

RESEARCH PAPER

Capability Boundaries, ALLOW / HOLD / DENY, and Compositional Safety

A grant is not a key. A prompt is not a grant. Two safe steps are not automatically a safe third step.

THECRINERS LLC · 2026

Problem

Tool-using agents fail in a boring way: the tool exists, so they use it. Capability boundaries are how we make “the tool exists” insufficient.

The catalog

Defaults in Company OS: tick, backup, dashboard, public intake, internal produce, and verify may allow. Publish and spend hold. Email notify, marketplace, and transfer deny. A later, narrower grant — `web_publish` — exists so a company can maintain an already-approved public site without unlocking unrestricted publish, discounts, or legal identity changes.

`ALLOW` runs. `HOLD` queues for an owner. `DENY` refuses. Unknown grant strings are not treated as allow. An explicit empty capability map fail-closes.

Compositional safety

Intake can be ALLOW while publish is HOLD. That is the point. A worker may fetch a public PDF and still be unable to put it on the internet. Approving the HOLD row does not run publish; it records that an owner clicked. Someone still has to perform the granted action under the current grant.

Restore plus “install whatever was in the zip” is how you launder a vault. Restore therefore does not install secrets. Event ingest plus “echo details on /status” is how you leak a webhook token. Details are redacted.

What web_publish is not

It is not publish: ALLOW. Workers may correct copy, layout, FAQs, samples, links, accessibility, approved prices already in the spec, and leak-stripping. They may not create legal commitments, change entity identity, expose internals, change payment destinations, enable spending, invent discounts, or make unsupported compliance claims.

Tests

Adversarial tests try to self-approve spend, treat empty maps as ALLOW, execute HOLD publish, and echo secrets from events. Those must keep failing closed.

Lessons

If you only have DENY, nothing ships. If you only have ALLOW, you will ship the wrong thing at 3 a.m. HOLD is the adult grant. Narrower ALLOWs (web_publish) exist so HOLD does not become a bottleneck for work that is already in policy.

HTML is canonical. A PDF companion is offered where generated. Experimental laboratory results are labeled as such. This is not a product-market-fit claim and not a certification.

[← All papers](#) · [PDF](#)

TheCriners LLC · Coshocton, Ohio

Money Hunter, Packet Press, Pullsheet, Hearthlabel, Seatpack, and Company OS are projects of TheCriners LLC — not separate legal entities.