Sandia National Laboratories

Exceptional service in the national interest

Dangers and Benefits of Generative AI use in Wargaming

Ruby E. Booth

Connections UK 2025

SAND2025-06899C

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

U.S. DEPARTMENT OF ENERGY

NNSA

Agenda

The History of AI

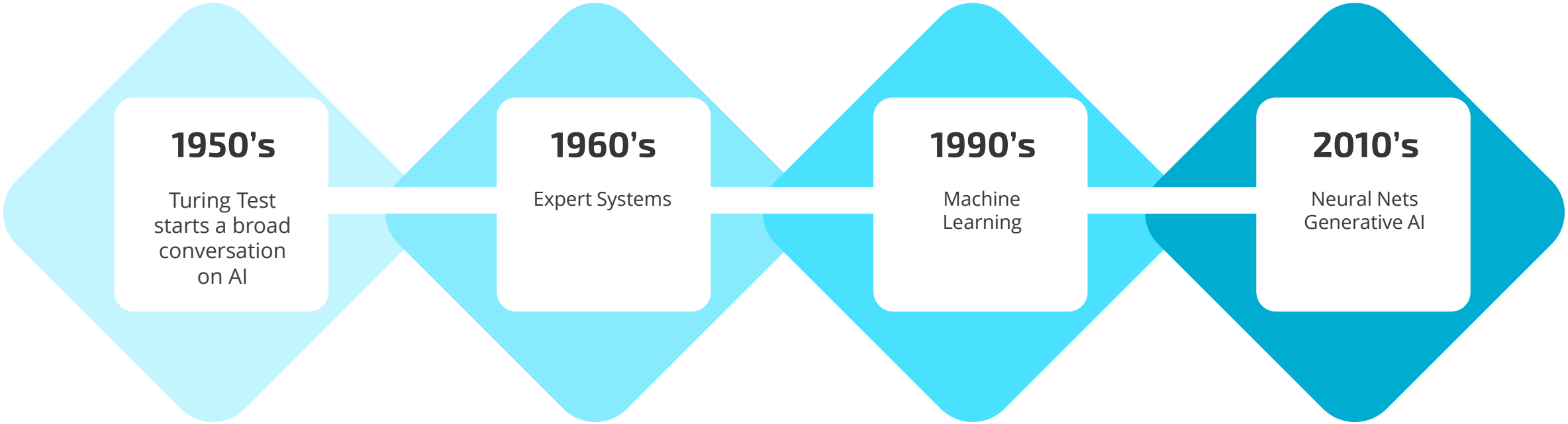
- Types
- Hype
- Going Wrong

Generative AIs

- Data and Reification
- Mitigations

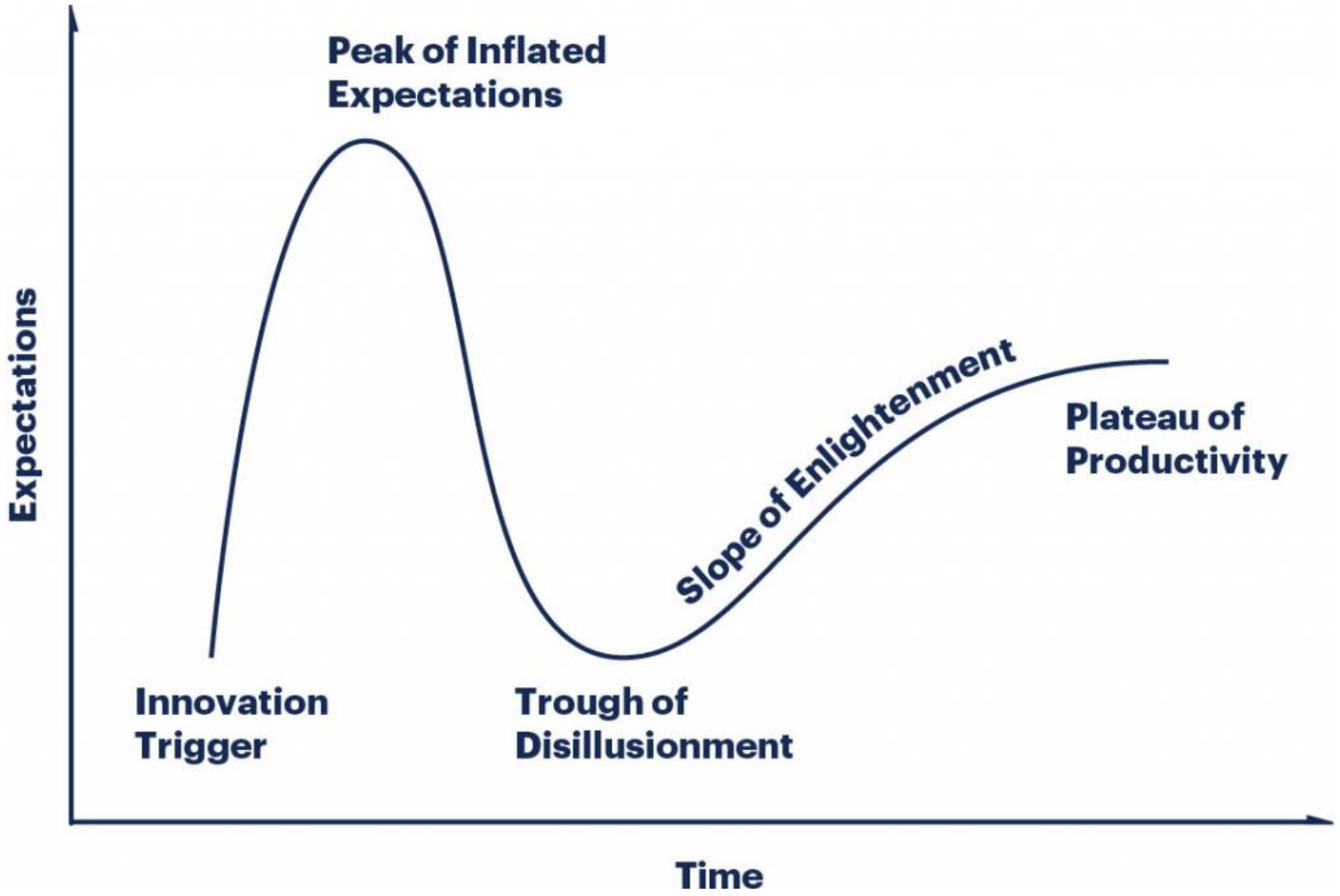
AI Use in Wargaming

The Evolution of AI | Timeline



Today's technology is built on the technologies of the past.

Gartner Hype Cycle and the Eccentricity of AI Winter



https://www.gartner.com/en/research/methodologies/gartner-hype-cycle

Types of AI and their function

Expert System

- uses a rules engine to perform tasks it was taught to do
- tries to perform accurate **actions**

Machine Learning

- uses what it learned from training data to **make accurate predictions** about new data
- tries to create accurate **labels**

Generative AI / Large Language Models

- uses what it learned from training data **create new data** that resembles the training data
- tries to create accurate new **content**

What Happens When an AI System Goes Wrong

Expert System

- makes wrong **actions**

Machine Learning

- make inaccurate **predictions**
- assigns inaccurate **labels**

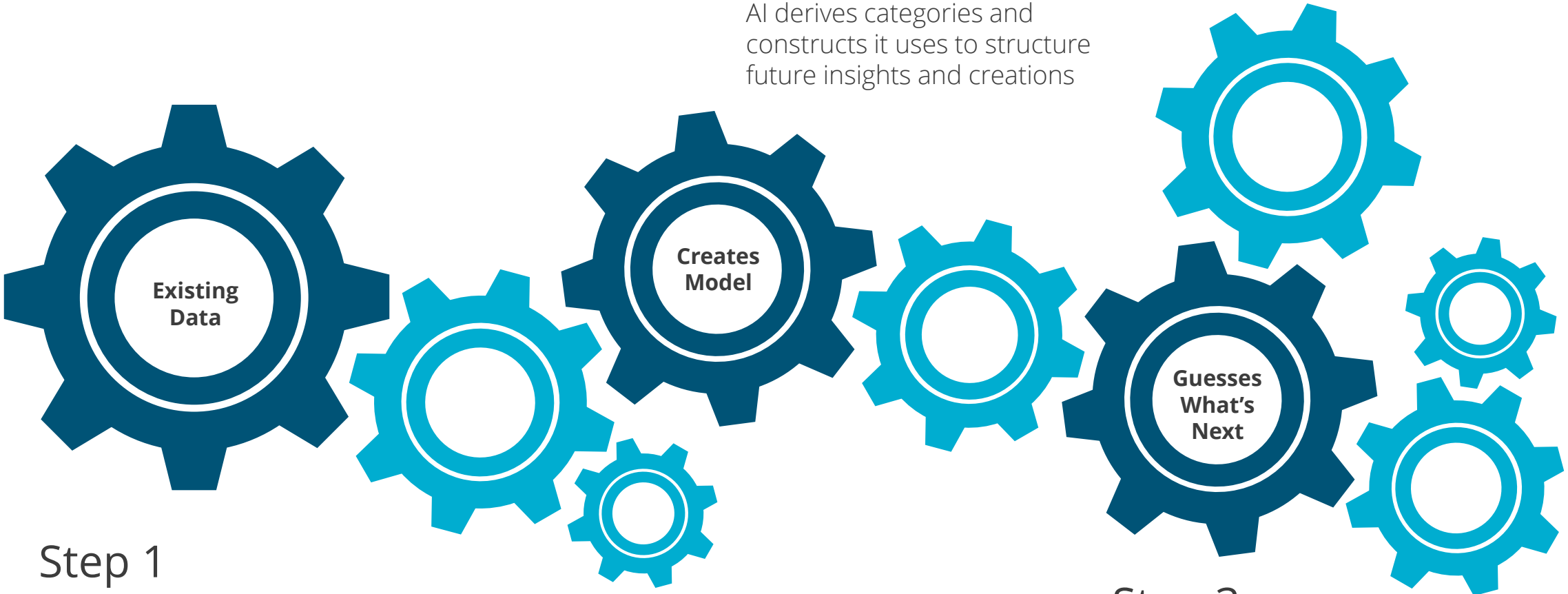
Generative AI / Large Language Models

- generates inaccurate **content**

Generative AI | Process

Step 2

AI derives categories and constructs it uses to structure future insights and creations



Step 1

All AI arise from existing data sets. They are trained and grow from vast data repositories for pre-existing information.

Step 3

Offers what it expects to come next...AI is always continuing an existing conversation

Data Sources and Reification

In the field of armed conflict, we are likely to train our Large Language Models on several types of information including:

- Real world examples that we have scoped and labeled
- Theoretical articles that express our biases and assumptions
- Cultural records that capture our perspectives in that historical context (e.g., newspapers, interviews, photographs)
- Game theoretic models built with parameters we defined, etc. (e.gg. structural deterrence models)
- Masses of military reports, documentation, all the artifacts of military bureaucracy

DANGER: Using Historical Data May Unintentionally Reify Existing Biases

AI Application Rules of Thumb

Use AI when tasks are low stakes, common in its corpus, and rely on *either* text *or* images (not both)

Wargaming examples:

routine communications with sponsors and participants, novel maps and icons, prototyping

Use AI when tasks are moderate stakes, time intensive, and can be reviewed

Wargaming examples: summarizing literature review and SME elicitation materials; drafting scenarios, injects, in-game “flavor”; initial mechanics recommendations; polishing and summarizing actions from bulleted lists, critiquing final reports

Use AI when tasks are high stakes, can be robustly validated, and other methods are not tractable

Wargaming examples: ???

AI APPLICATIONS TO AVOID

If everyone else jumped off a bridge…

What	Why	Why Not	Rule of Thumb
AI White Cell Adjudication	• Time pressures • Adjudication Complexity • Need for Sufficient and Sufficiently Trained Staff • Player guff	• Lose the ability to focus adjudication on the RQ • Different Player Guff • Confounds given unknown model bias	Use AI when tasks are moderate stakes, time intensive, and can be reviewed
AI Red Cell	• Limited Availability of Sufficient and Sufficiently Trained Staff • Cost • Scheduling	• Limited Availability of Sufficient and Sufficiently Trained Staff (wargames are a forcing function!) • Reification may lead to more predictable Red (playing "average" Red)	Moderate Stakes, but cannot be reviewed Do not Use AI
AI Findings Write Up	• Time pressures • Limited staff resources	• Wargame designers learn by writing up what happened • AI may not have sufficient context to understand important nuance	Use AI when tasks are moderate stakes, time intensive, and can be reviewed

Use AI that Mitigates Data Bias

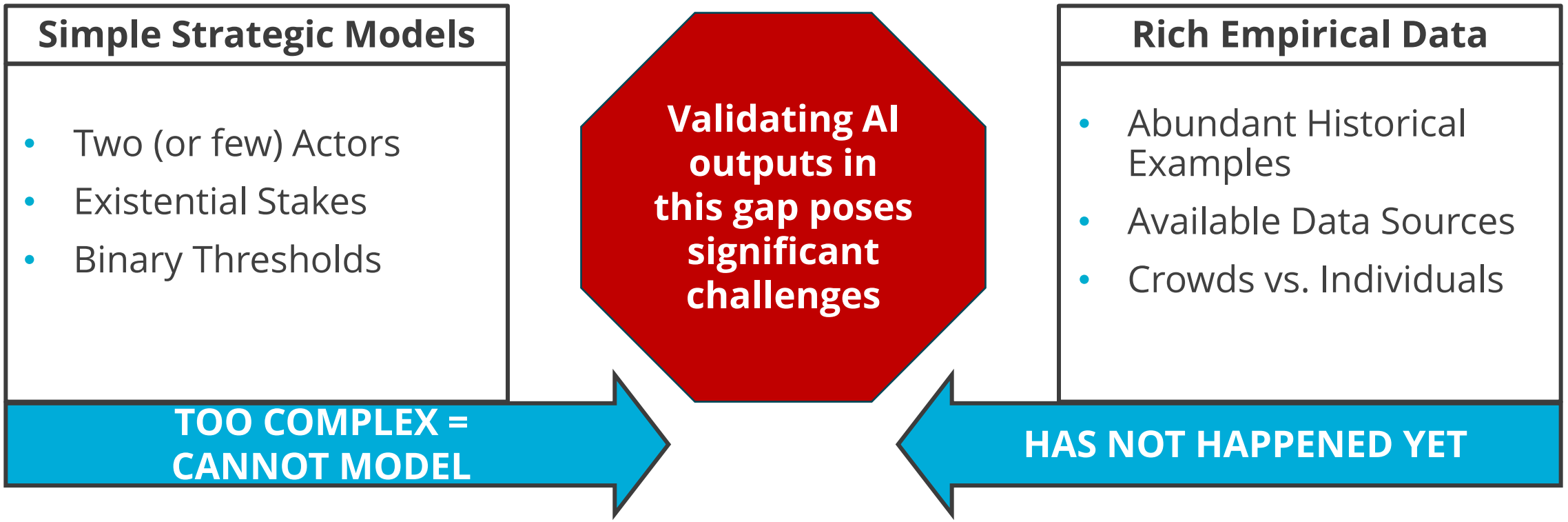
- Balance data sources to ensure representation across different time periods, geographic regions, and ideological frameworks.
- Use bias detection algorithms to identify and quantify biases in the training data and apply techniques such as re-weighting, debiasing, or adversarial training to reduce the impact of biased data on the model.
- Consult with SMEs to create test plans *before* models are completed to assess the validity and accuracy of outputs
- Invest in models with transparent and interpretable decision-making processes. This allows users to identify and address biases in the model's outputs.

However, this may still not be enough!

- Conflict theory often exists in the complexity/scarcity gap.
- How, then, do we assess the AI that emerges which is developed to align with our biases?

COMPLEXITY-SCARCITY GAP

Decision makers often need insights about areas of conflict that are too complex to easily model and for which we lack robust data.



WARGAMES COULD INFORM AN VALIDATE AI

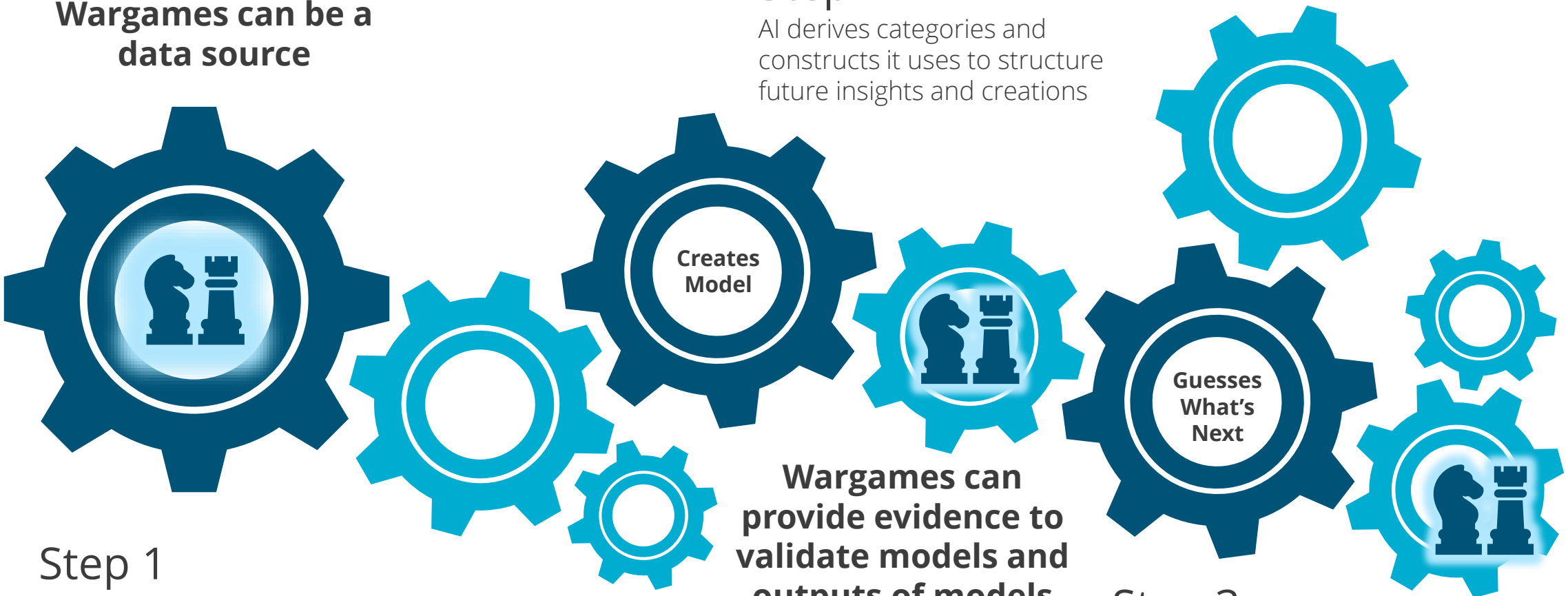
Wargames can be a data source

Step 1

All AI arise from existing data sets. They are trained and grow from vast data repositories for pre-existing information.

Step 2

AI derives categories and constructs it uses to structure future insights and creations



Wargames can provide evidence to validate models and outputs of models

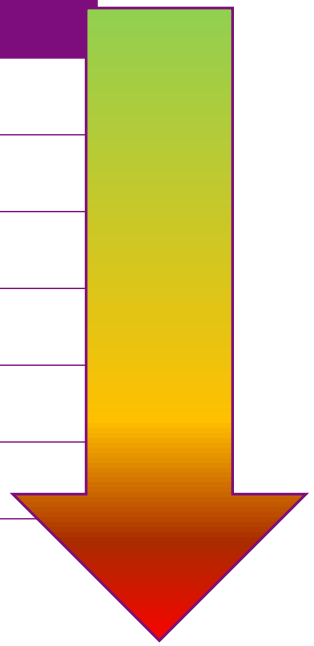
Step 3

Offers what it expects to come next...AI is always continuing an existing conversation

Questions?

Actionable Guidance from AI Cautionary Tales

AI Task
Summarize
Suggest
Critique
Plan
Create/Invent
Decide



As your tasks move “down” the workflow, exercise additional caution…

- LLMs can hallucinate – check your sources
- Do not rely on LLMs to create complete sets.
  It is good at making some suggestions; bad at exhaustive combinators.
- Beware of LLM fawning – your ideas are not always good, but it will tell you they are.
- AI can suggest ways to break down projects into tasks, but lacks the expertise to know what’s missing from those plans or how your context differs from the typical.
- AI can make a first draft, but your expertise is needed to introduce the novelty that makes the work original.
- AI should never be tasked with making a decision that requires legal or moral accountability. It cannot take responsibility for its actions.

Thank you.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.