Snow Globe Multi-Player AI System

Lessons from Human-AI Teaming in War Games

Andrea Brennan, Rachel Grunspan, Daniel Hogan, Jessica D. Smith, Elizabeth VanderVeen

As the US Intelligence Community continues to adapt to emerging threats and rapid technological change, human-AI collaboration is becoming a critical enabler of mission success. To meet this need, CIA's Directorate of Digital Innovation (DDI) Futures team and In-Q-Tel (IQT) have collaborated on a research project to explore how humans and conversational AI can work together more effectively and, ideally, achieve outcomes neither could accomplish alone. This project leverages Snow Globe, a multi-player AI system built by IQT's Applied Research Team, that uses large language models (LLMs) to play open-ended war games. Through a series of jointly designed games in which human participants play alongside (or against) simulated personas, the team has demonstrated how AI war games can serve as a testbed for human-AI teaming in intelligence work.

All statements of fact, opinion, or analysis expressed are those of the author and do not reflect the official positions or views of the US government. Nothing in the contents should be construed as asserting or implying US government authentication of information or endorsement of the author's views.

Snow Globe Multi-Player AI System

Our first jointly designed game, held in April 2025, was guided by several research questions:

- What new techniques and tradecraft emerge when conversational AI becomes a trusted partner for intelligence officers? How can this enable a continuous cycle of discovery and insight?
- How might the integration of conversational AI into intelligence workflows transform the way officers approach complex challenges and make decisions?
- In what ways can conversational AI enhance human-machine collaboration? How can it lead to more effective and efficient mission outcomes?[a]

By pairing six human players with AI agents, each customized to simulate a predetermined temperament and disposition, the team identified several preliminary findings that confirm the potential for conversational AI to complement and enhance intelligence officers' tradecraft. In this article, we summarize those findings, explain why the IC needs a rapid and adaptable way to test human-AI teaming, and argue that AI-enabled war games offer a uniquely compelling solution.

In 2025, the rapid evolution of AI capabilities makes it tempting to assume that every AI problem has a purely technical solution. US government and business leaders, however, make this assumption at their peril because it overlooks the very real challenges and opportunities of human-AI teaming or "the collaborative interaction between humans and artificial intelligence systems to achieve a common goal."[1]

While full automation is appealing, most organizations find that the benefits will be judged by how well AI integrates with existing workflows to augment human capabilities. Successful integration, however, is about more than technology; it requires "clear communication and mutual understanding between humans and AI systems."[2]

As leaders adopt AI, they must consider the training, skill development, and expectation-management critical to its success. They also need ways to assess the return on investment for new capabilities—determining, for example, how (or whether) a model's performance gains translate across different applications. Finally, they also need effective strategies to divide work between humans and AI agents so that organizations can fully leverage the strengths of both.

There are no off-the-shelf solutions to the challenges of human–AI collaboration. Organizations are still learning how to define the roles, processes, and expectations needed to team effectively with AI. This transformation is unlike any before it, and best practices are still emerging.

Collaboration in any form is inherently organizational and messy, and human–AI collaboration won't be "solved" by a technical quick fix. That said, the IC cannot afford to wait. AI capabilities are evolving too quickly, and the pressures to deploy them are too urgent. We must experiment, learn, and adapt as this chapter unfolds.

Intelligence officers need flexible frameworks that allow them to experiment safely and iterate now, but, in high-stakes intelligence applications, the cost of failure may be too high for this type of trial and error. For these use cases, AI-enabled war games offer a valuable path forward.

Our joint IQT-DDI team has designed and run war games that simulate future intelligence workflows, allowing participants to use curated AI assistants to augment their tradecraft and decisionmaking. These games provide an unclassified, low-risk environment where officers can test new capabilities, apply them

a. Snow Globe's architecture is described in a preprint [https://arxiv.org/abs/2404.11446v1], and the source code is publicly available on GitHub. [https://github.com/IQTLabs/snowglobe]

in different ways, and share feedback on what works best.

About Snow Globe

To support these games, IQT built Snow Globe, a proof-of-concept AI platform that uses LLMs to play open-ended war games. Its flexible, multi-player architecture is model agnostic and can be used to automate multiple aspects of war-gaming, including:

- Augmenting human players' abilities with customized AI assistants
- Simulating AI personas to serve as allies or adversaries
- Generating scenario inputs (or "injects") dynamically based on player decisions
- Evaluating game outcomes and behavioral dynamics
- Running multiple iterations of fully automated games to anticipate possible outcomes and second-order effects

War-game decisionmaking exercises that "include everything from small seminar exercises ... to large multi-day, multi-team war games"[3] are powerful tools for rehearsing real-world decision making and observing how people solve problems under simulated conditions.[45] Importantly, war games do not need to focus on armed conflict.[6] In prior work, IQT used Snow Globe to explore scenarios involving both AI incident response and geopolitical analysis.[7] Other researchers have shown that war-gaming can also support complex decisionmaking in areas like disaster response.[89]

The value of automating war games has long been recognized.[10, 11, 12] Most prior efforts, however, have focused on quantitative games, where players are limited to predefined sets of actions. By leveraging LLMs, Snow Globe enables the automation of open-ended, qualitative games, where players are free to take a wide range of creative actions.

Snow Globe can respond to any move a player makes and can dynamically introduce unexpected scenario injects based on gameplay. The ability to generate qualitatively novel responses is Snow Globe's central innovation.

Real life rarely follows predefined rules, which is why open-ended war games can more accurately reflect real-world complexities. These games offer unique value as tools for training, planning, and decision-support. By introducing AI into the design, game-play, adjudication, and analysis of open-ended war games, Snow Globe creates new opportunities to explore (and test) human-AI collaboration. We have observed Snow Globe contributing meaningfully to human players' problem-solving discussions and participants have reported that AI integration significantly enriched their overall experience.

Snow Globe also allows us to simulate a wide range of AI personas. In geopolitical games, we have experimented with generic characters as well as personas modeled after historical foreign leaders. In these cases, we generated Snow Globe personas based on publicly available information about historical figures' political leanings, leadership style, and decision-making preferences sourced from sites such as Wikipedia. Across multiple experiments, we found that the inclusion of personas can significantly shape the trajectory of game-play.

We have tested several types of AI assistant personas. Some are designed to reflect specific temperaments and dispositions; others aim to adapt to an individual's preferred learning style. Still others are grounded in curated document sets and guidance, enabling them to provide domain specific expertise such as legal reasoning or a particular intelligence tradecraft. It is worth noting that the goal of this research is not to replicate the expertise of human specialists, but rather to explore the potential of AI to augment and support human intelligence officers' capabilities.

What We've Learned

In April 2025, we conducted our first jointly designed, AI-enabled

war game; it was held at IQT with six human participants. During the game, participants collaborated with both their AI assistants and one another to develop a strategic response to a geopolitical scenario. Below, we summarize key findings from the game and the work leading up to it. Prior to the April war game, a preliminary "play-test" helped us refine Snow Globe's AI assistant personas and determine which LLM to use.

Align Capabilities with Participants' Expectations

The preliminary play-test was conducted at IQT using an on-premise version of Snow Globe, powered by a local instance of the open-source Mistral 7B model. Several participants attempted to use their assistants to retrieve information not only about the fictional world of the game, but also about real-world events and current data. Although we did not expect the Mistral 7B model (relatively small by LLM standards, with a 2022 knowledge cutoff) to behave like a search engine, many participants appeared to expect exactly that. When their AI assistants failed to meet those expectations, some participants become frustrated, which colored their view of Snow Globe and the game-play experience.

We suspect that prior use of ChatGPT likely shaped participants' expectations about how all AI tools should behave. The rapid evolution of consumer-facing models—especially free, highly capable tools like ChatGPT—continues to raise the bar for what users expect from enterprise AI systems.

To help align expectations in the April game, we took two steps. First, we provided a Snow Globe user guide with clear instructions and sample prompts. Second, we ran Snow Globe using the much larger GPT-4o model, which offered significantly better performance on information-retrieval tasks compared to the smaller Mistral 7B model.

Show the Possibilities

Initially, we envisioned personalized Snow Globe personas for each participant, that would respond to participants' preferences along with data about what helped them perform best. In advance of the preliminary play-test, we surveyed participants about what they (thought they) wanted in an AI assistant but, given that we were preparing for our first game, we had little performance data to incorporate.

During the preliminary play-test, we were surprised by how similarly our "personalized" personas behaved. Many factors probably contributed to this—we saw opportunities to refine both our survey instrument and our implementation of personas—but we also wondered how would participants know what they wanted from Snow Globe, before they saw what the tool could offer them?

With new capabilities, users may not know what they want until they see what is possible. Even then, building the features people want is not always the best way to address their underlying needs. Recognizing this, we changed our strategy for War Game 1 and offered participants a curated a set of personas that showcased a range of behaviors and dispositions. We intend to revisit the idea of personalization once our participants have more experience with Snow Globe and we have more data on game outcomes.

Conversational AI Could Enhance Intelligence Officers' Workflows

During War Game 1, Snow Globe's AI assistants provided facts, analysis, and suggestions that blended the fictional scenario elements with real-world context, offering detailed outputs that enriched the game-play experience. On our post-game survey, participants gave the experience an average rating of 4.25 out of 5. All respondents agreed that their AI assistant "seemed accurate," meaning that it provided outputs containing reliable information.

Although the six human participants brought varying levels of AI experience, all made frequent use of their AI assistants and were generally receptive to the recommendations they received.[a] Overall, feedback was strongly positive. Three-quarters of participants said their AI assistants helped them, or their team achieve their goals during the game. We observed occasional issues with truncated responses, but in general, response lengths varied appropriately with the prompt, from a single line to outputs exceeding 1,000 words.

Participants used their AI assistants for six different tasks during the game:

- Getting oriented to the scenario
- Retrieving fact-based information
- Requesting explanations or clarifications about key concepts, actors, and events
- Comparing multiple courses of action to support decisionmaking
- Organizing, structuring, and sequencing information
- Composing and refining recommendations and written outputs

As expected, participants' use of AI shifted over time, indicating potential to tailor assistant behavior to different stages of the analytical workflow. Half of survey participants reported that their AI assistant was able to handle ambiguous prompts. The remaining participants were neutral, neither agreeing nor disagreeing with that assessment. When asked whether their AI assistant suggested a course of action they had not previously considered, three-quarters of participants said yes.

Personas are not "One Size Fits All"; Neither are LLMs

Brevity Matters

Although we did not ask participants about this directly, their behavior suggested that long, overly detailed responses were often viewed negatively. LLMs can produce lengthy outputs with ease, but human bandwidth hasn't grown to match, nor has our appetite for reading repetitive or bloated content. In some cases, long AI responses overwhelmed or discouraged users, even when the content was valuable. Prompting can do a lot to improve clarity and conciseness, but not all war game participants have the experience to shape model output effectively. To bridge this gap, we recommend providing sample prompts that help users generate shorter, more readable responses from their AI assistants.

Personas

Before game-play began, participants selected AI assistants from a curated set of personas:

- Pacifist: Prefers the least aggressive course of action
- Aggressor: Prefers the most aggressive course of action
- Tactician: Focuses on immediate problems and short-term outcomes
- Strategist: Focuses on broad challenges and long-term outcomes
- Detail Oriented: Provides specific, detailed plans
- Big Picture: Emphasizes overarching goals and priorities

All six personas were used during the game and, as expected,

a. Notably, we observed differences in how players who had more hands-on experience with AI used their assistants. For example, during War Game 1, one participant had significantly more prompt engineering experience than the other players. Our analysis of chat logs showed that the more experienced participant demonstrated greater variation in prompt type, greater emphasis on refining AI-generated output, and was the only participant who explicitly asked the AI assistant to share its sources and show its reasoning. We view this as an indication of how improving officers' digital acumen will probably change the way they use AI tools.

different personas responded in distinctive ways.

In some cases, players prompted their AI assistants in ways that aligned with each AI persona's disposition by, for example, asking the Aggressor for military options. (It responded with a detailed list of aggressive military strategies.) However, we also observed cases where the AI assistant's behavior reflected its persona even without direct prompting. In one instance, the Tactician persona was asked about Ukraine. It addressed both short-term and long-term considerations but ultimately focused the player's attention on Ukraine's immediate role, consistent with its tactical orientation.

We see considerable value in giving participants the ability to choose the kind of advice and assistance they receive from AI. That said, we have also observed ways in which the use of a particular LLM might shape what choices are available. For example, in testing we observed that responses generated with Mistral 7B frequently leaned toward diplomatic negotiation, even when behavior was inconsistent with the assigned persona. While this could be mitigated through prompt engineering, a larger model we tested (GPT-3.5) exhibited this issue to a lesser degree. One advantage of using a relatively small LLM like Mistral 7B is that it can run locally with light hardware. However, we have found that larger models often capture persona nuances with greater fidelity.

While model selection plays a role in persona design, embedding written persona descriptions directly into prompts provides significantly more control. We are exploring how to further improve persona performance by equipping agentic assistants with tools like document retrieval.

One common criticism of LLMs is their proclivity to hallucinate, generating responses that sound plausible, but are inaccurate. Although this behavior is problematic in many applications, in war-gaming, it is actually a benefit. For AI to participate in war games effectively, it must be able to blend facts with the fictional world of the game. For example, our geopolitical scenario for War Game 1 involved countries with fictionalized names. With only a minimal amount of background context about these fictional nations, Snow Globe was able to generate plausible responses detailing their involvement in complex geopolitical dynamics.

Additionally, to adjudicate open-ended games, AI must be able to generate outcomes that no one planned for but that still make sense within the game's logic. The kind of grounded creativity required to introduce new events and evaluate their consequences is, essentially hallucination with purpose.

Next Steps

As we refine the Snow Globe platform and expand our library of AI personas, upcoming war games will explore more complex geopolitical scenarios and multi-domain crises to stress-test human-AI teaming under layered, dynamic conditions. We also plan to integrate curated, declassified datasets and synthetic intelligence data streams to create richer, more realistic operational environments that mirror the complexities of intelligence work. Our next war game will build on these efforts, and we invite collaboration from across the national security enterprise to help shape the next generation of human-AI collaboration.

About the authors: Andrea Brennan is senior vice president and deputy director of IQT Labs. Rachel Grunspan is the former director of Digital Futures in CIA's Directorate of Digital Innovation. Daniel Hogan is a senior data scientist in IQT Labs. Jessica D. Smith is a digital intelligence strategist in DDI Futures. Elizabeth VanderVeen is an AI strategist in DDI Futures. ■

Endnotes

1. Definition provided by GPT 4o, accessed via chat.iqt.org on July 10, 2025.
2. Ibid.
3. P. K Davis and P. Bracken, "Artificial intelligence for wargaming and modeling," *Journal of Defense Modeling and Simulation* (February 2022).
4. S. Burns, D. DellaVolpe, R Babb, N. Miller, and G. Muir, "War gamers' handbook: A guide for professional war gamers," tech. rep., Naval War College, November 2015.
5. M. F. Cancian, M. Cancian, and E, Heginbotham, "The first battle of the next war: Wargaming a Chinese invasion of Taiwan," report, Center for Strategic & International Studies, January 2023.
6. E. Lin-Greenberg, RB, C. Pauly, and J.G. Schneider, "Wargaming for international relations research," *European Journal of International Relations* 28, No. 1, (2022): 83–109.
7. To provide a library of off-the-shelf geopolitical games, IQT built an integration between Snow Globe and the ICB Project, a dataset of 496 historical geopolitical crisis scenarios. The ICB Project homepage can be found at https://sites.duke.edu / icbdata /data-collections/ A list of crisis scenarios is at https: //duke.app.box.com/s/ddtpvalv33dzyom0j7obk7whitmouzrl
8. Department of Homeland Security, "Homeland Security Exercise and Evaluation Program (HSEEP)." https://www.fema.gov/sites/default/files/
9. Department of Homeland Security, "Homeland Security Exercise and Evaluation Program (HSEEP)" January 2020. https://www.fema.gov/sites/default/files/2020-04/Homeland-Security-Exercise-and-Evaluation-Program-Doctrine-2020-Revision-2-2-25.pdf.
10. P.K. Davis and P.J.E. Stan, "Concepts and models of escalation," Report R-3235, RAND Corporation, May 1984.
11. J, Goodman, S. Risi, and S. Lucas, "AI and wargaming." arXiv:2009.08922, 2020.
12. Davis and Bracken, "Artificial intelligence for wargaming and modeling."

■