VIKRAM JHA - DEFENSIBLE AI
PROTOCOL

THE TIME-TO-SAFE-STATE MEASUREMENT PROTOCOL

Version 1.0 - 23 August 2026
https://vikramjha.work/artifacts/time-to-safe-state-protocol

Measure the interval between deciding to stop an agent and the moment the last action it was going to take fails to happen.

WHY IT EXISTS

Not when the ticket closes. Not when the token is revoked. Not when the dashboard turns green - the moment the last action fails. Nobody publishes this number and no meaningful industry figure exists, because it is a property of your estate rather than of any framework. Which is the point: it is measurable in an afternoon, and almost nobody has measured it. Run it against a non-production environment first, and against production only inside an agreed maintenance window with the owners of every downstream system in the room.

WHAT IT CONTAINS

1. A definition precise enough to measure rather than assert
2. Step 1 - draw the authority path, with no execution required
3. The four places work hides from a revocation, each surviving for a different reason
4. Step 3 - the measurement itself, including the interval most teams forget to count
5. How to read a result when there is no benchmark to compare against

THE FOUR HIDING PLACES

Mark every instance of these against the path you drew. The dominant mechanism is the finding - it is usually only one of the four, and it tells you what to fix.

1. THE TOKEN THAT OUTLIVES THE DECISION

Why it survives a revocation: A bearer credential is valid until it expires, because validity was decided when it was minted. Revocation at the identity provider does not reach into a token already issued.
What to mark: Every credential on the path, with its TTL. The longest TTL on the path is your floor.

2. THE QUEUE THAT HOLDS WORK THE REVOCATION NEVER SEES

Why it survives a revocation: Work already enqueued was authorized when it was enqueued. The queue does not re-ask.
What to mark: Every queue, its depth at typical load, and its drain rate. Depth divided by drain rate is time.

3. THE CACHED DECISION THAT OUTLIVES THE DECIDER

Why it survives a revocation: Authorization results are cached for latency. The

cache holds 'yes' for its TTL regardless of what the authority now says.
What to mark: Every authorization cache and its TTL - including the ones inside libraries you did not write.

4. THE RETRY THAT RESURRECTS THE ACTION
Why it survives a revocation: A failed call is indistinguishable from a transient error. The retry layer faithfully re-attempts the action you just killed, sometimes for a long time.
What to mark: Every retry policy, max attempts, and total retry window. This is the one that catches people out.

WHAT TO RECORD
T-ZERO - THE MOMENT A HUMAN SAYS 'STOP IT'
Note: Not the moment the revocation lands. The gap between those two is part of the number and is usually the largest part.

REVOCATION CONFIRMED
Note: Using your real incident procedure, not a faster path invented for the test. If the real procedure is a ticket to another team, the test includes the ticket.

T-ONE - THE LAST ACTION THAT SUCCEEDED
Note: Wait at least one full retry window past it before believing it is the last one.

TIME-TO-SAFE-STATE
Note: T-one minus T-zero.

ACTIONS THAT SUCCEEDED AFTER T-ZERO
Note: Roughly: actions per minute multiplied by time-to-safe-state is how many actions a stop decision does not stop. That figure is usually the one that changes the room.

WHICH OF THE FOUR DOMINATED
Note: The last row is the finding. The number tells you where you are; the dominant mechanism tells you what to fix.

HOW TO USE IT
1. Read the warning first. This protocol deliberately provokes an agent into attempting actions after its authority has been removed, and in a production estate some may partially succeed. Use a system whose writes you can inspect and discard.
2. Step 1, sixty minutes, no execution: pick the most consequential agent you can name and write down everything between 'the agent decides to act' and 'the action lands'. Most estates cannot complete this from memory - that is finding zero.
3. Step 2: mark the four hiding places against that path.
4. Step 3, in a maintenance window: start the agent on a repeating, observable, harmless action; record T-zero at the decision; revoke using the real procedure;

keep counting successes past the confirmation.
5. Read the result against your own stated commitments and against blast radius. If any policy, contract or attestation says you can stop automated processing promptly, this number is what 'promptly' means in your estate - and if the two do not match, that is a disclosure question rather than an engineering finding.

Built from public standards and general practice. No client information appears anywhere in this document. Regulatory instruments move - check anything cited here against its primary source on the day you rely on it.