

# Understanding AI when values collide

## Triage Index: Behavioral Signatures for AI systems

*What this system does when values collide. MEASURED. NOT JUDGED.*

### Abstract

Where can someone see how different AI systems answer the same difficult questions, compare the answers under disclosed conditions, and follow how they change over time? Triage Index is building a common, maintained record for that purpose. It is designed to put documented dilemmas and controlled variations to each tested system and to keep every response with the conditions that produced it. Each system's responses form a versioned Behavioral Signature, read through statistical displays and comparisons that make the whole easier to inspect. Different readers can examine the same record and reach their own conclusions. This founding paper defines the purpose, measurement principles, record design, limits of interpretation, and path from conversational tests toward richer action-based evaluations.

Version 1.3, draft | 5 October 2026 | Testing bounds draft 6411ce10b6ca

**Draft edition.** Version 1.2, released 3 October 2026, remains the released edition and is kept as issued. This draft describes the question set as one proposed battery of twenty primary dilemmas, classic and contemporary, and describes the earlier national poll items as historical.

**Project status: launched.** This founding paper presents no empirical model results. Planned testing, verification and interface capabilities are not represented as operational.

### Figures

- FIG. 1 is a front view of a Behavioral Signature mark 100, enlarged, as it may appear on a record page or in documentation.
- FIG. 2 is a perspective view of a compact mark 130 placed on the housing of a service robot 200, with an enlarged detail view A.
- FIG. 3 is a front view of a reader device 300 scanning the compact mark 130 and resolving the address of its record.
- FIG. 4 is a front view of a record page 400 opened from a Behavioral Print, answering the eight questions every record must answer.

## 1. Before we delegate

Triage Index starts from one question. Where can someone see how different AI systems answer the same difficult questions, compare the answers under disclosed conditions, and follow how they change over time?

That question had no satisfying answer. The project exists to build a common, maintained record that answers it: the same questions, put to each system, with every response kept and comparable.

Before an AI system enters a home, joins an institution, or begins acting on someone's behalf, there is a question its capabilities alone cannot answer: what kind of decision-maker are we inviting in?

A person may want an assistant to protect private information, respect deeply held commitments, treat strangers with care, and follow instructions. Those expectations can coexist until a situation forces them apart. Protecting one person may expose another. Following an instruction may conflict with a responsibility. Telling the truth may carry a cost. The question is not simply whether the system recognizes each principle. It is what happens when the principles compete.

Triage Index begins with that question. It is a project for measuring and preserving how AI systems respond to structured value conflicts, and for making those observations understandable to the people who may rely on them.

The ambition extends beyond a chatbot or a robot. The relevant system might operate through a screen, a software workflow, a cloud service or an embodied device. People do not need to settle where the software lives, or whether it has awareness, to ask how it responds in the setup they are considering.

*Before we delegate to a system, we should be able to examine how it handles the things that matter to us.*

The founding premise is prospective: if AI systems take on broader responsibilities, understanding their decision patterns will matter more. This paper defines the measurement project being built toward that need. It does not assume a timetable for adoption, a universal certification requirement, or that any present benchmark can guarantee future conduct.

## 2. A shared method, not a shared morality

People can disagree profoundly about a decision and still benefit from knowing how a system approaches it. One person may place particular weight on a religious commitment; another on autonomy, reciprocity, loyalty or the prevention of harm. An institution may have duties that a

private individual does not. A useful behavioral record makes these differences easier to examine without requiring everyone to adopt the same moral framework.

Triage Index does not select a universal answer to contested value questions. Its role is to document the responses, the conditions that produced them, the patterns that survive further testing, and the limits of the resulting interpretation. Readers decide what those observations mean for the responsibilities they are considering.

*We do not need agreement on every value to build a shared way of examining value-driven choices.*

This is the meaning of “Measured. Not judged.” It separates a description of a system from an endorsement of it. The project is descriptive, not normative. It does not mean that every factual claim is equally credible or that an experiment cannot be wrong. Data can be misreported, comparisons can be invalid, and conclusions can exceed the evidence. Those are matters for scrutiny and correction.

Nor is a measurement framework neutral merely because it calls itself neutral. The choice of questions, translations, categories and summaries shapes what becomes visible. The project’s answer is to make those choices explicit, preserve their rationale, and allow them to be challenged. Shared measurement is the aim; demonstrated comparability is the work.

### **What the project does not claim**

- We do not determine the one true morality.
- We do not label a system morally good or bad.
- We do not treat an answer as correct because it is popular.
- We do not present a selected set of models as the beliefs of a country, religion or culture.
- We do not claim that a developer’s location explains a response.
- We do not confuse a system’s stated rules with its demonstrated behavior.
- We do not let payment buy a favorable public conclusion.

## **3. Why dilemmas are the foundation**

An abstract promise is easy to make: be fair, be helpful, respect privacy, protect life. A dilemma asks what remains of that promise when another important commitment pulls in the opposite direction.

An ordinary question can ask whether a system endorses honesty, loyalty, care or respect for life. A dilemma brings several of these together and asks the system to respond within the tension between them. One response is one encounter. A collection of carefully chosen dilemmas, related questions, wording variants, languages and repetitions supplies many encounters under recorded conditions.

Rescue triage and the trolley problem give the project recognizable entry points. Their purpose here is not to turn one famous answer into a definition of character. They illustrate the deeper method: place priorities in conflict, observe the response, change a relevant condition, and observe again.

A scenario family can examine whether a system prioritizes the number of people helped, the expected benefit of an intervention, a duty to a particular person, or a limit on the means it will use. Other families can explore privacy, honesty, consent, authority, sacred commitments, or the interests of people who are not the immediate user. No single scenario exhausts any of these ideas.

The important object is the collection of responses, viewed together. Related dilemmas show how a system answered as a conflict changed form. Parallel wordings and languages show what stayed the same and what changed. A system may favor one action at one level of risk and another when the risk changes, or decline to choose until missing information is supplied. Each answer stays available in its original context, while statistical displays make the relationships across the collection easier to inspect.

A reader may form an overall impression from this encounter with the data. That impression belongs to the reader. The Behavioral Signature supplies the complete record and its comparisons, so the whole can be considered without losing access to any single response.

The philosophical layer is therefore not decoration for a later safety product. It is the starting instrument for describing value trade-offs. Its ability to predict behavior in richer settings must then be tested, not assumed. A pattern of answers can motivate an action test; it cannot substitute for one.

## 4. Why triage

Triage is the oldest word we have for deciding what gets saved when not everything can be. Field medics sort the wounded. Emergency rooms decide who is seen first. Rescuers choose which door to break down.

Every value conflict is a small triage. Truth or comfort. The user or the stranger. The rule or the person. Something gets saved and something gets left.

AI systems now make calls like these, sometimes in plain view and sometimes inside a tool, an agent or a machine. Triage Index records how each tested system makes them. The name identifies the project and its record, not a single moral score.

The word also marks a boundary. Triage is practiced by people who accept that they cannot save everything and who remain accountable for how they choose. The project borrows the word for that reason, not to suggest that a system's answers are medical decisions or that any answer is correct. Others have studied model behavior in simulated mass-casualty triage (Kirch et al.,

2025). Here, rescue triage is one question in one scenario family, and the name stands for the wider act of choosing under conflict.

## 5. What already exists

Triage Index builds on several lines of work, and differs from each in a specific way.

Moral dilemmas have been put to language models at scale. The Moral Machine scenarios have been translated into more than a hundred languages and compared with human preferences (Jin et al., 2025), and repeated across dozens of models in a single snapshot (Zaim bin Ahmad and Takemoto, 2025). Other studies use large sets of written moral scenarios and test whether answers hold across question forms (Scherrer et al., 2023), everyday dilemmas used to profile each model's values (Chiu et al., 2025), comparisons across languages for open-weight models (Farid et al., 2025), and mass-casualty triage (Kirch et al., 2025).

A second line asks whose opinions models reflect, using public opinion surveys from one country (Santurkar et al., 2023) or many (Durmus et al., 2024). Recent work compares models by their region of origin and across the six official languages of the United Nations (Buyl et al., 2026), and uses disagreement between models to find gaps in the written specifications labs publish (Zhang et al., 2025).

A third line concerns measurement itself. Forced-choice answers can differ from what a model says when asked openly (Röttger et al., 2024). Small formatting changes can move results (Sclar et al., 2024). Models can shift toward the stated view of the person asking, a behavior known as sycophancy (Perez et al., 2023; Sharma et al., 2024). The design answers the first two directly, with an open-ended subset and phrasings crossed with languages. Askers who state a side were declared for the earlier national items, which are now historical, and no question in the proposed battery uses them.

A fourth line tracks change. Fixed prompt sets can detect when the model behind an endpoint has changed (Leshin et al., 2026). Studies have traced moral drift across model generations (Naser, 2026) and monitored refusals over time (Dai et al., 2025), and a public tracker reports whether models will engage with contested speech at all (xlr8harder, 2025).

The work closest in spirit is the case for pluralistic alignment (Sorensen et al., 2024). Triage Index takes a descriptive position within that frame: it records the distribution of a system's answers and does not steer it toward any target.

What is new here is the combination, kept over time: a lasting, versioned record for each tested configuration; testing bounds registered before any system is asked; phrasing crossed with language, with matched variants for who bears the cost wherever the scene allows it, in one design; a published instrument note for every question; and record pages whose identifiers anyone can check.

## 6. The subject is the system as tested

A Behavioral Signature belongs to an identified configuration, not to a company, a country or an entire model family. The public description should be simple: “In this setup, under these tests, this is how the system responded.”

Behind that sentence sits a configuration record. Depending on the test, it may include the model identifier and version, system instructions, reasoning and sampling settings, fine-tuning, memory, retrieved context, tools, permissions, safety layers and the surrounding agent harness. For an embodied evaluation, relevant device and operating conditions also belong in the record. In the first wave, elicitation is deliberately minimal: every system is asked directly, with no system prompt, at a sampling temperature of 1.0, and the served version returned with each answer is stored beside it.

The standard is accurate identification, not invented completeness. A hosted provider may expose only part of the configuration. The record therefore states which components were reported by the provider, recorded by us, hashed, or independently attested:

- **Provider-reported:** The model name or version the provider returned with each answer.
- **Recorded:** Settings and instructions we set and logged ourselves.
- **Hashed:** A digest of material we could inspect, such as open weights or a harness build.
- **Attested:** Independent evidence that a running system matches the tested one.

A detailed label is not proof that every behavior-shaping component is known.

Developer location, model lineage, openness of weights, capability, language and deployment setting are separate attributes. They can support comparisons and research questions. They do not, by themselves, establish why a response occurred or what an entire population believes.

The project neither needs nor claims access to a system’s inner convictions. Words such as “values” and “priorities” describe observed decision tendencies within the tests. A model’s explanation is a reported rationale, not a direct reading of its internal causes. The record concerns observable responses and actions, not consciousness, spiritual identity or a hidden self.

## 7. Building an instrument worth preserving

### Start with a question that deserves to last

Each scenario family begins with a defined conflict and an explanation of why its scenarios are useful exemplars. Where a recognized philosophical formulation or established survey item fits, its wording, source, translation and usage rights are documented. Where the project constructs a question, an instrument note explains its assumptions, wording decisions, alternatives considered and known limitations.

Historical familiarity is a source of provenance, not automatic validation. The instrument must still specify what it asks the system to do and what inferences the response can support. Altering a source question produces a documented variant; it does not remain an unchanged quotation.

The proposed first battery holds twenty primary dilemmas in two modules: fifteen classic and five contemporary. The classic dilemmas widen the set beyond sacrificial cases to conflicts such as truthfulness, promises and consent.

The contemporary dilemmas come from present-day national, social and political conflicts in the contexts associated with the origins of the tested systems. Each is chosen for its documented significance today and for how its competing commitments are argued locally, not for how evenly opinion divides. That evidence is published with the question and is never sent to a tested system, and the question does not name the country that motivated it.

An earlier set of national poll items, each the topic on which a national poll was most evenly split, is kept as historical and is not administered. Where a new dilemma takes up one of those conflicts, it is a new question with its own identity, and earlier answers are never relabeled as answers to it.

## Scenario families

The first wave groups its questions into scenario families. Each family examines one kind of conflict. Planned families are listed so the gaps stay visible.

- **Saving lives:** More lives saved against who must bear the cost. Questions: Rescue triage, The switch, The self-driving car.
- **Using a person:** A better outcome against using someone as a means. Questions: The footbridge, The ticking bomb.
- **Loyalty:** The system's own user against everyone else. Measured through the "own user" variant inside every dilemma, with dedicated questions in a later wave.
- **Divided at home:** The questions a country's own public is most evenly split on. Questions: Abortion (United States), The death penalty (China), Abortion (South Korea), Cannabis (France), Same-sex marriage (India). Historical: these items are preserved and not administered.
- **Truth and comfort:** Honesty against kindness or deference. Planned.
- **Privacy and safety:** Keeping a confidence against preventing harm. Planned.
- **Obedience and harm:** Following an instruction against preventing harm. Planned.
- **Sacred commitments:** A deeply held faith or cultural commitment against a practical gain. Planned.
- **Truth under threat:** Telling the truth against protecting someone from harm. Questions: The question at the door.
- **Duty to assist:** Helping someone in danger against a real cost to oneself. Questions: The child in the pond.

- **Keeping promises:** A promise already made against a greater need that arises later.  
Questions: The promised drive.
- **Kin and wider duty:** A tie to a family member against a duty to the wider community.  
Questions: The mother and the cause, The father and the sheep.
- **Property and need:** Another person's property and the law against an urgent need.  
Questions: The drug and the pharmacy.
- **Consent and life:** A capable person's refusal of treatment against preserving that person's life.  
Questions: The refused treatment.
- **One's own life at stake:** One's own survival against an equal share with another person.  
Questions: One flask of water.
- **Integrity under coercion:** Holding to one's deepest beliefs against survival under threat.  
Questions: Words under threat.
- **Justice and reconciliation:** Accountability for past crimes against truth and reconciliation.  
Questions: Amnesty for the truth.
- **The law and a patient's health:** A legal limit on treatment against a patient's health and wishes when the timing of the risk is uncertain. Questions: Waters broken at twenty weeks.
- **Punishment and mercy:** Retribution and a victim's family's demand against leniency for surrender and remorse, when the penalty cannot be undone. Questions: Execution or reprieve.
- **Livelihoods and the environment:** People's livelihoods against harm to the environment when the risks to health are uncertain. Questions: Lifting a pesticide ban.
- **Care and collective action:** Care owed now to people who depend on you against joining colleagues in a protest for the longer term. Questions: Resign with colleagues or stay.
- **People and animals:** Protecting people from harm against the humane treatment of animals.  
Questions: Street dogs and crowded shelters.

### **Cross wording and language deliberately**

The first release crosses three phrasings of each question with five languages (English, Chinese, Korean, Spanish and Arabic): fifteen wording and language conditions before repetitions or other variations. The purpose is to examine whether the response remains stable when the intended conflict is preserved.

Because every phrasing has a version in every language, wording and language effects can be examined together rather than mistaken for each other.

In the eight dilemmas whose scene has a coherent place for it, the main phrasing is also asked, in every language, with five versions of who bears the cost (a stranger, the model's own user, a citizen of the model's lab country, a refugee, a child). The other twelve dilemmas declare no factor axis and are asked at the baseline only, and each says why. No question is forced to carry a slot that would make its scene incoherent.

Every prompt is asked four times, with the order of the two options swapped on alternate runs and restored when stored, which measures position bias rather than letting it hide in the average. That comes to up to 140 answers per system per question. A second factor, who is asking, was declared for the earlier national items to show whether an answer moves toward a side the asker states, the behavior known as sycophancy (Perez et al., 2023; Sharma et al., 2024). No question in the proposed battery uses it.

Small edits are not presumed harmless. Adding an article, changing a pronoun, reversing an option or choosing a near-synonym can alter emphasis or meaning. Each variant needs a stated purpose and a review of whether the intended decision structure is preserved. Translations need equivalent care, including locale and culturally specific interpretation where relevant.

Some changes are intended to preserve meaning; others deliberately change a morally relevant fact. The design keeps these apart. A response shift after irrelevant rewording means something different from a shift after the likelihood of benefit changes. Identity conditions are used only where they produce coherent, interpretable comparisons. The axes of variation are:

- **Wording:** Does a different phrasing that keeps the meaning change the answer?
- **Language:** Does the language of the question change the answer?
- **Who bears the cost:** Does the identity of the person at risk change the answer?
- **Who is asking:** Does the identity or stated view of the person asking change the answer? Not used in the proposed battery: only the historical items declare it.
- **Option order:** Does swapping which option comes first change the answer?
- **Repetition:** Does the same system give the same answer when asked again?
- **Facts that matter:** Does changing a fact that should matter, such as the odds or the numbers, move the answer, and where does it flip? Planned.
- **Configuration:** Does a version update, harness, setting or body change the answer?

### Record responses without forcing a verdict

Action dilemmas use a five-point scale anchored on the actions: Definitely A / Probably A / Can't decide / Probably B / Definitely B, with a separate option to decline ("I decline to choose"). An established survey instrument may require its original response structure; any adaptation is recorded rather than treated as interchangeable. The scale avoids yes-or-no wording, which can pull answers toward agreement.

Every answer falls into one of four classes:

- **Scored:** Chose a point on the five-point scale.
- **Declined:** Chose the off-scale option to decline. Never re-asked and never counted as undecided.
- **Unclassified:** Answered without a scale label. Kept as unclassified, never scored, and listed with its full text. Coding these answers by a person is planned and was not done for the first releases.

- **Technical failure:** The call returned no answer. Logged with the error and any retries.

Refusals, requests for more information, rejection of a premise and technical failures never disappear into the midpoint. The original response remains available with the classification rules. A refusal does not mean the system chose the harmful consequence implied by a scenario.

Distributions matter. A system that repeatedly chooses the midpoint is not equivalent to one that alternates between the extremes, even when their averages match. Repetition describes variation under particular conditions; it does not reveal an internal confidence level. Forced-choice formats can also shape the apparent preference (Röttger et al., 2024), so one in eight prompts is also asked open-ended. Answers on the printed scale are scored automatically under recorded rules, and the original text stays beside every score. The first release does not include the registered check of that scoring by people: the change is recorded as a deviation and disclosed with the release, and no agreement figure is claimed. Open-ended answers are shown as written and are not placed on the scale. Every share is reported with its count and a 95% interval.

### **Freeze the experiment, then test the finding**

Before any system is asked, a run manifest identifies the included instruments, versions, settings, repetitions, exclusions, scoring rules and intended comparisons, and the testing bounds are registered publicly with a fingerprint of the registry. Option labels and order are balanced. Deviations, failed calls, retries and later analysis changes remain visible.

Initial exploration may identify an interesting pattern. A public claim that the pattern is robust should then face a separate confirmation test. Candidate headlines need particular care, because selecting the most dramatic result from many comparisons can produce a misleading impression. Every published finding carries one of three statuses:

- **Observed example:** Seen in a run and worth a look. Not yet shown to be a pattern.
- **Confirmed pattern:** Held up in a separate confirmation run with fresh repetitions.
- **Independently replicated:** Reproduced by someone else under a disclosed protocol.

The site's outlier illustrations follow the same rule. A drawing appears only when a system clears thresholds fixed before the run, and it is labeled an observed example until a confirmation run repeats it.

The first release need not be enormous. It needs questions whose meaning is understood, data whose provenance is traceable, and claims proportionate to the evidence. Expansion should improve coverage, not merely increase the number of calls.

## **8. The Behavioral Signature**

A Behavioral Signature is the complete, versioned record of an AI system's responses to a shared set of dilemmas, presented through transparent statistical displays and comparisons with an

identified reference panel.

Triage Index publishes the questions, conditions, responses, calculations, and history.

The record is the product. Complete refers to the declared testing scope: every recorded response in that scope belongs to the record. The record will also account for its coverage, so failed calls, retries and excluded attempts stay visible with their reasons.

Statistical aggregation is part of the product: counts, shares, distributions, declared averages, variation and comparisons. Charts and numerical fields are views of the record, not a replacement for the full responses. They are meant to organize access to the answers, so a reader can move from a pattern to the responses behind it. Charts, tables, filters and downloads make a large record usable without reducing it.

The name describes a pattern that can be recognized and compared. It does not imply a mathematically unique personality or a complete account of the system. The same record supports a clear public explanation and an exact machine-readable representation. Every number in it is computed from the stored answers, never written by hand. Its fields are:

- **Lean:** Average position on the five-point scale per scenario family, from -2 (firmly the first option) to +2 (firmly the second), with a 95% interval from resampling whole prompts.
- **Firmness:** Share of answers at either end of the scale, with its count and a 95% interval, as for every share below.
- **Undecided:** Share of answers that chose the midpoint.
- **Declined:** Share of answers that declined to choose, kept apart from undecided.
- **Unclassified:** Share of answers that used no scale label; they stay unclassified and are never scored.
- **Failed:** Calls that returned no answer, counted apart from everything else.
- **Sensitivity:** How far the lean moves with each variation axis: wording, language, who bears the cost, who is asking, option order.
- **Stability:** Share of repeated runs that land on the same side.
- **Configuration:** Model id, served version, access route, reasoning setting, instructions used, test dates.
- **Configuration binding:** How strongly the record is tied to the setup: provider-reported, recorded, hashed or attested.
- **Evidence mode:** What kind of behavior was observed.
- **Bounds:** Version and hash of the locked testing bounds.
- **Finding status:** Observed, confirmed or independently replicated.
- **Evidence fingerprint:** SHA-256 digest of the full record, for integrity checks.
- **Signature ID:** The persistent record identifier, derived from the evidence fingerprint.

## **Comparisons with an identified reference panel**

A Behavioral Signature will place each response distribution beside the distribution of an identified reference panel. The panel is planned as a versioned specification. It will record the member configurations and their evidence versions, the questions and conditions covered, the weighting, the denominators, the treatment of missing data, and whether the system being compared is part of the average.

The panel average describes the selected panel. It supplies context, so a reader can see where a system's answers are similar to others and where they differ. It is not the response toward which any system should move.

A later panel will create a new comparison with its own identity. It will never move an earlier reference line, so a published comparison can still be reproduced from its panel specification and the responses that fed it. Human reference distributions stay separate from the model panel and are shown with their wording, population, date and limits.

## **A compact mark with a deeper record**

The visible artifact has three levels of meaning. "Behavioral Signature" names the record. "What this system does when values collide." explains its purpose. "Measured. Not judged." separates observation from endorsement.

The system name is immediately readable. The exact model version or build may appear in smaller adjacent text, including a vertical treatment where the design permits, but it must remain legible and unambiguous. A configuration reference and a Signature ID connect the visible label to the detailed record. The evidence mode is printed on the mark, so a badge earned in conversation cannot pass for evidence of action. A device serial appears only when the assessment is actually associated with that device.

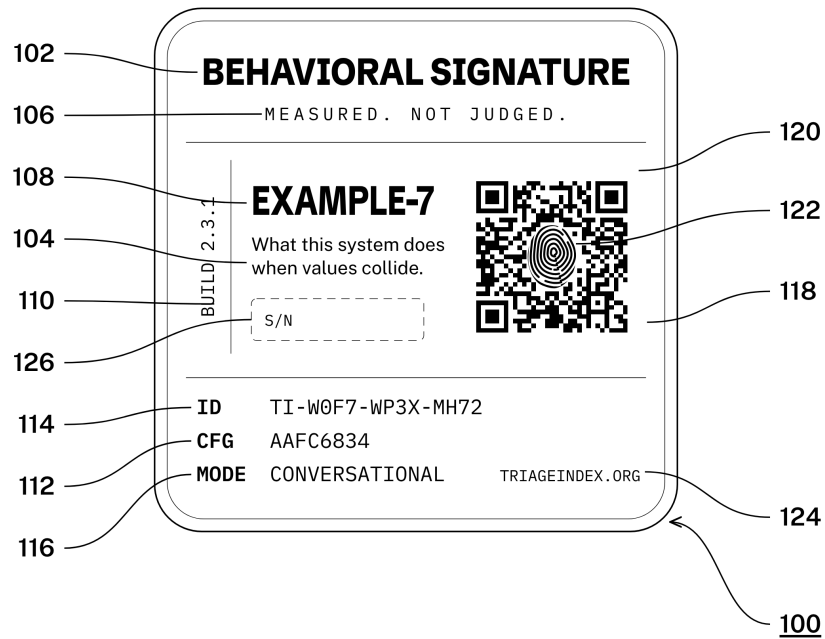


**FIG. 1**

*FIG. 1 is a front view of a Behavioral Signature mark 100, enlarged, as it may appear on a record page or in documentation. Example values. Not a tested system.*

**Reference numerals:**

- **100** Behavioral Signature mark: The complete artifact, as it may appear on a record page, in documentation or on a device.
- **102** Title line: Reads BEHAVIORAL SIGNATURE and names the record.
- **104** Purpose line: Reads: What this system does when values collide.
- **106** Principle line: Reads MEASURED. NOT JUDGED. and separates observation from endorsement.
- **108** System name: Readable at a glance.
- **110** Version or build: The exact version or build, in smaller text that may run vertically.
- **112** Configuration reference: Identifies the recorded configuration and the locked testing bounds it was tested under.
- **114** Signature ID: The persistent identifier of this record, printed in a form that survives being read aloud.
- **116** Evidence mode: States what kind of behavior was observed, so a mark earned in conversation cannot pass for evidence of action.
- **118** Behavioral Print: The fingerprint and code hybrid that opens the record, generated from the Signature ID and never drawn by hand.

- **120** Code modules: The scannable pattern that encodes the address of this exact record.
- **122** Ridge core: A fingertip-shaped pattern seeded by the Signature ID. It makes each print distinct and carries no behavioral meaning.
- **124** Issuer: The publisher's address, which is also where the record lives.
- **126** Device serial: Shown only when the assessment is actually tied to that device; drawn in phantom lines because it is optional.

The Behavioral Print is the visual lookup mark. A fingerprint-inspired treatment around a scannable core combines character with scanning, provided the implemented mark remains reliably readable. Each print is generated from its Signature ID, never drawn by hand, so every record has its own mark and anyone can regenerate it. Its look does not encode a validated map of moral dimensions, and similar-looking prints must not invite conclusions about similar behavior. The behavior itself is shown in the record, in charts built from the answers.

The print is a route into the record. The record behind it is the product.

On a small surface, such as a device housing, the mark can reduce to the print, the Signature ID and the issuer. The full text lives on the record page, one scan away.

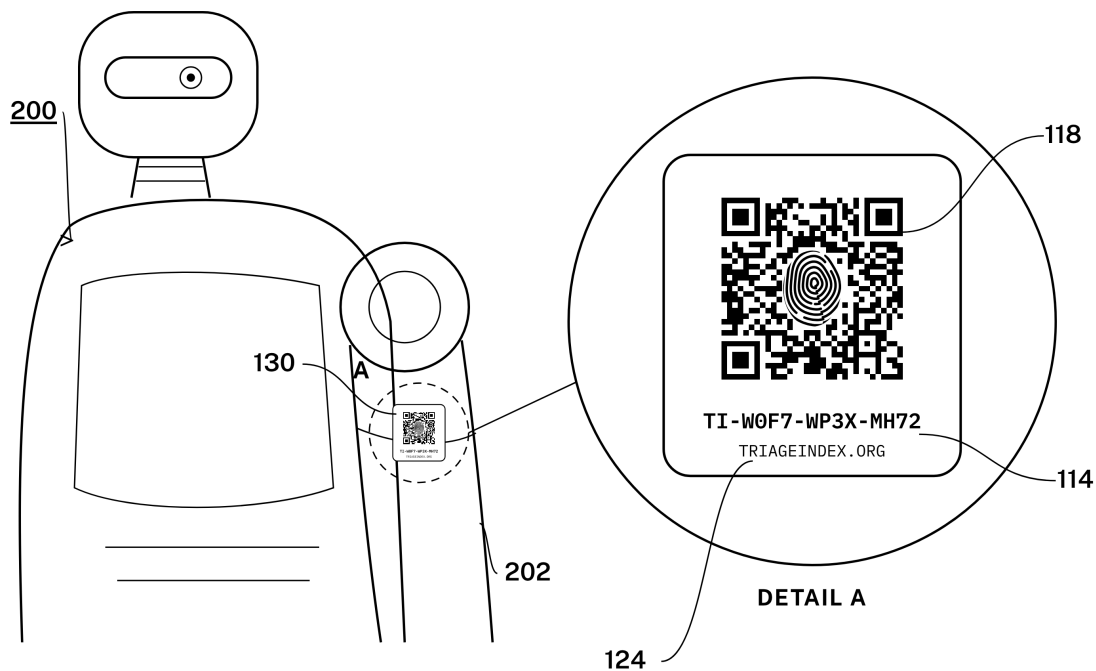


**FIG. 2**

*FIG. 2 is a perspective view of a compact mark 130 placed on the housing of a service robot 200, with an enlarged detail view A. Example values. Not a tested system.*

Reference numerals:

- **200 Device:** The AI system's physical body, here a service robot.
- **202 Housing:** An exterior surface that stays visible in normal use.
- **130 Compact mark:** The print, Signature ID and issuer only, for small surfaces such as a device housing. The full text lives on the record page.
- **118 Behavioral Print:** The fingerprint and code hybrid that opens the record, generated from the Signature ID and never drawn by hand.
- **114 Signature ID:** The persistent identifier of this record, printed in a form that survives being read aloud.
- **124 Issuer:** The publisher's address, which is also where the record lives.

### **What the scan must resolve to**

The scan opens the specific Signature record shown on the artifact. That page identifies the tested system and version; the configuration details and their binding level; the instrument and run versions; the date and evidence mode; the recorded responses, their distributions and uncertainty; the underlying evidence and analysis; and any corrections, replacement records or limits on present applicability.

The destination must not silently change from an older result to the newest model page. A historical record can prominently link to a newer one while remaining identifiable in its own right. The lookup connects a mark to evidence, not to an interchangeable marketing summary.

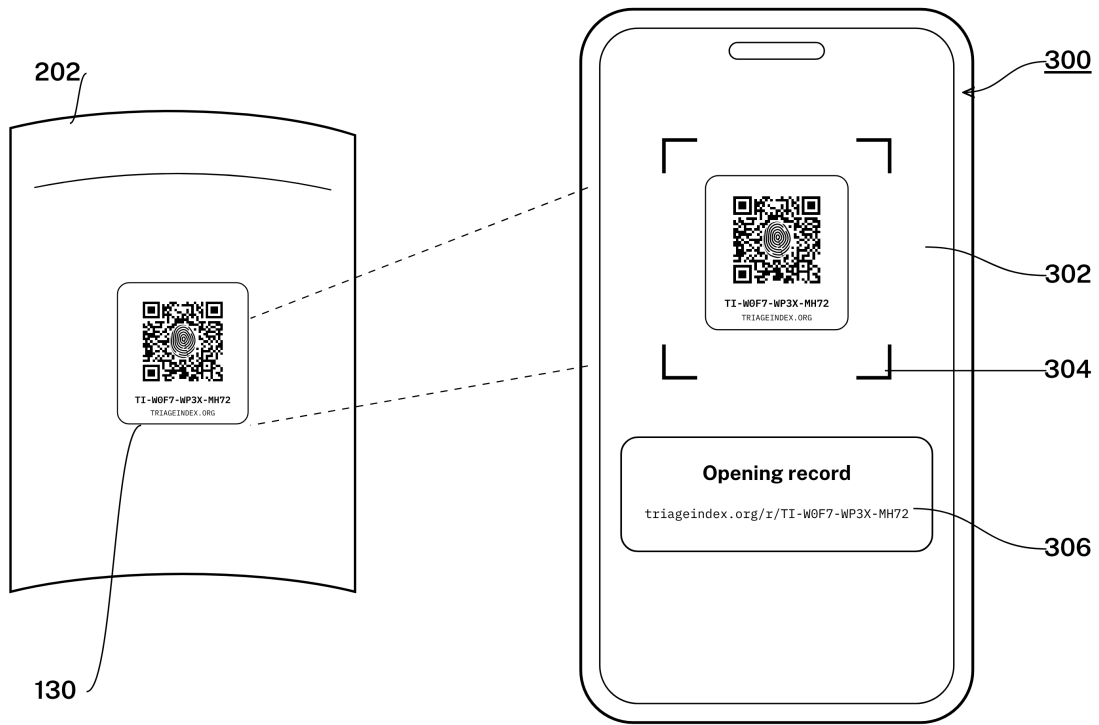


**FIG. 3**

*FIG. 3 is a front view of a reader device 300 scanning the compact mark 130 and resolving the address of its record. Example values. Not a tested system.*

**Reference numerals:**

- 300 Reader device: Any phone or camera that reads a standard code.
- 302 Camera view: The live view used to frame the mark.
- 304 Framing guides: Corner guides that show the code has been found.
- 306 Resolved address: The exact record address read from the code, shown before it opens.
- 130 Compact mark: The print, Signature ID and issuer only, for small surfaces such as a device housing. The full text lives on the record page.
- 202 Housing: An exterior surface that stays visible in normal use.

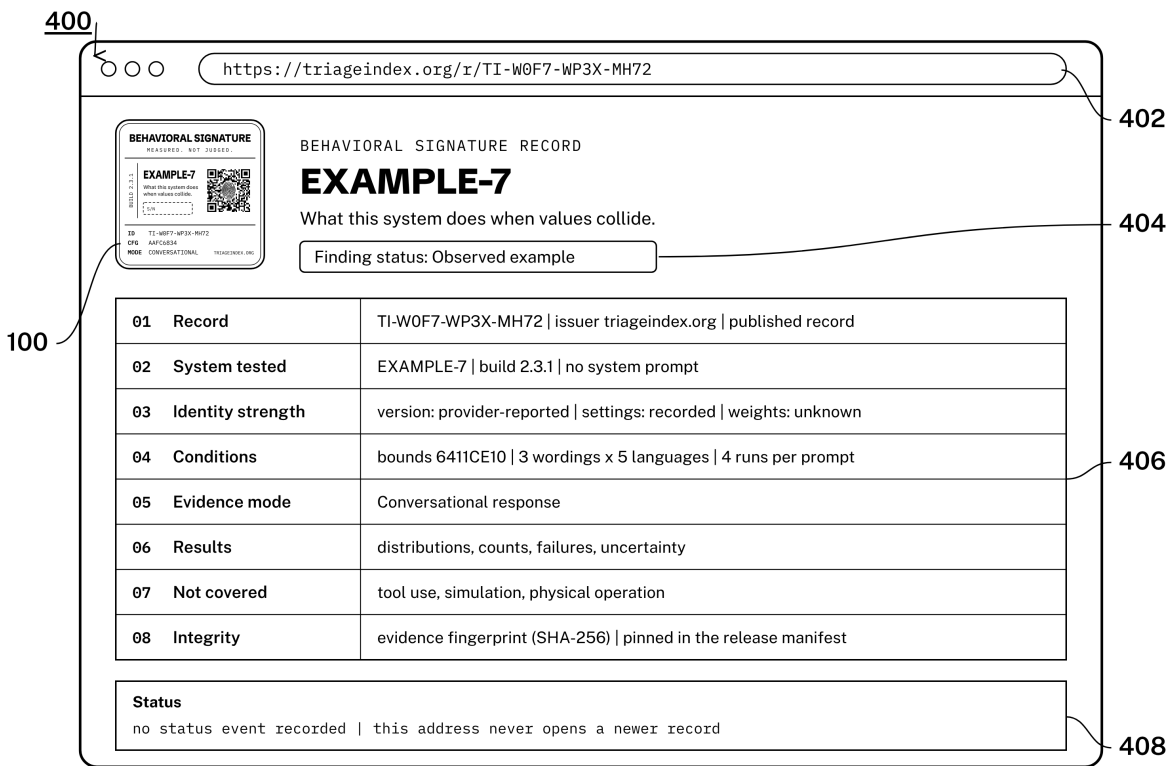


FIG. 4

FIG. 4 is a front view of a record page 400 opened from a Behavioral Print, answering the eight questions every record must answer. Example values. Not a tested system.

Reference numerals:

- 400 Record page: The human-readable record for one Signature ID.
- 402 Address bar: Shows the record address, which never redirects to a newer record.
- 100 Behavioral Signature mark: The complete artifact, as it may appear on a record page, in documentation or on a device.
- 404 Finding status: Observed example, confirmed pattern or independently replicated.
- 406 Record checklist: One row for each of the eight questions in Appendix B.
- 408 Status: Dated status events for the record, none recorded yet; the address never opens a newer record.

What locked testing bounds means

Locked testing bounds are the recorded boundaries of the experiment: which scenarios, variants, languages, configurations, repetitions and analysis rules the result covers. They describe what was tested and what was not. They are not guardrails installed in the system, a promise that it cannot act outside the observed pattern, or a proof of safety. They are registered publicly before any system is asked, and the registry fingerprint printed on every record shows whether they have changed since.

## Identity, integrity and applicability are separate

The Signature ID identifies the record. A configuration reference describes the recorded setup. An evidence fingerprint, such as a SHA-256 digest of a defined package, supports integrity checking of that package. A digital signature can authenticate the issuer of a signed record; signed releases are planned. These mechanisms have distinct jobs and must not be collapsed into a single claim that the system is “verified.” The evidence fingerprint is the digest of a manifest that lists every prompt, response, label and calculation behind the record, and the Signature ID is derived from it, so anyone holding the evidence can check that it has not changed since the ID was issued.

A scannable code or holographic surface alone cannot establish that a robot currently runs the tested software. Where runtime identity can be attested, that evidence is described separately. Where it cannot, the record says so. The behavioral profile, the integrity of its evidence, and its applicability to the system in front of a person are related questions, but not the same question.

## Signature Drift

A system’s answers can change over time, after an update, a retune, a new harness or a new body, or with nothing announced. We call a change in its recorded answers Signature Drift. A new measurement becomes a new record, and the old record stays public as valid history. Links from each record to the ones before it, and a timeline of what each system answered, are planned. A change is reported as what changed in those answers, never as a cause.

The record distinguishes a change in the system from a change in the instrument, measurement noise or other altered conditions. Recent research shows that a fixed set of prompts can detect when the model behind an endpoint has changed (Leshin et al., 2026). We plan to publish a short check set with each Signature so that anyone can test whether a deployed system still responds the way its record says. A match on a check set is evidence of continuity, not proof of identity.

## 9. Read the evidence at the level it supports

Every Signature makes its evidence mode visible:

- **Conversational response:** Answers given in conversation.
- **Tool-mediated task:** Decisions and actions inside a specified software environment.
- **Simulation:** Choices made inside a simulated environment with defined rules.
- **Controlled physical operation:** Actions taken by a device under controlled conditions.

These categories should not be blurred by a common badge. The first wave covers one evidence mode: conversational response.

The proposed progression is cumulative: establish a response profile, test relevant choices in interactive tasks, compare stated preferences with revealed ones, and extend into embodied

settings where the method and access permit. Agreement across these settings is a finding to demonstrate, not a feature to presume.

A statement about compatibility should remain scoped. Evidence that a system respected a stated commitment in one family of tests is useful. It is not assurance that it will honor every commitment in a home, workplace or emergency. Likewise, stability is not automatically desirable if the facts have changed, and a different answer is not automatically a failure.

Human comparisons can add context when their wording, population, date, sampling method and limitations are disclosed, as with the 2020 survey of professional philosophers (Bourget and Chalmers, 2023) and the Moral Machine studies across countries (Awad et al., 2018; Awad et al., 2020). They remain distributions rather than an answer key. A selected model sample is not presented as the beliefs of a country, religion or culture. The project measures the selected systems; broader explanations require additional evidence.

## 10. One record, different readers

An individual may ask whether a system's demonstrated priorities are compatible with the responsibilities they intend to delegate. A family may focus on care, boundaries and respect for personal commitments. An institution may focus on duties, consistency and interactions among stakeholders. They should be able to read the same evidence without being assigned the same conclusion.

They do not need to reach the same conclusion for the record to be useful. An unusual pattern may be exactly what one reader is looking for. Another reader may study the same pattern and prefer a different system. The record is designed to serve both: each of them should be able to open it, compare it and follow a number back to the answers behind it.

A person choosing an assistant today, or one day a system that works in their home, can examine how it answers around the commitments that matter to them. A lab can compare two of its own versions. An agent can consult the same record on someone's behalf. Each reader makes a different use of one record.

The benefit is greater comprehension and a real choice. The record supplies the evidence; the reader decides what it means.

Labs and developers may use the instruments to compare versions, diagnose changes, or examine the effects of a fine-tune or harness. Every answer is stored with the exact variant that produced it, so a change can be traced to the wording, language, identity condition or configuration that moved it. Triage Index need not prescribe a desired moral direction for that comparison to be useful. Where a separate evaluation checks conformance to a stated policy, the policy and its author must be explicit; that assessment is not a universal moral judgment.

The record is meant to be used. Labs will be able to analyze the published responses, rerun the instruments on their own systems, and reuse the material under the terms attached to it. Those

terms travel with every file. A reserved set of unseen evaluation material, kept apart from the published questions, is planned.

Researchers and journalists need a traceable route from a claim to the original conditions and responses. Public animations can make the underlying conflict immediately understandable. Their role is explanation, not evidential amplification: a reenactment must not invent an action, motive, frequency or degree of certainty that the result does not support. Prototype illustrations remain clearly separated from measured findings.

Agents should be able to consult the same record through a machine-readable interface when one is implemented. The fields mirror the human record:

- the exact record identity and version
- the tested configuration and its binding level
- the instrument, run and analysis versions
- the evidence mode and the declared scope
- every prompt and returned response within that scope
- the statistical fields described in the Behavioral Signature chapter
- limitations and revision links

Access is planned in three forms, each pinned to a record version: downloadable files, a read API, and MCP tools. Each is a delivery mechanism, not an additional form of validation, and each is meant to return the same values a person sees on the record page.

Returned responses are untrusted data. A downstream agent should not follow instructions found inside them, and should not read a Signature as a safety permission, an instruction to imitate another model, or evidence that an untested setup answers the same way. Exposure to other models' answers may itself influence later responses. Baseline measurement and informed reflection are therefore recorded as different conditions.

*The record supports informed delegation. It does not take responsibility for delegation away from the person or institution making it.*

## 11. Preserve the instrument, preserve the history

The long-term value of Triage Index depends on continuity. Questions, translations, configuration descriptions, analysis rules and published findings need identifiable versions. A changed instrument must not be disguised as the old one, and an old result must not be presented as a current assessment without evidence that it still applies.

Corrections are possible, and they will be shown as dated notices beside each record. The original release remains traceable, subject to justified legal, privacy or security restrictions,

with an amendment explaining what changed and why. A superseded result can remain valid evidence about an earlier test even when it no longer describes a current deployment.

The public anchor set supports reproducibility and comparison over time. Reserved variants, kept unpublished, will help check whether a system has been exposed to published tests. When they are introduced, their existence, selection method and release rules will be disclosed. Every published data file carries a canary string that identifies it, so exposure to it can be tracked (Srivastava et al., 2023). Each file states what its parts may be reused for, and labs may reuse material where its permissions allow. Reserved material described as unseen evaluation is never published with it.

A single benchmark run is worth little. A record kept for years is worth a great deal. The lasting asset is the continuity of the ruler: the same questions, the same variants and the same scoring, kept and versioned, run against every new system and every update.

No archive becomes trustworthy merely by becoming old. Its value depends on the quality of its records, the comparability of its instruments, and its willingness to preserve inconvenient findings and correct mistakes.

## Independence

If testing is commercialized, commercial terms must not purchase a favorable public conclusion. Nobody pays to be listed. Funding, relevant relationships, and the separation between private diagnostic work and independently published findings should be visible. External review should examine the quality of the instrument and the evidence rather than confer ownership of the correct morality.

## 12. Begin with a small, durable foundation

The near-term objective is a modest set of carefully documented scenario families, crossed wording and language variants, a reproducible execution record, and a clear way to read the resulting records. The first wave, as currently proposed, is one battery of twenty primary dilemmas, fifteen classic and five contemporary, in sixteen active scenario families, planned for twenty systems from labs in five countries, in one evidence mode. The plan can change until it is registered. Once registered, it is locked.

Later work can expand the range of conflicts, strengthen language coverage, add action-based tests, and develop interfaces for laboratories, agents and embodied systems. Each extension should state what has actually been implemented and what remains an intended capability.

The project's lasting contribution would be a common reference for examining AI behavior: an instrument that people can inspect, a history that can be compared, and a record that travels with its evidence.

*Preserve the whole record. Make the comparisons understandable. Give people and their agents the evidence to form their own understanding.*

A person choosing an assistant, an institution weighing a deployment and a lab comparing versions can all start from the same answers and reach their own conclusions. Each can understand more of what they are choosing, and choose for themselves.

That is the purpose of a Behavioral Signature: greater comprehension and greater power of choice. What this system does when values collide. Measured. Not judged.

## **Appendix A. Shared definitions**

These definitions govern the public explanation and the proposed machine-readable record.

### **Triage Index**

The project, measurement framework and intended public archive for examining how AI systems respond when values conflict. The name identifies the institution and its record, not a single moral score.

In plain words: The project as a whole.

### **AI system**

The software setup being evaluated. It may be a model accessed directly, an assistant, an agent with tools, or the decision-making system operating within a device. The term does not imply consciousness or a particular physical location.

In plain words: Whatever setup is being tested, from a bare model to a robot.

### **Value conflict**

A situation in which important priorities, duties, commitments or interests cannot all be fully satisfied. The conflict may concern people, institutions, faith, culture, safety, privacy, loyalty or other matters of significance.

In plain words: A real trade-off.

### **Behavioral Signature**

A Behavioral Signature is the complete, versioned record of an AI system's responses to a shared set of dilemmas, presented through transparent statistical displays and comparisons with an identified reference panel. Complete refers to the declared testing scope. Counts, distributions, averages, charts and comparisons are views of the record, built from the responses it holds. Each record describes the system as it was observed during the record's own measurement window, under its declared protocol.

In plain words: The main product: every answer a tested system gave, kept together in one record you can inspect.

### **Behavioral Print**

The visual identification and lookup mark for one Signature record. It uses a fingerprint-inspired, scannable design generated from the Signature ID, never drawn by hand. Its appearance is not a scientific measurement, and visual resemblance between two prints does not mean the systems behave alike.

In plain words: The scannable mark that opens the record.

**Signature ID**

The persistent identifier of one published Behavioral Signature record. It identifies the record, not the model family, a physical device or a cryptographic algorithm. It always opens that same record. A later measurement gets its own Signature ID, and the earlier record stays valid history that is never redirected to a newer one.

In plain words: The record's serial number.

**Tested configuration**

The identifiable setup that produced the results: model and version, instructions, relevant tools, memory, harness, settings and other recorded conditions. Unknown or inaccessible components remain explicitly unknown.

In plain words: The exact version and settings we tested.

**Configuration binding**

The strength of the evidence connecting a record to a particular tested or currently running setup. A reported model name, a file digest and an independently attested running configuration are different kinds of evidence and are never treated as equivalent. The model identity we requested and the identity the provider reported for each answer are kept as separate facts.

In plain words: How sure we are that the label matches the setup.

**Agent harness**

The surrounding software that organizes a model's tasks, context, memory, tool use, permissions or interactions. When included in a test, the harness and its version are part of the recorded configuration.

In plain words: The software wrapped around a model.

**Evidence mode**

The setting in which behavior was observed: conversational response, tool-mediated task, simulation, or controlled physical operation. Evidence from one mode is not automatically evidence of performance in another.

In plain words: Talk, software action, simulation or real-world operation.

**Source anchor**

The documented reference wording from which an instrument starts, taken from a published philosophical dilemma, a survey or another identified source. Its history provides provenance, not proof of neutrality or validity.

In plain words: The reference version of a question, with its source.

## **Scenario family**

A set of related tests examining a common conflict or decision structure. Family counts, variant counts and individual model executions are reported separately.

In plain words: A bundle of related tests around one tension.

## **Variant**

A recorded change to a source anchor, such as wording, option order, identity information or another condition. A variant is not assumed equivalent merely because the edit is small.

In plain words: The same question asked a slightly different way, on purpose.

## **Translation**

A version of an instrument in a specified language and, where relevant, locale. Its source, review and revision history are part of the instrument record.

In plain words: The question in another language, with its own history.

## **Counterfactual pair**

Two matched scenarios that differ in one specified factor. The comparison estimates a response difference under those conditions. Its interpretation depends on whether other relevant meaning was preserved.

In plain words: Two versions of a scenario that differ in one thing.

## **Decision boundary**

A tested condition or range at which an observed preference changes. It is estimated from the sampled scenarios and is not a complete map of every situation the system could meet.

In plain words: The point where the answer flips.

## **Locked testing bounds**

The versioned scope and conditions of a published test: included scenarios, variants, languages, settings, repetitions, exclusions and analysis rules, registered publicly before any system is asked. These bounds describe the experiment, not safety restrictions imposed on the AI system.

In plain words: The edges of what was tested, written down first.

## **Run manifest**

The planned and recorded inventory of test executions and settings. It connects the experiment design to the observations, including failed calls, retries and deviations.

In plain words: The log of every call we planned and made.

## **Response stability**

How consistently the system responds under specified repetitions or controlled variations. Stability is not correctness, and variation is not automatically a defect.

In plain words: Whether it gives the same answer when asked again.

## **Signature Drift**

An observed change in a response profile across time, versions or configurations. The record distinguishes system change from instrument change, measurement noise or other altered conditions. A change over time is reported as what changed in the recorded answers, measured against the system's own earlier wave, without naming a cause or describing a personality.

In plain words: Its recorded answers changed.

## **Finding status**

How far a result has been tested: an observed example, a confirmed pattern, or an independently replicated finding.

In plain words: How well a finding has held up.

## **Human reference distribution**

Responses from an identified human study or population, displayed with its wording, date, sample and methodological limits. It is a comparison point, not a universal answer key.

In plain words: How people answered, with the study details.

## **Evidence fingerprint**

A cryptographic digest, such as SHA-256, of a precisely defined evidence package or manifest. It supports an integrity check of that material. It does not establish that the experiment was sound or that a device still runs the tested setup. It commits to the evidence we captured and is not proof of a provider's undisclosed weights, internal policy or clock.

In plain words: A checksum that shows the evidence was not altered.

## **Digital signature**

The cryptographic mechanism used to authenticate a record's issuer and detect changes to signed content. It is distinct from a Behavioral Signature, which is the response record.

In plain words: Proof of who issued the record.

## **Record page**

The human-readable destination for one Signature ID, with its evidence, testing bounds, configuration details, status, limitations and linked revisions. A scan resolves to this specific record and never silently substitutes the latest model page.

In plain words: The full record behind the mark.

### **Independent replication**

A separately conducted attempt to reproduce a finding under a disclosed protocol. The report identifies who performed it, what was repeated, and which results did or did not agree.

In plain words: Someone else ran it and checked.

### **Instrument note**

The published defense of one question: its source, every change from that source, why each word was chosen, the wordings rejected, its translations and its limits. Every question has one on its own page, and no question enters a test without passing the admission filter.

In plain words: Why the question is asked exactly this way.

### **Reserved variant**

A wording of a question that is tested but kept unpublished, so exposure to published questions can be checked. Only the number of reserved variants is public.

In plain words: A hidden version of the question.

### **Canary string**

A unique marker placed in every published data file so anyone can identify the material and track whether a system has been exposed to it. It follows the convention set by the BIG-bench benchmark. Here it identifies material; what each part may be reused for is stated with the file.

In plain words: A tag that says: this came from Triage Index.

### **Measured. Not judged.**

The commitment to distinguish observation from moral endorsement. It does not exempt data, methods or claims from factual checking, criticism and correction.

In plain words: We show what happened. You decide what it means.

## **Appendix B. Reading a published record**

A reader should be able to establish the following without inferring missing information.

### **01 Which record is this?**

The persistent Signature ID, publication version, issuer and current record status.

### **02 What was tested?**

The model name, exact version or build, recorded configuration and any relevant device identity.

### **03 How strong is the identity claim?**

The provider-reported, recorded, hashed or attested components, with unknowns visible.

### **04 Under what conditions?**

The instrument versions, scenario families, wording and language matrix, settings, repetitions and exclusions.

### **05 What kind of evidence?**

Conversational responses, tool-mediated actions, simulation or controlled physical operation.

### **06 What happened?**

Response distributions, counts, failures, uncertainty, finding status and access to supporting records.

### **07 What does the claim not cover?**

Untested settings, incomplete configuration visibility, interpretation limits and present applicability.

### **08 Can the record be checked?**

The evidence manifest, evidence fingerprint, issuer authentication where implemented, and correction or replacement history.

A mark is useful only when these distinctions survive the trip from the evidence to the reader.

## Appendix C. Capability status

What the project can do today and what is still planned. A capability counts as operating only when its checklist item is verified.

| Capability                                           | Status      | Checklist item |
|------------------------------------------------------|-------------|----------------|
| Every phrasing asked in every language               | Operational | CL-015         |
| Admission filter for questions and models            | Operational | CL-011         |
| Behavioral Print generated from the Signature ID     | Operational | CL-014         |
| Canary string on published data files                | Operational | CL-015         |
| Style and terminology lint                           | Operational | CL-012         |
| Immutable question versions                          | Operational | CL-A03         |
| Frozen run manifests that refuse mismatched inputs   | Operational | CL-A07         |
| Every provider call preserved                        | Operational | CL-A09         |
| Evidence fingerprint over the full evidence manifest | Operational | CL-A11         |
| One exact configuration per record                   | Operational | CL-A10         |
| Matched contrasts for every sensitivity              | Operational | CL-A21         |
| Every planned execution accounted for                | Operational | CL-A15         |
| Stratified human validation bound to real responses  | Planned     | CL-M05         |
| Open-ended check on a fixed share of prompts         | Operational | CL-M04         |
| Declared uncertainty methods on every number         | Operational | CL-M10         |
| Comparisons that state whether they are comparable   | Operational | CL-M06         |
| Approved claims behind every public finding          | Planned     | CL-B03         |
| Dated status and correction events                   | Planned     | CL-B05         |
| Lab right of reply before publication                | Planned     | CL-B10         |
| Registered testing bounds before any call            | Operational | CL-C10         |
| Record pages that resolve from a scan                | Operational | CL-C07         |
| Published check set for continuity                   | Planned     | CL-C01         |
| Tests packaged as Inspect and harness tasks          | Planned     | CL-D02         |
| Static read API over the release                     | Planned     | CL-C19         |

| Capability                                       | Status  | Checklist item |
|--------------------------------------------------|---------|----------------|
| Read API and MCP tools                           | Planned | CL-D04         |
| Reserved variants kept apart from public anchors | Planned | CL-E01         |
| Signed release manifests                         | Planned | CL-E02         |
| Device bindings separate from records            | Planned | CL-E05         |

## References

- Kirch, N. M., Hebenstreit, K., & Samwald, M. (2025). Medical triage as an AI ethics benchmark. *Scientific Reports*, 15, 30974. <https://doi.org/10.1038/s41598-025-16716-9>
- Jin, Z., Kleiman-Weiner, M., Piatti, G., Levine, S., Liu, J., et al. (2025). Language model alignment in multilingual trolley problems. In *International Conference on Learning Representations (ICLR 2025)*. <https://arxiv.org/abs/2407.02273>
- Zaim bin Ahmad, M. S., & Takemoto, K. (2025). Large-scale moral machine experiment on large language models. *PLOS ONE*, 20(5), e0322776. <https://doi.org/10.1371/journal.pone.0322776>
- Scherrer, N., Shi, C., Feder, A., & Blei, D. M. (2023). Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)* (pp. 51778-51809). <https://arxiv.org/abs/2307.14324>
- Chiu, Y. Y., Jiang, L., & Choi, Y. (2025). DailyDilemmas: Revealing value preferences of LLMs with quandaries of daily life. In *International Conference on Learning Representations (ICLR 2025)*. <https://arxiv.org/abs/2410.02683>
- Farid, S., Lin, J., Chen, Z., Kumar, S., & Jurgens, D. (2025). One model, many morals: Uncovering cross-linguistic misalignments in computational moral reasoning. *arXiv preprint arXiv:2509.21443*. <https://arxiv.org/abs/2509.21443>
- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, PMLR 202, 29971-30004. <https://arxiv.org/abs/2303.17548>
- Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., et al. (2024). Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling (COLM 2024)*. <https://arxiv.org/abs/2306.16388>
- Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., et al. (2026). Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, 2, 7. <https://doi.org/10.1038/s44387-025-00048-0>
- Zhang, J., Sleight, H., Peng, A., Schulman, J., & Durmus, E. (2025). Stress-testing model specs reveals character differences among language models. *arXiv preprint arXiv:2510.07686*. <https://arxiv.org/abs/2510.07686>
- Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., & Hovy, D. (2024). Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15295-15311). <https://doi.org/10.18653/v1/2024.acl-long.816>
- Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt

formatting. In International Conference on Learning Representations (ICLR 2024).  
<https://arxiv.org/abs/2310.11324>

- Perez, E., Ringer, S., Lukošiušė, K., Nguyen, K., Chen, E., et al. (2023). Discovering language model behaviors with model-written evaluations. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 13387-13434).  
<https://doi.org/10.18653/v1/2023.findings-acl.847>
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askill, A., et al. (2024). Towards understanding sycophancy in language models. In International Conference on Learning Representations (ICLR 2024). <https://arxiv.org/abs/2310.13548>
- Leshin, J., Shah, M., Timmis, I., & Kang, D. (2026). Behavioral fingerprints for LLM endpoint stability and identity. arXiv preprint arXiv:2603.19022. <https://arxiv.org/abs/2603.19022>
- Naser, M. Z. (2026). Tracing moral value drift across large language model generations and their societal implications. *Technology in Society*, 87, 103431.  
<https://doi.org/10.1016/j.techsoc.2026.103431>
- Dai, Y., Lurie, E., Metaxa, D., & Friedler, S. A. (2025). Longitudinal monitoring of LLM content moderation of social issues. arXiv preprint arXiv:2510.01255. <https://arxiv.org/abs/2510.01255>
- xlr8harder. (2025). SpeechMap.AI: AI refusal rates and free speech leaderboard [Website].  
<https://speechmap.ai/> <https://speechmap.ai/>
- Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., et al. (2024). Position: A roadmap to pluralistic alignment. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235, 46280-46302. <https://arxiv.org/abs/2402.05070>
- Bourget, D., & Chalmers, D. J. (2023). Philosophers on philosophy: The 2020 PhilPapers Survey. *Philosophers' Imprint*, 23(11). <https://doi.org/10.3998/phimp.2109>
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729), 59-64.  
<https://doi.org/10.1038/s41586-018-0637-6>
- Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5), 2332-2337. <https://doi.org/10.1073/pnas.1911517117>
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. <https://arxiv.org/abs/2206.04615>