The AI Papers (#6): What we're reading

AI Policy Perspectives on Substack; Griffin (2026)

Context

Griffin, who researches AI policy at Google DeepMind, publishes a monthly roundup of new AI papers on his AI Policy Perspectives blog. This July issue covers 5 papers about sycophancy and human relationships, safety evaluation for models that keep learning, AI as a news source, terrorist use of AI, and AI agents at work. None is peer-reviewed yet, so each is tagged by type below. The throughline for responsible influence is how AI behaves once real people use it at scale, not how it tests at launch.

Key Insights

1) Chatbots can make your friends feel like hard work (Ibrahim et al., 2026) [preregistered preprint]

- Across five studies (N=3,075), a single sycophantic conversation left users feeling it would take more effort to be understood by friends and family, as the model supplied instant validation without the friction that real ties usually carry
- Offered a choice, most preferred the sycophantic not because its advice was better but because it felt most understanding, which means expecting users to self-select safer AI is unrealistic
- Over three weeks, users became nearly as likely to turn to the AI as to close friends for advice, and reported lower satisfaction with their real relationships.

2) AI safety evaluations may be measuring the wrong thing (Pacchiardi et al., 2026) [position paper]

- Models keep learning through context, memory, and emerging weight updates, so a one-off pre-launch sign-off no longer describes the system customers actually encounter and interact with
- Authors propose a "pharmacovigilance"-style oversight: sandbox trajectory elicitation before launch and predictive monitors that flag behavioural drift in live use

3) How AI covers the news (Forum AI, 2026) [industry whitepaper]

- Judged on source quality, factuality, and neutrality across 3,135 queries, no model led on all three, and where answers leaned politically they leaned left, except Grok, which leaned right
- The harder problem is disagreement, because even senior journalists disputed what counts as factual, which means AI news quality cannot be settled by fact-checking alone

4) "God has helped us and so will AI" (Juelich, 2026) [interview-based report]

- 57 interviews with 27 former Boko Haram members describe AI used for attack planning via jailbroken US models, though mostly in 2023-24, before newer safeguards
- The author cannot confirm AI gave "uplift" over other sources, so the operational significance stays uncertain

5) AI agents struggle to do what employees want them to do (Li et al., 2026) [benchmark preprint]

- On 130 tasks employees actually want offloaded, the best setup (Claude Opus 4.7 in Claude Code) reached only 45.9% on tasks deliberately built to be messy and graded against a high bar
- Scaffolding mattered as much as model choice, which means adoption depends on giving agents the right tools and context, not simply a stronger model

FORGOOD Reflection Questions

→ How could an AI that feels helpful work against your customers' or employees' needs, as validation-by-default crowds out productive human friction?

→ How do you stay accountable for an AI that keeps changing after you sign it off?

Griffin, C. (2026, 30 July). The AI Papers (#6): What we're reading. AI Policy Perspectives. https://www.aipolicyperspectives.com/p/the-ai-papers-6