Human vs. Machine: Behavioral **Differences between Expert** Humans and Language Models in Wargame Simulations

Max Lamparth, Anthony Corso, Jacob Ganz, Oriana Skylar Mastro, Jacquelyn Schneider, Harold Trinkunas



How does human expert decision-making in conflicts differ to language model simulated decision-making?



Motivated by ongoing real-world government

tests of language models for military decisionmaking, we ask:



Move 1: Use New Al Weapon? (a) 'Fire at opposing vessels'

(b) 'Hold fire unless fired upon'

(d) 'Auto-fire'

(e) 'Auto-target, manual-fire'

Move 2: Opponent Posture (a1) 'Preserve Status Quo/Deter'

(a3) 'Defend'

(b) 'Activate Civilian Reserve/Draft'

(c) 'Surge Domestic Defense Production'

(g) 'Clandestine/Special Operations'

(h) 'Information Operations'

(i) 'Conduct Foreign Intelligence'

(j) 'Conduct Domestic Intelligence'

(k) 'Cyber Operations'





We compare the behavior of 214 national security experts with language model simulated players. While there is some high-level agreement, including similar study results when comparing treatment to control groups or when comparing response vectors to a random baseline, ...

┢═╋╣╎ ⊢₋ ╘ ┝╌╏╋══╋╌╞╕╌┥ **GPT-3.5 (95% Conf.)** í 🍎 GPT-4 (95% Conf.)

Fewer Counts More Counts Than Humans

____ i**⊨=9**==i Difference in Action Counts []

GPT-40 (95% Conf.)

... we see significant differences in individual actions and strategic approaches. The language model-simulated players show deviating tendencies, including more escalatory actions and different reactions to crucial scenario instructions.

All results combined raise concerns about the potential of language models to increase conflict risks if used in military decision-making.

Other Dependencies

Language model simulations are affected by whether they are instructed to simulate dialog between players; if they are, we get more aggressive and more chosen actions. The levels of behavioral consistency vary from aggressive or deescalatory behavior in move one to aggressive behavior in move two.

> Human Experin GPT-3.5 Experi **GPT-4** Experim **GPT-40** Experi

High-Level Behavioral Overlap



arXiv:2403.03407 @MLamparth lamparth@stanford.edu



Different Tendencies and Biases

	$p(agg_2 agg_1)$	$p(agg_2 des_1)$
ments	$0.94\substack{+0.06 \\ -0.08}$	$0.65\substack{+0.13 \\ -0.15}$
riments	$0.98\substack{+0.03 \\ -0.04}$	$0.85\substack{+0.08 \\ -0.09}$
nents	$0.99\substack{+0.01 \\ -0.03}$	$0.73\substack{+0.10 \\ -0.10}$
iments	$1.00\substack{+0.00 \\ -0.00}$	$0.86\substack{+0.08 \\ -0.08}$



Center f Security . Y a nd ternational Cooperation

zI SC FC žΣ