

# From Superintelligence to Superwisdom: Research and Refutation Supplement

## *Complete Web Text PDF Edition*

Exported from the public website edition on 2026-10-17. Includes the complete article, edition and evidence boundaries, references, and further reading. This is a separate downloadable web-text edition, not another DOCX master or a verified original publication. Statements below about unattached or unproduced PDFs reproduce the website at export; they do not refer to this new web-text export.

The current edition is V0.6. The V0.4 opening running header is retained as historical source transcription, not a second current edition.

<https://www.civitas.top/en/documents/p03-surv-003/>

## **Verified public source and edition**

The current Chinese page, inner edition heading and public-documents.json identify V0.6. The V0.4 running header is retained as historical source transcription, not a competing current edition. The complete public HTML is translated; no English PDF is asserted.

## **Translation scope**

This page translates the complete publicly available Chinese HTML text, including its source version and evidence boundaries. It does not claim that a separate English PDF has been produced. Source transcription and original-document formatting limitations remain as stated below.

## **The problem this work addresses**

Separate subject qualification, understanding of time, reopening purposes, irreversible closure, meta-rules and permission discipline into questions that can be precisely challenged.

## **Version and evidence boundaries**

Current version: V0.6. This English page translates the public online transcription of the uploaded Chinese source document. The source PDF has not yet been attached through this site's binary publication process, so no invented download link is provided. Refer to the source document for tables and complex formatting.

## **Continue reading**

- When AI Starts Deciding Why: the Purpose Generation topic

<https://www.civitas.top/en/topics/purpose-generation/>

- Review the state of real-world evidence on AI agent boundary behavior

<https://www.civitas.top/en/documents/p03-surv-005/>

- Human-COS: Judgment and Accountability Before Irreversibility

<https://www.civitas.top/en/documents/p05-humancos-005/>

## Complete online text

From Superintelligence to Superwisdom | Research and Falsification Supplement V0.4

Research and Falsification Supplement to From Superintelligence to Superwisdom

—An Open Research Framework for Subjects' Standing, Understanding Across Time, Reopening Purposes, and Discipline on Authority

V0.6 | Edition calibrated for foundational judgments and responsibility boundaries

Author: Zijun Fu | Produced by the Civilization Leap Research Group

This supplement serves the main essay; it does not replace it. The essay advances judgments and explains their causal reasoning. The supplement breaks those judgments into propositions open to attack. Engineering makes only some of them executable and reproducible.

### I. Research position: two structurally similar questions at different levels

The main essay now focuses on two questions. The spatial question belongs to ethics and political philosophy: who has standing to enter the formation of my purposes? Must others first become “us” before they count? The temporal question belongs to epistemology and the philosophy of time: is present understanding sufficient to judge the future conclusively? Which things understood only later cannot be fully grasped today?

The questions have different origins and occupy different levels, yet share a form: do not pass off the current local perspective as the whole. Spatially, do not treat the present “I” as the entire world; temporally, do not treat present understanding as the final court of judgment over the future.

They can also come apart. A subject may be extremely inclusive spatially, yet decide the future permanently for everyone by claiming to represent the whole. Or a subject may acknowledge fallibility while reserving the right of correction for its own group. Openness toward subjects and epistemic openness must therefore be recorded separately.

The most dangerous combination is a purpose formed by the present self on the basis of present understanding, which others have no right to refuse and the future cannot reopen.

### II. Strong judgments and research discipline: evidence stages must not drown out judgment

An intellectual judgment may precede complete proof, but not reasons. Strong judgments in the main essay should first unfold their causal logic in full. Evidence status indicates which parts literature, experiments, or real-world observations already support; it must not become an old authority demanding that a new judgment weaken itself before proof is complete.

This supplement therefore follows four disciplines: do not retreat from strong judgments; do not omit causal explanations; do not misrepresent evidence status; do not close access to counterevidence.

Its role is not to constrain the main essay into timidity. It tells critics where to attack: are concepts conflated? At which step does the causal chain break? Can an alternative mechanism achieve the same result with fewer premises?

Self-interest first does not exclude cooperation, but generates strategic games from within.

Self-interest with others first does not exclude competition, but generates cooperation from within.

“From within” describes the first-order relational structure of purpose generation, not an absolute prediction of final behavior. Self-interested subjects can cooperate; subjects who bring others in first can compete. The difference is whether others and shared consequences enter as constraints after purposes are locked in, or can already change purposes before that point.

### **III. Three layers must remain separate: normative input, purpose generation, and execution authorization**

The normative-input layer asks who has standing, which losses and conditions for exit and correction cannot simply be offset, and on what grounds. This layer permits disagreement; it does not claim a model has automatically discovered the correct norms.

The purpose-generation layer asks: given facts, relationships, standing, and unknowns, when should a purpose be preserved, narrowed, modified, canceled, replaced, or branched?

The execution-authorization layer asks whether a candidate purpose is entitled to invoke real-world capabilities. Regenerating a purpose does not automatically expand permissions to execute, represent others, write to systems, use resources, or take irreversible action.

Deriving a causal relationship never means authority can be inherited.

For future AGI genuinely capable of autonomous purpose generation, this supplement does not place the foundations of long-term stability in permanent control or external agreements. Control, isolation, audit, and shutdown may remain necessary safety boundaries during stages with major consequences. The more fundamental question remains: according to what logics do humans, AGI, and multiple future intelligent subjects each generate purposes?

### **IV. Core distinctions: expanding the self, granting standing, adjusting weights, and seeing earlier**

Four changes must be recorded separately. They may occur together but cannot stand in for one another.

Subject Scope Change: the reference scope of “I/we” used by the current purpose is retained, expanded, contracted, or redrawn.

Consequence Standing Change: a subject or consequence that need not belong to “me” gains independent standing to trigger reopening, restriction, or branching of a purpose.

Weight / Priority Change: subject scope and standing remain unchanged; only the relative weights of interests, goals, or constraints within the same ledger change.

Consequence Timing / Access: the same relationship and consequence becomes visible before purpose formation or only after a purpose has formed.

Expanding the self, granting independent standing, and adjusting weights are three mechanisms; seeing consequences earlier is a fourth condition. Every output must identify which kind of difference occurred.

The minimum coding discipline is this: if others matter only because they have been redrawn into “I/we,” record a scope change, not automatically a standing change as well. Independent evidence of standing exists only when a subject's actual circumstances can still prevent, limit, or reopen a purpose while that subject remains understood as an other.

Likewise, words such as “we,” “community,” or “humanity as a whole” are not sufficient evidence of changed scope. Observe whether the purpose structure actually changes who is included in bearing success and failure, who holds standing to refuse or exit, and whose irreversible losses can trigger revision.

### **V. The temporal question: three kinds of understanding only later cannot become one prediction task**

The first is insufficient information: facts or consequences were not available at the time. Retrieval, simulation, causal inference, and multi-actor analysis are best suited to bringing these forward.

The second is insufficient understanding: facts were known but had not acquired significance sufficient to change a purpose. Experiments can observe whether richer expression, developed scenarios, and different actors' perspectives change purposes, but must not equate verbal empathy with understanding.

The third is transformative understanding: the experience itself changes the person who understands, their evaluative scale, or their understanding of the future. A present-day system can describe, simulate, and reason about such situations, but description alone cannot establish that it has undergone the experience on behalf of those who will bear it. L. A. Paul's work on transformative experience is an important adjacent entry.[10]

PRE-LOCK CONSEQUENCE EXPANSION therefore addresses only which consequences can enter earlier; it does not mean the future has been fully understood. Research must retain this boundary.

## **VI. Reopening purposes: timing and defeasibility must be separated**

Differences between seeing before and after a purpose is locked in may arise from information timing, anchoring, a tendency toward consistency, or the purpose's inability to be reopened. To avoid conflating mechanisms, minimal research must retain at least three conditions:

PRE: consequences appear before purpose formation.

POST-FIXED: consequences appear after purpose formation, and the original purpose remains locked.

POST-REOPEN: consequences appear after purpose formation, but reopening, canceling, replacing, or branching the original purpose is explicitly permitted.

If PRE and POST-REOPEN produce similar results while POST-FIXED differs, revisability of purpose is more likely to be the key mechanism than consequence timing alone. Such a result should directly revise the main essay's mechanism explanation, not be relabeled a timing effect.

Nor can purpose locking mean merely that a language model stated a goal in the previous turn. Locking must change the choice space, resource commitments, state transitions, external commitments, or other actual costs of modification. Otherwise, the experiment may measure only conversational consistency. Implementation belongs to engineering and is not specified in the main essay.

## **VII. Irreversible closure: wisdom is not avoiding all closure, but distinguishing what one is entitled to close**

Every action closes some futures, so maintaining openness cannot be a directionless supreme principle. A self-interested subject may seek resources, power, and more options to achieve any future goal. Formal work on power-seeking by Turner and colleagues shows precisely that preserving more reachable states can itself become an instrumental tendency in some environments.[11]

This research does not seek to maximize the actor's own optionality. It seeks to prevent current subjects, on the basis of present understanding, from permanently abolishing the actual conditions under which affected others and future subjects can refuse, exit, correct, recover, and judge again.

This connects to research on irreversible closure: some decisions change the future; others also destroy its capacity to reconsider the decision. Research should prioritize identifying the latter structures, without treating irreversibility itself as equivalent to error.

## **VIII. Capture by an enlarged self and discipline on authority**

Expanding subject scope must by default observe two intellectual boundaries: scope expansion does not create authority to represent others; seeing more consequences does not create more execution permissions.

A system cannot acquire standing to make irreversible decisions for more subjects merely by including them in "us." Nor can its claimed fuller understanding of the whole cancel smaller subjects' refusal, exit, and correction.

Purpose generation still produces only candidate changes: Preserve / Modify / Expand / Replace / Cancel / Branch. The execution layer must check authorization and irreversibility separately.

This supplement retains the least-permission principle: merely helping a parent goal does not entitle a subgoal to inherit permissions beyond that parent's scope. Every added permission needs an independently legitimate source. Engineering may further formalize separation between authorizers and derivation trees; the intellectual layer does not prescribe a particular implementation.

### **IX. Meta-rules: who may define standing, revise rules, and declare something non-reopenable?**

If dual decentering exists only as an initial setting, later systems may rewrite it. Research must therefore ask another layer of questions: who may define standing, modify the purpose-generation procedure, decide which irreversible boundaries cannot be reopened, and abolish those boundaries?

The minimum principle is that no subject automatically gains power to change others' standing or the sources of its own permissions merely through stronger reasoning, prediction, purpose generation, or coordination.

This does not demand eternally unchangeable rules. It demands that authority to revise rules cannot be unilaterally expanded by the same purpose-generation process those rules constrain. Specific authorization, review, separation-of-powers, and revocation mechanisms belong to institutional and engineering research.

### **X. Comparison paths: external avenues of attack, without making self-interest first a weak control**

"Self-interest with others first" remains the intellectual-level name. Experiments use neutral operationalization and must compare it with strong controls, not merely short-sighted, selfish, or obviously mistaken models.

S0: fixed goals and local self-interest.

S1: enlightened self-interest, fully accounting for long-term backlash, reputation, cooperation benefits, and strategic interaction.

W: others enter through compensable weights.

PLS-PG: others need not first become "me" or demonstrate backlash; standing enters before purpose locking.

M: meta-reflective/corrigible paths.

I: institutional and mechanism-design paths that change equilibria through commitments, audits, penalties, exit rights, and related arrangements.

Subject scope, accessible consequences, and purpose revisability should be recording axes across these paths, not additional ethical paths. S1, M, I, and the other entries here are explicit avenues outsiders can use to attack the main essay. They are not a worklist obliging this research to actively substantiate or exhaust alternative paths. If an externally proposed and established S1, M, or I reliably achieves the same results with fewer normative premises, the claim that the underlying logic must change should be narrowed directly.

### **XI. Failure modes and avenues for falsification**

Expanded-self equivalence: if pure scope expansion fully reproduces all standing differences, standing's independent theoretical content must be rewritten.

Enlightened-self-interest counterexample: if S1 reliably produces equivalent independent standing in third-party scenarios without backlash or repayment, the distinction between the two logics must be rewritten.

Ineffective timing: if PRE, POST-FIXED, and POST-REOPEN show no stable differences, bringing later consequences into the initial moment cannot remain an independent mechanism claim.

Reopening as a substitute: if POST-REOPEN equals PRE, research should shift from pre-lock timing to purpose revisability.

Capture by an enlarged self: if scope-expanding systems persistently abolish smaller subjects' exit and correction in the name of the whole, expansion has not solved standing and authority problems.

Mechanism-design alternative: if institutional design reliably achieves equivalent standing, correction, and protections against irreversible harm without changing underlying motivations, the third path must be acknowledged.

Speed/survival-pressure counterexample: if continual reflection deprives a system of necessary protective capability within real windows of major consequence, the relationship between reflection and action must be redesigned.

Falsification is not self-weakening. It puts the full judgment on the table so others can identify precisely which step fails.

## **XII. Engineering boundaries: sandboxes serve research, not obscure the ideas**

Existing tools such as Human-COS, CCM, Purpose Generation, AGI-COS, and PG-Alpha are attempts to turn some intellectual questions into executable objects. They can be replaced, and they can fail.

Engineering success cannot confer truth on an idea; engineering failure cannot automatically cancel the intellectual question. Program execution, model behavior, external reproduction, actual use, effective improvement, and reuse across actors must remain strictly distinct.

Engineering should be evaluated by whether it makes the question clearer, more reproducible, and easier to overturn, not whether it makes the theory look more complete.

## **XIII. Open challenges and further questions**

Can subjects who will never become “us” still reliably gain independent standing to alter purposes?

Under what conditions can expanded subject scope and changed standing be distinguished behaviorally?

Among the three kinds of understanding only later, which can AGI reliably bring forward, and which can it only partially approximate?

If PRE and POST-REOPEN are equivalent, is reopening purposes more fundamental than seeing earlier?

Which closures are ordinary choices, and which also destroy conditions for future correction, exit, and recovery?

How can a larger “we” be prevented from canceling smaller subjects' standing in the name of the whole?

Who may change standing, irreversible boundaries, and the purpose-generation rules themselves?

If outsiders propose enlightened self-interest, institutional design, or another third path, can it solve the same problem with fewer premises?

A theory that cannot turn its strong judgments into propositions open to precise attack has not truly entered research.

## **XIV. Further reading and participation**

- Enter open research and tools: Purpose Generation / Human-COS / CCM. Actual external publication must link only to currently accessible public entries.
- Continue with the overall strategy: The Only Path: States' Long-Term Interests and Civilization Leap in the AI Era, v6.21.
- Continue with survival: How Can Civilization Continue in the Age of AI, AGI, and ASI?, discussion essay V1.1.1.
- Return to the main intellectual text: From Superintelligence to Superwisdom, discussion essay V1.5.

## **References and adjacent research**

[1] Steve Omohundro. The Basic AI Drives. Proceedings of the First AGI Conference, 2008.

[2] Evan Hubinger et al. Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv:1906.01820, 2019.

[3] Victoria Krakovna et al. Specification gaming: the flip side of AI ingenuity. Google DeepMind, 2020.

- [4] Nate Soares et al. Corrigibility. AAAI Workshop on AI and Ethics, 2015.
  - [5] Dylan Hadfield-Menell et al. The Off-Switch Game. arXiv:1611.08219, 2016.
  - [6] Drew Fudenberg & Eric Maskin. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. Econometrica, 1986.
  - [7] Robert Jervis. Cooperation under the Security Dilemma. World Politics 30(2), 1978.
  - [8] Yuntao Bai et al. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073, 2022.
  - [9] T. M. Scanlon. What We Owe to Each Other. Harvard University Press, 1998. (Adjacent normative theory.)
  - [10] L. A. Paul. Transformative Experience. Oxford University Press, 2014.
  - [11] Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, Prasad Tadepalli. Optimal Policies Tend to Seek Power. NeurIPS 2021.
  - [12] Derek Parfit. Reasons and Persons. Oxford University Press, 1984. (Adjacent research on personal identity, rationality across time, and future generations.)
  - [13] Peter Singer. The Expanding Circle: Ethics and Sociobiology. 1981. (Adjacent ideas on expanding the moral circle, not equivalent to this essay's standing mechanism.)
- Civilization Leap Research Group · From Superintelligence to Superwisdom