Projektnummer/Project no
E38568
Uppdragsgivare/Client
Försvarsmakten
FoT-område
Ledning och MSI

Författare/Author
Joel Brynielsson

Datum/Date
2025-12-31
Memo nummer/Number
FOI Memo 9138

Teknik för framtidens AI-tillämpningar inom försvar och säkerhet 2025

(AT.9220222 Ledning och MSI FOI 25)

1 Inledning

Den snabba utvecklingen inom AI-området ger upphov till smart möjliggörande ingenjörskonst där kunskap och tillämpning är nära förknippade, och där kunskapsuppbyggnad och förmågeuppbyggnad ofta är samma sak. Stora tekniska genombrott har skett de senaste åren till följd av framgångsrik utveckling och tillämpning av maskininlärning. Utvecklingen går fort och det är den civila forskningen som är drivande. För försvar och säkerhet är det följaktligen centralt att följa med i den snabba utvecklingen och fånga de relevanta trenderna för att kunna tillämpa inom relevanta försvars- och säkerhetsområden. Försvarsmakten har därför identifierat ett behov av att på djupet analysera aktuella tekniker och algoritmer med bäring på nydanande AI-tillämpning för försvars- och säkerhetsändamål, där valet av ämnen bestäms och utvärderas på årsbasis med fokus på nytta för svenskt försvar. I möjligaste mån säkerställs analysernas kvalitet och spårbarhet genom publicering på konferenser och i tidskrifter med extern referentgranskning. I detta memo redovisas publikationer utgivna under 2025, där arbetet helt eller delvis utförts med stöd från denna satsning.

2 Kostnaden för osäkerhet i självspelande förstärkningsinlärning och framåtblickande sökning

Oscarsson, M., Brynielsson, J., Cohen, M., Kamrani, F., & Limér, C. (2025). The cost of uncertainty in self-play reinforcement learning and search. I *Proceedings of the 2025 IEEE International Conference on Intelligence and Security Informatics (ISI 2025)* (s. 113–120). IEEE. https://doi.org/10.1109/ISI65680.2025.11201174

Kombinationen av förstärkningsinlärning och framåtblickande sökning, som introducerades i AlphaGo, har revolutionerat vår förståelse av taktik och strategi i klassiska strategispel som Go och schack. Fram till nyligen har detta banbrytande angreppssätt varit begränsat till spel med perfekt information, där spelarna har full insyn i spelets aktuella tillstånd. Denna studie undersöker den senaste generaliseringen av förstärkningsinlärning med framåtblickande sökning till spel med ofullständig information, såsom poker, där delar av spelläget – exempelvis motståndarens hand – är dolda för spelaren. I studien utforskas hur väl detta angreppssätt skalar när mängden dold information ökar. För att göra detta reproduceras och utvärderas ett nytt resultat inom förstärkningsinlärning med framåtblickande sökning över tre varianter av ett specialanpassat pokerspel som skiljer sig åt avseende antalet dolda kort som delas ut till spelarna. Resultaten visar som väntat att spel med mindre mängd dold information lärs in mer effektivt, men att de beräkningsmässiga kraven skalar sublinjärt när den dolda informationen ökar.

3 Strategisk styrning av stora språkmodeller genom optimering av det spelteoretiska handlingsutrymmet

Lavebrink, S., Brynielsson, J., Cohen, M., Kamrani, F., Limér, C., Lindström, M., & Vangeli, M. (2025). Strategic steering of large language models via game-theoretic action space optimization. I *Proceedings of the 2025 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2025)* (s. 508–528). Springer. https://doi.org/10.1007/978-3-032-14107-1_41

I denna studie undersöks hur stora språkmodeller kan styras att agera mer strategiskt i textbaserade förhandlingssituationer. Två promptbaserade utformningar av handlingsutrymmet jämförs, dels känslomässig ton, dels explicita erbjudanden, inom en förhandlingsmiljö, och utfallen jämförs i simulerade dialoger. Resultaten visar att båda angreppssätten förbättrar de strategiska utfallen relativt att inte använda strategisk styrning, där känslomässig ton ger högre överenskommelsegrad medan explicita erbjudanden erbjuder mer stabila avvägningar. Dessa resultat visar hur utformningen av handlingsrummet påverkar agenters beteende och ger insikter för användning av stora språkmodeller i strategiska förhandlingsscenarier för att skaffa sig ett övertag till exempel i samband med påverkansoperationer.

4 Positionspåverkan vid användning av stora språkmodeller för kritiskt beslutsstöd: en fallstudie om triage vid stora skadeutfall

Wickenberg-Bolin, U., Cohen, K., Björnesjö, H., & Tegen, A. (2025). Position bias in LLMs for critical decision support: A case study on multiple casualty triage. I *Proceedings of the 2025 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2025)* (s. 529–541). Springer. https://doi.org/10.1007/978-3-032-14107-1_42

Stora språkmodeller används allt mer inom högriskområden såsom räddningstjänst, medicinsk prioritering och i säkerhetskritiska domäner. Utgående från ett kontrollerat fiktivt beslutsstödsscenario för prioritering av skadade i ett masskadescenario undersöks i denna studie effekterna av språkmodellers positionspåverkan, det vill säga tendensen hos språkmodeller att prioritera information baserat på var den står i en lista snarare än baserat på hur relevant den är. I en lista med patienter ändrades placeringen av den mest allvarligt skadade personen för att kunna bedöma systematiska avvikelser från vad som vore korrekt enligt medicinskt etablerade protokoll för triage vid olyckor med många skadade. De erhållna resultaten uppvisar en tydlig och återkommande tendens att prioritera det mest aktuella, där den mest kritiskt skadade patienten mer sällan prioriterades när personen stod först i listan än när den stod senare i listan. Denna effekt var extra tydlig då patientlistorna var kortare, vilket står i kontrast till föreställningen om att korta instruktioner till språkmodeller är mindre känsliga för positionspåverkan. Resultaten väcker viktiga frågor om språkmodellers operativa tillförlitlighet i tidspressade och riskfyllda sammanhang där mycket information måste hanteras snabbt. Studien bidrar därmed med ytterligare bevis och kunskap om att användning av stora språkmodeller i säkerhetskritiska tillämpningar kräver rigorös validering, testning och modellanpassning.

5 Att överlista uppsåtligt tänkande motståndare: Bayesiansk uppdatering av motståndaruppfattningen vid användning av stora språkmodeller

Lindström, M., Brynielsson, J., Cohen, M., Kamrani, F., Lavebrink, S., Limér, C., & Vangeli, M. (2025). Outsmarting willful-thinking opponents: Bayesian belief revision for adversarial reasoning in large language models. I *Proceedings of the 2025 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2025)* (s. 559–578). Springer. https://doi.org/10.1007/978-3-032-14107-1_44

Vid resonemang om uppsåtligt tänkande motståndare beror framgång ofta på att förstå inte bara vad motståndaren vet, utan även vad de tror på och hur de reviderar sina övertygelser. Denna studie undersöker hur stora språkmodeller kan göras mer motståndskraftiga och strategiskt kapabla genom att modellera motståndarens resonemang med hjälp av Bayesiansk revidering av den aktuella lägesförståelsen. Genom att formalisera förhandlingar som Bayesianska spel med ofullständig information visas det att modeller utrustade med revidering av lägesförståelsen är bättre på att motverka vilseledande tänkande motståndare. Resultaten understryker rollen av andra ordningens resonemang i sammanhang med bedrägliga agenter, med konsekvenser för social påverkan i samband med till exempel onlinekommunikation och underrättelseinhämtning.

6 Upptäckt av nya cyberhot med hjälp av aktiv inlärning

Brynielsson, J., Carp, A., & Tegen, A. (2025). Detection of emerging cyberthreats through active learning. I U. Onyekpe, V. Palade, & M. A. Wani (Red.), *Recent Advances in Deep Learning Applications: New Techniques and Practical Examples* (s. 123–144). CRC Press. https://doi.org/10.1201/9781003570882-9

Inom cybersäkerhetsområdet är maskininlärning en lovande teknik för att kunna förbättra förmågan att upptäcka hot. Ändå utgör den stora mängden omärkt data ett utmanande hinder för effektiv datahantering. Detta bokkapitel fördjupar sig i effektiviteten hos aktiva inlärningsmetoder i syfte att reducera arbetet med att märka upp data manuellt. Via användning av olika frågestrategier identifieras i studien de mest informativa omärkta datapunkterna som är lämpliga för märkning. Prestandan hos de olika frågestrategierna testades genom att undersöka en maskininlärningsmodells förmåga att urskilja tweets som refererar till avancerade långvariga cyberhot (eng. *advanced persistent threats*). I scenarier där märkta träningsdata är knappa tyder resultaten på att den diversitetsbaserade k-means-frågestrategin överträffar både den osäkerhetsbaserade metoden och ett slumpmässigt urval av datapunkter. Vidare undersökte studien kostnadseffektiv aktiv inlärning (eng. *cost-effective active learning*), som integrerar datapunkter med hög säkerhet i träningsdatamängden. Överraskande framstod denna metod som den minst effektiva strategin. Sammanfattningsvis förklarar resultaten inte bara potentialen med aktiv inlärning inom cybersäkerhet, utan understryker också vikten av strategiskt dataurval för att optimera modellprestanda.