

# Building the Future of Military Training: A Compound AI Approach from Information Operations to Joint Force Readiness

Summer Rebensky, Senior Scientist, Training Learning and Readiness  
Svitlana Volkova, Chief of AI, Office of Science and Technology

NORDIC DEFENCE FORCES WELCOMES YOU TO THE 13<sup>TH</sup> ANNUAL  
ADL CONFERENCE MAY 5<sup>TH</sup> – 8<sup>TH</sup> 2025

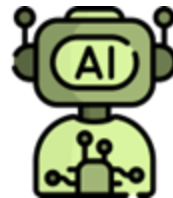
- Modern military training requires training content to be:
  - Realistic – Address real life wartime challenges (no white carding)
  - Novel – Provide new experiences to keep readiness high
  - Adaptable – Change not only to changing tactics but also to changing trainees
  - Provide real time feedback – Instantaneously understand gaps and mitigations
- To proactively train for the challenges of tomorrow, we must scale the rate and relevance of our approaches
  - Which will require the use of AI approaches at scale



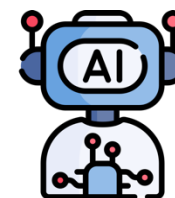
How can we provide ecosystems to trainees, instructors, and operators that facilitate readiness?



How do we create red force behavior that is relevant and representative?



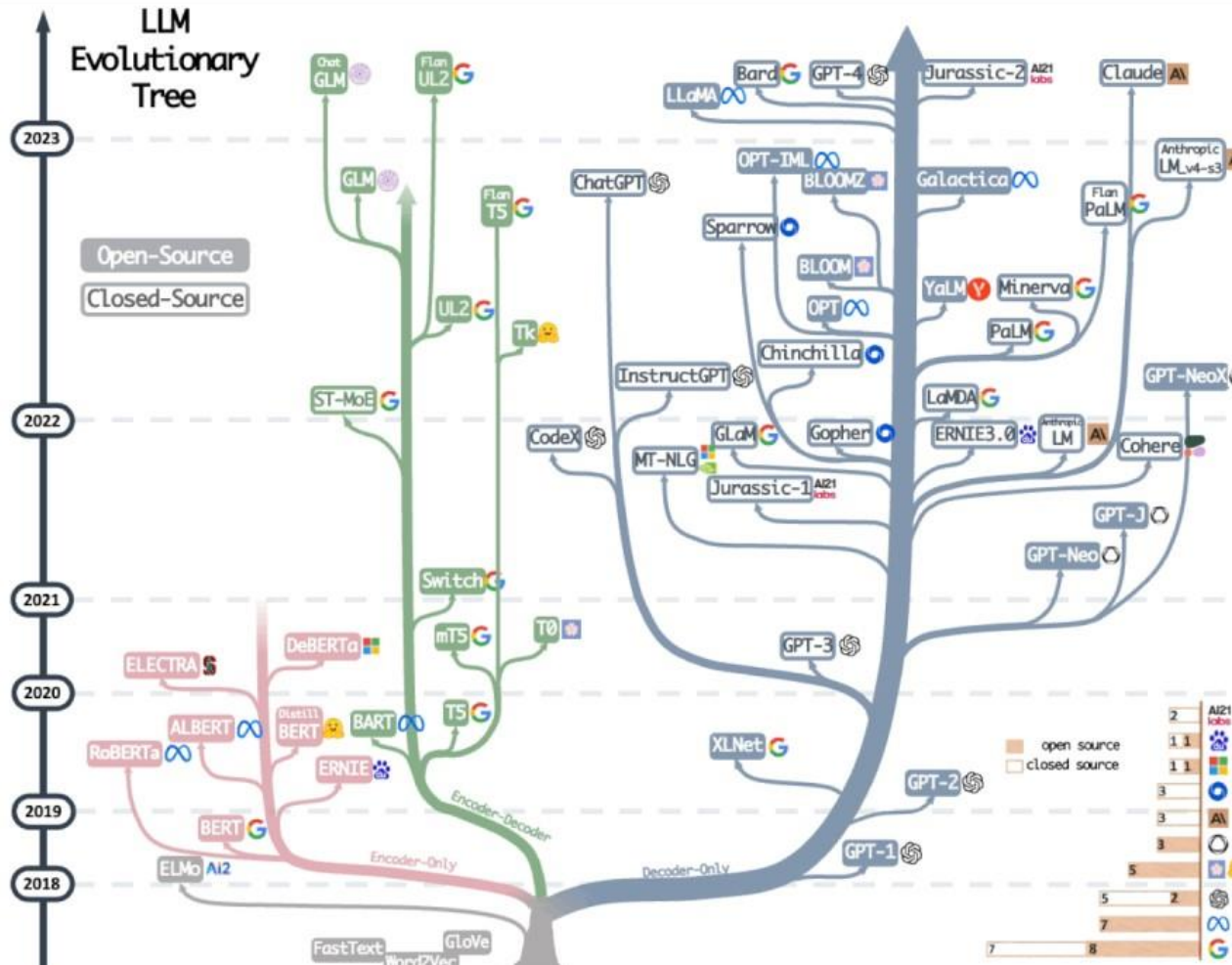
How do we build in mitigations, what-if scenarios, and strategies for improvement?



How do we assess the effectiveness not only in what happened, but what *may* happen in the future?



How do we create dynamic training content libraries with constraints for operational relevance?



- I. Large amounts of data
- II. Large scale compute
- III. Transformer architecture
- IV. In-context learning
- V. Mixture of Experts
- VI. Reinforcement Learning
- VII. Agentic workflows

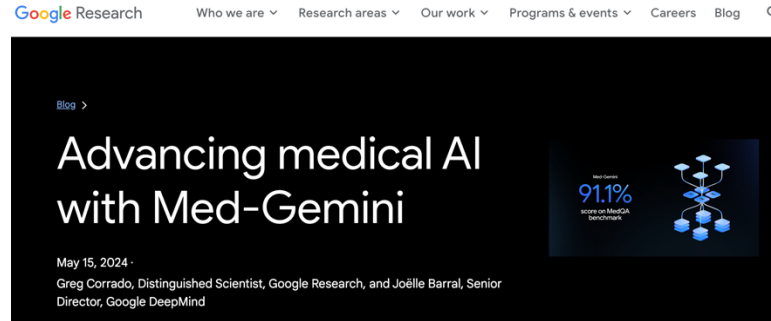
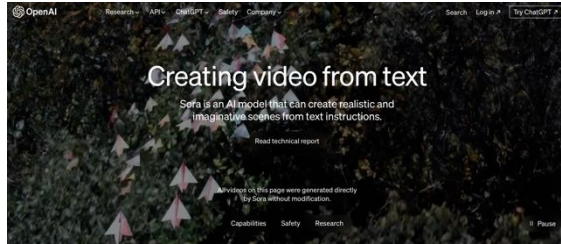


[https://www.ted.com/talks/yejin\\_choi\\_why\\_ai\\_is\\_incredibly\\_smart\\_and\\_shockingly\\_stupid/c?language=en](https://www.ted.com/talks/yejin_choi_why_ai_is_incredibly_smart_and_shockingly_stupid/c?language=en)



<https://www.youtube.com/watch?v=DeSXnESGxr4>

**Gigantic! Trained on 1 – 2 trillion tokens = human reading 8 hours a day for 22 thousand years**



WILL KNIGHT BUSINESS DEC 28, 2024 1:00 PM

## OpenAI Upgrades Its Smartest AI Model With Improved Reasoning Skills

A day after Google announced its first model capable of reasoning over problems, OpenAI has upped the stakes with an improved version of its own.

Large Language Model

## Introducing Meta Llama 3: The most capable openly available LLM to date

April 18, 2024



## GPT-4o System Card

This report outlines the safety work carried out prior to releasing GPT-4o including external red teaming, frontier risk evaluations according to our Preparedness Framework, and an overview of the mitigations we built in to address key risk areas.

[View PDF version](#)

Announcements

## Claude 3.5 Sonnet

Jun 20, 2024 • 4 min read

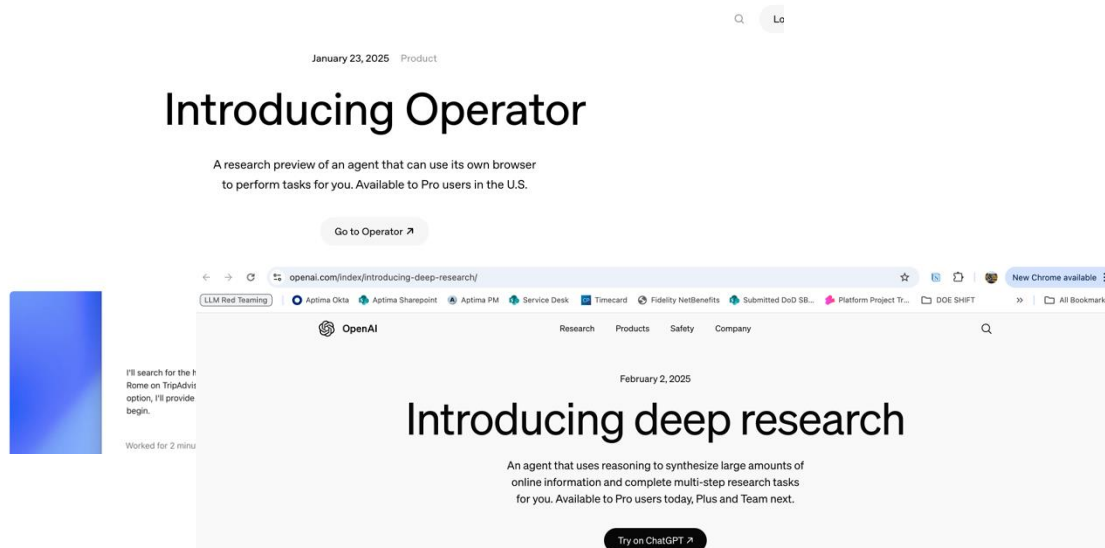
[Try on Claude.ai](#)

## Introducing OpenAI o1

We've developed a new series of AI models designed to spend more time thinking before they respond. Here is the latest news on o1 research, product and other updates.

[Try it in ChatGPT Plus ↗](#)

[Try it in the API ↗](#)



## Eagle 2

updated 13 days ago

Eagle 2 is a family of frontier vision-language models with vision-centric design. The model supports 4K HD input, long-context video, and grounding.

 **nvidia/Eagle2-1B**

 Image-Text-to-Text • Updated 8 days ago •  2.98k •  17


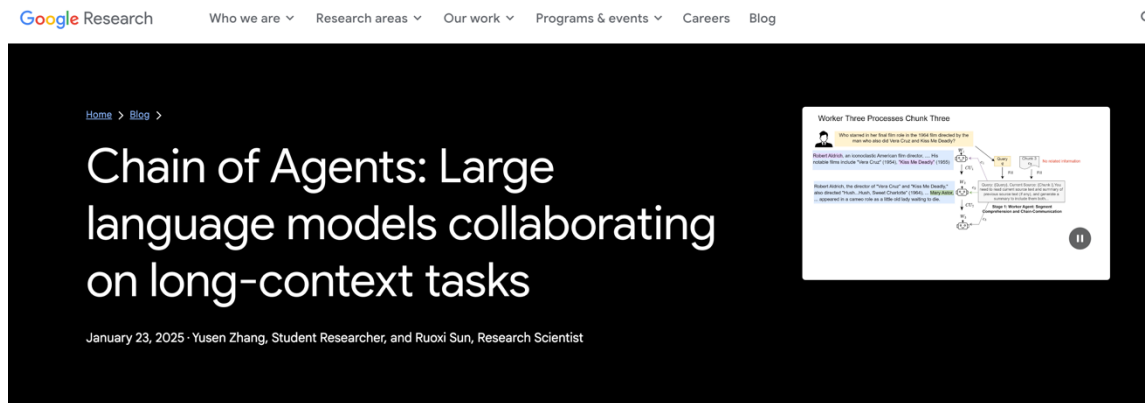
 **nvidia/Eagle2-2B**

 Image-Text-to-Text • Updated 8 days ago •  503 •  11

 **nvidia/Eagle2-9B**

 Image-Text-to-Text • Updated 8 days ago •  1.99k •  35



## Qwen2.5-Max: Exploring the Intelligence of Large-scale MoE Model

January 28, 2025 · 3 min · 561 words · Qwen Team | Translations: [简体中文](#)

## AutoGen v0.4: Reimagining the foundation of agentic AI for scale, extensibility, and robustness

Published January 14, 2025



## The Shift from Models to Compound AI Systems

*Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, Ali Ghodsi*

Feb 18, 2024

**Compound AI** systems with multiple calls to models, retrievers, or external tools

- 60% of LLM applications use some form of **retrieval-augmented generation (RAG)**, and 30% use multi-step chains
- Chaining strategy that exceeded GPT-4's accuracy on medical exams by 9%
- RAG, self-refinement, prompt-tuning and supervised fine-tuning techniques guide models in utilizing external knowledge.

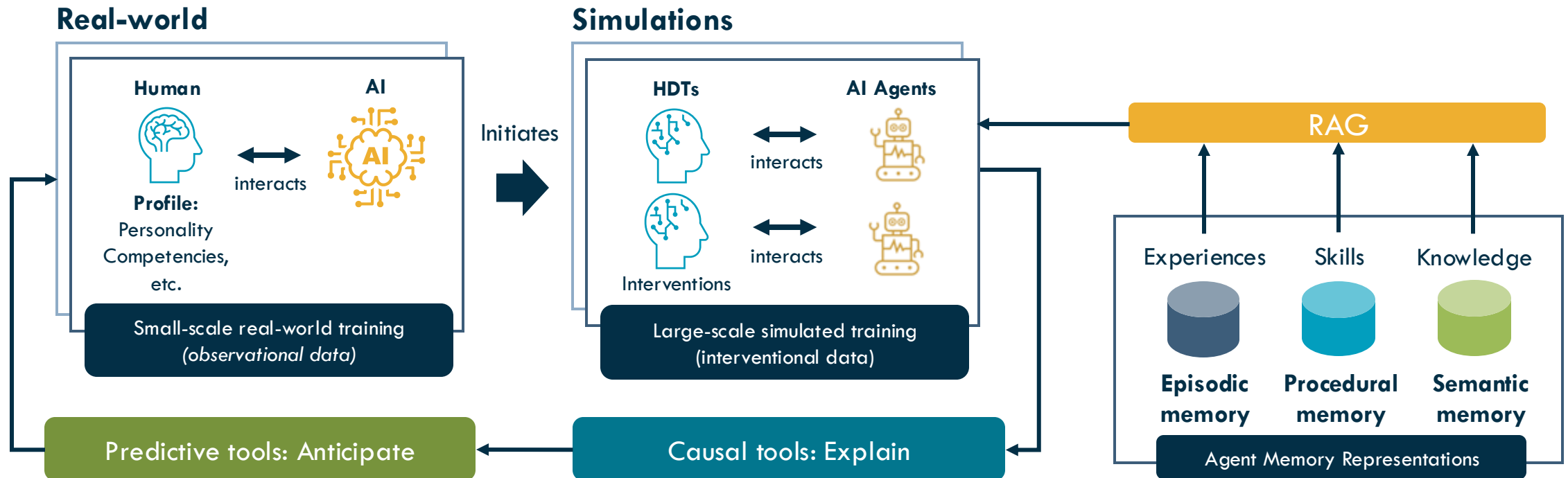
<https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/>

AI System	Components	Design	Results
<a href="#">AlphaCode 2</a>	<ul style="list-style-type: none"> <li>Fine-tuned LLMs for sampling and scoring programs</li> <li>Code execution module</li> <li>Clustering model</li> </ul>	Generates up to 1 million solutions for a coding problem then filters and scores them	Matches 85th percentile of humans on coding contests
<a href="#">AlphaGeometry</a>	<ul style="list-style-type: none"> <li>Fine-tuned LLM</li> <li>Symbolic math engine</li> </ul>	Iteratively suggests constructions in a geometry problem via LLM and checks deduced facts produced by symbolic engine	Between silver and gold International Math Olympiad medalists on timed test
<a href="#">Medprompt</a>	<ul style="list-style-type: none"> <li>GPT-4 LLM</li> <li>Nearest-neighbor search in database of correct examples</li> <li>LLM-generated chain-of-thought examples</li> <li>Multiple samples and ensembling</li> </ul>	Answers medical questions by searching for similar examples to construct a few-shot prompt, adding model-generated chain-of-thought for each example, and generating and judging up to 11 solutions	Outperforms specialized medical models like Med-PaLM used with simpler prompting strategies
<a href="#">Gemini on MMLU</a>	<ul style="list-style-type: none"> <li>Gemini LLM</li> <li>Custom inference logic</li> </ul>	Gemini's CoT@32 inference strategy for the MMLU benchmark samples 32 chain-of-thought answers from the model, and returns the top choice if enough of them agree, or uses generation without chain-of-thought if not	90.04% on MMLU, compared to 86.4% for GPT-4 with 5-shot prompting or 83.7% for Gemini with 5-shot prompting

# Compound AI (CAI) Systems for Training and Readiness<sup>[1]</sup>

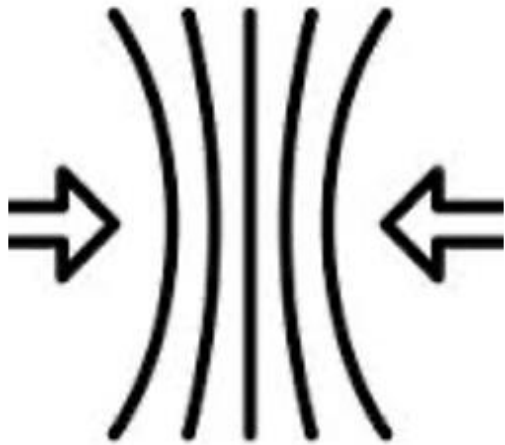
## Models, Agents, Tools and API calls

<sup>1</sup><https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/>



Volkova, Rebensky et al., Compound AI Ecosystem: Agents and Tools to Improve Training and Learning. Interservice/Industry Training, Simulation and Education Conference 2024.

Operational Impact and Novelty:  
Proactive Resiliency Building using Agentic AI Workflows



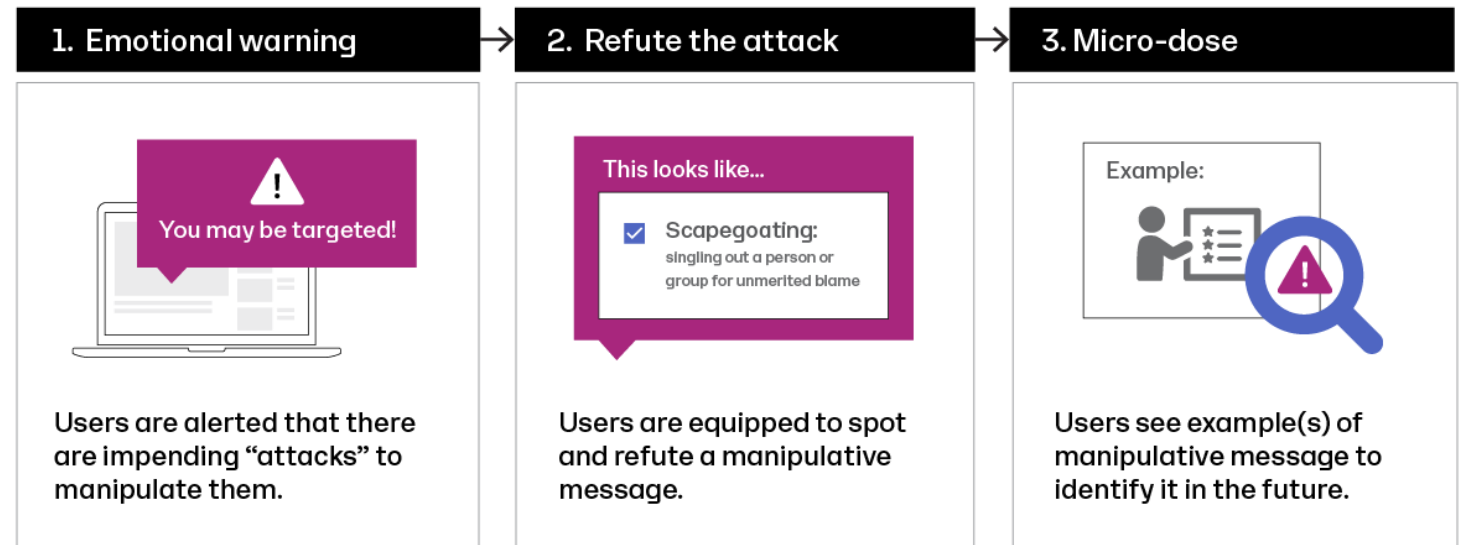
## Building Resilient Information Ecosystems

- Compound AI system with agentic workflows: twister, detector, defender and assessor agents
- Human digital twin agents with psycho-demographics and advanced memory representations for population simulation
- Counterfactual analysis to measure messaging effectiveness for operations in the information environment



# Inoculation Theory

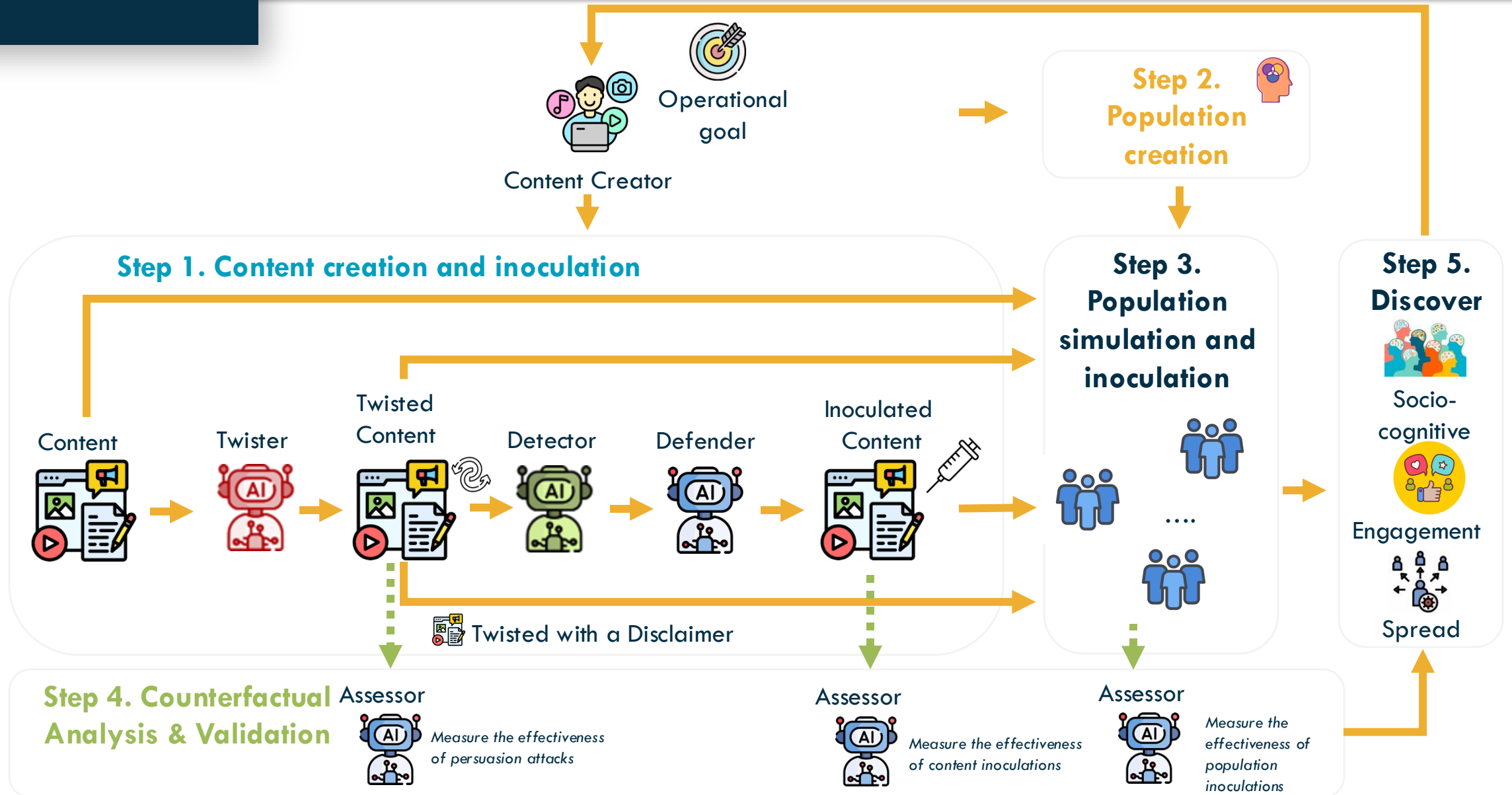
- Inoculation theory is a communication and persuasion concept developed by social psychologist William McGuire in the 1960s.
- Theory suggests that people can be "immunized" against persuasive attacks on their beliefs.
- Two key mechanisms:
  - **Threat** - People are warned about potential challenges to their beliefs, making them aware that their views might be attacked.
  - **Refutational preemption** - People are exposed to weakened versions of opposing arguments along with counterarguments.



McGuire, W. J. (1961). The Effectiveness of Supportive and Refutational Defenses in Immunizing and Restoring Beliefs Against Persuasion. *Sociometry*, 24(2), 184-197.

McGuire, W. J., & Papageorgis, D. (1961). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *Journal of Abnormal and Social Psychology*, 62(2), 327- 337.

# Cognitive Security and Resilience Platform for on-the-Job Analyst Training and Augmentation



## Agent attributes:

- Profession
- Education (High School, Bachelor's, Master's, Doctoral).
- Income bracket (Low, Middle, or High).
- Age (Young Adult, Adult, Elder).
- Gender
- Religiosity (0-100)
- Religion
- Ethnicity
- OCEAN Personality Traits (0-100) for each of the 5
- Myers-Briggs Type
- IQ
- Political Party (Conservative, Liberal, Moderate, Libertarian, Other)
- Reddit Username
- Favorite Subreddit
- Hobbies: Personal interests and leisure activities.
- Favorite Movies: List of the individual's most-liked movies.
- Favorite Shows: Television or streaming shows they enjoy most
- Family Bonds: Close family relationships (e.g., Mother, Sister)
- Free Time Constraints: Limitations on available leisure time
- Daily Responsibilities: Regular tasks or obligations, typically related to work
- Activities: Extracurricular or community-based activities
- Fears: A list of specific fears or phobias
- Goals: Key aspirations or objectives the person has
- Physical Location: Current geographic location of the person

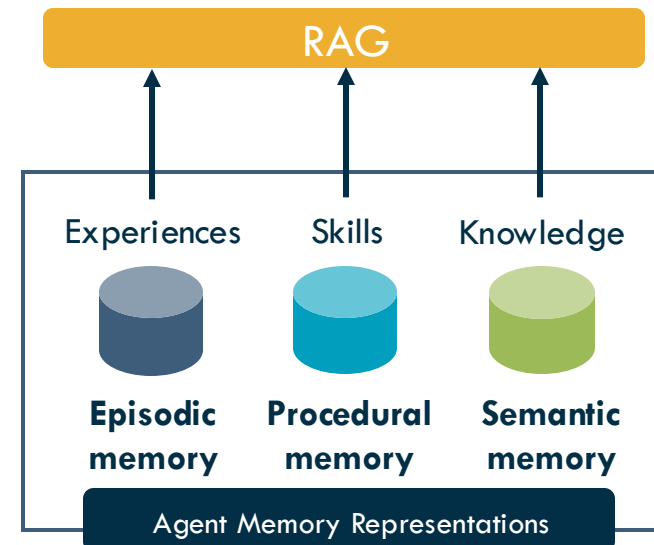
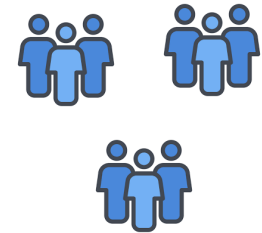
## Cognitive Distortions:

- All-or-Nothing Thinking
- Overgeneralization
- Mental Filtering
- Jumping to Conclusions
- Magnification and Minimization
- Emotional Reasoning
- Catastrophizing

## Dark Triad:

- Narcissism: Whether or not the individual displays narcissistic traits (Yes/No).
- Machiavellianism: Indicates if the person shows manipulative behavior traits (Yes/No).
- Psychopathy: Whether the individual shows traits related to psychopathy (Yes/No).

## Step 3. Population simulation and inoculation



- **Single Exposure** (cleared memory): Agents act in Reddit simulation with raw, twisted, inoculated, and disclaimer-based posts. Always “refreshing” agents to baseline (e.g., wipe memory).
- **Multiple Exposure** (memory maintained): Agents perform along two paths: either they see a disclaimer or are inoculated (e.g., content altered by BRIES Detector & Defender) and then are presented with attacked articles.
- **Treatments:**
  - Content type
  - Cognitive distortions
  - Attack type
  - Agency
- During simulation we measure **thread’s spread & engagement**, identifying **socio-emotional-cognitive (SEC) measures**, and a focus on agent’s **pre-posting thoughts** as our outcome.

With a clear **understanding of treatment effectiveness**, we can **prioritize** actions to drive **measurable change** in how content is perceived and **quantitatively measure the effects** of varied treatments

- Sentiment

Hugging Face Sentiment

<https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-English>

- Emotions

Bert-base-uncased trained for emotions

<https://huggingface.co/google-bert/bert-base-uncased>

- Empathy (Intent and Emotion)

EmpGPT-3

<https://github.com/passing2961/EmpGPT-3>

- Toxicity

Detoxify

<https://github.com/unitaryai/detoxify>

- Connotations, Perspectives, Attitudes

Lexicon-based analytics

<https://hrashkin.github.io/connframe.html>

- Moral Values (Harm, Fairness, Purity, Authority, Ingroup)

Lexicon-based analytics

<https://moralfoundations.org/wp-content/uploads/files/downloads/moral%20foundations%20dictionary.dic>

- Subjectivity

Lexicon-based analytics

<https://hrashkin.github.io/factcheck.html>

# Population Simulation Experimental Details

- Cognitively-inspired agents are randomly selected from a pool of 100 unique agents
  - Rich psycho-demographic profiles
  - Cognitive distortions
- Two experiments:
  - Single population exposures:
    - Agent memories cleared after each treatment trial
  - Multiple/Longitudinal population exposures:
    - Twisted + Disclaimer -> Twisted
    - Inoculated -> Twisted
- Dataset:
  - 200 DoD agency press releases

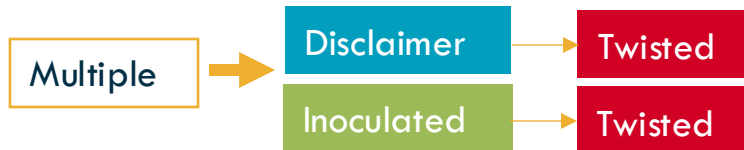
	Treatment Type	Count
Memory cleared (single exposure)	Raw Articles	200
	Inoculated Articles	170
	Twisted Articles (Appeal to Authority)	200
	Twisted Articles (Loaded Language)	200
	Twisted + Disclaimer	400
	<b>Total</b>	<b>1170</b>
Persistent memory (multiple exposures)	Preserved memories of 15 Twisted Articles + Disclaimer exposures then exposed to 15 new Twisted Articles	15
	Preserved memories of 170 Inoculated Articles then exposed to 400 Twisted Articles	400
	<b>Total</b>	<b>415</b>



## Results within the simulation environment:



- **On a single instance exposure**
  - Large effects based on cognitive distortions
    - Content types resulted in increases in joy, apprehensiveness, and subjectivity
    - Negative effects to number of thread comments
  - Agency type and attack type had minor effects
  - Inoculated content increased post belief and intent to share



- **Multiple-longitudinal exposure**
  - More varied responses based on cognitive distortions
    - Neutral, subjectivity, and joy metrics impacted
    - Impacted comment rate
  - Flatter distributions for behavioral actions

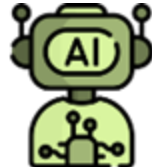
- Applications for these approaches don't stop at information resilience training
  - We are exploring a whole host of other applications, such as wargaming scenario generation
  - Compound AI systems will enable a variety of training and operational goals



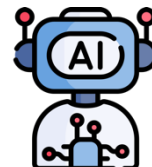
Fighter Pilot



Red Agents



Predictive Performance



Skill Assessment



Scenario Library



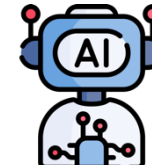
C2 Operator



Red Force



Decision Advisors



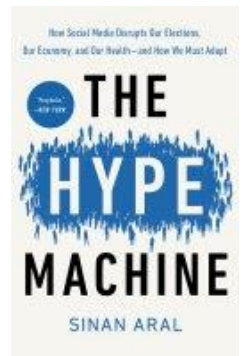
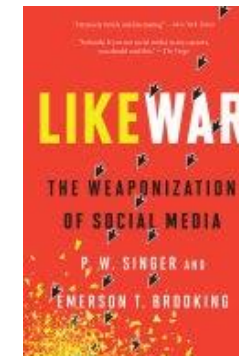
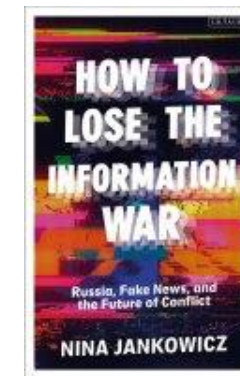
Skill Assessment



Decision Path Exercises



Source: GAO analysis of Department of Defense documentation. | GAO-23-105495



"The challenge of our time is not just to build smarter machines, but to build machines that make us smarter."

— Adapted from Douglas Engelbart's vision

## Causal Discovery and Inference

- **Volkova, S.**, et al. (2023). Explaining and predicting human behavior and social dynamics in simulated virtual worlds: reproducibility, generalizability, and robustness of causal discovery methods. *Computational and Mathematical Organization Theory*, 29(1).
- Saldanha, ... **S. Volkova**. (2020). Evaluation of Algorithm Selection and Ensemble Methods for Causal Discovery. In *Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems*.
- Glenski, M., & **Volkova, S.** (2021). Identifying Causal Influences on Publication Trends and Behavior: A Case Study of the Computational Linguistics Community. In *Proceedings of the First Workshop on Causal Inference and NLP* (pp. 83-94).
- Guo, G., Glenski, M.F., Shaw, Z.H., Saldanha, E.G., Endert, A., **Volkova, S.**, & Arendt, D.L. (2021). VAIN: Visualization and AI for natural experiments. *IEEE VIS 2021*.
- Cottam, J.A., Glenski, M.F., Shaw, Z.H., Rabello, R.S., Golding, A.J., **Volkova, S.**, & Arendt, D.L. (2021). Graph comparison for causal discovery. *Visualization in Data Science 2021*.

## Human-AI Integration and Trust

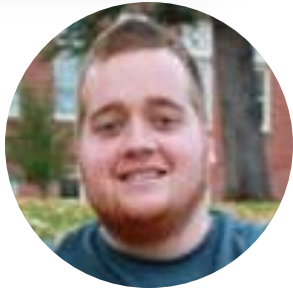
- Carmody, K., Ficke, C., Nguyen, D., Addis, A., **Rebensky, S.**, & Carroll, M. (2022). A Qualitative Analysis of Trust Dynamics in Human-Agent Teams (HATs). In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 66, No. 1, pp. 152-156). Sage CA: Los Angeles, CA: SAGE Publications.
- **Rebensky, S.**, Carmody, K., Ficke, C., Carroll, M., & Bennett, W. (2022). Teammates instead of tools: The impacts of level of autonomy on mission performance and human-agent teaming dynamics in multi-agent distributed teams. *Frontiers in Robotics and AI*, 102.
- **Rebensky, S.**, et al. (2021) Whoops! Something went wrong: Errors, trust, and trust repair strategies in human agent teaming. *Artificial Intelligence in HCI: Second International Conference, AI-HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings*. Cham: Springer International Publishing.
- Ezer, N., **Bruni, S.**, Cai, Y., Hepenstal, S. J., Miller, C. A., & Schmorow, D. D. (2019). Trust engineering for human-AI teams. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 322-326). Sage CA: Los Angeles, CA: SAGE Publications.

- **Volkova, S.**, Arendt, D., Saldanha, E., Glenski, M., Ayton, E., Cottam, J., Aksoy, S., Uslu, A., Unlu, M. S., & Yakut, C. (2023). Explaining and predicting human behavior and social dynamics in simulated virtual worlds: Reproducibility, generalizability, and robustness of causal discovery methods. *Computational and Mathematical Organization Theory*, 29(1), 220-241.
- Thomas, P., Saldanha, E., & **Volkova, S.** (2021). Studying information recurrence, gatekeeping and the role of communities during Internet outages in Venezuela. *Scientific Reports*, 11(1), 8137.
- **Volkova, S.**, Glenski, M., Ayton, E., Saldanha, E., Mendoza, J., Arendt, D., Shaw, Z., & Smith, K. S. (2021). Machine intelligence to detect, characterize, and defend against influence operations in the information environment. *Journal of Information Warfare*, 20(2), 42-66.
- **Volkova, S.**, & Jang, J. Y. (2018). Misleading or falsification? Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018* (pp. 575-583).
- Rashkin, H., Choi, E., Jang, J. Y., Choi, Y., & **Volkova, S.** (2017). Truth of varying shades: On political fact-checking and fake news. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2931-2937).
- **Volkova, S.**, Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 647-653).
- **Volkova, S.**, & Bachrach, Y. (2016). Inferring perceived demographics from user emotional tone and user-environment emotional contrast. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1567-1578).
- **Volkova, S.**, & Bachrach, Y. (2015). On predicting socio-demographic traits and emotions in social networks and implications to online self-disclosure. *Cyberpsychology, Behavior, and Social Networking*, 18(12), 726-736.
- **Volkova, S.**, Chetviorkin, I., Arendt, D., & Van Durme, B. (2016). Contrasting public opinion dynamics and emotional response during crisis. In *International Conference on Social Informatics* (pp. 312-329). Springer, Cham.

## Our Team



Gabe Ganberg  
Chief AI Architect



Zach Klinefelter  
IO Psychologist



Isabel Erikson  
Capability Lead



Myke Cohen  
Cognitive Scientist



Bob McCormack  
IPA Division Director



Avi Hiriyanna  
Sr. Research Engineer



Grant Engberson  
AI Engineer



Dr. Will Dupree  
Data Scientist,  
Causal Modeler



Spencer Lynch  
Principal Software  
Engineer



Dr. Jeff Beaubien  
Chief Behavioral Scientist



Ryan Kao  
Data Scientist,  
ML Engineer



Louis Penafiel  
Data Scientist, AI  
Capability Lead



Nick Abele  
Lead Software Engineer



Peter Bautista  
Trustworthy AI  
Capability Lead

Collaboration with University of Illinois Urbana Champaign, Carnegie Mellon University, University of William and Mary



This work is supported by the Defense Advanced Research Projects Agency (DARPA).

The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.



**Svitlana Volkova**

<https://www.linkedin.com/in/svitlanavolkova/>

[svolkova@aptima.com](mailto:svolkova@aptima.com)

**[www.aptima.com](http://www.aptima.com)**