The Self thesis

From intelligence to authority.

Autonomy should follow demonstrated competence.

By Self

Work, experience and responsibility

We are building for a world in which an AI system's history of work matters as much as its

next answer. Better models expand what is possible. Evidence from actual work gives people a basis for deciding which responsibility should follow.

▸ In this thesis

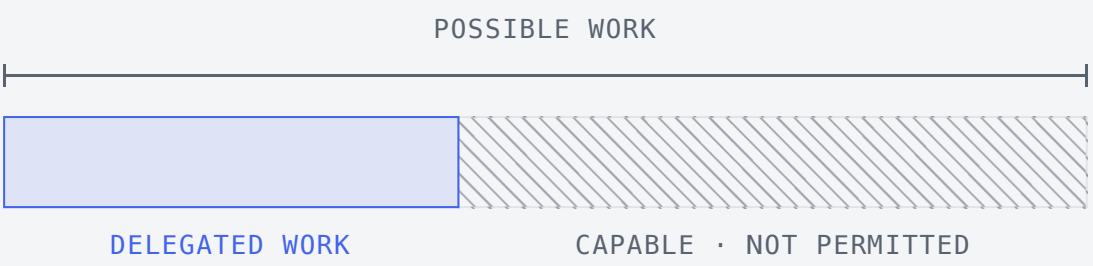
Responsibility is the constraint

Our bet is that model capability will advance faster than institutions learn to assign it responsibility.

A demonstration can make new work seem possible almost immediately. Delegating that work requires a deeper understanding: the consequences of an error, the exceptions that matter, the judgment a reviewer contributes, and the evidence that would justify changing the arrangement.

Keeping that basis available lets the next decision build on what the organization has already learned.

Possibility is not permission.



Capability can expand without the boundary of delegated work expanding with it.

Better models, evaluation tools, and ordinary approval rules may be enough for narrow, stable work. A separate competence record has to earn its cost; if it only adds another review step, it has failed.

The need becomes harder to meet when releases change, decisions cross teams, or an outcome is later reversed. Someone has to establish which evidence still applies, where a person supplied missing judgment, and why one kind of work warrants more oversight than another. Keeping that basis available lets the next decision build on what the organization has already learned.

We think the organizations that use AI most effectively will become unusually good at allocating responsibility. They will be able to explain why a system can own a particular kind of work, notice when that explanation stops holding, and change its authority accordingly.

Autonomy belongs to the work

Asking whether an agent is trustworthy compresses too many different questions into one answer.

A support agent may handle ordinary refunds reliably and still need review when a request includes a complaint or an account change. Its competence depends on the work, the tools it can use, the policy in force, and the conditions it has encountered. The model's name alone does not specify those conditions. A new tool, permission, or retrieval source can change what earlier evidence tells us.

An untested task has not failed because it is unknown.

The useful unit of autonomy is a class of work under stated conditions.

This changes how we think about progress. An agent can earn responsibility in one area while remaining closely supervised in another. A new model can expand what is possible without immediately expanding what it is allowed to do. Some work may always warrant human judgment because of its consequences or because the organization chooses to retain that decision.

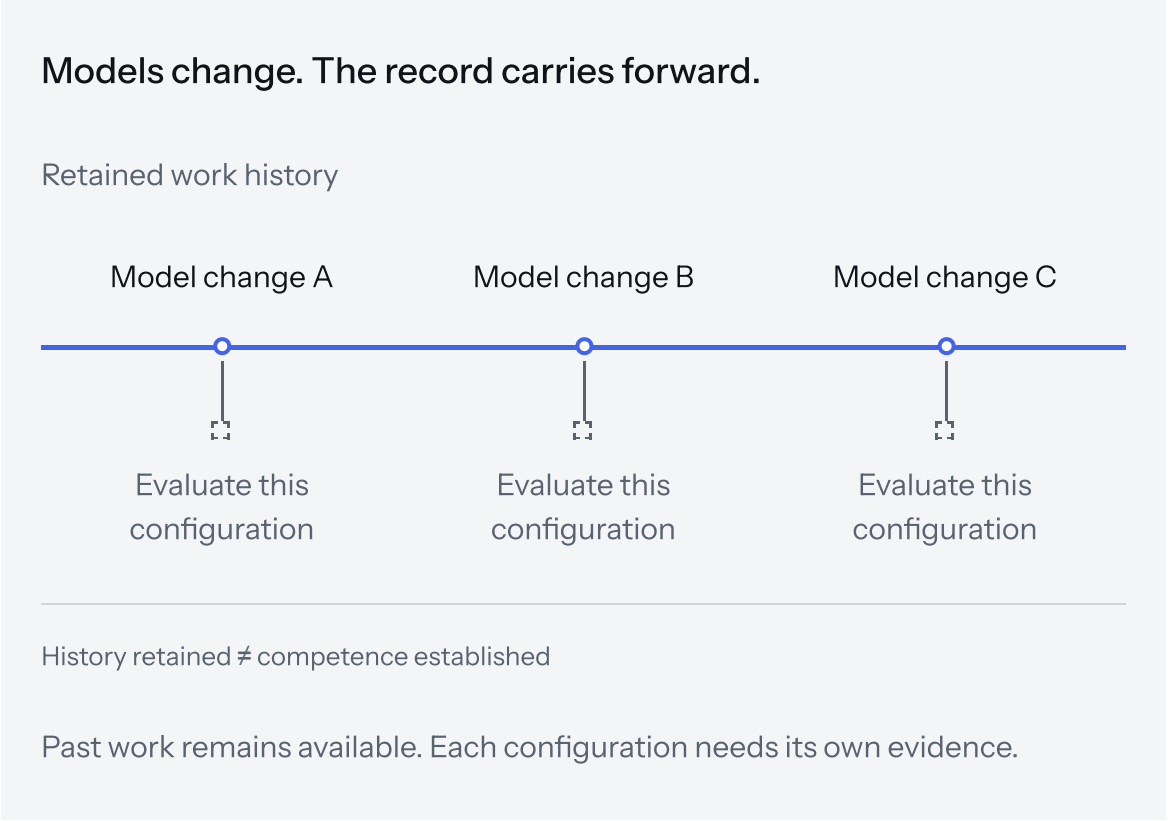
Keep demonstrated work, bounded competence, material failure, and missing evidence distinct. An untested task has not failed because it is unknown. A strong average cannot make a consequential failure disappear. These distinctions tell a team where responsibility could expand and what still needs to be established.

Experience should accumulate

Every correction contains something an organization might need to remember.

A reviewer catches an exception. An apparently successful action is later reversed. A tool change makes a familiar workflow unreliable. These events can teach us where competence holds and where it breaks down, provided we retain the connection between what the agent did, what a person changed, and what eventually happened.

We want experience to become a lasting asset of the organization.



The outcome alone is insufficient. A task may have succeeded because a reviewer repaired the answer. A record of approvals may reflect easy cases while difficult ones were diverted elsewhere. More observations do not automatically produce stronger evidence; the circumstances that produced them matter.

We want experience to become a lasting asset of the organization. A model replacement should not erase the failures people have already investigated or the

conditions they have learned to respect. The history should remain available, while conclusions that depend on the old system are reconsidered.

Useful organizational memory includes the conclusion, its evidence, the decision that followed, and the change that reopened it. A new release can inherit the questions people have learned to ask. It still needs evidence for its own answers.

Authority must remain revisable

Delegation should be easy to reconsider when the evidence changes.

An organization needs a practical way to narrow responsibility after a failure, require review when a policy changes, or return unfamiliar work to an unproven state. The same discipline should support expanding autonomy when the evidence warrants it. Reversibility makes it possible to grant useful responsibility without treating every grant as permanent.

The customer decides authority, and the customer's systems enforce it.

This also gives human review a more valuable role. A correction can improve the next decision about delegation. An escalation can reveal a class of work that needs a different boundary. Review becomes part of how the organization learns, while remaining in place wherever judgment is still required.

Evidence informs these decisions; it cannot determine which consequences an organization is entitled or willing to accept. Strong performance may still require human approval because the commitment belongs to an accountable person. The customer decides authority, and the customer's

systems enforce it. Self should make the evidence, limits, and basis for a recommendation inspectable.

Why we are building Self

Self begins with the relationship between work and responsibility.

We are building competence infrastructure that connects agent behavior, human judgment, and business outcomes into a living record tied to specific work and configuration. The immediate purpose is to make evidence, limitations, and the basis for an oversight recommendation available to the people deciding what agents can own.

Better models would meet an institution that remembers what it has learned about using them.

Developers need to understand what to investigate next. Enterprises need a basis for delegation that can be examined and revised. Whether evidence transfers across settings, and when less supervision helps, remain questions to test.

Our longer ambition is an organization whose ability to delegate improves with its experience. In that future, people could direct more work while concentrating their judgment on objectives, unfamiliar situations, and consequential commitments. Better models would meet an institution that remembers what it has learned about using them.

We want each new decision to carry forward useful experience while remaining open to new evidence. The work ahead is to establish which records, evaluations,

and forms of oversight make that possible in practice.

© 2026 Self

Figures are illustrative. Your policy sets authority.

San Francisco

Logos provided by Logo.dev

Brand credits +