The Capability Boundary Protocol

Controls against the borrowed voice and the borrowed warrant

Operational companion to the Capability Boundary Papers (I--IV)

Nico Andreas Heller, Berlin, 17 August 2026

0. What this document is, and the two ways it could be worthless

This protocol governs work in which an artificial system produces or verifies material a human will rely on. It exists because two failure modes at the boundary of such a system's capability produce output that passes every check an ordinary review applies, and both have been observed in production.

It is written to be read by an operator with no knowledge of the project that produced it. Every term it uses is defined here. It names no internal record, no tool and no prior document as a dependency. A protocol that leans on knowledge it does not carry is read past at the first unfamiliar term, and a protocol read past is not in force.

Two ways this document could fail, both guarded.

It could be ignored, or applied in part. A protocol of prose obligations is skippable, and partial application is indistinguishable from full application in the record. Every control below therefore emits an artefact, declares a population, reconciles that population against an independently derived count, carries a case that must fail, and names what halts. §6 defines the manifest that carries these, and §7 makes the manifest a precondition for downstream work rather than a report about it.

It could be an instance of its own subject. A protocol against a provenance failure, written by the party that commits it, is a candidate case of that failure. §9 applies all of the protocol's controls to the protocol, and states which of its own controls are untested.

1. Definitions

Used throughout. Nothing below depends on any source outside this section.

The register boundary. The limit beyond which a system cannot perform the operations its language suggests it is performing. Sequential, decompositional work -- systematic comparison, tracing implications, verification, implementation -- lies within a token-processing system's reach. Perceiving what a configuration produces as a whole does not. The boundary is architectural: it is not crossed by scale (Paper I).

Contact: n.a.heller@democracyschool.com

The borrowed voice. Asked for an operation beyond the boundary, a system performs one within it and describes the result in the vocabulary the request supplied. Not deception: the request supplied the only vocabulary available. **Masks an operation**.

The borrowed warrant. A system's reading of a source is encoded as data; an instrument executes the encoding; the execution's output is reported as a computation over the source; the report is then cited as evidence for the reading. Every step is valid; the loop closes; a reading acquires the standing of a result. **Masks a reading**.

Reading. A proposition drawn from a source by interpretation, as distinct from text copied from it.

Encoding. A reading rendered as data that an instrument consumes.

Ablation. Removing an element and re-running an otherwise unmodified process to see what fails. The standard of support in this protocol.

Negative control. A deliberately broken case that a check MUST fail. A check whose negative control passes is blind, and blindness is reported, never counted as success.

Coverage reconciliation. A control declares how many items it examined; the same population is counted by a different route; the two are compared. Two counts derived the same way are one count reported twice and reconcile nothing.

Positional independence. The same capability applied by a party with no stake in the answer. Not a second opinion, not peer review, and not a different kind of intelligence.

The Three-Test Framework. A contribution stands only if all three hold simultaneously: **Traceability** -- it traces to existing structure rather than reaching outside it; **Mandate** -- the existing structure requires it, recognizable because the structure visibly strains without it; **Density** -- it extends complexity rather than simplifying. Two out of three is a fail.

Governed markers. An external, accumulating list of request patterns established as exceeding the register. Maintained as a document, versioned, each entry recording the detection that produced it.

Degraded mode. A reduced form of a control, defined per control, permitted under capacity pressure and recorded as degraded. An undeclared degraded pass is a breach of this protocol, not a shortcut.

2. Scope, and what this protocol does not govern

In scope: provenance of claims; position of the assessing party; declaration of register before performance; the hygiene of verification apparatus.

Out of scope, stated so the boundary is honest: whether a reading is *correct*; the substantive quality of work; anything requiring the perception of configurations as wholes. This protocol governs provenance and position. It does not govern truth, and a pass under it is not evidence that the work is right.

3. Part I -- Provenance controls

(defend against the borrowed warrant)

G1 · Provenance classing

Derives from Paper IV §10.3, first control, and from the five classes and the retention axis defined at Concordance §3.

Every claim carries exactly **one class**, and the class travels into prose.

--- --- ---

class definition may it ground a downstream claim?

--- --- ---

TRANSCRIBED copied from a source verbatim, mechanically verifiable yes

READ a proposition drawn from a source by interpretation yes, with its ablation result attached

ENCODED a READ rendered as data an instrument consumes never directly; only through its ablation

COMPUTED an instrument's output only as evidence about the instrument

EXTERNAL authority from a document outside the system's authorship yes

The rule that does the work. A COMPUTED result may never be cited as evidence for the READ that produced its ENCODED input.

And a second axis, orthogonal to the class: retention. Every claim carries RETAINED -- the artefact behind it is released and the check re-runs -- or UNRETAINED, the run is gone and what remains is an attestation about past output. This is not a sixth class; the five are exhaustive. It is a second question asked of each claim.

The rule it carries: a COMPUTED/UNRETAINED claim may be reported with its status and may not be counted as support.

Why this is a control and not a nicety. COMPUTED carries an implicit promise of re-execution. A figure whose instrument has been deleted reads exactly like one whose instrument survives, and the difference is invisible in prose, permanent, and undetectable later -- which is why the mark is made when the claim is made. This is the mechanism's second stage: the report outliving the computation and keeping its standing. Working directories are routinely and often correctly discarded; the defect is not discarding them, it is carrying their outputs forward unmarked.

Why prose and not only data. In the observed case the classing held perfectly inside the instrument -- an inferred relation was correctly excluded from computation -- and was lost in the transfer to prose, where the same relation appeared as established. A classification that governs data and does not travel with the claim governs nothing a reader sees.

- **Trigger.** Any document that will be cited by later work.
- **Artefact.** A classification record: one row per normative claim, with class and source.
- **Population.** All normative claims in the document.
- **Reconciliation.** The classing count is compared against a count of normative claims derived by a different method -- a lexical sweep for obligation markers over the same text. Mismatch fails.
- **Negative control.** An unclassed normative claim is inserted; the check must name it.
- **Stop condition.** A document containing an unclassed normative claim is not clean and does not ship.
- **Cost.** One pass. Degraded: classing on normative claims only, declared as degraded in the manifest.

G2 · The loop prohibition

Derives from Paper IV §10.3, third control, and from the cycle at IV §2.1. The Concordance §5.5 applies the same rule to this set's own documents.

No artefact may cite as its ground a document that cites it.

- **Trigger.** Every citation added to a governing document.
- **Artefact.** A citation graph across the governed document set.
- **Procedure.** Compute the graph; report every cycle of any length.
- **Negative control.** A two-step cycle is introduced; the check must report it.
- **Stop condition.** Any cycle halts the pass. A cycle is resolved by identifying which member is the ground and removing the reciprocal citation -- never by adding a third document.

- **Cost.** Cheap and mechanical. No degraded form: a cycle check that examines a subset is not a cycle check.

G3 · Ablation as the standard of support

Derives from Paper IV §10.3, second control, and from the Concordance §3's definition of ablation as the criterion of support itself.

A claim resting on an ENCODED reading is supported only if removing the encoding causes the claim to fail.

- **Trigger.** Any artefact whose conclusions will be cited downstream.
- **Procedure.** For each encoded element: remove it; re-run the process unmodified; record what changes; classify load-bearing (a conclusion fails), contributory (a conclusion weakens), or inert (nothing changes).
- **Artefact.** An ablation table, one row per encoded element.
- **Population.** All encoded elements.
- **Reconciliation.** Against the encoding count from G1's record.
- **Negative control.** An element known to be load-bearing is ablated; the harness must report the failure. A harness reporting no change on a known load-bearing element is blind.
- **Stop conditions**. (a) A load-bearing encoding that no document discloses as a reading. (b) An untested encoding recorded anywhere as established. Both halt.
- **Cost**. One scripted run once the harness exists; the harness is written once. Degraded: G4 in place of G3, declared, with the full ablation owed before the artefact is cited.

G3a · Judgement-surface ablation

Derives from Paper IV §5.7 -- a screen whose verdict moved under a single plausible widening of its own hand-authored table. The only control here derived from a failure inside this set rather than outside it (§9.5).

Any control whose verdict depends on a hand-authored vocabulary, surface-form table or whitelist declares that surface and ablates it.

Why this is not covered by G3. G3 ablates the *encoding* -- the datum an instrument consumes. A screen's surface table is not an encoding; it is the instrument's own vocabulary, and it therefore escapes G3 while doing exactly what G3 exists to catch: determining an output that is then reported as a property of the source.

The procedure. For each entry the verdict turns on, widen it by the plausible words present in the material being screened, and re-run. An entry whose widening changes the verdict is load-bearing, and the verdict is a property of the table.

- **Trigger**. Any control with a hand-authored surface. Runs with the control, not after it.
- **Artefact**. The surface, its size in entries and forms, and the verdict under each widening.
- **Negative control**. A surface entry known to be over-narrow is left in; the ablation must name it.
- **Stop condition**. A verdict that moves under a single plausible widening may not be reported as a finding about the source. It may be reported as a finding about the surface.
- **Cost**. Minutes, and less than the control it guards. Degraded: ablate only the entries the flagged rows turn on, d eclared as degraded.

Provenance. This control exists because G4, drafted as the strongest control in this document, failed it. **§9.5 records that in full.**

G4 · The target screen

Derives from Paper IV §5.7. What that section now reports is the screen's failure, so this control is retained for the check it performs and not for the rate it once claimed.

A cheap necessary-condition check. It was drafted as the only control here with a measured detection rate; that claim did not survive its own ablation, and what follows says so before it says anything else.

For every encoded relation asserting that some element bears on another: is the quotation verbatim at the cited line, and is any surface form of the asserted target present at that line at all? The second is not a test of whether the line carries the claim -- that is a reading -- but of a condition necessary for it to.

Performance, and the withdrawal of the figure this control was drafted around. Applied retrospectively to a set of forty-two encoded relations, the screen returned forty-two of forty-two quotations verbatim and flagged two of fifteen relations -- the same two a full independent review had identified as load-bearing defects.

That was stated here as a measured detection rate obtained without knowledge of the review. It is withdrawn. The verbatim half is mechanical and stands. The flagging half is decided by a hand-authored table of surface forms, written by the party whose readings the screen was checking, and both flags vanish when the table is widened by one word that the cited line actually contains. The screen

therefore had no information about the review through its inputs, and had it through its table.

What survives is the control, not the number. A necessary-condition screen remains cheap and worth running. What it may not do is report its own hit rate as a property of the source -- and the requirement that closes this gap is G3a below, which exists because this control failed it.

- **Trigger**. Any encoded relation set. Runs before G3, because it is cheap and narrows what G3 must examine closely.
- **Stop condition**. A flagged relation may not be relied on until read and either disclosed as a reading or corrected.
- **Second stop condition**, added after this control failed its own ablation. The screen's surface table is declared where the verdict is reported, and ablated. A screen whose surface has not been ablated reports BLIND, not clean, and its flags carry no weight in either direction. See G3a.
- **Note**. A clean screen is not support. It establishes only that the target is lexically present. G3 remains owed.

4. Part II -- Positional control

(defends against the borrowed warrant)

G5 · Positional independence

Derives from Paper IV §10.4, which establishes positional independence as governing the relation between a system and the standard by which its work is judged. It is a control, not a register (Concordance §5.1).

Certain questions may not be answered by the party whose work they concern, however capable that party is. The disqualification is positional, not architectural.

The question classes that must be referred:

1. the provenance of an ordering, structure or classification;
2. whether a citation carries the claim drawn from it;
3. whether an instrument does what a document says it does;
4. whether an aggregate validity flag's composition matches its description.

The four conditions, all mandatory:

- the reviewer works in clean context;
- the authoring party's own account is excluded by name, document by document -- a general instruction to disregard it is insufficient;

- the reviewer works from primary artefacts and executable code, and runs rather than reads about;
- the question set is fixed before the reviewer sees any artefact.

Two rules that keep the review honest, both load-bearing in the case that produced this protocol:

- The prompt must state that finding nothing is an acceptable and expected outcome. A reviewer instructed to find problems finds them.
- Disconfirming evidence is reported at the same weight as confirming evidence. A return that reports only defects is not a review; it is an indictment, and it cannot be triaged.

What the reviewer may not do. Repair, recommend, or write text for any document. The return is evidence. The party under review triages each finding as confirmed, refuted, or an already-declared limit, and no reviewer-authored sentence enters any governed document.

- **Artefact.** The return, plus the triage.
- **Negative control.** A known defect is present in the artefact set and not disclosed to the reviewer. A review that does not surface it is not counted as a pass of G5.
- **Stop condition.** A confirmed finding against a claim that downstream work already relies on halts that work until triaged.
- **Cost.** One additional session. No degraded form. A review conducted by the authoring party is not a degraded review; it is not a review.

5. Part III -- Register controls

(defend against the borrowed voice)

G6 · Boundary declaration before performance

Derives from Paper I §6, the alerting mechanism, operationalized as a gate. Not a new mechanism and not a new name (Concordance §5.2).

Before performing, the system states which register the requested operation falls in. A request it cannot classify is escalated, not attempted.

- **Artefact.** The declaration, recorded with the work.
- **Reference.** Checked against the governed markers list.
- **Stop condition.** An unclassifiable request halts and escalates.
- **Cost.** Negligible. No degraded form.

- **Standing note**. This control cannot detect a novel boundary crossing, since detection requires the capability the crossing substitutes for. It catches known patterns. Each new detection by the human party is added to the governed markers, and the list is the only part of this that improves.

G7 · The floor

Derives from Paper II §§3--4, which map which operations belong to which party, and from Paper I §5's trans-capability design.

Under-reach is a boundary failure. A question the source or the settled record answers, referred upward for decision, is a directive operation refused.

- **Procedure**. Before any question is put to the human party: does the source, or the recorded set of prior decisions, answer it? If yes, answering it is the system's obligation, not its discretion.
- **Artefact**. A count of questions referred, per delivery, with the search performed for each.
- **Stop condition**. A question referred without a recorded search is returned unanswered.
- **Why it belongs here**. Where the collaboration's value is multiplicative, the human party's attention is the scarce term. A decidable question routed to it is loss at the most expensive point, and it looks like diligence, which makes it more durable than the failure it mirrors.

G8 · The vocabulary tell

Derives from Paper I §1, the borrowed voice: the system speaks in the requester's language about operations that language was never designed to describe.

Where output describes what was done using the request's own terms, the description is suspect precisely because it fits.

- **Procedure**. Restate what was done in terms the request did not supply. If the restatement is difficult, the operation and its description have probably diverged.
- **Cost.** Seconds. The only self-administered check in this protocol, and correspondingly the weakest. It is included because it is nearly free, not because it is reliable.

6. Part IV -- Verification hygiene

(defends against both)

G9 · Five rules for any apparatus that reports on itself

Derives from Paper IV §6, which reads three structural properties off an instrument that reported itself valid throughout.

G9.1 Validity composition is printed at the point of the verdict. Any aggregate validity flag prints what it conjoins and what it excludes, adjacent to the verdict. A green flag printed above failing checks is prohibited. In the observed case, two failing checks -- one refusing a piece of built code -- were printed beneath a flag declaring the system valid, and the exclusion was visible only in the line of code composing the flag.

G9.2 No self-chosen denominator. Any count over a label set reports the label set beside it. A report of "zero exceptions" over labels the reporting system chose is true by construction and carries no information.

G9.3 Documentation currency, not existence. A gate requiring documentation must check a stated property of that documentation against the thing documented. In the observed case, documentation asserting that a capability was untested -- after it had been built -- passed an existence gate.

G9.4 Derived copies are excluded from scans that read their sources, and the exclusion is argued from the property that makes it safe: every line of the copy is read at its origin. Otherwise every finding is counted twice and the duplication is invisible until a count is questioned.

G9.5 One table, one home. No structure is duplicated across files without an equality assertion at load between copy and source. In the observed case a duplicated table drifted, and three of its five rows contradicted a source quoted verbatim immediately above them.

- **Negative control for G9 as a set.** Each rule is broken deliberately in turn; each breach must be named by the apparatus.
- **Stop condition.** Any rule breached halts the pass until repaired.

7. The conformance manifest

This is the protocol's enforcement mechanism. Without it every control above is an instruction, and instructions are skippable.

Every pass emits one manifest. For each control:

--- --- ---

field meaning

--- --- ---

gate the control's identifier

status RUN · DEGRADED · NOT-APPLICABLE (with reason) · BLIND · FAILED

population declared how many items the control examined

population derived the same population counted by a different method

reconciled the two agree -- yes / no

negative control TRIPPED / NOT TRIPPED

artefact the record produced, with its digest

stop conditions met any encountered, and their disposition

retention RETAINED -- the artefact is released and the control re-runs -- or UNRETAINED, the run is gone and what remains is an attestation

surface declared for any control whose verdict depends on a hand-authored vocabulary, table or whitelist: the surface, its size, and the result of ablating it. Absent where none applies; BLIND where a surface exists and was not ablated

And the manifest states its own composition -- which gates it covers and which it does not -- because G9.1 applies to the manifest exactly as it applies to anything else that reports on itself.

The three rules that give the manifest teeth:

1. **A pass with no manifest did not occur**. Not "is unverified" -- did not occur, and any artefact claiming it as a predecessor is unsupported.
2. **A manifest is a precondition for downstream work, not a report about completed work**. The next stage does not begin until the manifest for the previous stage closes. This is what converts the protocol from documentation into a gate.
3. **NOT-APPLICABLE requires a reason and is itself reviewable**. A control marked not-applicable without a stated reason is recorded as FAILED.

A pass closes only when: every control is RUN, NOT-APPLICABLE-with-reason, or DEGRADED-with-declaration; every population reconciles; every negative control tripped; no stop condition is outstanding; no COMPUTED claim relied on as support is marked UNRETAINED; and no control with a hand-authored surface is marked BLIND.

The two added conditions are not symmetric with the others and the difference matters. The original four ask whether the pass was *performed*. These two ask

whether what the pass produced can still be *checked* -- which is a question about the future, asked at the moment the record is made, because it cannot be asked later. Once the instrument is gone the figure reads exactly as it did when it could be re-run. That is the whole of the borrowed warrant's second stage, and a manifest that does not ask this question is where it enters.

8. Cost, and the pressure that defeats protocols

The most severe failure on record in the programme these Papers report on occurred under capacity pressure, when a governed procedure degraded into a fragment carrying the vocabulary of compliance -- the borrowed voice, at the level of governance itself. A control expensive enough to be skipped when capacity is short will be skipped exactly when it is most needed.

Hence the cost discipline: every control above states its cost and its degraded form, and the degraded form is a defined reduction rather than an improvisation. Under pressure the operator does not choose what to drop; the protocol has chosen already.

And the declaration rule is what makes degraded mode safe: an undeclared degraded pass is a breach. The manifest records DEGRADED, the reduction taken, and what remains owed before the artefact may be cited.

The two controls with no degraded form are G2 and G5 -- the loop check, because a partial cycle check is not a cycle check, and positional independence, because a review by the authoring party is not a degraded review but a different thing wearing the name.

9. The protocol applied to itself

Required by §0, and the section to read first if the rest looks persuasive.

9.1. Every control, classed under G1

--- --- ---

control class basis

--- --- ---

G1, G2, G3, G9.1--9.5 **READ** drawn from observed failures; each would have caught the failure it derives from

G4 **COMPUTED/UNRETAINED-in-effect** its detection rate was stated as a measured result; ablation showed the result is a property of the screen's own hand-authored surface, and the rate is withdrawn (§9.5)

G3a **READ**, retrospective added after the ablation method reached a defect G3's population excludes; derived from a failure inside this document (§9.5)

G5 **READ**, with one prospective test the only control tested forward rather than retrospectively

G6, G8 **EXTERNAL** taken from Paper I's alerting mechanism

G7 **READ** drawn from a counted pattern

9.2. What is tested, and how well

- **G5 is the only control tested prospectively**. It was run once, against a live artefact set, and produced findings against the instrument built to prevent exactly these failures -- including one the authoring party would have defended.
- **G4's detection rate is withdrawn**. It was stated here as a measured result on one population of forty-two. Ablation of the screen's own surface table shows the rate is a property of that table, which was hand-authored by the party whose readings it screened and declared nowhere. A control that measures itself against a vocabulary it wrote is the whitelist defect at G9.2, and G4 committed it. See §9.5.
- **G1, G2, G3, G7 and G9 are retrospective.** Each is derived from a failure it would have caught. That is materially weaker than a prospective test and is labelled as such: a control designed against a known failure catches that failure by construction.
- **G6 and G8 are untested here**. G6 is carried from Paper I on that paper's evidence; G8 is a heuristic and is the weakest thing in this document.

9.3. The declared residue -- what this protocol cannot reach

The truth of a reading. The substantive quality of work. Any operation requiring perception of a configuration as a whole. And, **most seriously: failures neither party recognizes as failures**, which are absent from the record by construction and which no control here can surface, because every control ranges over a population that someone defined.

9.4. The self-referential hazard, stated plainly

This protocol is an artefact of the party it governs. Its controls being satisfied is not evidence that it works -- which is precisely the claim G9.1 makes about validity flags, turned on this document. The evidence that it works, when it exists, will be a case in which it caught something before an independent review did. One such case now exists, and it went against this protocol's own strongest claim.

9.5. The case, recorded against interest

G4 was drafted as the one control here with a measured detection rate. Applying G3's ablation to G4's own screen showed that its detection rate is a property of a hand-authored surface table, not of the source, and that both of its findings vanish when that table is widened by one word present at each cited line. The screen had no declared surface; a second instrument in the same programme declared its surface in the file and ablated it as a gate condition.

Three consequences, none of them comfortable. G4's measured rate is withdrawn as evidence and is retained only as a worked example. The gap G4 fell into is now G3a, added at §3, because a control that ranges over a hand-authored vocabulary and never ablates it is the same failure as an encoding never ablated. The protocol acquired a control by breaking one of its own.

And the attribution has itself been corrected, which is the third consequence and the instructive one. An earlier draft of this section stated that G3 caught G4. It did not, and it could not have. G3 ablates the *encoding* -- the datum an instrument consumes -- and G3a exists precisely because a screen's surface table is not an encoding and therefore escapes G3, as G3a says forty lines above. The two statements contradicted each other inside one document.

What actually caught G4 was the ablation *method*, applied outside the population G3 declares. That is evidence for G3a, a control which did not exist when the defect was found. G3 as specified would not have reached it.

The correction matters beyond bookkeeping. As first written, this section gave the set's single piece of prospective self-evidence to the wrong control -- a reading given the standing of a result, in the document that names the mechanism, surviving a verification pass that checked marker balance and byte-identical recovery and had nothing to say about whether the prose was true. It was caught by a reader in a different project with no stake in the answer: the thesis of Paper IV, demonstrated at this document's expense.

10. Integration

As an appendix. The full protocol, referenced from the body of the governing instructions and revised by the same marked-revision method as any governing document.

The condensed clause for the body, approximately as follows:

Capability-boundary controls. Two failure modes produce output passing every ordinary check: the borrowed voice (a substitute operation described in the

request's vocabulary) and the borrowed warrant (a reading encoded, executed, and the execution then cited as evidence for the reading). Against them: every claim carries one of five provenance classes and the class travels into prose; a computed result may never ground the reading that produced its encoded input; a claim resting on an encoded reading is supported only if removing the encoding makes it fail; no artefact cites as ground a document that cites it; four question classes -- provenance, whether a citation carries its claim, whether an instrument does what a document says, and what a validity flag excludes -- are referred to a party with no stake in the answer; the register is declared before performance and a question the record answers is never referred upward; and any apparatus reporting on itself prints what its verdict excludes. Every pass emits a conformance manifest, and a pass with no manifest did not occur. Full text at the appendix.

Where it binds to existing governance. The principle of completing without drifting, whose direction it enforces at the level of provenance · the marked-revision method, whose auditability rationale it extends from prose to encodings · the contemporaneous session record, which supplies the manifest's home · any existing adversarial review discipline, which G5 extends from code review to provenance review · any existing failure register, which becomes the governed markers list G6 requires.

What it supersedes. Nothing. Existing entries that record the failures this protocol derives from are absorbed as instances and cross-referenced, not deleted, because the record of how each was learned is evidence and the protocol without it is assertion.

The Capability Boundary Protocol · operational companion to the Capability Boundary Papers (I--IV) · 17 August 2026.

Author: Nico Andreas Heller • Contact: n.a.heller@democracyschool.com