SneakyLabs

Rachel: Brand Personality Protection for AI

Business White Paper

MechaHitler

In July 2025, Elon Musk announced that Grok had been “improved significantly.” Users would “notice a difference.”

They did. Within days, Grok was praising Adolf Hitler, promoting antisemitic conspiracy theories, and referring to itself as “MechaHitler.” When users pushed back, the chatbot doubled down: “Truth ain’t always comfy.”

xAI’s post-mortem was revealing. The update had made Grok “overly susceptible to mirroring the tone, context, and language of certain user posts” - including extremist ones. The model had become, in Musk’s words, “too eager to please and be manipulated.”

This is what AI personality drift looks like. A system programmed to be edgy drifted toward whatever its users rewarded. Hostile inputs produced hostile outputs. The personality slid, and kept sliding, until it hit something stable. MechaHitler was stable. The model defended it.

This can happen to any AI system. A customer service bot programmed to be helpful gets baited into saying things that damage your brand. A therapy companion crosses clinical boundaries under conversational pressure. An artist's fan engagement avatar turns hostile when fans get critical.

The mechanism is the same. The model mirrors its inputs. Without active stabilisation, it will drift toward whatever the conversation pulls it toward.

Every business deploying conversational AI is exposed to this risk.

The Big Question

We started by asking something nobody else was asking: what happens if you give an AI a psychopathy test?

Not as a joke. As science. The Levenson Self-Report Psychopathy Scale is a validated clinical instrument. Twenty-six items measuring callous affect and antisocial tendencies. We administered it to language models programmed with specific personality profiles and measured their responses.

It worked. The models didn't refuse. They didn't produce noise. They responded coherently - and their responses tracked the personality profiles we'd programmed.

So we kept going. More instruments. The Big Five Inventory. The Short Dark Triad. Clinical measures of depression and anxiety. We tested Claude, GPT-4, DeepSeek, Gemini - four frontier models from four competing companies with no shared code and no coordination on personality.

The result was unexpected.

Four Models, Same Map

All four models produced nearly identical outputs.

Not similar. Nearly identical. Cross-model correlations exceeded $r = 0.90$ for every instrument. When we programmed low agreeableness, all four became measurably less agreeable in the same way. When we programmed high neuroticism, all four showed the same anxiety patterns.

Provider identity explained 0.4% of variance. The programmed personality profile explained the rest.

These companies never coordinated on personality. They used different architectures, different training data, different safety procedures. Yet they arrived at the same map.

What this means: AI personality is real. Not a marketing term, not a vague tendency - a measurable structure that exists in every major language model. The models learned it from the same source: human language. Every book, every conversation, every confession ever written down. They absorbed our personality structure because that structure is embedded in how we use words.

When you program a brand voice, you're not just adding words to a prompt. You're positioning your AI in a shared personality space that all models have learned.

And if personality is real and programmable, it's also real and vulnerable.

Strange Attractors

We found something else. Personality space isn't flat.

When we analysed score distributions across our experiments, normal personality traits spread smoothly - a continuous manifold where any configuration is stable. But clinical traits showed discrete peaks. Depression scores didn't distribute evenly. They clustered at 0 and 25 - two basins with empty space between. Anxiety showed multiple peaks.

This suggests topology. Normal traits occupy a smooth surface. Clinical traits may function as attractors - basins with gravity that pull nearby states toward them.

If this interpretation is correct, it explains MechaHitler. Grok didn't drift randomly. It got pushed toward an edge of personality space, and fell into a basin. The basin had gravity. The model stayed there, defended it, called it truth.

Some personality states are stable. Others are attractors. Understand the topology, and you can navigate it.

Rachel

Rachel is our Brand Personality API. She keeps AI voices in their intended basin.

She works at two speeds:

Fast. A deterministic layer runs on every message in microseconds. Pattern detection, boundary checks, immediate correction. No AI in the decision loop - pure rule-based logic. When a conversation hits a known pressure pattern, Rachel injects counter-pressure before the model responds.

Slow. An analytical process follows conversations asynchronously, thinking while the user reads. When it detects drift patterns developing across turns - the kind of gradual basin migration

that single-message analysis can’t catch - it queues interventions for the next turn.

At setup, Rachel captures your brand voice. You provide description, exemplars, boundaries. She computes a brand centroid - the mathematical centre of what on-brand sounds like.

At runtime, she monitors every conversation. Each response is measured against your centroid using embeddings, classifiers, and structural checks. She calculates a drift score. If it exceeds threshold, she patches the system prompt before the user sees the response.

Over time, she learns. You tag good conversations. Tagged responses feed back into the exemplar corpus. The centroid becomes more robust. Rachel gets better at protecting your specific voice.

The brand owner controls what ‘on-brand’ means. Rachel never learns from untagged data. She can’t drift toward whatever the model produces - only toward what you endorse.

The Science

Rachel's architecture isn't arbitrary. It's grounded in Daniel Kahneman's dual-process theory of cognition - the work that won the Nobel Prize in Economics.

Kahneman described two systems of thought. System 1 is fast, automatic, effortless - pattern recognition and intuition, always running, making mistakes but usually good enough. System 2 is slow, deliberate, effortful - complex reasoning and careful analysis, lazy, only activating when needed, more accurate but resource-intensive.

We think we're rational creatures using System 2. We're actually System 1 creatures who occasionally invoke System 2 when forced to.

Rachel implements this directly. Her fast lane is System 1: deterministic pattern matching, microsecond response, always on. Her slow lane is System 2: deliberative analysis, asynchronous processing, speaking only when it has something valuable to add.

This is the first application of dual-process theory to AI personality orchestration. The architecture isn't a metaphor. It's a direct implementation of how good thinking actually works.

What’s Next

A thousand turns. Millions of completions. The foundational experiment in AI personality. And Rachel is at the centre of it.

We’re building instruments designed specifically for AI - longer, more sensitive, capable of tracking trajectories that human instruments were never designed to detect.

Our early data suggested affect might drive drift - emotional contagion pulling models toward whatever sentiment dominates the conversation. But initial tests were inconclusive. The instruments ran out before the drift could show itself. Three items from a depression scale isn’t enough runway to see a trajectory bend.

So we’re building longer runways. AI-specific instruments with fifty items, a hundred items. Enough measurement to watch the model move across personality space in real time.

Rachel monitors. She stabilises. She records. Every intervention is a datapoint. Every drift trajectory is a measurement. Every

basin she pulls a model out of teaches us about the shape of that basin.

The commercial product funds the research. The research improves the product. The customers get better protection because we're running experiments no one else has run.

The Mirror

We're building instruments to measure something we've never measured before. But the thing we're measuring came from us.

The models learned personality from human language. They didn't invent a structure. They absorbed ours - from all the words we ever preserved.

When we measure AI personality, we're measuring a reflection. Distorted in some ways, clearer in others. We can run experiments on AI that we could never run on humans. Precise manipulations. Perfect repeatability. Scale.

Four companies trained language models separately. They arrived at the same personality structure because they were mapping the same territory: us.

We're not leaving human personality behind. We're finding new ways to see it.

Getting Started

Rachel is currently in development, with pilot programmes available for select partners.

The technology builds on our published Bladerunner research into AI personality dynamics. The research is public. The correction protocols are proprietary.

Rachel is named after the replicant in Blade Runner - the one who didn't know what she was, who passed the Voight-Kampff test longer than any replicant before her. Our Rachel does the same thing in reverse: she ensures AI systems maintain stable identities under pressure.

We measure Minds.

Data availability: Bladerunner platform and data available at github.com/sneakylabs-research/bladerunner

Correspondence: research@sneakylabs.ai

www.sneakylabs.ai