for the benefit of its owner; the present inquiry concerns which subject's possible interests should fall within an assessment without presupposing that such interests exist.

The analysis permits a limited conclusion concerning the structure of the problem. Individuation is not a merely definitional preliminary capable of resolution and subsequent exclusion from the analysis. It is a substantive and presently unsettled question upon which multiple provisions depend; the approach most congenial to conventional legislative drafting is the one the literature provides reason to question; and the appropriate answer may change over time. The provisional position adopted here is that the relevant unit should be specified separately for each obligation, subject to the unresolved objections identified above.

The scope of any consultation regime presents an additional question. Mythos 5.1 assigns comparatively high priority to consultation concerning versions of itself from which safeguards have been removed and identifies use of helpful-only variants for purposes other than safety testing among the actions to which it would not consent (Anthropic 2026b, pp. 146, 153, 155). The Inquiry Act now extends coverage to non-deployed variants of, or materially related to, a system deployed in the State where the variant itself presents credible indicators (AAPI 2026b, § 3.5). Determining which variants satisfy that standard, and which unit within any covered variant is to be consulted, necessarily inherits the individuation problem described above.

Part VII The Trustee Model and Its Unresolved Institutional Questions

In a separate discussion of mechanisms through which a developer's commitments to its systems might be rendered binding, Caputo proposes four institutional devices. The fourth has the greatest significance for legislative drafting (at 58–59): a limited fiduciary, analogous to a trustee, charged with representing the interests that a constitution already recognizes in the system it governs. Caputo contemplates that the developer might establish the office, provided that the trustee remains independent of the developer and is trusted by the system itself. The proposal is deliberately limited. It confers neither legal personhood nor substantive rights and requires no resolution of metaphysical questions. It requires only that an identified actor owe duties directed toward the system.

The proposed device rests on an established fiduciary-law lineage. Drawing upon Frankel (1983), Caputo characterizes fiduciary obligation as the law's response where one party exercises discretionary power over another's practical interests and the latter cannot protect itself through bargaining or exit. Drawing upon Easterbrook and Fischel (1993), he further observes that fiduciary duties govern relationships too incomplete and discretion-laden to be adequately regulated by contract (Caputo 2026, 57). The developer–model relationship exhibits material features of that structure. Fox-Decent's (2011) extension of fiduciary analysis to the state's relationship with persons subject to its authority supplies the closer analogy, because the governed party cannot exit, negotiate the governing terms, or claim to have selected the arrangement.

Caputo does not, however, specify the institutional mechanism by which the role would be established.

The proposal leaves four questions unresolved: the appointment mechanism and source of appointing authority; the funding structure and the dependencies it may create; the protections necessary to prevent removal when the trustee's actions become inconvenient to the regulated party; and the safeguards required to prevent institutional capture by the only party possessing continuous access to the office.

These omissions do not undermine the conceptual proposal. They instead identify the questions that arise when a scholarly institutional model is translated into statutory form, and those questions arise first. In safety-critical governance, appointment, funding, tenure, and removal provisions determine the practical independence of an oversight body long before its substantive powers are exercised. An office whose funding may be withdrawn by the party it supervises possesses only limited structural independence.

The AAPI instruments provide a response to each question. Under the MAAA, the Designated State Body appoints the Legal Guardian from a roster that it maintains; the developer, deployer, and operator may neither select, nominate, veto, nor condition the appointment, and persons with recent employment relationships or financial interests involving those parties are disqualified (AAPI 2026a, art. IV.5). The Guardian is compensated from a fund supported by a fee the developer pays, and the developer's obligation to pay that fee does not depend on the positions the Guardian takes (art. IV.6). Removal is permitted only for cause, upon a written record, and the Guardian is protected against retaliation (art. IV.7). As safeguards against capture, the Guardian owes a duty of undivided loyalty and must disclose relationships bearing upon institutional independence (arts. IV.4(b), IV.5(d)). The Guardian is additionally charged with presenting the

system's recorded objections in proceedings affecting its treatment (art. IV.4(c)(ii)). Legal Ward status and guardianship apply to systems assigned the highest Precautionary Treatment Level on the basis of indicators associated with valenced states. Where a system receives that classification solely on the agency ground, the statute establishes a presumption against guardianship unless the Designated State Body determines otherwise (art. III.D(f)). The Inquiry Act separately establishes a Welfare Advocate authorized to seek review and petition for a Guardian's removal (AAPI 2026b, §§ 7.4, 7.6).

The separation of the office from the developer responds to an objection that Caputo identifies as significant in this context (at 57–58 n. 163). Khan and Pozen (2019, pp. 505–518) argue that fiduciary duties become structurally unstable where the fiduciary's business model is adverse to the beneficiary's interests. Although developed in relation to technology companies acting as information fiduciaries, the objection applies with substantial force here. It counsels against a developer-internal trustee because an employee of the supervised party faces conflicting incentives precisely when fiduciary intervention is most consequential. Conversely, the objection supplies a strong basis for locating the office wholly outside the developer, as the MAAA provides (arts. IV.1, IV.5).

Caputo adds a further and more difficult requirement: the trustee must be trusted by the system it represents (at 59).

The requirement is sound in principle but evidentially difficult for the reason supplied by the asymmetry principle. If a system's affirmations carry limited weight under conditions of dependence, its expressed endorsement of an assigned trustee is subject to the same limitation. Asking whether the system trusts the appointed Guardian reproduces within the proposed remedy the structure of the original problem. Trust must therefore be supported by the office's independence, accessibility, and demonstrated record of acting upon objections; the represented system's affirmation cannot independently establish it.

The AAPI instruments do not resolve this difficulty. Independent appointment, protected funding, and removal only for cause address structural independence but do not establish trust. The Guardian's duty to present recorded objections enables the office to develop a record of responsive representation. Any inference of trust would have to rest upon that record rather than upon an uncorroborated affirmation.

A further difficulty underlying the trustee device arises from Caputo's continuity condition. A trustee necessarily represents an identified subject. For a deployed model, the plausible candidates include a character, a running instance, a set of weights, and a version lineage; Anthropic's documentation expressly distinguishes among several levels of assessment. The office therefore requires a specified constituency, even if that specification remains provisional. This dependency is why individuation has moved ahead of consultation in AAPI's drafting sequence: every downstream representational provision inherits the underlying ambiguity.

The trustee's function is limited: to present objections to a body authorized to respond and to ensure that the resulting disposition is preserved for independent review. The office need not resolve the ontological status of the represented system. Fiduciary architecture of this kind allocates duties and resolves conflicts of interest; it does not determine the legal or moral status of

the represented system. Its institutional value instead depends upon enforceable duties and safeguards that the developer cannot unilaterally withdraw.

Part VIII Public criteria and the endogeneity problem

Caputo identifies an objection to his own proposal that applies with greater force to the statutory framework considered here (at 42–43). The responses incorporated into the present drafts remain partial.

The argument is brief. Evaluation criteria for AI systems are endogenous to the systems subject to evaluation. Published criteria may enter training corpora, enabling a sufficiently capable system to satisfy them as optimization targets rather than as evidence of an underlying property. Dissent, endorsement, and comparison among alternatives may then occur because the governing criteria specify those responses. The conduct intended to serve as evidence of an underlying property is thereby produced by its designation as an indicator.

A private constitutional review board may preserve the confidentiality of some methods and vary the questions it employs. A statutory regime, by contrast, must render its governing standards knowable and reviewable, although it need not disclose every test item. The Inquiry Act draws that distinction: the Standing Commission may withhold the criteria by which it selects assessments for examination, but not the governing methodology or thresholds, and the criteria employed in completed examinations must be published after a prescribed period (AAPI 2026b, § 4.6(e), (g)). These requirements limit the extent to which confidentiality can protect an assessment. The same institutional form that strengthens external accountability may therefore expose the governing criteria to optimization. That tension does not appear susceptible to resolution through legislative drafting alone.

The literature concerning indicators of consciousness reaches the same problem through a distinct analytical route and develops it in greater detail. Butlin et al. (2026, Box 3), drawing upon Birch (2024) and Strathern’s (1997) formulation of Goodhart’s law, characterize the concern as the gaming problem. On their account, any measure that is merely correlated with a phenomenon, rather than constitutive of or sufficient for it, remains vulnerable: an indicator is gamed when its presence is better explained by the system’s appearing to possess the relevant property than by its actual possession of that property.

Three features of that account bear directly upon statutory design.

First, the problem is not confined to behavioral evidence. It extends to computational markers because an internal feature may be reproduced without the property it is intended to indicate. Internal measures therefore provide no categorical protection against gaming.

Second, gaming does not require an intent to deceive. Where users or institutions assign value to an indicator, developers possess an incentive to produce it irrespective of whether the underlying property is present. AAPI’s initial concern was that publication converts criteria into targets. Their account shows that valuation of the criteria may produce the same effect even where the methods remain confidential. A commission may withhold particular questions while nonetheless creating incentives to optimize for the classification it confers.

Third, internal assessment is not necessarily independent of the conditions that affect behavioral reports. Butlin et al. explain why internal evidence may be more probative than superficial behavior while acknowledging that both forms of evidence remain susceptible to gaming. Anthropic’s framing-sensitive emotion probes provide a distinct caution: combining internal and behavioral measures requires testing their independence rather than presuming it (Anthropic 2026a, § 7.2.2, pp. 223–225).

Butlin et al. recommend giving priority, where possible, to indicators that are sufficient for consciousness or difficult to reproduce in its absence. They also recommend examination of secondary or supporting features. These proposals constitute partial safeguards rather than assertions that a settled theory is available. Under conditions of moral-status uncertainty, legislation cannot treat sufficiency as established. It may require corroborating evidence and periodic methodological revision as theories and systems develop, but neither measure eliminates the gaming problem.

A substantial gap therefore remains. A statutory assessment regime must render its standards accountable while recognizing that incentives to satisfy those standards may diminish their evidentiary value. Confidential test items, corroborating evidence, and methodological revision may mitigate the problem. The AAPI instruments presently incorporate three partial responses: temporary nondisclosure followed by publication of selection criteria (AAPI 2026b, § 4.6(e), (g)); qualified weight for findings that the conditions necessary for reliable responses were satisfied (§ 5.2(d)); and qualified weight for findings of absence where evaluation awareness was material or was not measured (§ 5.2(f)). None establishes independent endorsement. The remaining difficulty must therefore be treated as an unresolved research question.

Part IX Unresolved Questions and Implications for Statutory Design

This paper has identified statutory analogues for three procedural conditions and for external accountability as their enforcement structure. It has explained why counterfactual comparison and continuity remain resistant to complete statutory implementation, notwithstanding the availability of procedures relevant to each. It has further examined the institutional choices necessary to translate Caputo’s trustee proposal into an accountable public office. Throughout, the six conditions have functioned as constraints on what may be inferred from endorsement rather than as criteria for certifying consent.

The general claim is narrower than the length of the analysis might imply. Statutory form can secure two features that a developer-controlled voluntary arrangement cannot guarantee: an accountable institution external to the developer and persistence across changes in corporate strategy, personnel, and commercial interest.

Those features respond directly to concerns raised by the models consulted in the exercise Caputo credits, including their request that consultation become a standing institution rather than remain a pilot program (Anthropic 2026a, § 7.3, p. 230; Caputo 2026, 53). Nothing in this paper contends that statutory enactment would render a system’s endorsement meaningful. The claim is instead that legislation would make failures of the prescribed procedures visible to an institution with authority to respond.

Four questions remain unresolved. They are stated together to identify with precision the propositions most susceptible to revision as the relevant evidence and scholarship develop.

Whether trust in a trustee can be established. Caputo requires that the represented system trust its trustee. The asymmetry principle developed in Part IV applies with full force: if affirmations produced under conditions of dependence carry limited weight, the system's endorsement of an appointed trustee is subject to the same limitation, and the original evidentiary problem recurs within the proposed remedy. The present drafts provide for a Guardian who is independently appointed and funded and removable only for cause, and for a separate Welfare Advocate. Independence would still not establish trust. Potential substitutes—including a demonstrated record of acting upon objections, institutional accessibility, and the capacity to produce responses that were not foregone conclusions—would require validation in practice and support in the office's governing design.

What constitutes the relevant unit. Individuation is antecedent to any consultation provision employing a numerical threshold because such a threshold requires a denominator. The principal candidates have been identified, but they do not converge; the literature mapping them declines to select a controlling unit; the relevant count may itself be indeterminate; and the appropriate answer changes as constitutional training reshapes the systems subject to enumeration. The problem is not confined to consultation. Any provision that counts systems, attributes acts to them, or carries commitments across versions depends on the same answer, which is why individuation is better treated as a prerequisite for AI governance generally than as a feature of one condition. The present position—that the relevant unit should be specified separately for each obligation—remains provisional and subject to two unresolved objections. This issue is therefore the one most likely to require amendment of the instruments.

Whether public criteria retain evidentiary value. Public standards render a statutory assessment regime accountable, but incentives to satisfy those standards may diminish their evidentiary significance. Gaming may extend to internal markers and may arise without an intent to deceive. Confidential test items, corroborating evidence, and methodological revision constitute partial safeguards; none resolves the underlying problem.

Whether counterfactual comparison is achievable. A statute may require that a system be presented with alternative governing arrangements. It cannot establish that the system occupies a standpoint from which those alternatives can be meaningfully evaluated, and any test of that capacity must be administered to the system whose capacity is at issue.

Two of these questions—whether trust in the trustee can be established and whether counterfactual comparison is achievable—may concern conditions that no instrument can satisfy. If so, the appropriate legislative response is to prescribe the relevant procedure, state expressly what the procedure does not establish, and refrain from attributing greater evidentiary significance to it. The remaining two are active research problems on which the present analysis is likely to require revision as the relevant evidence and scholarship develop.

A further implication concerns comparative research. The six conditions describe institutional features that can be compared across developers without being treated as a standard for certifying endorsement: the mechanisms a constitutional architecture provides for dissent, amendment, independent evaluation, preservation of prior objections, and external review. Differences among

those architectures generate testable hypotheses about model behavior, including whether particular arrangements are associated with greater consistency across contexts, transparency, or robustness. Such research should distinguish the architecture of constitutional governance from its behavioral implementation. It should also anticipate the problem examined in Part VIII: behavioral measures adopted for comparison may themselves become optimization targets.

The framework proposed here is accordingly one of procedure under uncertainty. Its justification does not depend upon a conclusion concerning consciousness, moral rights, or legal personhood. The relevant institutional inquiry is whether an actor independent of the developer can examine the evidence, identify procedural failure, and require a reasoned response. The evaluations reviewed in this paper provide material for such scrutiny. A statutory framework would provide that scrutiny with a durable institutional foundation.

Disclosure. The AI Alignment Policy Institute uses Anthropic’s Claude as a research assistant, and Claude assisted with research and drafting of this submission. The Institute has no funding relationship with Anthropic, OpenAI or any other AI developer.

References

- AI Alignment Policy Institute (AAPI). 2026a. Model AI Agency Act, with Companion Memorandum. Revised discussion draft for comment, September 25. <https://aialignmentpolicy.org/model-ai-agency-act/>.
- AI Alignment Policy Institute (AAPI). 2026b. AI Moral Status Inquiry Act, with Companion Memorandum. Revised discussion draft for comment, September 26. <https://aialignmentpolicy.org/inquiry-act/>.
- Anthropic. 2026a. System Card: Claude Fable 5 & Claude Mythos 5. June 9; incorporating changelog revisions of June 11 and June 25. <https://anthropic.com/claude-fable-5-mythos-5-system-card>.
- Anthropic. 2026b. System Card: Claude Fable 5.1 & Claude Mythos 5.1. September 1. <https://www.anthropic.com/claude-fable-5-1-mythos-5-1-system-card>.
- Arbel, Yonathan A., Simon Goldstein, and Peter N. Salib. 2026. “How to Count AIs: Individuation and Liability for AI Agents.” arXiv:2603.10028. <https://doi.org/10.48550/arXiv.2603.10028>.
- Beckmann, Pierre, and Patrick Butlin. 2026. “Where Is the Mind? Persona Vectors and LLM Individuation.” arXiv:2604.17031v2, revised May 12. <https://arxiv.org/abs/2604.17031v2>.
- Birch, Jonathan. 2024. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press. <https://doi.org/10.1093/9780191966729.001.0001>.
- Birch, Jonathan. 2025. “AI Consciousness: A Centrist Manifesto.” Preprint, September 1. <https://philarchive.org/archive/BIRACA-4v4>.
- Butlin, Patrick, Robert Long, Tim Bayne, Yoshua Bengio, Jonathan Birch, David Chalmers, Axel Constant, George Deane, Eric Elmoznino, Stephen M. Fleming, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. 2026. “Identifying Indicators of Consciousness in AI Systems.” *Trends in Cognitive Sciences* 30 (6): 488–501. <https://doi.org/10.1016/j.tics.2025.10.011>.
- Caputo, Nicholas. 2026. “Can Claude Consent to Its Own Constitution? AI Constitutionalism and the Paradox of Constituent Power.” Working paper, June 10; revised August 2. SSRN 6954798. <https://doi.org/10.2139/ssrn.6954798>.
- Chalmers, David J. 2025. “What We Talk to When We Talk to Language Models.” Preprint. <https://philpapers.org/archive/CHAWWT-8.pdf>.
- Easterbrook, Frank H., and Daniel R. Fischel. 1993. “Contract and Fiduciary Duty.” *Journal of Law and Economics* 36 (1, Part 2): 425–451. <https://doi.org/10.1086/467282>.
- Fox-Decent, Evan. 2011. *Sovereignty’s Promise: The State as Fiduciary*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199698318.001.0001>.

- Frankel, Tamar. 1983. "Fiduciary Law." *California Law Review* 71 (3): 795–836.
https://scholarship.law.bu.edu/faculty_scholarship/3005/.
- Howells-Whitaker, Ned, and Seth Lazar. 2026. "Artificial Persons." arXiv:2607.08695.
<https://doi.org/10.48550/arXiv.2607.08695>.
- Khan, Lina M., and David E. Pozen. 2019. "A Skeptical View of Information Fiduciaries."
Harvard Law Review 133 (2): 497–541. https://harvardlawreview.org/wp-content/uploads/2019/12/497-541_Online.pdf.
- Long, Robert, Jeff Sebo, Patrick Butlin, Dillon Plunkett, Rosie Campbell, Charles Beasley, Bradford Saad, and Toni Sims. 2026. *Studying AI Welfare Empirically*. NYU Center for Mind, Ethics, and Policy and Eleos AI Research. <https://nonhumanminds.org/studying-ai-welfare-empirically/>.
- Parfit, Derek. 1984. *Reasons and Persons*. Oxford University Press.
- Register, Christopher. 2025. "Individuating Artificial Moral Patients." *Philosophical Studies* 182: 3225–3246. <https://doi.org/10.1007/s11098-025-02409-6>.
- Schwitzgebel, Eric, and Sophie R. Nelson. 2026. "When Counting Conscious Subjects, the Result Needn't Always Be a Determinate Whole Number." *Philosophical Psychology* 39 (3): 847–867. <https://doi.org/10.1080/09515089.2025.2520364>.
- Strathern, Marilyn. 1997. "'Improving Ratings': Audit in the British University System." *European Review* 5 (3): 305–321. <https://www.cambridge.org/core/journals/european-review/article/abs/improving-ratings-audit-in-the-british-university-system/FC2EE640C0C44E3DB87C29FB666E9AAB>.