Nari and Responsible Scaling

What Anthropic's RSP v3.0 changes, and how Nari uses it for serious safeguarding

Anthropic's Responsible Scaling Policy v3.0 (24 February 2026) is a quiet but important shift in how a frontier lab signals seriousness. It is not a marketing flourish. It is an admission that safety cannot rest on slogans, nor on neat capability "thresholds" that everyone can agree on in public. Model capability is moving faster than evaluation certainty, politics is drifting toward competitiveness, and some mitigations cannot be delivered by one company acting alone.

Anthropic's response is to make safety look more like engineering and governance: publish a Frontier Safety Roadmap with concrete goals, publish regular Risk Reports that lay out threats and mitigations, and sometimes invite third-party review. In other words: show working, leave a trail, accept scrutiny. That is what mature safeguarding looks like when a system is powerful enough to matter and messy enough to be contested.

Nari is not a model lab. We do not train frontier models, we do not control weights, and we do not claim to solve state-level threats. But this shift is directly relevant to us because it clarifies a point many people still miss: at population scale, harm is often driven less by what an AI system can do in principle and more by how ordinary people interact with it in practice. As tools become more capable, more agentic, and more persuasive, the interface is not cosmetic. It is part of the risk surface.

Safeguarding professionals recognise the pattern immediately. Most failures are predictable human failures: over-trust, poor verification, misunderstanding of context, disclosure of sensitive information, delegation of judgement, and gradual dependence. These are not exotic edge cases. They are what happens when you put a blank prompt box in front of a tired parent, a teacher under pressure, a civil servant making decisions, or a teenager experimenting, and then call that "empowerment".

Nari exists to close that gap. We treat onboarding as a safeguarding discipline: not because we want to sound cautious, but because mass adoption without competence creates preventable risk. The product is designed to replace the blank box with structured starting points and to install basic habits that reduce predictable misuse-by-confusion.

Practically, we take advantage of the same logic that sits behind RSP v3.0, scaled to our reality:

First, we separate commitments from recommendations. We publish what we will do regardless (our commitments) and what we advise schools, parents, and employers to do around us (our guidance). That split is a form of honesty: one organisation cannot control the whole ecosystem, but it can be rigorous about its own obligations.

Second, we run a lightweight Risk Report rhythm. We publish short Trust Notes on a regular cadence: what we shipped, what we observed in real user journeys, what broke, what we tightened, and what risks remain. Safeguarding is not a certificate you wave once. It is an operating system you maintain.

Third, we design guardrails that teach judgement rather than create dependency. Early task packs explicitly train verification micro-routines: state assumptions, ask for uncertainty, request a counterargument, demand sources, and confirm with a second method when stakes rise. This is competence-building, not content-filtering theatre.

Fourth, we make the onboarding red-teamable. Task packs are written so that they can be systematically tested for the failure modes that matter: where novices freeze, where users over-trust, where prompt-injection patterns slip through, where people disclose too much, and where the system encourages brittle automation.

The result is a simple claim we can stand behind: as AI becomes more powerful, the first-use interface becomes part of safeguarding. Nari is the layer that turns AI abundance into competent use for ordinary people, with evidence, process, and accountability that safeguarding stakeholders can interrogate.

References

Anthropic. “Anthropic’s Responsible Scaling Policy: Version 3.0.” 24 February 2026. https://www.anthropic.com/news/responsible-scaling-policy-v3

Anthropic. “Responsible Scaling Policy (RSP) v3.0 (full policy).” https://anthropic.com/responsible-scaling-policy/rsp-v3-0