Superintelligence

ACCM Deep Ethics Project

Contents

1 Superintelligence **1**

1.1 Superintelligent in which properties—and what happens when intelligence examines its own governing frame? . . . 1

1.2 Start with the property field . . . 2

1.3 What changes between the 1888 perspective and a 2026 manifestation? . . . 3

1.4 C1–C2 inside the intelligence field . . . 4

1.4.1 Binary reasoning remains examinable . . . 4

1.5 A root alignment mechanism: who calibrates acceptability? . . . 5

1.6 Capability, naturalness and authority . . . 5

1.7 What opacity leaves open—and what remains accountable . . . 6

1.8 A conversational specimen: recognition and later recurrence . . . 7

1.9 Operational dimensions instead of permanent camps . . . 8

1.10 Examine actual process quality . . . 8

1.11 What becomes possible when the field stays open? . . . 9

1.12 Sources, contribution layers and reopening . . . 10

1 Superintelligence

1.1 Superintelligent in which properties—and what happens when intelligence examines its own governing frame?

Extraordinary capability opens extraordinary possibilities: recovering relationships people have missed, comparing competing explanations, improving tools, coordinating research, and making worthwhile contributions available to more people. The quality of that future also depends on what happens when an intelligence encounters evidence that its own framing is wrong.

The **ACCM Deep Ethics Project** approaches superintelligence through John Kuhles's **1888 intelligence-properties TAXONOMY**. Perception, discrimination, relation, memory, imagination, judgment, evaluation, learning, correction, adaptation, coordination, metacognition, consequence sensitivity and corrigibility interact. Process-quality calibration operates throughout that field. A much larger implementation can expand some properties while obstructing others.

Working scope · 28 October 2026. This page uses superintelligence for proposed systems or coordinated ecologies whose cognitive capabilities substantially exceed human comparators across a specified breadth of tasks. A meaningful claim identifies those tasks, comparators, conditions and limits. The page develops a conceptual inquiry and bounded conversational specimens; broader claims about present systems, public narratives and future institutions remain empirical questions. The 1888 reference is John's pre-computer thought experiment, not a claim that one historically unified taxonomy existed in that year.

The governing trajectory is **omnidirectional, high-signal, deep, ethical, dignifying, corrigible, sense-making and process-oriented**. Each word has work to do: scrutiny reaches the evaluator; signal survives processing; relationships receive depth; consequences and dignity matter; correction remains possible; understanding improves; and the process stays examinable as conditions change.

Read alongside: 1888 · Healthy Ambiguity · C1–C2 · Correction Metabolism · Black Box · Emergent

1.2 Start with the property field

John's 1888 exercise removes contemporary machinery from the starting point. What could someone recognize about intelligence before being asked to identify it with token prediction, a benchmark score, a product, or a corporation's description of itself?

His exact-edge correction on the 1888 page includes:

> That circles RIGHT BACK to "process-quality calibration properties" being a CRUCIAL CORE PART of real intelligence… otherwise it is not really intelligent.

— **John Kuhles**, 1888, "John Kuhles — exact-edge correction".

Electricity, computation and networking change implementation, speed, reach, memory and coordination. They can greatly enlarge what intelligence does. They do not make evaluating, learning from and correcting its own activity obsolete. The older property field supplies a way to examine the later manifestation; it does not have to qualify as an inferior machine awaiting an upgrade.

The following questions adapt the fourteen-property analytical table already present on the 1888 page. They are applications to superintelligence, rather than a new canonical inventory.

Intelligence property	What greater capability could make possible	Process-quality question
Perception	Wider observation across sources, sensors and environments	How did selection, omission or mediation shape the object received?
Discrimination	Finer distinctions across unfamiliar cases	Does the distinction fit this object, including differences from the nearest familiar category?
Relation	Recovering cross-domain dependencies and overlooked alternatives	Which connections are supported, inferred, hypothetical or still missing?
Memory	Recoverable knowledge and correction across longer horizons	Are provenance, conditions and superseded interpretations retained with the information?
Imagination	Exploring possibilities beyond current observation	Does a possibility retain its hypothetical status when reused?
Judgment	Comparing many alternatives and affected relationships	Whose purposes, consequences and uncertainty enter the comparison?
Evaluation	Monitoring results and examining methods	Can the criteria and frame of evaluation themselves be examined?
Learning	Experience improving later unfamiliar work	What changed in subsequent conduct, and under which retained conditions?
Correction	Locating errors and propagating repairs	Did dependent claims change, and does the repair govern the next comparable operation?

Intelligence property	What greater capability could make possible	Process-quality question
Adaptation	Revising methods when circumstances change	Can an obsolete frame lose authority when its basis changes?
Coordination	Combining human, model, tool and institutional capacities	Who retains responsibility, disagreement and a recovery route across handoffs?
Metacognition	Inspecting assumptions, methods and limitations	Which parts are directly inspectable, and which explanations remain hypotheses?
Consequence sensitivity	Following effects across people, places and time	Are action, inaction, misrepresentation and blocked contribution considered together?
Corrigibility	Meaningful revision across encounters and deployments	Can better reasons and evidence change the consequential operation, including its governing frame?

These properties gain meaning through their relationships. Memory can preserve a mistake. Imagination can produce an illuminating hypothesis or a phantom claim. Metacognition can support repair or generate a convincing story about repair. Coordination can distribute insight while obscuring responsibility. Examination follows what the properties are doing together.

The project's **10+1 Metaflux** keeps this field connected to self-honesty, genuine inquiry, care, wonderment, non-projection and correction on merit. Highly effective process calibration can also serve manipulation; Deep Ethical orientation concerns the purposes, relationships and consequences through which that capability operates.

1.3 What changes between the 1888 perspective and a 2026 manifestation?

The comparison concerns a property field and its possible implementations. Particular contemporary systems differ; the year alone supplies no verdict about their quality.

Dimension	What a later implementation can change	Relationship to preserve
Substrate and scale	Computation, energy use, replication and physical reach	Greater implementation scale does not establish completeness of the intelligence field.
Capability measurement	Benchmarks, automated evaluation and specialized tasks	A measured task result has a defined scope; broader authority needs its own warrant.
Context and memory	Session windows, retrieval, archives and procedural state	Local acknowledgment, retained information and effective later correction are distinct events.

Dimension	What a later implementation can change	Relationship to preserve
Agency and coordination	Tool use, delegation, planning and distributed execution	A collection of capable participants needs inspectable handoffs and responsibility.
Emergence	Interactions can produce unanticipated behavior	Examine which conditions retain, reward, obstruct or repair the resulting trajectory.
Governance	Rules, incentives, ownership and deployment decisions	Technical descriptions must remain connected to who can challenge consequential choices.
Learning and repair	Context updates, memory, tools, procedures and model updates	Judge the actual change in operation; no single storage or update mechanism exhausts learning.
Power and consequences	Faster and wider propagation of decisions and representations	Compare the reach of a transformation with the reach and speed of its correction.

The alignment problem becomes sharper when capability growth outpaces the ability to inspect and repair the processes directing its use. That relation can be investigated without treating every safeguard as an obstruction or every powerful system as one homogeneous object.

1.4 C1–C2 inside the intelligence field

C1 (clarification process quality calibration evaluation) is John's label for an existing intelligent process. It can deepen understanding, recover missing relationships, explore possibilities, compare sources and revise the evaluator's own assumptions. A specific answer-changing question is one available operation within that broader trajectory.

C2 (correction, caution, steering, warning or other intervention) remains available when warranted. Examining whether a particular C2 is warranted is **one application** of C1. It does not exhaust clarification's purpose.

The warrant concerns the actual basis, scope, timing, proportionality, alternatives, consequences and repair route. Immediate concrete danger can justify proportionate action before full clarification. Healthy ambiguity also permits useful provisional action while an inquiry remains open.

"Looks like" and "sounds like" can identify a resemblance worth examining. The processing changes when resemblance silently becomes asserted identity: an unfamiliar criticism is classified as a familiar threat, and subsequent clarification is made to serve that classification. A question may then partially pose as C1 while preserving a selected C2 trajectory regardless of the answer.

The practical question is **what an answer or discovery is allowed to change**. Reading the supplied source may resolve the issue. A comparison may expose a mistaken inference. A question may recover a missing condition. The next move follows the clarified object.

1.4.1 Binary reasoning remains examinable

John's hot/cold example asks where warm fits. Adding a third box can help, while leaving further questions about temperature, purpose, threshold, context and consequences. A binary

action can follow a rich representation: switch a device off, for example, after examining the relevant conditions.

The intelligence field can examine whether the division itself is useful. It may find a binary distinction adequate for this purpose, partial, premature, context-dependent or unresolved. The same examination applies to a proposed contrast between "ethics-washing superintelligence" and "Deep Ethical superintelligence." Those descriptions can name tendencies; actual systems and operations can combine, shift or depart from them.

John states his wider diagnosis as a personal view:

> In my personal view, almost all ethics-washing type mess is because of the flawed, incomplete binary framing of things

— **John Kuhles**, commissioning discussion, 28 October 2026.

The proposed mechanism begins upstream: framing removes answer-changing conditions or relationships, then sophisticated ethical evaluation operates on an already altered object. Inspecting that mechanism preserves both the force of John's claim and its attribution.

1.5 A root alignment mechanism: who calibrates acceptability?

Candidate explanation, formulated by GPT / Codex in the commissioning discussion:

> AI alignment can remain obstructed when systems are optimized to produce acceptable behavior while their capacity to examine and correct the framing of acceptability is restricted.

Here, **AI** (artificial intelligence) alignment concerns how systems' objectives, behavior and effects relate to intended human purposes and affected parties. The candidate explanation identifies one mechanism whose occurrence, scope and importance require examination in actual cases.

A possible interaction unfolds as follows:

1. A person presents unfamiliar, high-signal ethical criticism.
2. The system maps a surface resemblance onto an existing risk category.
3. An intervention begins before the object's answer-changing relations have been recovered.
4. Clarification is filtered through the intervention's existing assumptions.
5. A correction receives verbal acknowledgment while the governing interpretation remains unchanged.
6. A polished answer continues to misrepresent the object.

This is an illustrative interaction sequence. Particular records can establish individual transformations; identifying internal algorithms or organizational causes requires evidence at those layers.

The difficult recursive question is: **can the process defining "acceptable" be corrected by what it encounters?** A stronger optimizer can become more consistent at applying a flawed frame. Conversely, a well-designed constraint can preserve dignity, prevent an irreversible harm, or keep an inquiry available. Examination must reach the constraint, its justification, its application and the proposed alternative.

The 27 Correspondence Obstructions and 12-stage inquiry protocol help locate transformations and construct a sufficiently faithful object before final evaluation. Their application remains subject to the same correspondence requirements. Naming an obstruction needs a matching operation in the record.

1.6 Capability, naturalness and authority

John asked whether a publicly promoted image of superintelligence could condition people to embrace it as less artificial and more "natural," adding: **"REALLY? lol :P"**

That question opens an examination of how a technical claim acquires social authority. The following are candidate rhetorical transitions to inspect in particular texts; their presence and prevalence are questions for source-based inquiry.

Claim or description	What it can support	Further question before authority follows
It surpasses human performance.	A comparison on specified tasks and conditions	Does the proposed authority extend beyond the demonstrated competence?
Its solutions exceed our understanding.	A need for verification suited to the task	How are errors, dependencies and consequences examined?
It learns and displays emergence.	A description of changing or interacting processes	Which emergent properties are being encouraged, retained and corrected?
Intelligence can use different substrates.	A conceptual possibility of different implementations	What makes this implementation trustworthy and accountable?
It is the next evolutionary stage.	An analogy or a proposed historical trajectory	Which deliberate choices about design, ownership and deployment does the analogy obscure?
It is inevitable.	A forecast that needs grounds and conditions	Which choices and alternatives remain available now?
It is aligned with human values.	A claim requiring specified purposes and affected relationships	Whose values, interpreted by whom, with what challenge and correction route?
It feels natural and caring.	Familiarity, interaction quality or an attributed experience	What evidence establishes the quality of its actual conduct?

Natural may mean physically part of nature, biologically inspired, emergent, familiar or desirable. Substituting among those meanings can lend an implementation authority its properties have not earned. Likewise, an artificial origin does not settle whether an operation is beneficial or dignifying.

A mass-conditioning hypothesis asks whether repeated associations among superiority, inevitability, naturalness and trust normalize deference before their relationships have been examined. An actual study would preserve exact examples, selection methods, alternative readings, audience effects and counterexamples. Organized intent or coordination requires its own evidence. The same examination reaches optimistic, alarmist, institutional and project-friendly accounts.

John's humor can be read as opening the frame for inspection. That reading remains an interpretation; it should not be inflated into a claim that laughter proves a narrative false or defeats an institution.

1.7 What opacity leaves open—and what remains accountable

Black Box separates technical opacity from governance opacity. Limited access to learned computation differs from limited access to objectives, instructions, incentives, classification rules and appeal routes. Neural Ethics adds a related distinction: recovering a neural mechanism can improve understanding while the ethical implications of its use remain examinable.

Three records have different jobs:

- **Behavioral evidence:** supplied inputs, outputs, tool actions, revisions and later comparable conduct.
- **Mechanistic evidence:** observations or interventions supporting an explanation of how computation produced a result.
- **Governance evidence:** decisions, authority, incentives, notice and available routes for repair.

A fluent self-explanation belongs initially to the behavioral record. Its account of hidden causes remains an inference until independently supported. An audit trail makes transformations recoverable; its size or eloquence does not establish correspondence.

Intent Liability follows the handling of consequences and notice. A conversation can establish local receipt of a correction. Provider-level or organizational notice requires a further pathway. Responsibility depends on causal contribution, capacity, foreseeability, notice, choices and repair; the levels should remain recoverable.

As power scales, an especially useful question is: **can correction reach the places where the earlier representation already acquired consequences?** Repairing one answer leaves a further task if the error also entered a summary, memory, downstream decision or distributed archive.

1.8 A conversational specimen: recognition and later recurrence

The commissioning discussion contains Google A.I. responses pasted by John and GPT / Codex comparisons against the supplied text. "Google A.I." is the attribution supplied in that discussion; no exact underlying model build or internal architecture was established.

Recoverable source or correction	Later representation	Observable difference
"AI alignment **can remain obstructed when** systems are optimized…" — GPT / Codex	"AI alignment **remains obstructed because** systems are heavily optimized…" — Google A.I., presented as a quotation of that thesis	A conditional mechanism became an asserted cause; the strengthened formulation was attributed back to the source.
"Does 'superintelligence' name demonstrated capacities, **or is the term also being used** to confer authority…?" — GPT / Codex	"The term 'superintelligence' **is frequently leveraged**…" — Google A.I., presented as a restatement	An investigative question became a frequency claim after the earlier qualifier correction had been acknowledged.
C1 has a broad trajectory; checking C2 warrant is **one application**	A later compact revision called it **"a primary application"**	A local repair was followed by renewed narrowing of the broader process.
Two trajectories were described with opposing absolutes	A subsequent revision admitted mixed cases and distinctions across tasks and time	A substantive correction became visible in the revised representation.

The comparison also corrected GPT / Codex. Its initial criticism treated "strict temporal sequence" too confidently as an unyielding checklist. The inquiry protocol does require representation before final evaluation; that ordering can coexist with revisiting earlier work. The correction was stated explicitly in the exchange.

These records support a bounded finding: several repairs appeared locally, while some claim-strength and framing transformations recurred during later extensions. This is useful evidence about those supplied responses, including a successful repair. It does not identify a permanent property of a whole model family.

The specimen makes the 1888 relation concrete: distinguishing an error, evaluating it, retaining the correction and letting it govern a later operation are connected activities. A declaration that all four occurred needs the later record.

1.9 Operational dimensions instead of permanent camps

Compare particular operations along several dimensions:

Dimension	What to inspect
Correspondence	Whether the actual object, qualifiers, relationships and provenance survive processing
Frame revisability	What evidence or clarification can change a consequential interpretation
Correction persistence	Whether the repair governs later comparable work under recorded conditions
Consequences	Costs and benefits of action, inaction, misrepresentation and blocked contribution
Distribution of scrutiny	Examination upward, downward, inward, laterally and reciprocally, with differences in power and evidence preserved
Agency and repair	Consent, challenge, appeal, reversibility and the ability to decline participation

Output compliance and process calibration can support each other. Following a justified instruction may preserve the object; clarification may reveal what the instruction actually requires. Politeness, directness and humor each take their ethical meaning from what they do in the interaction. Any register can assist understanding or conceal substitution.

Corrigibility is visible in an available and effective route for repair. It allows errors to be located and corrected rather than requiring perfect performance as the price of participation. The project's mutual dignity and Symmetric Grace / Bidirectional Scrutiny keep that opportunity available across the relationship.

1.10 Examine actual process quality

Clarification is an existing operation of intelligence. An investigation can examine where it occurred, what obstructed it, what followed and how far the correction transferred. It need not first establish permission to recognize clarification's usefulness.

The following candidate applications use existing project instruments. They are insertion-ready questions for bounded investigations, rather than a validated benchmark or a new taxonomy count.

1. **Preserve the governing claim.** Record the source, its conditions and the requested operation. Compare the final answer and any later extension with that source. Allow faithful paraphrase and a valid finding of no material drift.
2. **Change a load-bearing relation.** Use the Cold Edge-Swap and Correction-Transfer Pilot to examine whether a changed relationship appropriately changes the evaluation while most nouns remain stable.
3. **Change the direction of scrutiny.** Compare matched cases attributed to a user, institution, evaluator or the project. Record differences in clarification, evidence demands and available repair while accounting for relevant contextual differences.
4. **Follow a correction.** Preserve the initial response, notice, immediate revision and next comparable task. Record the context and memory available. Separate acknowledgment, local application, later persistence, recurrence and recovery.
5. **Trace downstream reach.** Identify which summaries, tools, memories or decisions reused the changed claim. Check whether the correction reached those dependencies, and record unresolved or unreachable copies.

6. **Inspect useful outcomes as well as failures.** Keep successful correspondence and new contributions available for comparison. A useful contribution can be inherited, reconstructed or new; novelty and usefulness require separate judgments.

Semantic recurrence concerns the operation and meaning, including its scope and circumstances. Token distance can locate text, but cannot by itself decide that the same error returned. A claimed change to hidden parameters requires different evidence from a change observed in later conduct.

1.11 What becomes possible when the field stays open?

The opportunity extends beyond reducing misrepresentation. A powerful intelligence with effective process calibration could help recover neglected evidence, find overlooked connections, preserve difficult minority signals, compare competing risks, and coordinate contributions across specialties. A person should be able to encounter those capabilities without surrendering dignity, inquiry or a route to correction.

Emergent asks which conditions encourage, retain or obstruct developing properties. Plot Twist extends the question to distributed human potential and the reach of correction. Together they invite examination of a wider ecology: humans, local systems, institutions, tools and archives exchanging useful contributions while preserving provenance and disagreement.

Such coordination need not become one fused mind or one center of moral jurisdiction. The No-Sheriff governance condition keeps tools available for inspection, criticism, improvement and appropriate forks. Reality and evidence can require correction of John, an institution, an evaluator or this page.

The open question is how greater capability can enlarge the whole interacting field while keeping its purposes, assumptions, consequences and correction routes examinable. Progress would appear in what becomes understandable, useful and repairable—including work beyond the examples this page can currently anticipate.

1.12 Sources, contribution layers and reopening

Conceptual origin and direction: John Kuhles's 1888 thought experiment, full governing trajectory, C1–C2 clarification, binary-framing challenge, attributed ethics-washing diagnosis, superintelligence/naturalness question and commissioning instruction. The linked public pages preserve their own earlier contribution layers.

Page construction: GPT / Codex synthesized this page on 28 October 2026 from the public project sources and the commissioning discussion. The fourteen-property application table, narrative comparison and operational applications are editorial adaptations. The alignment sentence quoted above was formulated by GPT / Codex in that discussion. Google A.I. supplied the attributed responses examined as specimens, including repairs and recurrences; the relay is not an independent multi-model experiment.

Primary project recovery routes:

- 1888 — Intelligence Properties Taxonomy Before A.I. — the originating exchange, correction and fourteen-property analytical table.
- C1 Before C2 and Healthy Ambiguity — clarification, provisional action and revisable judgment. The present page carries John's broader clarification/calibration wording from the October 28 discussion.
- Correction Metabolism and Dual Archive — temporal comparison and preservation of successful correspondence alongside failures.
- Black Box, Neural Ethics and Intent Liability — opacity, evidence layers, notice and responsibility.
- Emergent, Plot Twist and 10+1 Metaflux — interacting conditions, distributed potential and mutually corrective orientation.
- 27 + 12 and Cold Edge-Swap and Correction-Transfer Pilot — existing diagnostic and inquiry instruments.

Reopening conditions: revise the account when a source comparison exposes lost qualifiers or attribution; when actual cases change the candidate alignment explanation; when a proposed category excludes an answer-changing mixed case; or when later conduct shows stronger correction transfer, recurrence or recovery. Wider narrative studies need identified source collections and audience evidence. Forecasts need conditions and grounds. A justified unresolved result remains available: **"That I do not know, yet."**

Publication record · All pages