AI Trends

When Bounded Minds Meet: Human and Machine Judgment in Government Cybersecurity

2026. 09. 08. 09:00 | MBH Bank Visionary Stage | ENG · 20 perc

Martin L. Skorczynski (US)
Feltörekvő technológiák és kiberbiztonsági kutatások igazgatóhelyettese · U.S. Government Accountability Office

Martin L. Skorczynski, az Egyesült Államok Számvevőszékének felmerülő technológiák és kiberbiztonsági kutatásokért felelős helyettes igazgatója szerint az intézményi irányításnak figyelembe kell vennie az emberi és mesterséges okoskodók sajátos módját, ahogyan a saját vakfoltjaikat kombinálják.

MI VOLT TÉMA

Martin L. Skorczynski bemutatja a korlátozott racionalitás fogalmát, megjegyezve, hogy mind az emberek, mind a nagy nyelvi modellek specifikus korlátokkal működnek. Az embereket az idő, a figyelem és az információ korlátozza, míg a modelleket a képzési adataik és a kontextus ablakai határolják. Amikor ezeket a két okoskodót egy kormányzati intézménybe helyezik, egy harmadik korlátozott racionalitási réteg kerül bevezetésre az eljárások és felülvizsgálati kapuk révén, amelyek tárolt ítéletként szolgálnak a magabiztos tévedés kockázatának csökkentésére.

A prezentáció kiemeli, hogy egy ember és egy modell közötti egyetértés inkább erősítheti a hibákat, mint biztosíthatja a helyességet, egy olyan jelenséget, amelyet interferenciának neveznek. Martin L. Skorczynski hangsúlyozza, hogy az észlelhetőség—az emberi és a modell különálló ítéleteinek rekonstruálásának képessége—elengedhetetlen ezeknek a hibáknak a felismeréséhez. Például egy tanulmány kimutatta, hogy az

AI kódsegítséget használó résztvevők kevésbé biztonságos kódot írtak, mert megbíztak a modell összefüggő, de hibás kimeneteiben. Ezek az elemek azt sugallják, hogy a fő kihívás nem maga a technológia, hanem az azt szabályozó intézményi határok tervezése.

FONTOS TÉNYEK

- Egyeztetés egy ember és egy modell között megerősítheti a helytelen következtetéseket, ami azt jelenti, hogy a konszenzus nem bizonyíték a pontosságra.
- Az intézményi eljárások szükségszerű korlátként szolgálnak, hogy kiszűrjék azokat a hibákat, amelyek kívül esnek mind az emberi, mind a modell képességein.
- Az észrevehetőség megköveteli az emberi és a modell egyedi ítéleteinek külön-külön megőrzését, nem pedig csak a végső kombinált kimenet naplózását.

ÜZLETI HATÁS

Ezeknek a megállapításoknak a gyakorlati alkalmazása a technológia bevezetése előtti irányítási keretrendszerek kialakításában rejlik. A szervezetek egy egységes kijelölt adathozzáférési útvonal létrehozásával olyan felelős környezetet teremthetnek, ahol a kompetens alkalmazottak által létrehozott „árnyék” eszközök nem kerülik meg a biztonsági ellenőrzéseket. Ez a megközelítés biztosítja, hogy az emberi erőfeszítéseket magas értékű feladatokra, például ötletelésre és architektúrára irányítsák, miközben szigorú felügyeletet tartanak fenn dokumentált szabványokon és független felülvizsgálati kapukon keresztül.

KÖVETKEZTETÉS

Martin L. Skorczynski a hallgatóságot arra ösztönzi, hogy lépjen túl a folyamatos emberi figyelemre támaszkodó egyszerű felügyeleten, és egy intézményesített tudásrendszert hozzon létre. Azt javasolja, hogy dokumentálják és őrizzék meg a szervezeti szakértelmet, hogy azt a döntéshozatal során felhasználhassák. A cél olyan megoldások kialakítása, amelyek lehetővé teszik az új technológiák gyorsaságát, miközben tiszteletben tartják az emberi és intézményi korlátokat.

AI Trends

When Bounded Minds Meet: Human and Machine Judgment in Government Cybersecurity

2026. 09. 08. 09:00 | MBH Bank Visionary Stage | ENG · 20 perc

Martin L. Skorczynski
Assistant Director for Emerging Technologies & Cybersecurity Research · U.S. Government Accountability Office

Martin L. Skorczynski, Assistant Director for Emerging Technologies & Cybersecurity Research at the U.S. Government Accountability Office, argues that institutional governance must be intentionally designed to manage the unique blind spots of both human reasoners and large language models.

WHAT WAS DISCUSSED

Martin L. Skorczynski introduces the concept of bounded rationality, noting that both humans and large language models operate with specific limits on time, attention, and information. While humans are limited by cognitive fatigue and fixed content windows, models are bounded by their training data and framing. These different shapes of limitations mean that when the two are combined, they can create interference where errors cancel out or, more dangerously, reinforce one another to produce a confident but incorrect result.

The presentation emphasizes that institutions act as a third bounded reasoner, using procedures and review gates to store judgment and manage risk. Martin L. Skorczynski highlights that the goal of governance is to ensure observability, which allows for the reconstruction of how a decision was reached by keeping human and model judgments separate. For example, a model might identify a technical anomaly across a massive dataset, but a human must provide the institutional context to determine the actual severity of that finding.

KEY TAKEAWAYS

- Agreement between a human and a model does not prove correctness because amplification can occur in the wrong direction.
- Observability requires preserving separate records of human and model judgments rather than just logging the final exchange.
- Governance must precede technology to ensure that all data access and experimentation occur within accountable, designated paths.

BUSINESS IMPACT

The practical application involves moving from "shadow" tools built by well-meaning employees to sanctioned environments where experimentation is governed. By establishing clear boundaries, organizations can increase capability and move work toward ideation and architecture without losing accountability. This approach ensures that while models can be responsible for producing code or data, humans remain accountable for the final outcomes by providing the necessary institutional knowledge that models cannot access.

CONCLUSION

Martin L. Skorczynski concludes by urging the audience to focus on what institutional knowledge can reach the reasoner at the moment of decision. He recommends shifting the human role from constant oversight to providing documented, preserved knowledge that the model cannot see. The final call is to build arrangements that allow for the detection of errors sooner by intentionally redrawing institutional boundaries.