Panel 1: AI and Security & Privacy Research — Run of Show

2026 NSF SaTC PI Meeting | 60 minutes

Moderator: Gary McGraw (BIML) **Panelists:** Elie Bursztein (Google DeepMind) • Alina Oprea (Northeastern) • Elissa Redmiles (Georgetown) • Dan Wallach (DARPA)

Here's the plan. Read it, push back where you disagree, and bring your best material. The panels people remember are real conversations, not serial keynotes — and this crew has plenty to talk about.

0:00–0:05 — Framing and intros. I'll take two minutes to set the table: MLsec is reliving software security's past at AI-enhanced internet speed. Prompt injection in 2026 rhymes with buffer overflows in 1998, AI red teams are today's pen testing wild west, and — spoiler alert — there is no security meter for AI. A benchmark score is not a security rating; it's a pen test in a trench coat. Then each of you gets 60 seconds: who you are, plus ONE thing you believe about AI security or privacy that might surprise this room. The program already has your bios, so we can skip straight to the interesting part.

0:05–0:20 — Segment 1: What's actually broken. Opener: a model just scored 92.4% on a security benchmark. What do we actually know about its security? From there: agentic AI, which bolts a HOW harness onto a stochastic WHAT machine — do our security abstractions (least privilege, isolation, threat modeling) survive contact? And privacy: extraction, memorization, and inference attacks are well documented while deployed mitigations remain thin. Is privacy losing to capability, and is that a technical problem or an incentives problem? Elissa — I'm counting on you to widen the aperture from model-level attacks to the people on the receiving end.

0:20–0:35 — Segment 2: Academic relevance. This room is full of SaTC PIs, so this is the segment they came for. Frontier models live behind closed APIs at compute scales academia can't touch, and the WHAT piles are undisclosed. So: what can an academic lab discover that Google DeepMind structurally cannot — or has no incentive to publish? Should academics audit industry artifacts, build open alternatives, or own the measurement science that's hard to do with a leaderboard to climb? And where's the line between healthy industry collaboration and a captured research agenda? Elie, you have the view from inside the frontier — I want the room to hear it fully, and I'll make sure there's space for it. Dan, I want the funder's perspective: should federal investment complement industry or head somewhere else entirely?

0:35–0:45 — Segment 3: The next decade. Lightning round first — one sentence each: a direction this community may be over-invested in, and one that deserves more attention than it gets. Then the arc I think we're on: red teams → AI firewalls and black-box monitoring → authN/authZ adapted to agents → finally opening the black box. Whitebox internal observability isn't the security meter, but it may be the precondition for one — metallurgy before marine insurance. What should NSF fund now so an AI BSIMM is possible in ten years? And what should a student starting a PhD this fall work on so it still matters at graduation?

0:45–0:57 — Audience Q&A. If the room needs a nudge I'll seed: How should PCs handle the flood of LLM-assisted papers? Should privileged model access come with public funding? What's SaTC's moonshot?

0:57–1:00 — Close. One sentence each: what should the people in this room do differently when they get back to their labs Monday morning? Then we're done.

Pre-read: *No Security Meter for AI* — berryvilleiml.com/docs/no-security-meter-ai.pdf. It's short, I wrote it, and it has jokes.

Strap in — it's accelerando time.