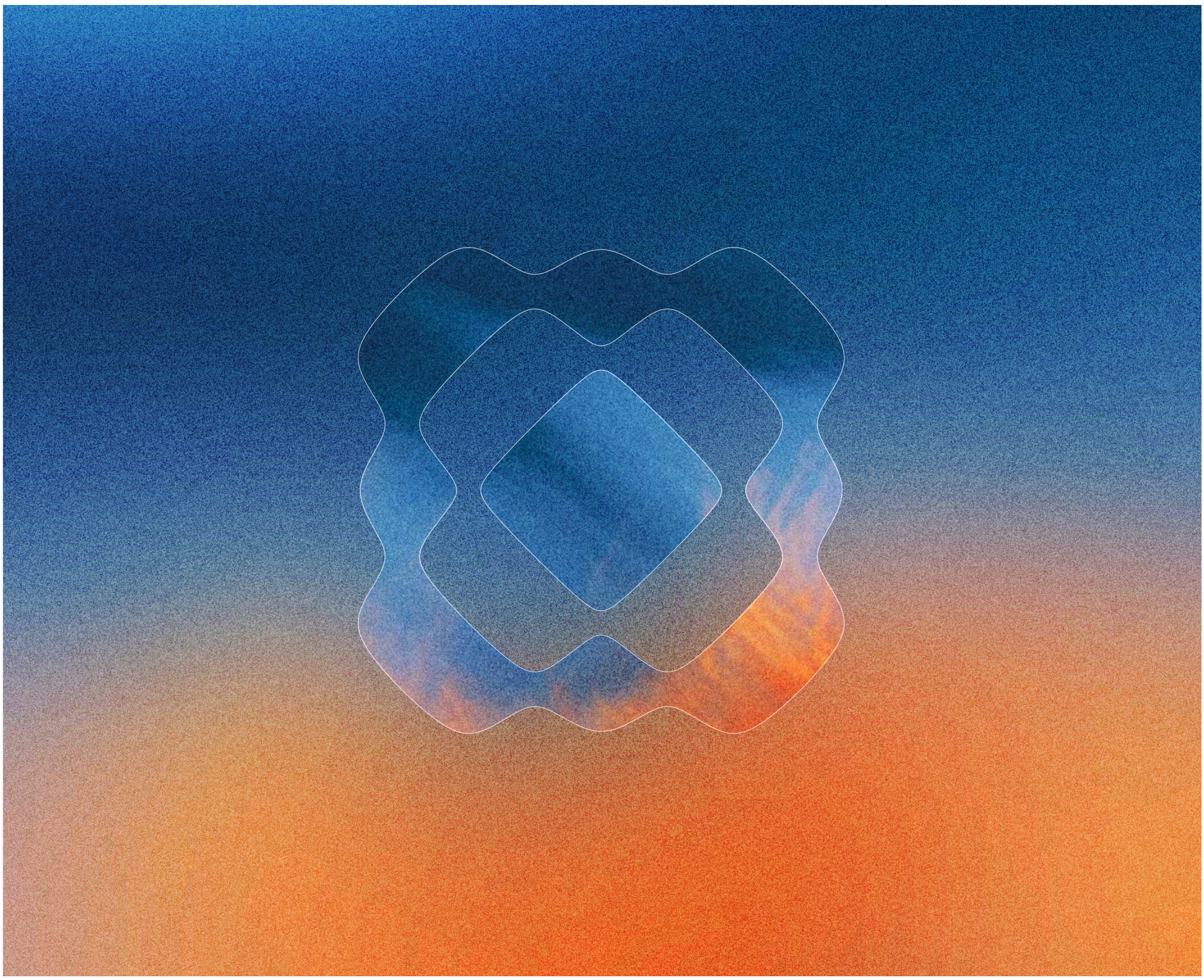
IIElevenLabs

CASE STUDY

ElevenLabs Leverages the NIST AI Risk Management Framework for Voice AI



What is the NIST AI RMF?

In collaboration with private and public sectors, the National Institute of Standards and Technology (NIST) developed a framework to better manage risks to individuals, organizations, and society associated with artificial intelligence. The NIST AI Risk Management Framework (AI RMF) is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.

Snapshot

ElevenLabs is an AI research and product company. It builds AI voice and audio models, creative tools, and conversational agents across 70+ languages. Millions of consumers, developers, governments, and businesses use ElevenLabs - including employees at two-thirds of Fortune 500 companies. Safety is inseparable from innovation at ElevenLabs, and we deploy comprehensive, multi-layered safeguards to prevent and address potential misuse of our technology.

Why we benchmarked against AI RMF

Much of that infrastructure already lived inside our ISO/IEC 42001-certified AI Management System, and across separate Security, Safety, Legal, and Policy workstreams, backed by a broader compliance portfolio that includes SOC 2 Type II, ISO/IEC 27001, and GDPR alignment. As our products deal with sensitive customer data and have moved into regulated industries (e.g. finance, healthcare, government) we wanted to benchmark our work against a widely adopted standard built for the full AI lifecycle: holistic enough to cover governance, deployment context, measurement, and continuous risk management end to end, and approachable enough to actually put into practice. The NIST AI RMF gave us that playbook, and a shared vocabulary for AI risk that our teams, customers, and external partners could all work from.

Mapping our controls to the four functions

Our Security, Risk & Compliance team reviewed all 72 subcategories of the AI RMF across the four functions (Govern, Map, Measure, and Manage) and matched each to evidence already in place. Here's what that mapping found, function by function:

Function	What we found
Govern	A defined AI Governance working group and AI Safety team hold accountability for AI risk, operating under our ISO/IEC 42001-certified AI Management System. Roles, escalation paths, including direct escalation of high-severity incidents to the AI Safety and Security lead, and sometimes to CTO/CEO, and annual role-based training are documented; vendor and subprocessor obligations are set contractually and reviewed annually.
Map	Every model family (our proprietary TTS, STT, Music and voice-cloning models) has documented intended use, known limitations, and a completed AI Impact Assessment with impact/likelihood scoring, reviewed by an interdisciplinary team spanning technical, operational, and policy expertise.
Measure	Models are evaluated against defined quality metrics (e.g. accuracy, reliability, performance), combining internal red-teaming, independent third-party evaluation, and continuous production monitoring. Findings are captured in model cards, benchmarking evaluations, and impact assessments.
Manage	Risk treatment is prioritized by impact and likelihood, backed by documented incident response, and change-management procedures. Third-party and pre-trained models are monitored under the same governance as systems we build ourselves.

Framework in practice: voice cloning example

Professional Voice Cloning is where the framework's abstractions turn into product decisions fastest; a cloned voice can authenticate, impersonate, or defraud in ways other types of synthetic media cannot. Mapped against the same four functions, one product looks like this:

- **Before a clone can be created (Govern, Map).** Users of the tool must acknowledge our policies and confirm they have all necessary rights or consents to clone a voice. Uploaded samples are also checked against our "No Go Voices" (NGV) safeguard, which is designed to block the creation of voices belonging to high-risk individuals such as public officials, political candidates, celebrities, and other widely recognized figures. We continually refine and expand NGV based on internal monitoring, current events, user reports, and external intelligence.

- **At the point of cloning (Map, Measure, Manage).** We treat unauthorized cloning (someone cloning a voice without necessary rights or consent) as a high impact misuse path for this category of the product, which is why for our more advanced capabilities, through Professional Voice cloning, we leverage Voice CAPTCHA as a safeguard: it tests the uploaded sample against a live recording of the user reading a dynamically generated script. Certain voices are either blocked outright or evaluated using a CAPTCHA flow. If the check fails the CAPTCHA flow, the clone is blocked from being created, including an additional manual verification process.

- **After a clone exists (Measure, Manage).** Outputs are monitored by automated classifiers and human review, in real time and after the fact; enforcement ranges from warnings to account bans based on the severity and volume of Prohibited Use Policy violations. Our self-serve systems are designed to trace generated content back to the account that created it, and our free AI Speech Classifier allows anyone to check whether certain pieces of audio came from our platform.

Governance evidence that stood out

One governance team and review process remains consistent throughout Govern subcategories, there is not a different committee per team, and a direct escalation path to the Head of Safety & Security for high-severity AI incidents exists.

Outside feedback arrives through several independent channels (reporting mechanisms, support tickets, an appeals process) and has dedicated owners to address incidents or feedback.

Red-teaming is graded across benign, manipulative, and adversarial use, not scored pass/fail, and repeats quarterly for agentic products as part of an external third-party certification.

Sustainability and third-party risk (EcoVadis, vendor due diligence) already sit inside the same AI Impact Assessment used for privacy, safety and security, rather than living on a separate track.

Because our ISO/IEC 42001 Statement of Applicability already lists controls clause by clause, mapping them to AI RMF subcategories was mostly cross-referencing, not new documentation, which is crosswalk work we intend to extend across more global frameworks.

> *“When you hear someone speak, you know it is them. Voice carries identity in a way that other AI modalities can't, and that means it has a unique risk profile. NIST's AI Risk Management Framework helps us cover a robust set of risks and thus ensure our systems are used safely and prevent misuse of voice AI.”*
>
> **Marco Mancini, Safety and Security Lead at ElevenLabs**

Where We're Taking This Next

As AI legal requirements and governance frameworks continue to mature and expand, the NIST AI RMF will remain a core benchmark for how we build and evolve our AI governance practices. Next, we plan to run crosswalks between the AI RMF and the other governance frameworks we track, so we can continue to find gaps and improve our program. Our goal is to keep applying that same lens as we release new products and make significant updates to existing ones.

Learn More

1. ElevenLabs Safety, our safety principles and multi-layered approach to misuse prevention
2. ElevenLabs Trust Center, certifications, audit reports, and security documentation
3. Prohibited Use Policy, what our products cannot be used for, and how we enforce it
4. AI Speech Classifier, our free public tool for checking whether audio was AI-generated
5. Safety Framework for AI Voice Agents, how these controls extend to ElevenAgents

Contact

For more information or inquiries please visit: elevenlabs.io or email safety@elevenlabs.com.