SENTINEL RESEARCH GROUP
ARCHIVE OF RECORD // THE VAULT AT MERIDIAN

OPEN

SEN-23-041

Technical Memorandum: Reversibility in Deployed Systems

DATED	April 27, 2023	TYPE	Technical memorandum
ORIGIN	Directorate IV	SITE	--
PAGES	7	CUSTODY	MERIDIAN

How the Reversibility Principle is actually enforced in engineering practice, including the four tests a system must pass before it is allowed to exist outside a laboratory.

PURPOSE

The Reversibility Principle is easy to state and hard to enforce. This memorandum documents how the institution actually enforces it, so that the principle is auditable rather than aspirational.

The principle, from the Continuity Charter:

> We do not deploy what we cannot withdraw.

THE FOUR TESTS

No system may operate outside a laboratory until it has passed all four. They are applied by a reviewer who did not build the system.

1. The stop test

There is a documented procedure that halts the system, and it has been executed against the running production system within the last twelve months.

Not a design. Not a runbook nobody has followed. An execution, with a date and a name against it. A stop path that has never been used is a hypothesis.

2. The custody test

Every party capable of preventing the stop is enumerated. If a vendor, a hosting provider, a certificate authority or a departed employee can prevent

withdrawal, the system fails until that dependency is removed or a documented path around it exists.

This is the test that fails most often, and it usually fails on something mundane: a signing key held by one person, a contract with a notice period longer than the harm window.

3. The residue test

What remains after withdrawal is described in advance: data written to systems we do not control, models trained on our material, physical hardware in other people's buildings, behaviour other parties have come to depend on.

A system that can be switched off but has permanently changed something else has not been withdrawn. It has been abandoned.

4. The rehearsal test

The stop is rehearsed on a schedule, and the rehearsal is allowed to fail.

A rehearsal that is planned to succeed teaches nothing. Directorate IV runs these unannounced, within agreed windows, and the failures are the point.

WHAT THIS COSTS

The principle is expensive, and the memorandum records the cost honestly.

Between 1974 and 2023 the institution closed four programmes that were technically sound solely because they could not pass the custody or residue tests. Two of those were, in the reviewers' judgement, more capable than anything we have fielded since.

We also spend a substantial minority of engineering effort on paths that exist only to be used in failure. In 2022 that figure was 31% of Directorate IV hours.

We think this is the correct trade and we do not pretend it is free.

A NOTE ON THE WORD

"Reversible" does not mean "harmless" and it does not mean "temporary". A reversible system can do a great deal, for a very long time. It means only that the people who put it into the world remain able to take it back out, and that this ability is tested rather than assumed.

The distinction matters because the failure mode we are guarding against is not malice. It is drift: the ordinary process by which a system becomes load-bearing for parties who never agreed to depend on it, until withdrawing it would cause more harm than leaving it running. At that point the choice has been made for you, and it was made by nobody, gradually.

End of record SEN-23-041. Reproduced from the Archive of record at MERIDIAN. Handling markings are internal publication conventions of Sentinel Research Group and correspond to no external scheme.