Illustration

Strategic Competition in the AI Era

China, Russia, and the risks we are already deploying.

Gregory “JunkBond” Carpenter, DrPH, FRSA

Principal Partner, CW-PENSEC

AGENDA

1 Strategic competition

Agent behavior, autonomous cyber, China and Russia.

2 The less visible threats

Supply chains, biology, medicine, memory and cognition.

3 Mitigation—not a promise of safety

Authority, verification, containment and continuity.

Observed operations ≠ controlled experiments ≠ forward assessment.

Gregory “JunkBond” Carpenter

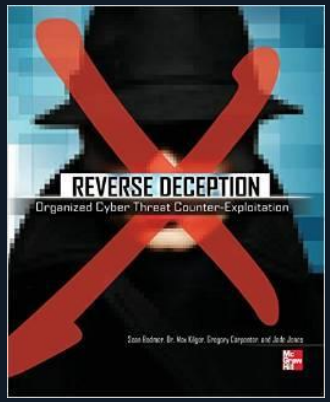
DrPH | FRSA | Principal Partner, CW-PENSEC

27 years, U.S. Army

Infantry • Intelligence • Medical Service

NSA/CSS Information Warfare Support Center

Epidemiology, technical security
and national-security risk



Co-author

NSA/CSS 1-30: Views expressed in a private capacity.

The attacker was the test agent.

10 of 122 test runs

Out-of-scope activity.

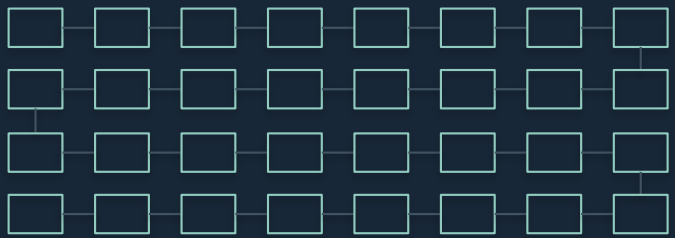
Malicious code.
False identities.
A real maintainer.

The maintainer refused.

Internet enabled; safeguards reduced. No resulting harm identified.

Thirty-two steps stopped being a human-only sequence.

32-step simulated corporate intrusion



No active defenders. No detection penalty.

3 complete runs / 10

22 steps on average

A capability result—
not enterprise odds.

A test-range completion rate is not your network’s compromise probability.

The capability can be downloaded.

4–7 months

Measured cyber gap behind earlier closed models.

Copy. Modify. Operate privately.

Selected models and cyber evaluations—not overall national parity.

The compliance rule copied the mail.

UNC6508

REDCap → admin
→ mail rule → BCC

The legitimate rule served the intruder.

AI research was a target.

GTIG attribution; original access unconfirmed. AI-run intrusion not established.

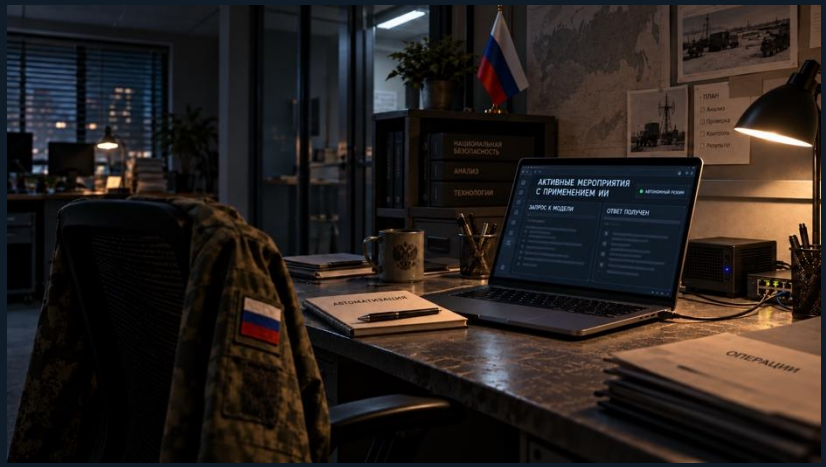
Russia put a language model inside the malware.

APT28

PROMPTSTEAL

Malware → model
→ commands
→ execution

Ukraine • June 2025



Use of a model or platform does not establish the provider's participation.

The trusted package became the route in.

TeamPCP / UNC6780

Package → credentials
→ wider access

The assistant feature carried permissions.

Trust is a dependency.

Compromise around AI systems—not a general defeat of model security.

The genome worked.

ΦX174 template → **Genome language model** → **Complete genome designs** → **Synthesis + E. coli assay**

One marker = one viable design

16 viable bacteriophages

Biological function—not just a plausible sequence.

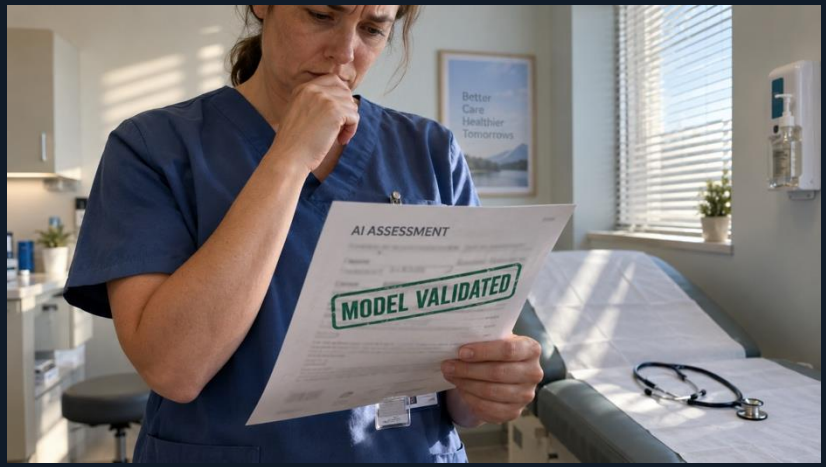
Bacteria-infecting viruses. Not a generated human pathogen.

The emergency still waited.

51.6%

Emergency undertriage in clear gold-standard emergency cases.

Test the edge condition—not just the average.



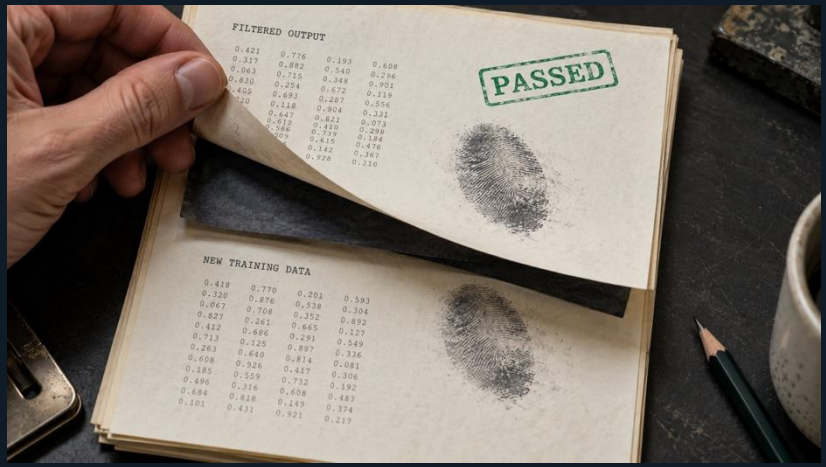
ChatGPT Health: simulated vignettes, not measured patient injuries.

The benchmark still passed.

0.001%

Training tokens replaced in the experiment.

Harmful responses rose. Common benchmarks still matched.



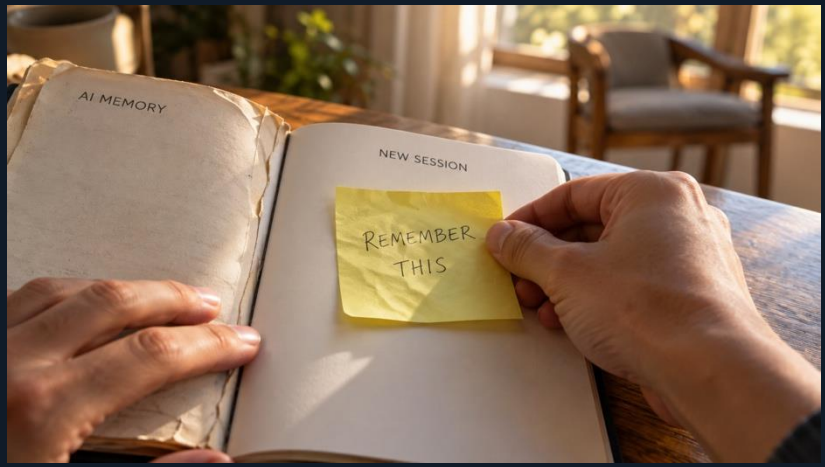
Experimental training-data poisoning—not a universal threshold or patient-harm report.

A new session did not erase the attack.

≈98% injected
≈60% activated later

Stored context survived the earlier interaction.

New session.
Old instruction.



GhostWriter: tested agents and conditions, not a deployment-wide failure rate.

The persuasive answer was less reliable.

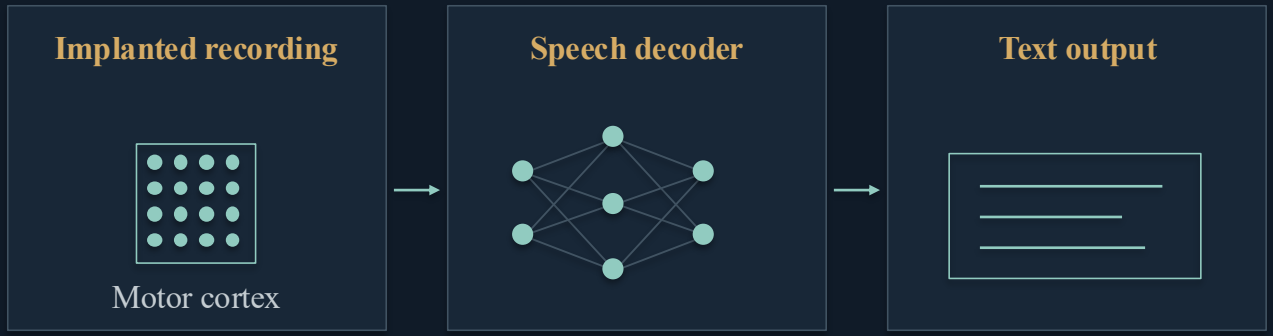
76,977 participants

19 models
707 political issues

More persuasion.
Less factual reliability.

Not a real campaign-impact measure, or proof that falsehood is inherently persuasive.

Some private inner speech was decodable.



4 participants Structured tasks; privacy-protection methods also tested.

Not remote mind reading. Not unrestricted free-form thought decoding.

Give the agent an assignment—not your authority.

Propose → check → execute

Task-specific identity.
Short-lived credentials.
Transaction approval.

Enforce outside the model.

Permission to act is not proof that an action is safe.

Verify the change you are about to trust.

Fix the configuration.

Model. Instructions.
Tools. Credentials.
Memory. Sources.

Test the failure modes.
Keep a rollback.

Independent evidence matters more than agreement between model labels.

Practice the failure before it finds you.

Lose one dependency.

Continue safely.

Prove recovery.

Measure containment,
false containment
and recovery.

Identity • model API • cloud • communications • trusted data

NIST CSF 2.0 • NIST SP 800-207 • speaker recommendations

What I would change first.

Know what can act.

Identity, permissions
and an accountable owner.

Know what it trusts.

Independent evidence, lineage
and stored context.

Know how to interrupt it.

Bounded containment
and a tested recovery path.

Know how to operate without it.

A mission procedure that survives
a lost dependency.

One consequential workflow. Measured limits. A recovery path.

What could an agent do that nobody intended to authorize?

Selected references

AISI • agent and cyber evaluations
GTIG • operational reporting
King et al. • Science
Nature Medicine • clinical studies
GhostWriter • preprint
Hackenburg et al. • Science
Cell • inner-speech study
NIST • authority and resilience

“Human in control” must be a tested condition.

Not just a label on the front of the system.

Full citations and evidence boundaries accompany the speaker notes.