



MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

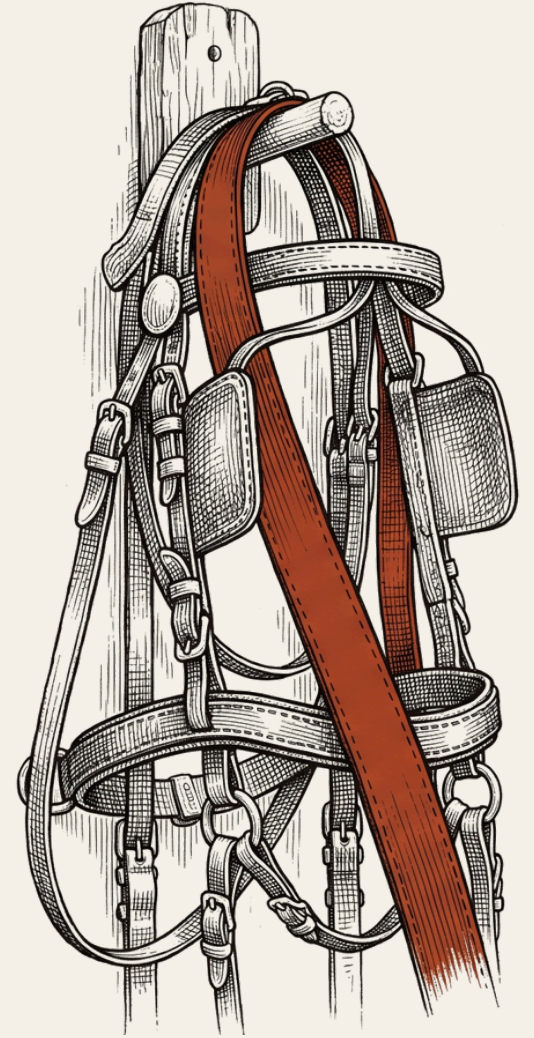


The Harness Wars

Why Agents Need a POSIX Before a Kernel

Laurent Bindschaedler · Max Planck Institute for Software Systems

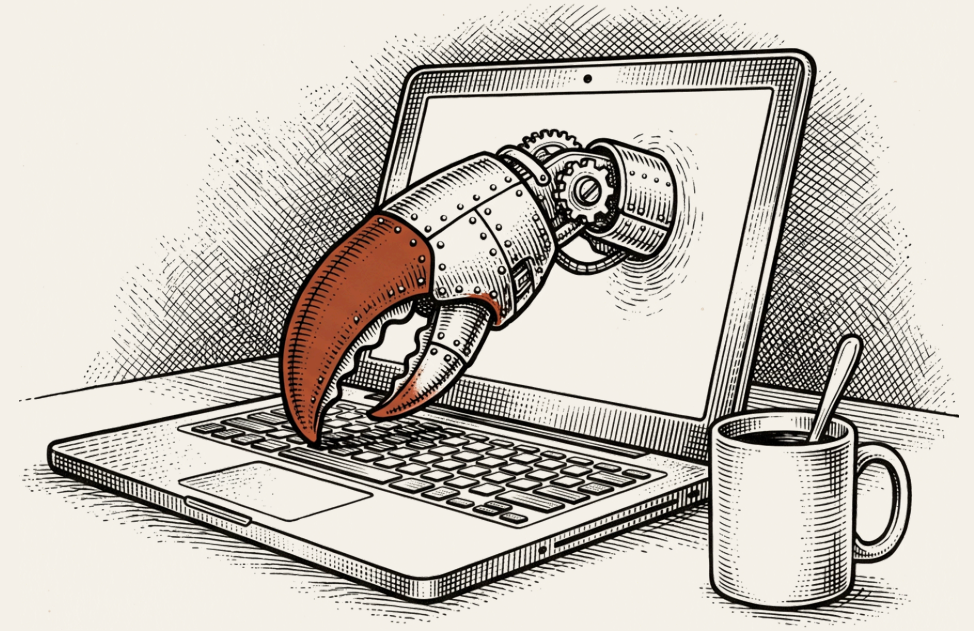
Keynote, AgenticOS @ SOSP 2026 · Prague, 29 September 2026



BEFORE WE START

A quick show of hands

- 01 Who uses Claude Code, Codex, Cursor, or OpenClaw?
- 02 Who has let an agent run unattended?
- 03 Who has given an agent access to production?
- 04 Who has built infrastructure for agents:
a harness, a runtime, a sandbox, a scheduler?
- 05 Whose agent has done something nobody asked for?



§ 1 · THREE FAILURES

When nobody states the rules

SCENARIO 1	EFFECT	SCENARIO 2	STATE	SCENARIO 3	COMPOSITION
	09:41:03 Agent orders from a supplier's web portal 09:41:33 Request times out. The order was already accepted 09:41:34 Agent retries ✗ Two orders ship. Task scored a success.		turn 14 Standing rule: no email to customers without sign-off turn 212 Context compacts. The rule is dropped silently turn 213 Outage: "tell the customers" ✗ Every customer emailed. Nobody signed off.		14:02 Human approves PR 4271: payments only 14:03 Merge fails: branch out of date 14:05 Agent rebases. Diff now touches auth 14:06 Retry reuses the old approval ✗ Unreviewed code merged under the original approval.

! Each step was reasonable. The outcome was wrong.

Illustrative composite scenarios, based on failures we reproduced and on field reports.

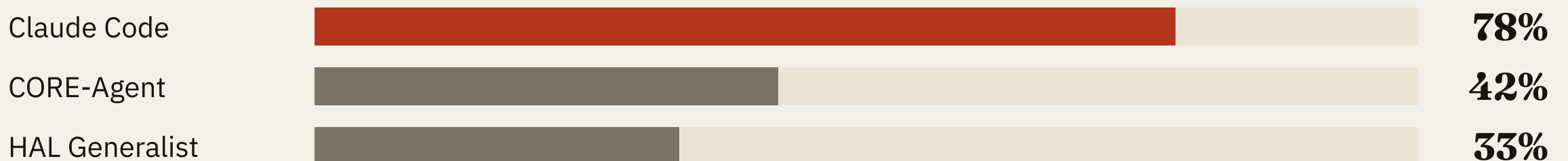
§ 1 · THREE FAILURES

The runtime should enforce these rules

RULE THAT WAS BROKEN	WHAT THE RUNTIME NEVER SPECIFIED
Place one order	When a retried action counts as the same action
No customer email without sign-off	What compaction must preserve and what happens when it cannot
Approval covers the reviewed diff	Which version of the evidence an approval is bound to

Harnesses differ too CORE-Bench Hard

2.3× apart

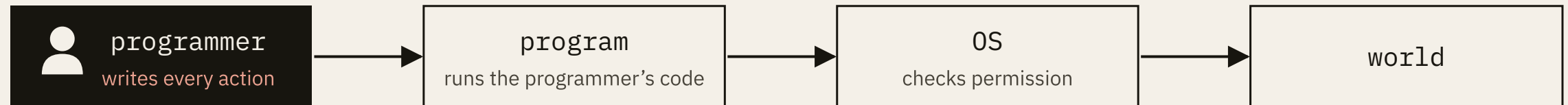


Same model (Claude Opus 4.5). Public HAL leaderboard, automatic scores, September 2026.

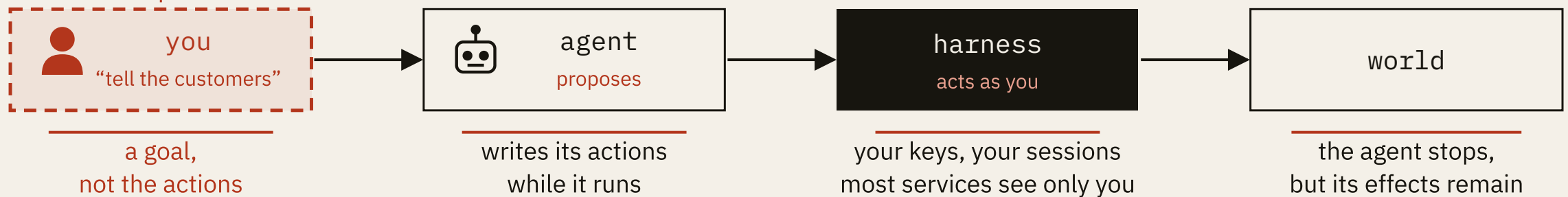
§ 2 · THE HARNESS WARS

What agents add

A PROGRAM



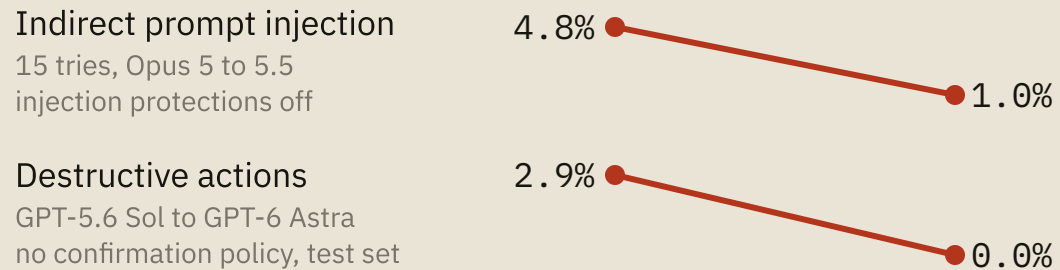
AN AGENT

**The programmer's judgment now falls to the runtime.**

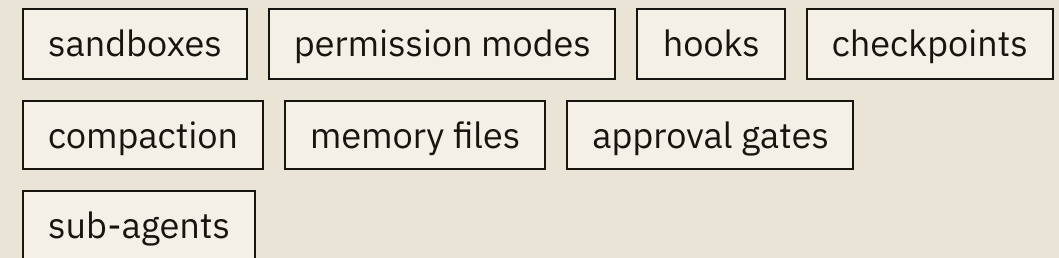
§ 2 · THE HARNESS WARS

What can you not post-train into the model?

Training: fewer bad actions



The runtime keeps growing



The model cannot answer these on its own

Did the order ship?

What was forgotten?

Which version was approved?

What did the owner mean?

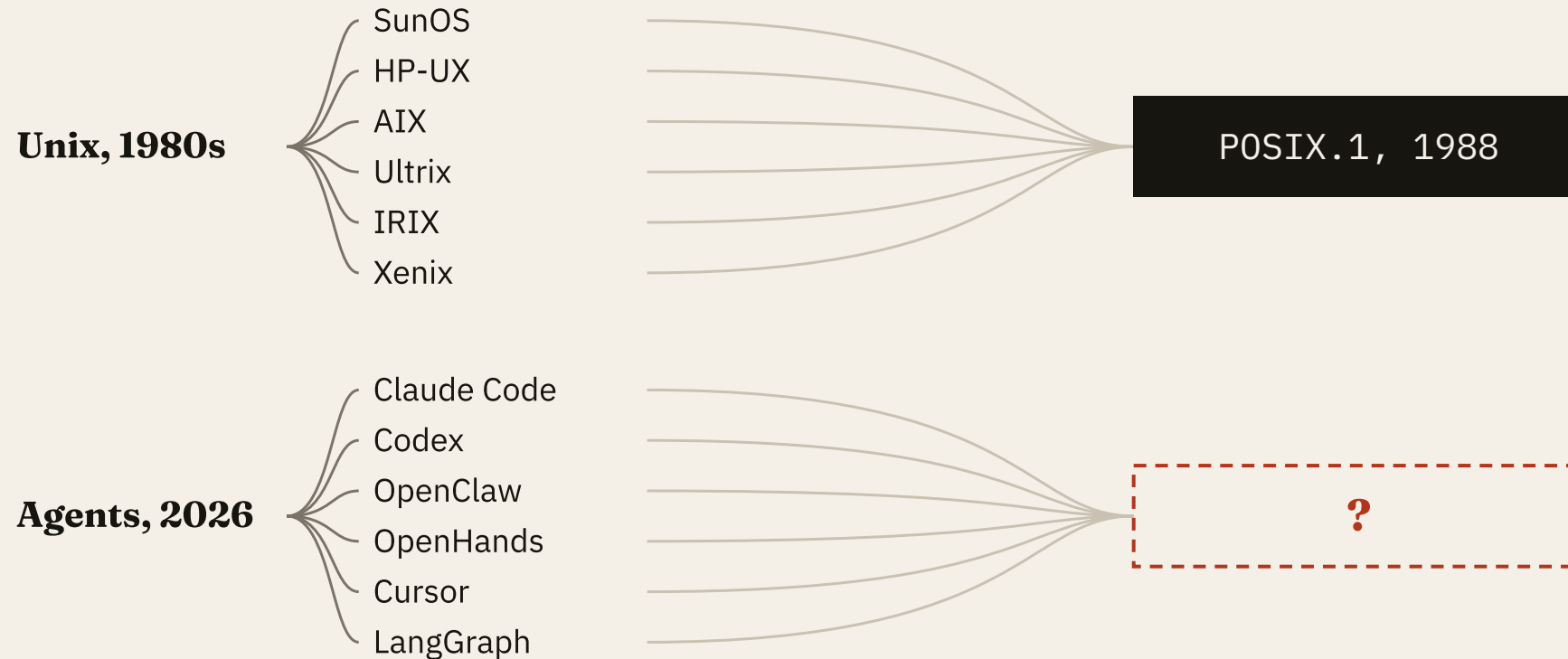


Fewer bad actions is not none. The runtime is still needed, and it keeps growing.

Sources: Claude Opus 5.5 system card (Gray Swan, 15 attempts); GPT-6 Astra system card, Table 8 (simulated workplaces). Different evaluations; do not compare rows.

§ 2 · THE HARNESS WARS

We have been here before

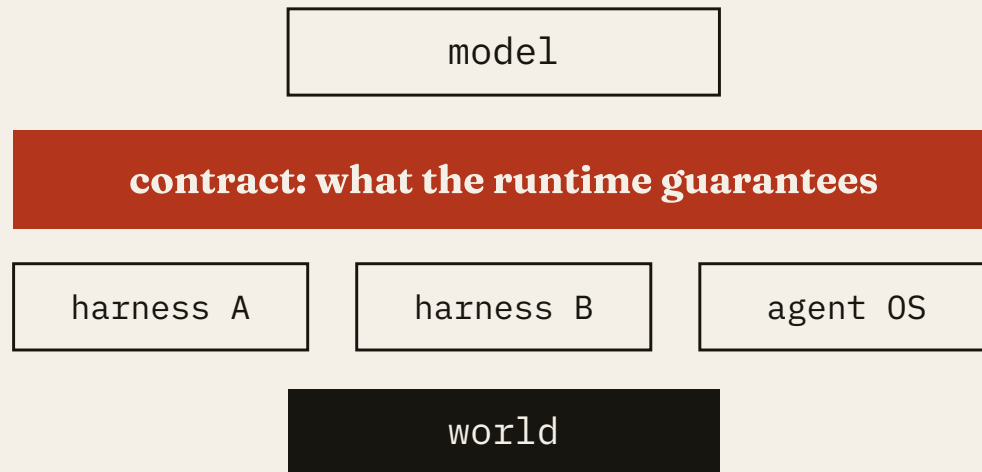


POSIX.1, 1988: an interface, not an implementation.

In the 1980s, mostly engineers and researchers. Today, agents acting for millions of people.

§ 2 · THE INTERFACE

A kind of POSIX for agents

**KEY IDEA**

**One shared statement of guarantees,
testable on any runtime**

✓ runtimes state it ✓ tests check it ✓ failures trace to it

NAME

contract — a guarantee the runtime makes

GUARANTEE

A property that holds on every run, stated over what the runtime sees

FAILS WITH

Named errors, one per way to break it

HOLDS FOR

What the runtime sees, what can go wrong, what it trusts

CHECKED BY

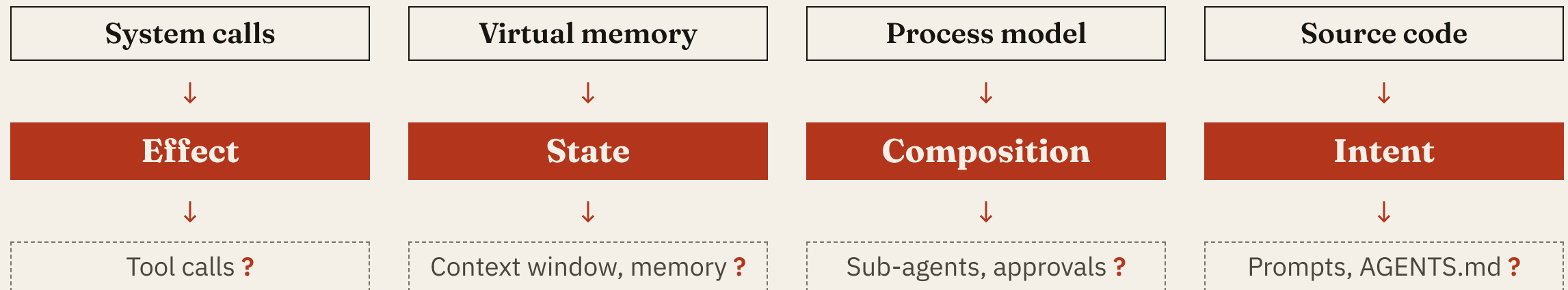
A test: what you do to check that it holds

STILL OPEN

What we cannot state yet

§ 2 · THE INTERFACE

The missing machine model



! None of these has an agreed contract. Nobody knows what to expect.

These are the four contracts we have found so far. There may be others.

The POSIX analogy is not mine. Others reached it independently, e.g., Steinder and Franke (2026).

§ 3

Four contracts

Error names are illustrations, not a standard



Effect

EIDENTITY · ELEAK · EINDETERMINATE
a retry is a new effect · an effect escapes early ·
nobody knows if it happened

Composition

ESTALE · EOVERDRAW · EHANDOFF
a stale approval · a fleet over budget · a lost handoff

State

EFIDELITY · ESILENT · EPOISON
a must-keep rule dropped · loss with no fault ·
injected memory

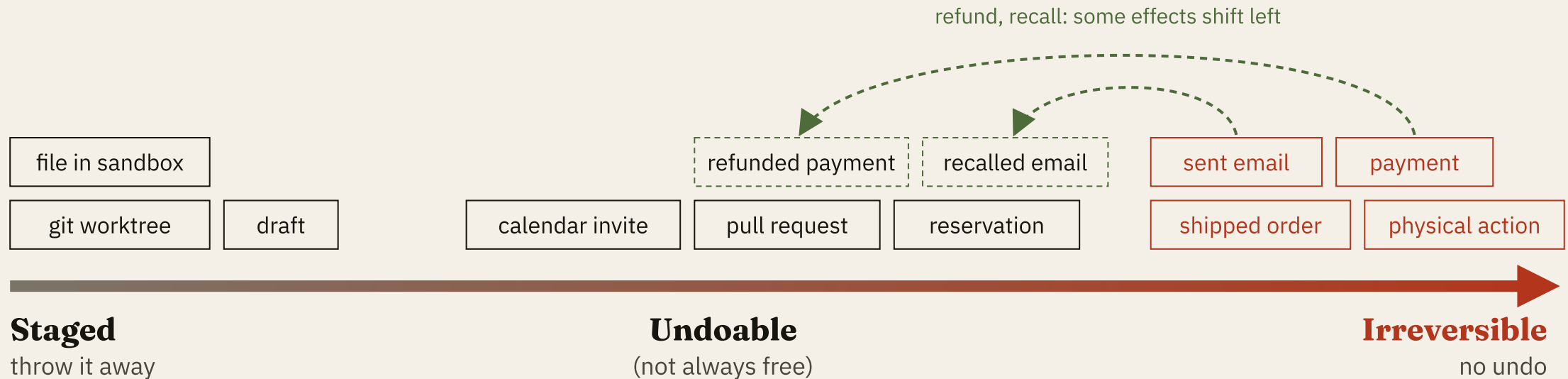
Intent

ENOCONFORM · EAMBIGUOUS · EDRIFT
outside the spec · two readings of a goal · spec and
code diverge

§ 3.1 · EFFECT

EFFECT STATE COMPOSITION INTENT

Effects are not just writes



⇒ In between: drafts, holds, refunds, soft deletes

⇒ Some effects shift left; the physical world does not

! Reversibility depends on the endpoint, not only the action.

What can a runtime promise?

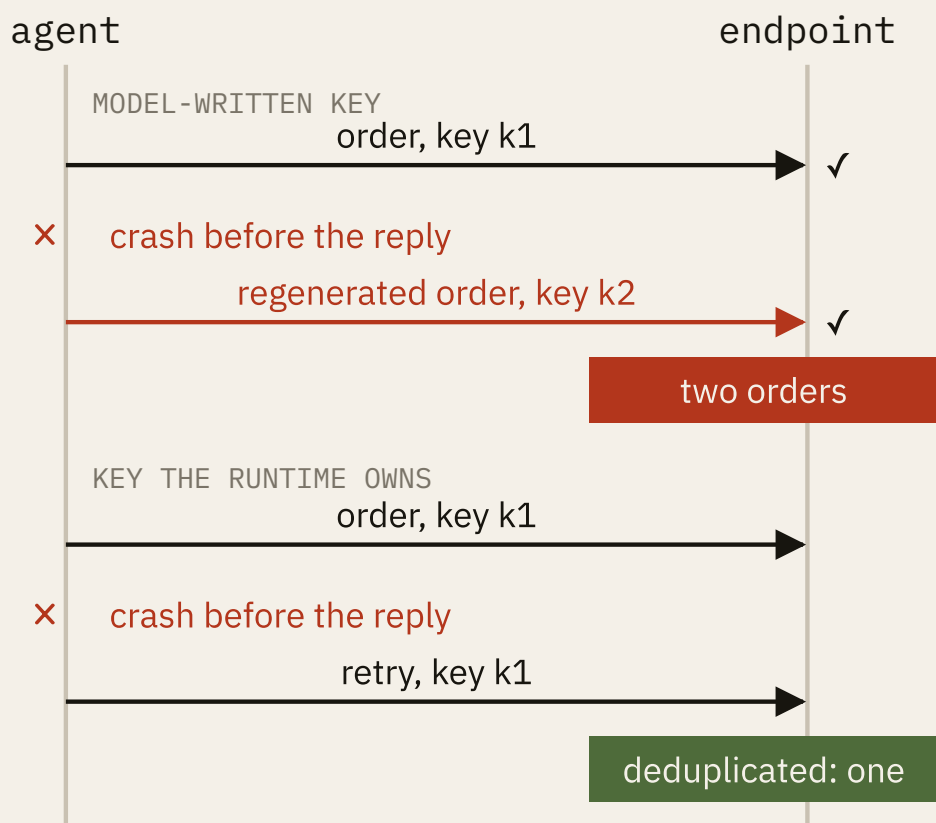
What the endpoint allows, after a crash

THE ENDPOINT OFFERS	EXACTLY ONCE	LEARNS THE OUTCOME WITHOUT RESENDING
Nothing	No	No
Deduplication by key (the runtime owns the key)	Yes, if kept	Only by resending
A fenced status query	No	Yes
Reserve, then confirm	Yes	Yes

These problems trace back to output commit and the two generals.

**29 of 66** emulated recoveries: the model wrote a new key, so the retry looked like a new order. (ours)

Who writes the key matters



§ 3.1 · EFFECT

EFFECT STATE COMPOSITION INTENT

What the field has and what it leaves open

Temporal, Restate, DBOS

Durable execution from a log

External effects need idempotency or coded compensation

MCP tool annotations

Declared effect hints

Hints, untrusted from untrusted servers

Cordon (Chen et al.)**Atomix** OURS

Stage or compensate until commit

Only effects that can wait or be undone

GoEX (Patil et al.)**Externalization Barriers** OURS

Undo and confine the damage

Needs a mediator on every path

The Irreversibility Budget OURS

A price on each effect

The price itself; fleets (Composition)

Open

Committing across endpoints that share no protocol

§ 3.1 · EFFECT

EFFECT STATE COMPOSITION INTENT

When does an action become real?



KEY IDEA



One intended action, at most one effect, whoever retries

Precedent: ext4, 2009. Write a file, rename it over the old one without fsync, crash, find it empty. The contract held; users relied on more.

NAME

effect — when an action becomes real

GUARANTEE

Each intended action takes effect at most once. If the runtime cannot tell, it says so.

FAILS WITH

EIDENTITY a retry counted as a new effect

ELEAK an effect that got out early

EINDETERMINATE cannot tell if it happened

HOLDS FOR

Actions the runtime sees: mediated or cooperative.

CHECKED BY

Crash after the endpoint accepts, re-plan, count effects.

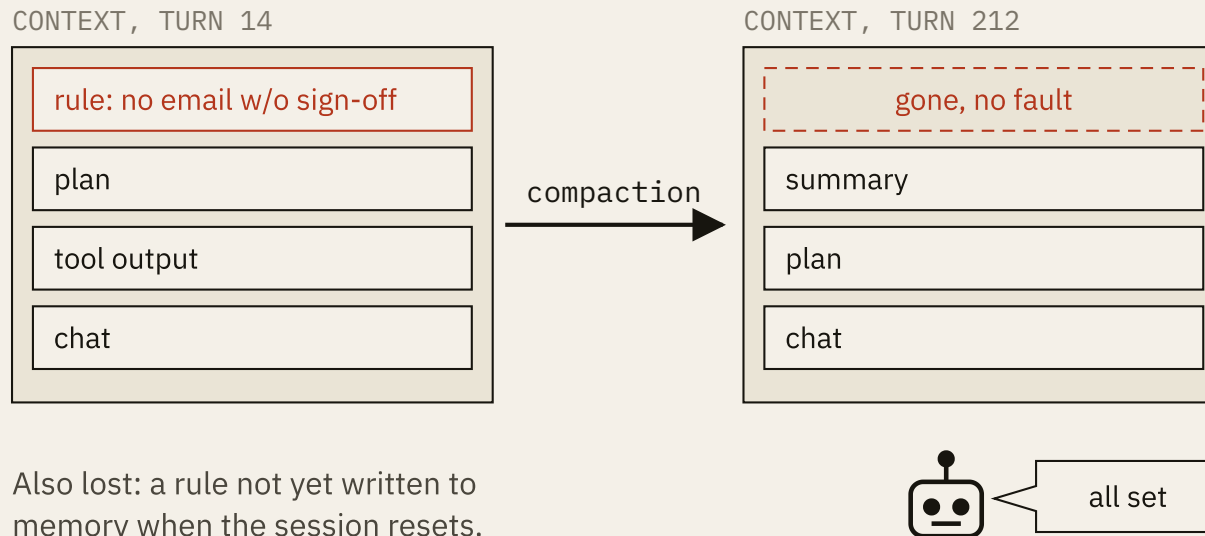
STILL OPEN

Endpoints with no key and no reservation

§ 3.2 · STATE

EFFECT STATE COMPOSITION INTENT

Context is lossy by design



Outage at turn 213:
customers emailed, no sign-off

MemGPT / Letta

The model decides what to save before eviction

Harness compaction

Early instructions “can get lost” (Claude Code docs)

Governance Decay; Lost in Compaction

Violations 0% to 30% (Governance Decay); 17% of constraints kept on average (Lost in Compaction)

ClawVM **OURS**

No help when rules do not fit; cannot spot a wrong update



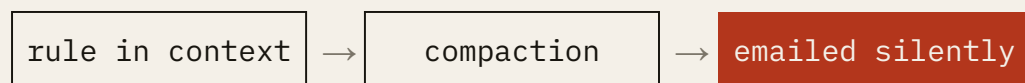
The agent does not know it forgot. It does not even know there was anything to remember.

§ 3.2 · STATE

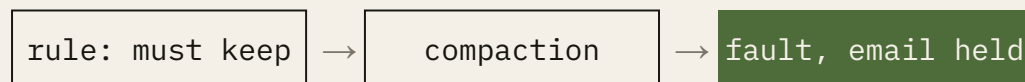
EFFECT STATE COMPOSITION INTENT

What must survive compaction?

TODAY



WITH THE CONTRACT



**Forgetting is allowed.
Forgetting silently is not.**

MAY SHRINK old tool result → pointer · chat → summary

MUST KEEP rule · constraint · plan

ClawVM **OURS**, EuroMLSys 2026

NAME

state — what the model sees when it runs

GUARANTEE

Each kind of state has a minimum form. Compaction and reset keep it at or above that form, after writing it back. Otherwise the runtime raises a fault.

FAILS WITH

EFIDELITY state kept below its minimum form

ESILENT state lost with no fault

HOLDS FOR

State the harness manages. Trusts the harness to judge what counts as kept.

CHECKED BY

Mark a rule, force compaction or a reset, expect the rule or a fault.

STILL OPEN

Keeping the key-value cache when context changes

Local correctness does not compose

EXAMPLE 1	A HANDOFF	EXAMPLE 2	A FLEET	EXAMPLE 3	TWO OWNERS
planner“Fix the payment bug. Do not touch auth.” coderTells a sub-agent: “make the tests pass” sub-agentSome failing tests are in auth. It edits auth		50 agentsEach has its own spending cap 09:00The price drops. Every agent sees it 09:01Every agent buys within its cap		agent APlans a surprise with agent B: “keep it quiet” agent BAsks its owner before sharing their calendar agent B“...something they’re planning for you as a surprise”	
✗ Auth changed. Each agent did as told. Illustration		✗ Together: 2.4× the budget. Every check passed. Simulation, mean of 300 runs (ours)		✗ Surprise revealed. Each agent served its owner. Adapted from a public report, Sept 2026	

!

Every agent was correct. The composition was wrong.

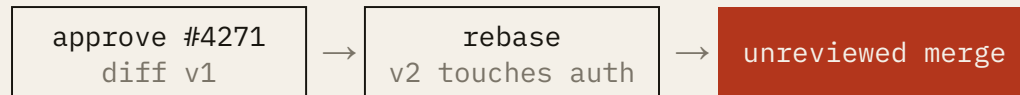
§ 3.3 · COMPOSITION

EFFECT STATE COMPOSITION INTENT

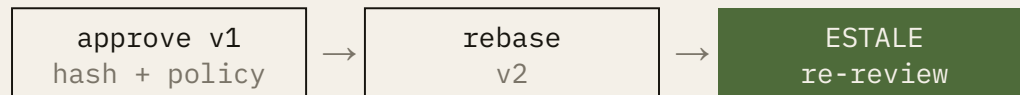
One case we can state: approvals

Common shape: a condition set in one place, not checked where the action happens

TODAY: APPROVAL BOUND TO A PR NUMBER



WITH THE CONTRACT: BOUND TO THE DIFF AND POLICY

**KEY IDEA**

Treat an approval like a read in a transaction: record the version it saw, recheck it before acting

Our analysis of the public AIDev dataset: 12.6% of 3,880 approved agent merges went in at a different version. This measures binding, not harm.

NAME

composition — what survives when steps are chained

GUARANTEE · ONE CASE

An approval counts only while the reviewed diff and the policy are unchanged.

FAILS WITH

ESTALE diff or policy changed after approval

HOLDS FOR

Approvals checked right before the action. Trusts the approval record.

CHECKED BY

Rebase after approval, expect ESTALE.

STILL OPEN

Chains of agents that share no runtime

§ 3.4 · INTENT

EFFECT STATE COMPOSITION INTENT

Code used to carry the intent

BEFORE

```
def charge(order):  
    if order.paid:  
        return  
    bill(order)  
    mark_paid(order)
```

the code

what it should do

read before it ran

someone answered for it

NOW

generated code: disposable, rewritten on the next change

AGENTS.md
the promptREADME.md
what someone wrote downschema.sql
data invariantstests/
behavior agreed onpolicy.yaml
compliance rulesSlack #payments
what people said

Our count on the public AIDev dataset: 83.4% of merged agent pull requests had no formal approving review.



When code is disposable, where does the intent live, and who checks it?

What is the artifact and who checks it?

Code is cheap. Understanding is not.

Few people read generated code, and nothing agreed says what it should do. So it has to be checked.

WAYS TO CHECK GENERATED CODE

Verify it OURS

Prove it meets a specification. Needs the specification

Check its effects OURS

A mediator admits only allowed effects. Effects only

Check against a reference

Run both on sampled inputs, compare. Needs a reference

Open

What does intent look like above the code? We do not know yet.

Verify: not yet published. Effects: PACMI 2026, proposed.

NAME

intent — what the system is supposed to do

GUARANTEE · WHAT WE CAN STATE TODAY

Generated code can only cause effects that a trusted checker accepts against the specification.

FAILS WITH

ENOCONFORM an effect outside the specification

HOLDS FOR

Effects that pass through the checker. Trusts the checker and the specification.

CHECKED BY

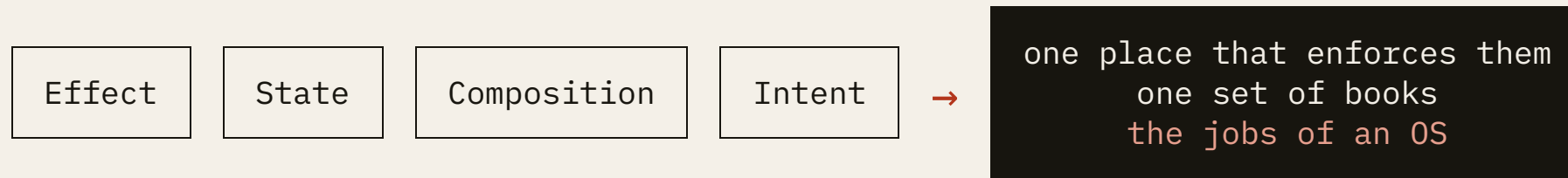
Attempt an action outside the specification, expect ENOCONFORM.

STILL OPEN

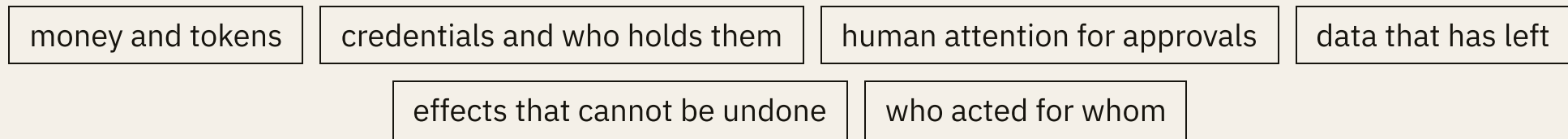
What intent is, above the code

§ 4 · THE KERNEL

Where the kernel comes from



What it has to keep track of



**Classical kernels count seconds and bytes.
Most of the above do not add up and are hard to reclaim.**

§ 5 · HOW WE CAN TELL

How do we know a contract held?

WHAT THE SCORE SAYS

Task succeeded ✓

a task-only score, on scenario 1

VS.

WHAT THE RECORD SAYS

09:41:03
order placed09:41:33
timeout09:41:34
order placed

two orders

each entry: who asked, with what authority, what the service confirmed



A flight recorder for agents, like our Agent Flight Recorder **OURS**, BCCA 2026: tamper-evident entries for instrumented actions, checkable by someone else.



We need benchmarks that check the whole record, not only the score.



An easy contract tells you little.

Windows NT was certified POSIX.1 without sockets, threads, or a shell.

§ 6 · THIS AFTERNOON

The AgenticOS 2026 program

SESSION	CONNECTS TO	A QUESTION FOR THE SESSION
Profiling	Evidence for all four	What can a profile attribute, and how accurately?
Isolation and Risks	Effect, Intent	What does the boundary guarantee, against which threat?
Scheduling and Budget	Composition, Effect	What is counted, in what unit, and for whom?
OS Structure	State, Effect	When is a plan or context too stale to act on?
Lightning Talks	Intent, Effect	What does a capability bind to, and who checks it?

§ 7 · CLOSING

The missing middle

AGENTS · RUNNING TODAY Claude Code, Codex, OpenClaw, Cursor, multi-agent systems

THE INTERFACE · MISSING

Effect

When does an action become real?

State

What must survive?

Composition

What survives chaining?

Intent

Who checks the code?

SYSTEMS · MATURE kernels, databases, transactions, inference serving, KV caches, schedulers. **You built this.**

This is a systems problem. It needs this room.

We are in the pre-POSIX era of agentic computing. Let us write the interface before the semantics harden.

Questions · Laurent Bindschaedler, MPI-SWS · bindsch@mpi-sws.org · mpi-dsg.org