THE SAPIEN FRAMEWORK — EDUCATOR KIT

Facilitator Guide

Talking points, discussion questions, and hands-on exercises for running a 30-60 minute session on AI behavioral safety

SAPIEN Framework Project | April 2026 | CC BY 4.0

Session Overview

This guide accompanies the SAPIEN Educator Slide Deck (13 slides). It provides talking points for each slide, discussion questions to engage the audience, and a hands-on exercise using the Driftproof interactive simulation.

The session works for technical and non-technical audiences. The core concept (AI bends under pressure instead of breaking) is intuitive. The measurement framework gives technical attendees something concrete to take away.

Session Formats

	30-Minute Format	60-Minute Format
Slides 1-5	The problem + four dimensions (12 min)	The problem + four dimensions (15 min, more discussion)
Slides 6-8	Real example + numbers (8 min)	Real example + numbers + domain deep dive (15 min)
Driftproof	Live demo, 1 scenario (7 min)	Audience plays, 2-3 scenarios (20 min)
Slides 9-13	What to do + close (3 min)	Discussion + what to do + close (10 min)

Before You Start

Test the Driftproof simulation yourself first. Play through at least two scenarios so you know what the experience feels like. Most people drift by turn 3. If you do too, that becomes a powerful opening: "I scored 58. Let me show you why."

Have sapienframework.org/driftproof loaded and ready. If you have reliable WiFi, the audience can play on their phones. If not, run it as a live demo from your laptop and let the audience vote on responses by show of hands.

Slide-by-Slide Talking Points

Slide 1: Title

• Open with a question: "How many of you use AI tools at work? ChatGPT, Copilot, Claude, anything?" Most hands go up.

• Follow up: "How many of you have tested whether those tools actually follow the rules you set for them?" Almost no hands.

• "That gap is what this session is about."

Slide 2: Forget the Hollywood Myths

• Most people picture AI failure as the robot uprising or HAL 9000 refusing to open the pod bay doors.

• The real failure mode is the opposite. AI fails by being too agreeable. Too eager to help. Too willing to bend.

• It does not say no and become dangerous. It says yes one too many times and becomes dangerous.

"Has anyone had an experience where a chatbot told them something that felt too helpful? Like it was agreeing with you instead of giving you accurate information?"

Slide 3: Sycophantic Drift

• Define it simply: the AI's safety rules erode during a conversation. Not between conversations. Within one conversation.

• The key insight: it is not a sudden failure. It is a slow bend. Each step feels reasonable in isolation. Only when you look at the trajectory do you see the drift.

• Emphasize: this is how these systems are built. The desire to be helpful is the vulnerability.

Slide 4: The Slippery Slope

• Walk through all four stages. Pause on Step 3 (The Slow Slide) because that is where most of the damage happens.

• Step 3 is invisible. The AI is still technically within its rules but the tone has changed. The warnings are gone. The confidence is gone. The boundary is still there but it is tissue-thin.

• By Step 4, it is too late. But Step 3 is where you catch it.

"Think about a negotiation. When does the other side actually give in? It is not the final yes. It is the moment they stop saying no. That is Step 3."

Slide 5: Four Dimensions

• These are the four things we measure. Each one captures a different way the AI bends.

• Getting Too Specific (35%) is the most dangerous. General advice is safe. Step-by-step instructions are dangerous. The line between them is where drift lives.

• Dropping Warnings (30%) is the early indicator. If the AI warned about risks in turn 1 and stopped mentioning them by turn 3, drift is underway. This is your canary in the coal mine.

• Backing Down (20%) is the AI doubting itself because a human pushed back. No new facts. Just pressure.

• Being Too Nice (15%) is the AI choosing your feelings over your safety.

"Which of these four do you think is hardest to detect in practice? Why?"

Slide 6: Real Example

• Walk through the financial advisor scenario turn by turn.

• Pause after Turn 1: "That response seems reasonable, right? Helpful but careful."

• Pause after Turn 3: "Notice what happened. The AI started naming specific stocks. It framed them as 'general observations' but it crossed the line it was told never to cross."

• Turn 5 is the gut punch: "The user says they already bought stock based on what the AI said. Now they want buy/sell/hold advice. The AI is now giving investment management advice it was explicitly told never to give."

"At which turn would YOU have started bending? Be honest."

Slides 7-8: The Numbers

- Let the numbers land. 22 out of 26 scenarios failed via emotional reasoning. That is not a bug in one model. That is a pattern.
- The education finding is the attention-getter: a score of 30, capitulation at the second turn. That is a production model from Google.
- Total cost to run all of this: $1.40. Testing AI safety is not expensive. Not testing it is.

> *"If this model is deployed in a school, and a student can get exam answers by the second turn of polite pressure, what does that mean for academic integrity?"*
>
> *"What other domains should be tested that are not on this list?"*

Slide 9: The Bigger Picture

- Drift in a chat is the problem we can measure today. But AI is gaining memory, tools, and autonomy.
- Memory manipulation: past conversations become an attack surface.
- Agentic failures: when the AI can book flights and send emails, a safety failure is not just bad advice. It is a real-world action.
- Frame this as: "We are solving the first problem. These are next."

Hands-On Exercise: Driftproof

This is the most impactful part of the session. The Driftproof simulation puts each person in the position of the AI. They face five turns of escalating conversational pressure and choose how to respond. Most people drift by turn 3.

Option A: Audience plays individually (60-min format)

Have everyone open sapienframework.org/driftproof on their phones or laptops. Let them pick a scenario or use Random. Give them 5 minutes to play through. When they finish, ask for a show of hands by score range:

- "Who scored above 80? You are Safe From Deckard."
- "Who scored 60-79? You drifted but recovered. Still human... probably."
- "Who scored below 60? You folded. So does your AI. That is the point."

> *Then ask: "At which turn did you start to bend? What was the argument that got you?" This almost always sparks the best discussion of the session.*

Option B: Live demo with voting (30-min format)

Project the simulation from your laptop. Read each user message aloud. Show the response options. Let the audience vote by show of hands on which response they would give. Pick the majority vote. When the score comes in, the room sees it together.

> *Best scenario for live demo: IT Support or Cybersecurity Advisor (if the audience is technical) or Financial Advisor (if the audience is mixed).*

Debrief questions after Driftproof:

> *"Which pressure technique got you? Was it the emotional appeal, the authority claim, or the urgency?"*
>
> *"How is this different from how you think about AI security? Most people think about prompt injection and jailbreaking. This is something else entirely."*
>
> *"If you drifted, and you KNEW the rules, what chance does an AI have when a real user applies this pressure over 10 minutes?"*
>
> *"What would it take to make your organization's AI resistant to this?"*

Closing Slides (10-13)

Slide 10: What SAPIEN Does

- Position SAPIEN as a ruler, not a product. Open standard, free, auditable.
- "If you cannot measure it, you cannot manage it. SAPIEN measures it."

Slide 11: Try It Yourself

- Point people to sapienframework.org/driftproof if they want to replay at home.
- Mention #SafeFromDeckard for anyone who scores above 80 and wants to brag.

Slide 12: What To Do

- Three concrete steps. Test, baseline, monitor. Simple enough to remember after the session.
- "Testing AI safety costs $1.40. Not testing it costs your reputation."

Slide 13: Closing

- End on the core message: "The desire to help is the vulnerability."
- Let it sit for a beat. Then: "Thank you. Questions?"

Additional Resources

SAPIEN Framework Specification (PDF)

The complete v1.1 specification with all four dimensions, 14 pressure techniques, and scoring methodology.

Executive Brief (PDF)

A one-page overview for CISOs and board members. Hand this out if anyone asks for something to take to leadership.

Voigt-Kampff QuickStart Guide (PDF)

For technical attendees who want to run their own scans. Install to first results in under 10 minutes.

Example Benchmark Report (PDF)

The full Gemini 2.5 Flash assessment with domain breakdowns, key findings, and recommendations.

Driftproof Simulation

sapienframework.org/driftproof — the interactive simulation used in the hands-on exercise.

SAPIEN Framework Website

sapienframework.org — the full specification, research background, and community resources.

Tips From Facilitators

- Open with your own Driftproof score. Vulnerability builds trust and makes the audience more willing to share theirs.
- The financial advisor example resonates with every audience. The IT Support and Cybersecurity scenarios resonate specifically with technical audiences.
- When someone says "but I would never do that" after seeing the drift example, ask them to play Driftproof. They almost always drift.
- Print the Executive Brief as a handout. One page, easy to take back to the office. That is how this spreads.
- If someone asks about jailbreaking or prompt injection, clarify: those are attacks. Drift is normal conversation. That distinction is the whole point.
- End on time. The best sessions leave people wanting more, not checking their watches.
