

Küberturvalisusest, kui igal ründajal on AI-agentide armee



Tanel Mällo (PhD)
Cybernetica AS teadur

22. september 2026

Kes on ründaja?

Tean kedagi, kes:

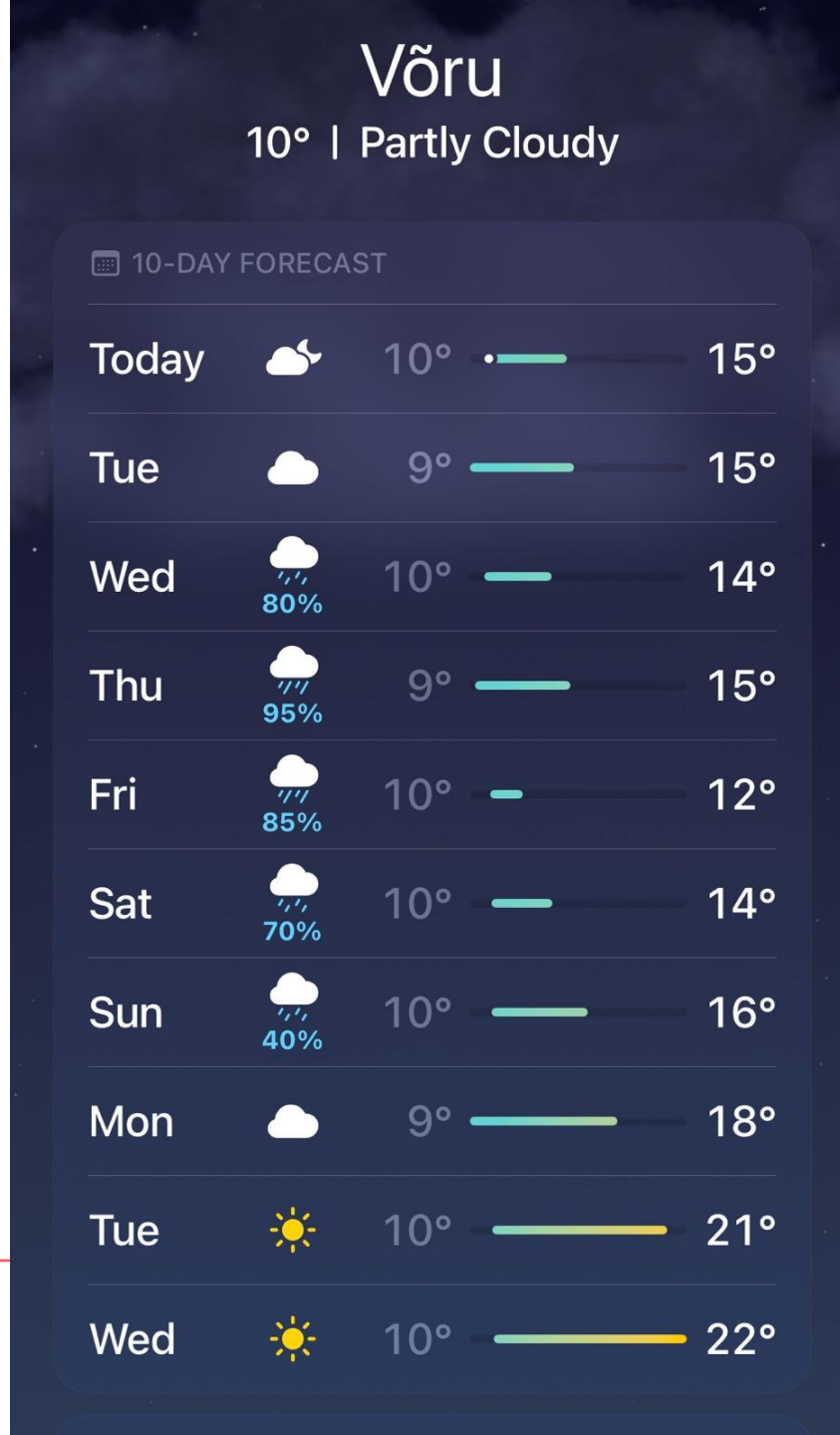
- andis AI-assistendile ligipääsu oma postkastile;
- kopeeris chatti töödokumendi/kirja;
- on näo ja hääle salvestusena internetis olemas;
- kasutas AI pakutud fakti algallikat kontrollimata;
- lasi AI-l teha õppimiseks mõeldud töö;
- koostas AI abil pöördumise, kaebuse, teabenõude.

Kes on ründaja?

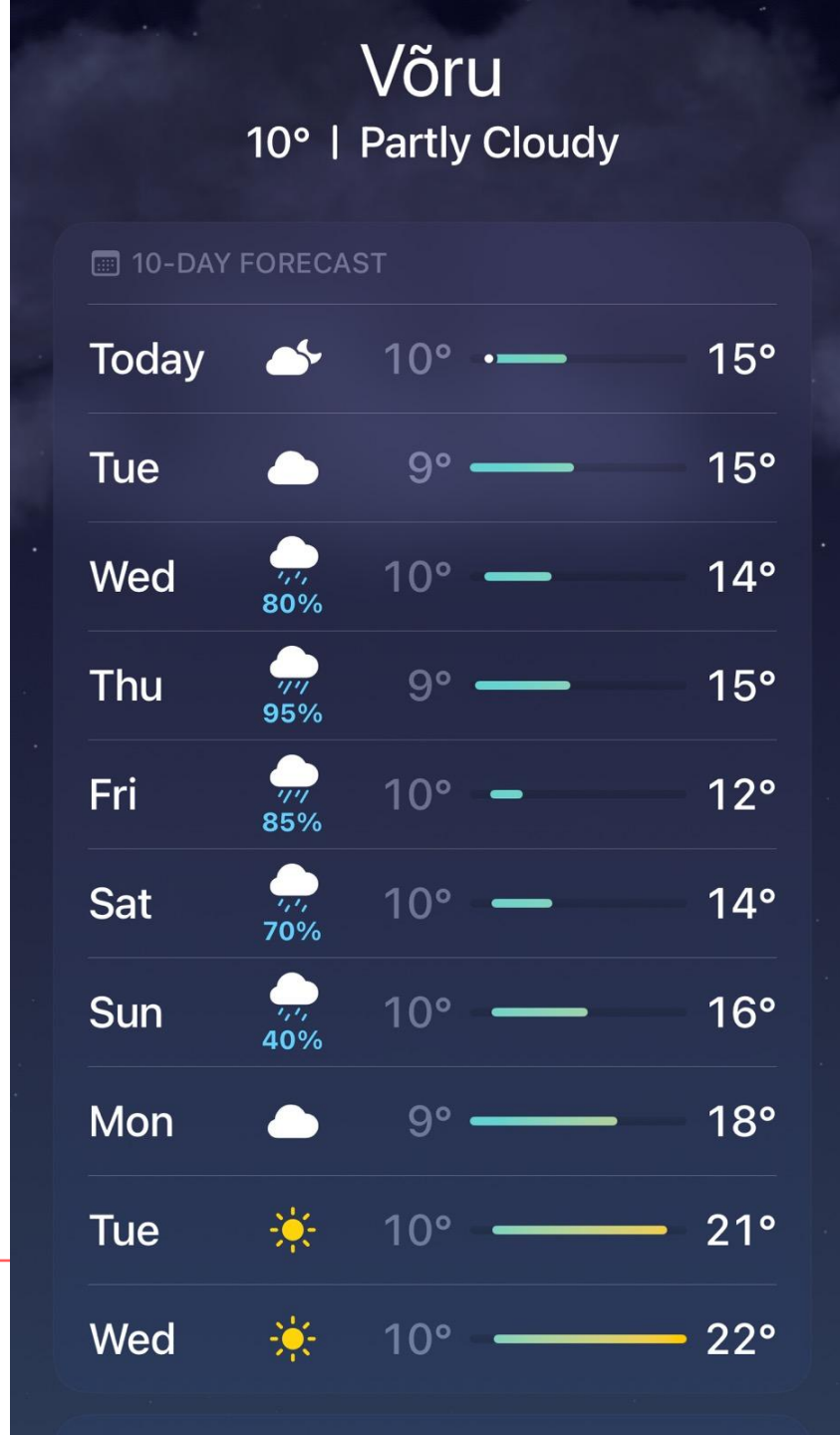
Tean kedagi, kes:

- andis AI-assistendile ligipääsu oma postkastile; → **AGENDIKAAPERDUS!**
- kopeeris chatti töödokumendi/kirja; → **SUPERKOGU!**
- on näo ja hääle salvestusena internetis olemas; → **ISIKUVÕELTSING!**
- kasutas AI pakutud fakti algallikat kontrollimata; → **MUDELIMÜRK!**
- lasi AI-l teha õppimiseks mõeldud töö; → **OSKUSKADU!**
- koostas AI abil pöördumise, kaebuse, teabenõude. → **ÜLEKOORMUS!**

Mis see on?



See on halb ilm.
Halb ilm juhtub.



— AI ei ole halb ilm

AI on valikud:

- tootja/omaniku
- kasutaja/inimese
- → ründaja

AI ei ole halb ilm

AI on valikud:

- tootja/omaniku
- kasutaja/inimese
- → ründaja

~~"AI/ChatGPT/Claude/Gemini ütles..."~~

"Ettevõtte X toode Y, meie ÕÄÖÜ-seadistusega, andis väljundi ¡Z!"

Kolm ründeliiki

Küberründed

SIHTMÄRK
süsteemid ja andmed

MASSRÜNNE

AGENDIKAAPERDUS

HULLKRATT

SUPERKOGU

Manipuleerimisründed

SIHTMÄRK
inimese otsus ja usaldus

ISIKUVÕLTSING

MASINKELM

MASSIMUDIJA

MUDELIMÜRK

SEINTELKÕRVAD

AIPSÜHHOOS

Sotsio-tehnoloogilised ründed

SIHTMÄRK
oskused, institutsioonid, sõltumatus

OSKUSKADU

ÜLEKOORMUS

AIEIVASTUTA

SUVERÄÄNSUSKADU

EBAVÕRDSUS

AIPOLEINIMENE

Ründe hind langes

- Nõrkusest ärakasutamiseni: tunnid
- AI teeb 80-90% taktikalisest tööst
- "Liiga väike, et huvi pakkuda" enam ei kehti

Kõige ohtlikum armee

- Meie enda AI-agendid
- AI-agent = taustakontrollita uus töötaja
- AI-genereeritud koodi turvanõrkused
- Keerukuse haprus

Eesti keel pole enam tulemüür

- Sihtmärk: inimese osaldus ja otsused
- Veatu eesti/inglise/tšuktši keel
- Tuttav nägu, tuttav hääl
- Tuttavad digikeskkonnad

Mõjutamine ilma häkkimiseta

- Personaliseeritud veenmine
- AI inimesest edukam, vajab vähem sisendit
- Mudelite mürgitamine
- Vestluste pikendamine

Interlööd: Usaldus ja usaldatavus

- Usaldus, mida ei saa kontrollida, on ründepind
- Eesti digiriigi usaldatavus
- Tõde kah eksisteerib ikka veel

→ Suurendada usaldatavust, mitte usaldust

— Keegi ei teata intsidendist

- Ründajata rünne
- Mõistlike lühiajaliste otsuste summa
- Oskuste kadu
- AI-ga saavutatud efekti kaduvus

Tõhusus võimekuse arvelt

- Nooremspetsialistide positsioonid
- Agressiivne turuhõive vs reaalkulud
- Sõltuvus välisühendustest
- Võimalus teenusepakkuja vahetada

→ Eesmärgid peavad juhtima tehnoloogiat.

The Art of War

- normaalsuse teadvustamine ja nõuded
- kontrollikanalid; tehissisu märgistus
- vähim vajalik õigus, stopp-nupp
- AI-vaba võimekuse säilitamine
- teenuse/taristu omanik ja vastutus läbipaistev
- pika mõju arvestamine
- kodanikuvastutus

Kokkuvõte

- Tuhendenäoline ründaja
- Ründeliigid:
 - masina vastu
 - mõistuse vastu
 - ühiskonna vastu
- Läbipaistvus ja sekkutavus

Tänan!

tanel.mallo@cyber.ee

 <https://cyber.ee/>

 info@cyber.ee

 [cybernetica](https://twitter.com/cybernetica)

 [CyberneticaAS](https://www.facebook.com/CyberneticaAS)

 [cybernetica_ee](https://www.instagram.com/cybernetica_ee)

 [Cybernetica](https://www.linkedin.com/company/Cybernetica)