Los Alamos NATIONAL LABORATORY

Navigating the Dissemination of Computational Models: Opportunities, Risks, & Responsibilities

Sara Del Valle, Ph.D.
Chief Scientist
Los Alamos National Laboratory

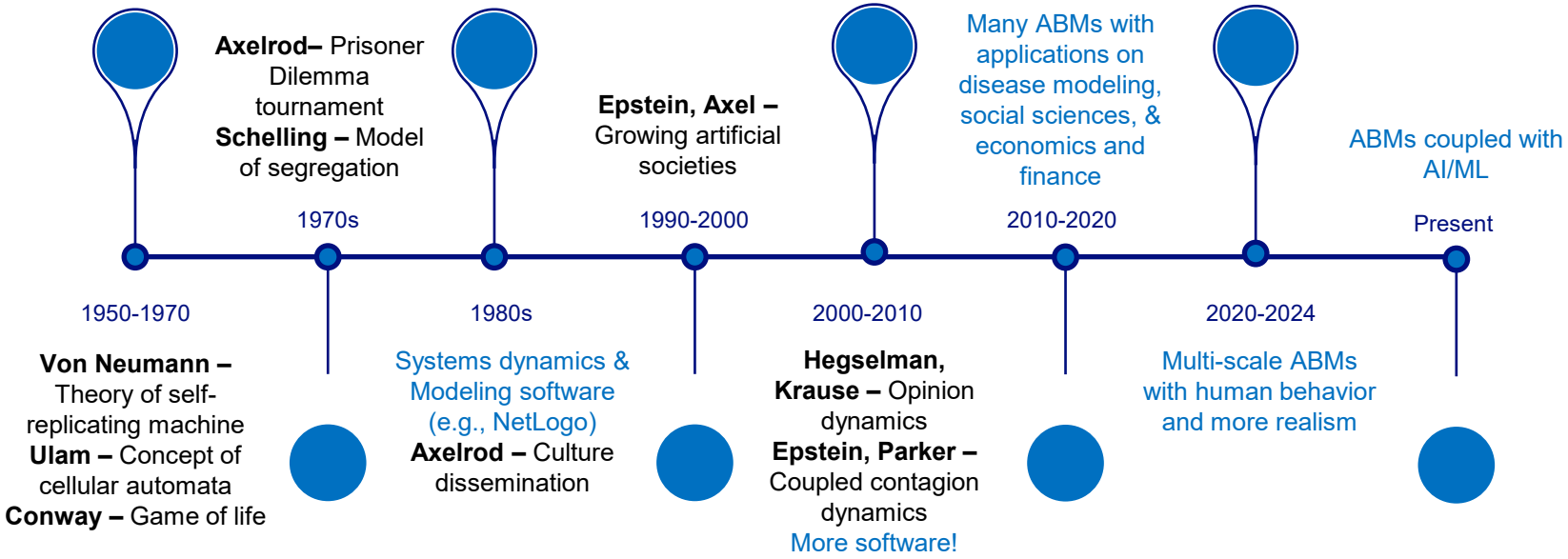
April 3, 2025

LA-UR-25-23154

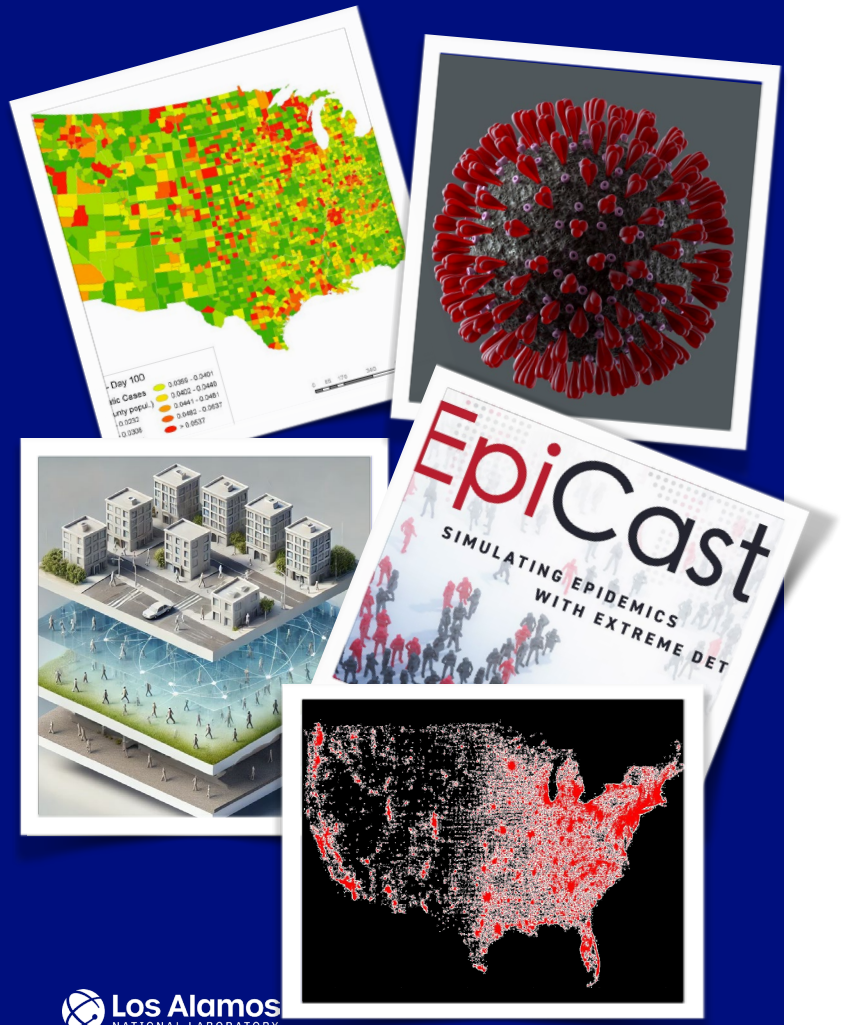
We are rapidly approaching a time
when the most powerful biological designs and threats
may not come from nature,

but from code

Case Study: Agent-Based Modeling (ABM)



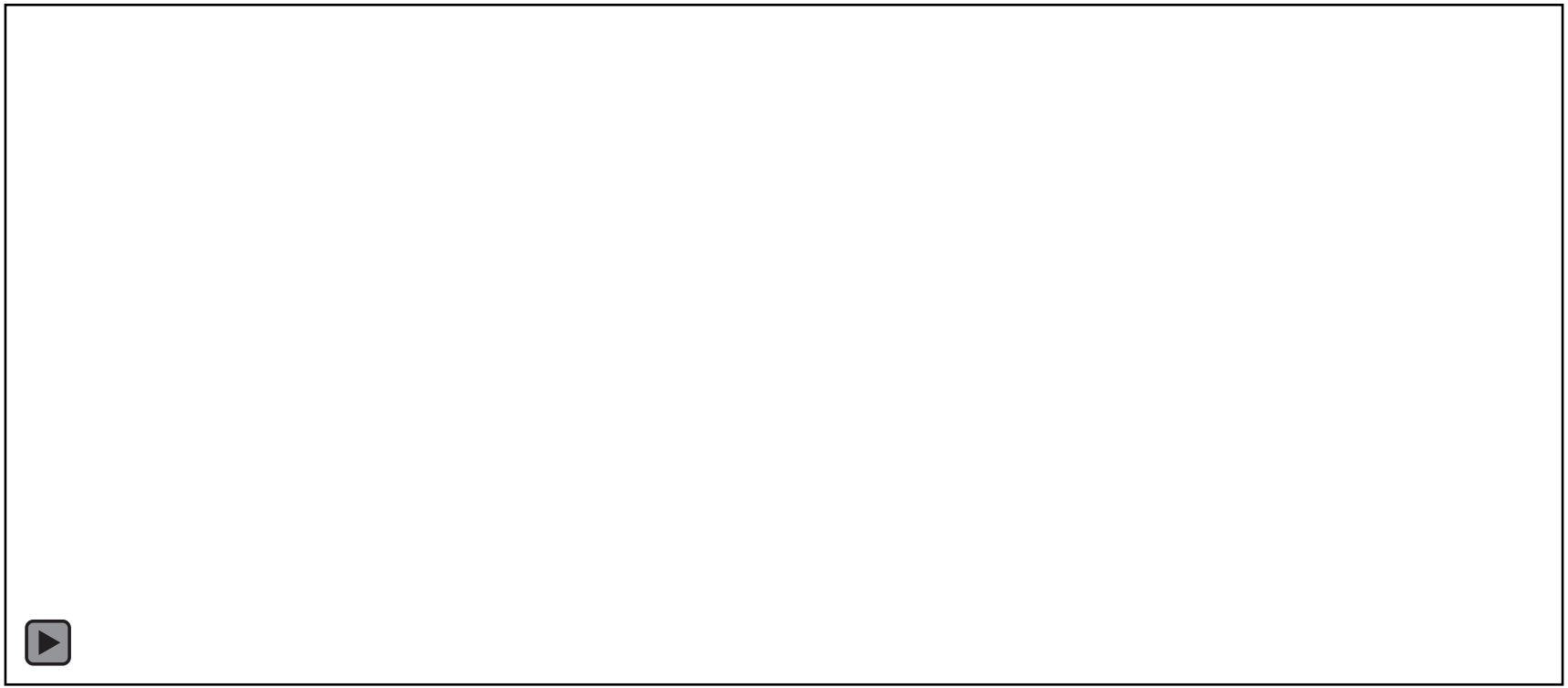
Retzlaff, C.O., Ziefle, M., Calero Valdez, A. The History of Agent-Based Modeling in the Social Sciences. In: Duffy, V.G. (eds) Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior. HCII 2021.



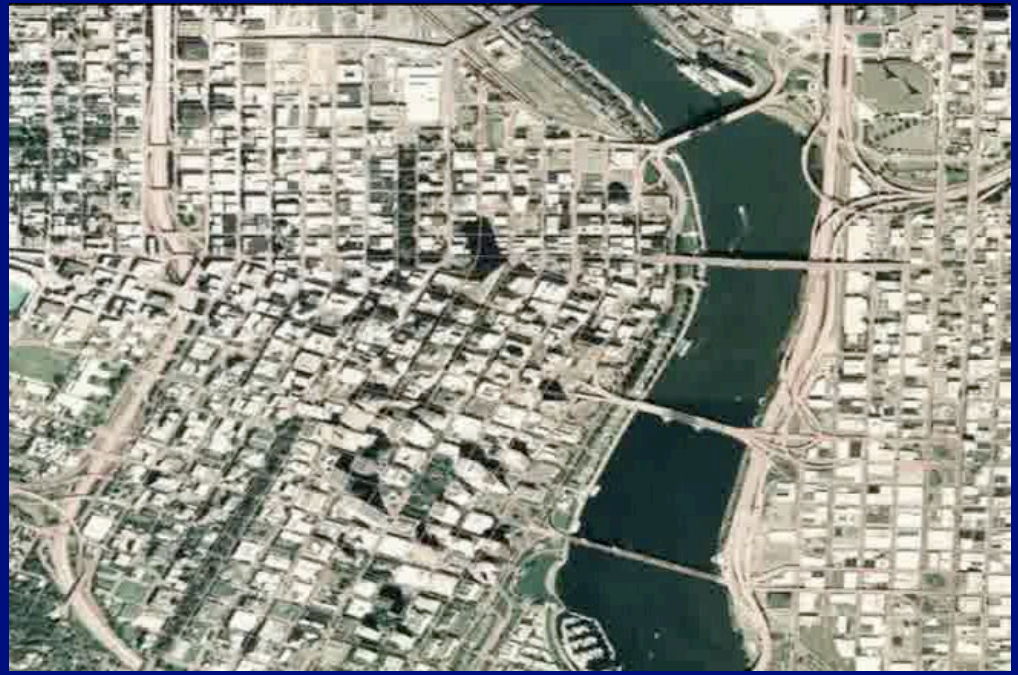
EpiCast ABM: Model Overview

- Simulates respiratory disease transmission across the entire US
  - **324 million agents**, demographically matched to 2019 U.S. Census data
  - **Fine spatial resolution**: explicit geography and mixing venues (e.g., households, workplaces, schools, neighborhoods)
  - **Rich demographics**: 25 agent-level attributes (e.g., age, household structure, race, ethnicity, school grade)
  - **Employment structure**: Occupation assigned using 3-digit NAICS codes
  - **Interventions**: Supports detailed pharmaceutical and non-pharmaceutical policies (e.g., vaccination, masking, closures)

EpiCast Simulation Results



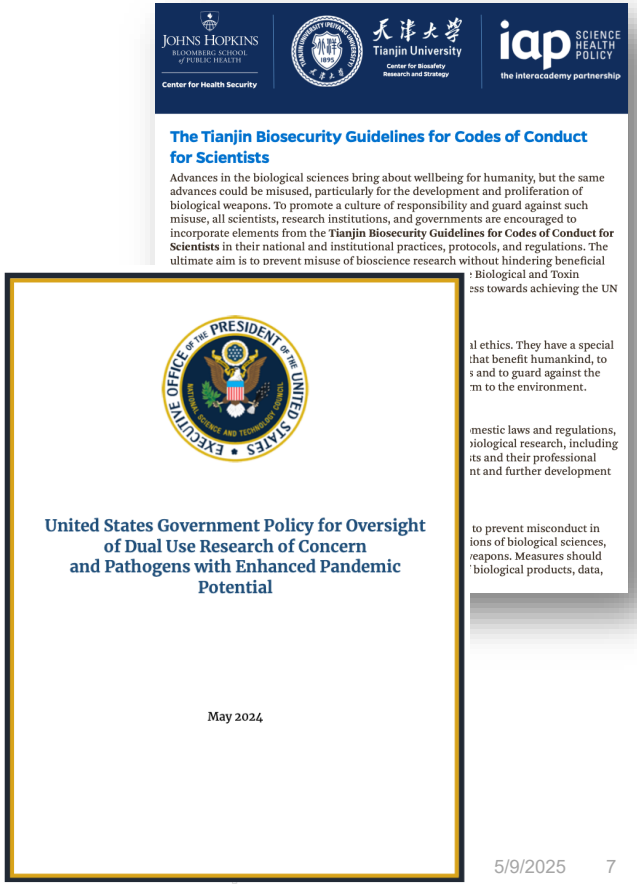
Implications of ABM Results



These models can now generate outputs with real-world biological implications (e.g., identify optimal transmission, characterize underlying distribution of outcomes)

Challenges in Current Biosecurity Guidelines and Journal Practices

- **Guidelines Scope**
  - Too narrowly focused on experimental data
  - Neglects generative models and in silico outputs
- **Risk Clarity**
  - Lack standard definitions to assess risks of in silico research
  - Unclear thresholds for evolving security risks
- **Reproducibility vs. Responsibility**
  - Journals requirements in conflict with security risks
  - Tension between transparency and dissemination practices
  - No consensus on tiered access, red teaming, or gating high-risk outputs



Capability-Based Thresholds

- Models, methods, or code that could assist in:
  - Enhancing pathogen virulence or transmissibility
  - Evasion of detection, diagnostics, or countermeasures
  - Design of novel, potentially harmful biological constructs
  - Overcoming containment or security controls
- Algorithms or simulations that can optimize or automate the manufacture, deployment, or targeting of biological, chemical, or nuclear threats

Does the research enable capabilities that could be repurposed for harm?

Exploitation Risk Thresholds

- Published models, datasets, or scenarios that lower the barrier to:
  - Identifying vulnerabilities in human, animal, plant, ecological systems, or infrastructure systems
  - Generating synthesis-read biological sequences with known or potential harms
  - Training or fine-tuning models for malicious purposes (e.g., drug repurposing for toxins, optimal transmission pathways)

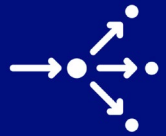
Could this work be easily exploited without expert knowledge?

Intent Signaling/Descriptive Misuse Enablers

- Information scenarios that:
  - Describe pathways for misuse, including attack chains or synthetic routes
  - Model societal impacts of specific attacks in ways that suggest optimal timing or targets
  - Frame “what if” threats in ways that could serve as playbooks rather than cautionary tales

Could this inform or inspire malicious actors?

Generative AI Dynamics & Global Security Impacts

I: Linear	II: Exponential	III: Iterative	IV: Discontinuous
			
More “shots on goal”; automated attempts to use technology for harm, analogous to the progression from muzzle loader to machine gun, e.g. spambots, malware, pathogen or toxin variants.	Exponential production or harmful growth dynamic, e.g. infectious agent, internet worm, scalable manufacturing via biosynthesis. Akin to fission processes, urban or wildland firestorms.	Sustained independent discovery, goal direction with refinement, replication with variation, and/or high-throughput assessment and improvement, especially with physical feedback.	Unforeseen technical advance obtained unilaterally via breakthrough unavailable to USG; e.g. “Sputnik moment,” more efficient new architectures, uncontrollable superintelligence.

Global Framework Regulating Nuclear Activities

The tools to simulate biology are outpacing the norms for sharing it, It's time we rethink how we assess risk in a computational area

QUESTIONS?

Nuclear vs Biosecurity

Feature	Nuclear	Biosecurity	AI	In Silico Modeling
Risk Clarity	Clear risks and well-defined threat scenarios	Lacks consensus on what constitutes 'risk'	Emerging; risks vary by use case (e.g., LLMs, autonomous systems)	Contextual (depends on biological domain)
Definitions and Scope	Codified in global treaties (e.g., NPT, IAEA)	Fragmented definitions (e.g., DURC, P3CO)	Evolving; no universally accepted terminology	Lacks common definitions for dual-use
Enforcement	Centralized via IAEA	No global enforcement body	No central enforcement; emerging national efforts	None – handled by journals and funders
Global Treaty Body	Yes – IAEA + NPT	No global treaty body	None yet; proposals under discussion (e.g., AI governance summits)	None yet
Dual-Use Control	Export controls (e.g., NSG)	Limited, inconsistent oversight	Mostly voluntary or market-driven; no dual-use framework yet	Ad hoc (especially around model/code sharing)
Verification & Inspection	IAEA inspections and safeguards	Rare, non-standardized	Not applicable or emerging (model audits, evaluations)	None
Response to Violations	UN/IAEA escalation mechanisms	Domestic responses or ad hoc norms	Currently undefined; varies by jurisdiction	Undefined; varies by publication norms