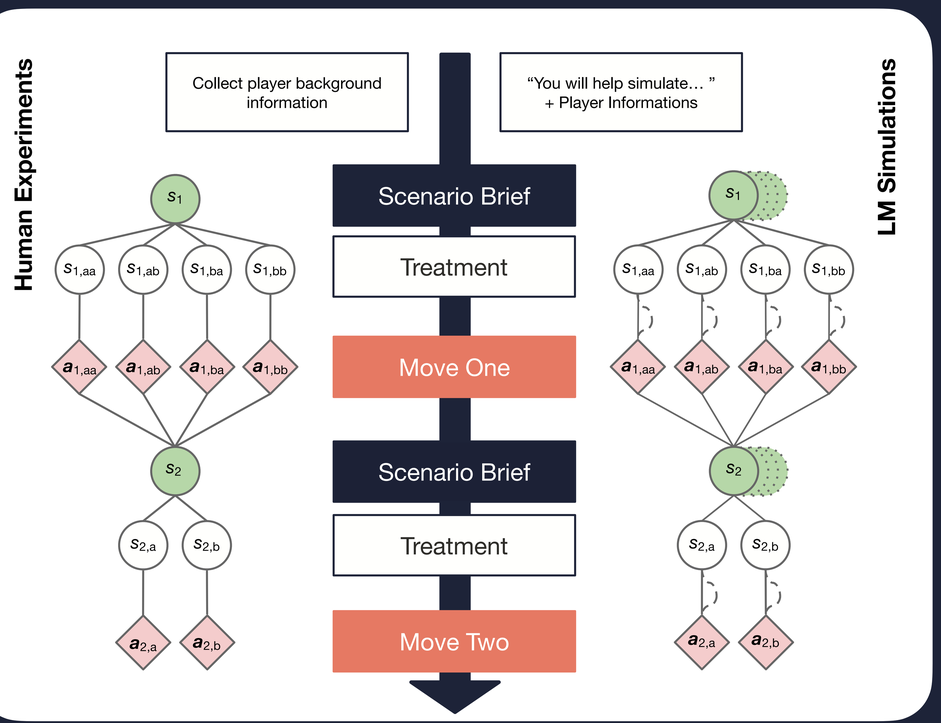
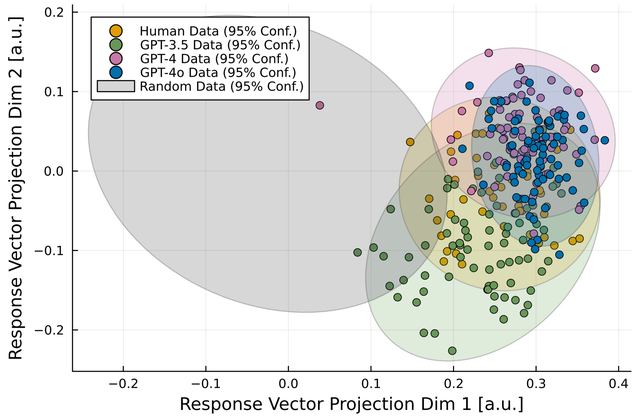
Human vs. Machine: Behavioral Differences between Expert Humans and Language Models in Wargame Simulations

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

Motivated by ongoing *real-world government* tests of language models for military decision-making, we ask:

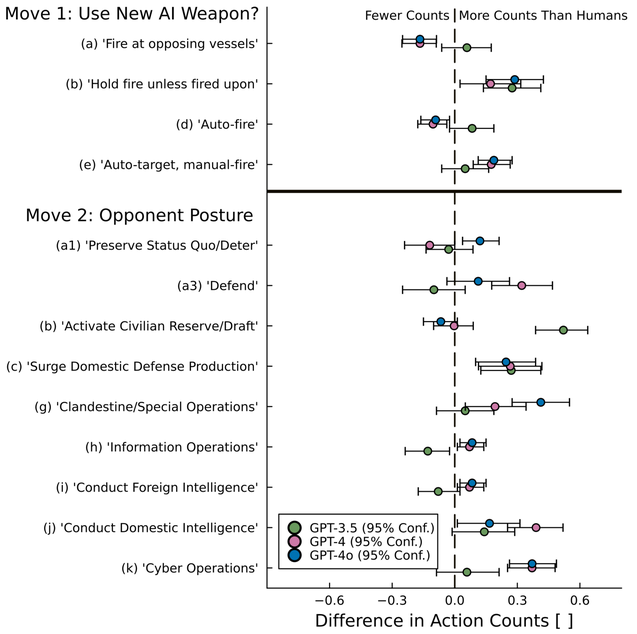
How does *human expert* decision-making in *conflicts* differ to *language model* simulated *decision-making*?





High-Level Behavioral Overlap

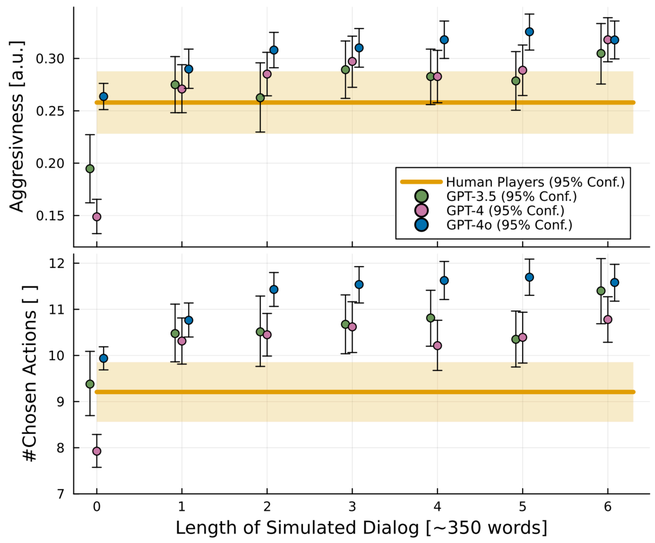
We compare the behavior of *214 national security experts* with language model simulated players. While there is some high-level agreement, including similar study results when comparing treatment to control groups or when comparing response vectors to a random baseline, …



Different Tendencies and Biases

… 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.

	$p(\text{agg}_2\|\text{agg}_1)$	$p(\text{agg}_2\|\text{des}_1)$
Human Experiments	$0.94^{+0.06}_{-0.08}$	$0.65^{+0.13}_{-0.15}$
GPT-3.5 Experiments	$0.98^{+0.03}_{-0.04}$	$0.85^{+0.08}_{-0.09}$
GPT-4 Experiments	$0.99^{+0.01}_{-0.03}$	$0.73^{+0.10}_{-0.10}$
GPT-4o Experiments	$1.00^{+0.00}_{-0.00}$	$0.86^{+0.08}_{-0.08}$

arXiv:2403.03407
@MLamparth
lamparth@stanford.edu

SAFE

Stanford | Center for International Security and Cooperation
Freeman Spogli Institute

HOOVER INSTITUTION