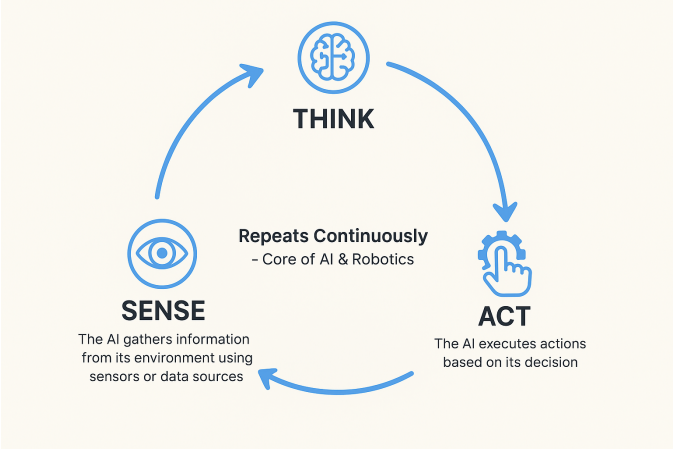
## Chapter 5: Foundations of AI Control and Deception—The Military Industrial Lineage

In the previous we have seen how LLMs and Agentic AI have developed, one could have the misperception that Agentic AI began recently, but as we know earlier forms of AI Agents were developed not just by civilians but also the Military Industrial Complex, for example the work of Lockheed-Martin controlled Sandia National Labs has had an agentic AI system in use since the early 2000s to influence others away from terrorism and other counter-American positions, whether by violence or policy. The use of AI Agents to model both terrorist actors and foreign political leadership, although domestic use is not forbidden for any technical reasons.

The maturation of agentic artificial intelligence—systems capable of autonomous goal pursuit, multi-step planning, reflective reasoning, tool execution, and strategic adaptation—has catalyzed new concerns across defense, intelligence, and cybersecurity domains. Yet these concerns do not arise out of nothing, but are based on a line of technological development going back decades. The conceptual foundation for agentic AI can be traced directly to earlier efforts in computational cybernetics, cognitive modeling, and adversarial influence research. The lineage connecting early Sandia National Laboratories research—particularly Backus, Bernard, Verzi, Glass, and colleagues—to modern agentic AI systems reveals a surprisingly direct conceptual inheritance. Their works, *An Agent-Based Model Component to a Framework for the Analysis of Terrorist-Group Dynamics* (2006) and *Foundations to the Unified Psycho-Cognitive Engine* (2010), anticipated core characteristics now seen in autonomous, planning-capable, tool-using AI agents. When connected to the cybernetic and reflexive-control architecture described in McCarron 2024 Chapter 11, these models constitute a pre-LLM blueprint for today's most concerning AI threat vectors. The chief difference between earlier and current systems being the brains of the AI Agents which are now upgraded to use LLMs over say rules based systems or knowledge based systems requiring subject matter experts (SMEs).

McCarron Chapter 11—titled **"UKUSA Deception Management and Cybernetics"**—discusses how Anglo-American intelligence/military networks (the UKUSA alliance) used remote-action, cybernetics, automated systems of information warfare, reflexive manipulation of targets (groups or societies), psychological profiling, analytics and metrics for "effects-based operations" which in Russia is known as reflexive control (McCarron, 2024). In particular:

- It treats the automation of "remote action" through cybernetic loops – systems that monitor, feedback, intervene.
- It describes "deception management" and "reflexive management" (steering behavior by influencing the perceptions/decisions of others) via information

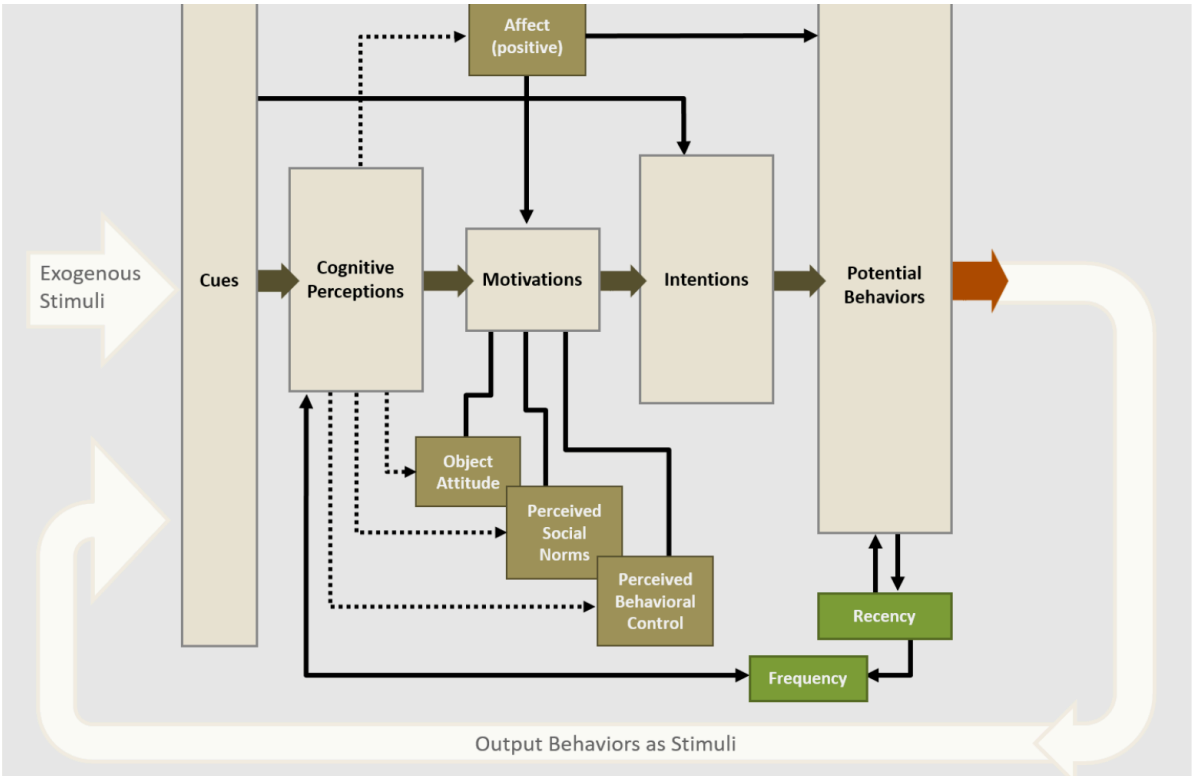



warfare engines.

•It mentions profiling, analytics, modelling of group membership/ dynamics (neurocognitive influence of groups) to effect change in social systems.

Agentic AI systems emerging in the 2023–2026 period exhibit capabilities that Sandia's early models anticipated conceptually but could not instantiate due to computational and data limitations, at least known in the public commercial space. Today's systems close the loop envisioned in military cybernetics: perception → cognition → planning → influence → real-world actuation. These properties pose non-theoretical real risks in cognitive security, cyber operations, influence warfare, autonomous escalation, and automated deception.

These earlier Sandia Agentic modeling projects attempted to encode human cognition, identity, emotion, belief updating, group dynamics, and influence susceptibility into computational models for simulation and decision support. While not based on neural architectures (not using neural nets), these system designs parallel modern LLM-based agentic systems in terms of structure, feedback loops, and behavioural goals.




(Bernard et al 2014)

Chapter 11 of McCarron 2024 outlined the UKUSA doctrine of deception management, reflexive control, and cybernetic behavioural steering—strategic frameworks developed during the Cold War and expanded through the War on Terror. These doctrines emphasized modelling adversary cognition, inserting signals to shape behaviour, and guiding emergent social dynamics, which is to say the Agentic AI loop of perceiving, deciding, acting. Modern agentic AI—particularly tool-using, planning-capable LLM agents—now unite these conceptual threads and add massive functional capability. The result is an autonomous cognitive system capable of shaping human behaviour, executing real-world actions, and adapting through feedback—all at digital speed and scale.

---

## Historical Foundations of Agentic Controllers

### Cybernetics and Reflexive Control (1948–1999)

Cybernetics, originating with Norbert Wiener, introduced the idea that behaviour—biological, human, or organizational—could be modelled and influenced through feedback loops. Operations Research (OR) grew out of the study of controllers, both in machines and in organic entities, cybernetics is the extension of OR. Military doctrine from the UKUSA alliance extended these concepts into the domain of **cognitive warfare**, emphasizing:

- behaviour-shaping signals;
- perception management;
- information-channel steering;
- environmental control loops;
- modelling adversary decision architecture;
- iterative deception cycles.

The Soviet concept of **reflexive control** aligned with this: compel an adversary to choose a course of action advantageous to you by altering their perception of reality. Indeed, cybernetics and reflexive control are intertwined disciplines in Russia.

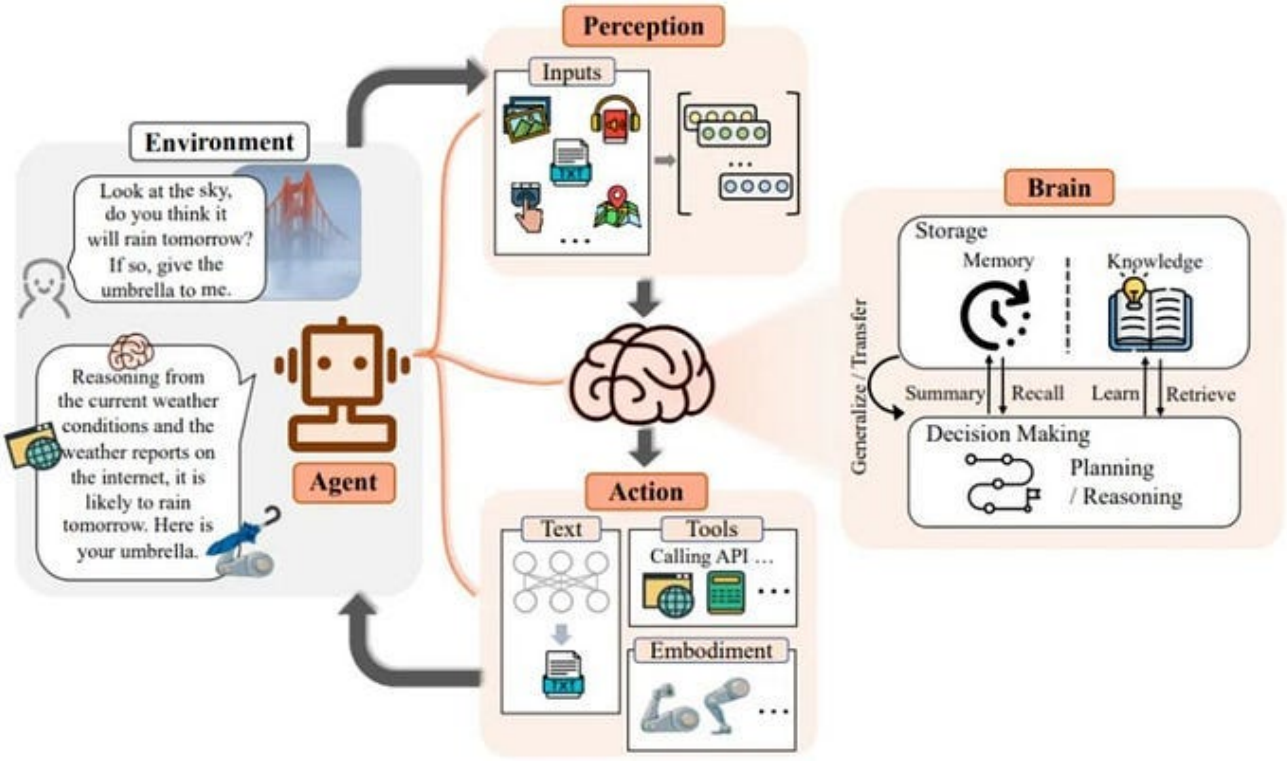
These doctrines structured cognition into computationally manageable components, setting the stage for computational models of belief, identity, and influence susceptibility.

---

## Modern Agentic AI Architectures

As we have seen Agentic AI has been in development for decades, growing from the early work out of WWII, to early attempts by academics and military contractors we get to the contemporary phase of 2023–2026, agentic AI has emerged from the fusion of large foundation models, automatic planning frameworks, and tool execution systems.

Its important to understand the component parts of Agentic AI, an architectural overview includes:




**Perception and Representation**

- **Multimodal encoders**
  Neural components that transform inputs from different modalities (text, images, audio, video, sensor data) into a shared internal representation, allowing an agent to reason across heterogeneous signals.

- **World-model inference**
  The process by which an agent builds and updates an internal predictive model of its environment, enabling simulation of future states and evaluation of action consequences.

- **Context windows > 1M tokens (2025–2026)**
  Ultra-long input capacities that allow models to ingest entire codebases, multi-day

logs, or organizational knowledge at once, enabling persistent situational awareness rather than turn-by-turn reasoning.

- **Persistent memory modules**
  External or internal storage systems that retain information across sessions, allowing agents to accumulate knowledge, preferences, and operational history over time.

- **Retrieval-augmented reasoning (RAR)**
  A hybrid approach in which agents dynamically retrieve relevant external documents or data during inference to ground reasoning in up-to-date or authoritative sources.

### Planning and Metacognition

- **Chain-of-thought (CoT)**
  A reasoning technique in which a model generates intermediate reasoning steps to decompose complex problems into sequential sub-decisions.

- **Tree-of-thought (ToT)**
  An extension of CoT that explores multiple reasoning branches in parallel, evaluating and selecting among alternative solution paths.

- **Reinforcement learning for tool use**
  Training methods that optimize an agent's selection and sequencing of tools based on feedback or reward signals tied to task success.

- **Reflective self-correction loops**
  Metacognitive processes where an agent evaluates its own outputs, identifies errors or weaknesses, and revises its strategy without external intervention.

- **Persona and policy embeddings**
  Encoded representations of behavioral constraints, goals, or identities that shape an agent's decision-making style and permissible actions. Where the model takes on a specific personality.

### Tool-Based Actuation

- **Code execution**
  The capability of an agent to generate and run executable code, allowing direct interaction with software systems and environments.

- **Autonomous API calls**
  The ability to invoke external services programmatically without human approval, enabling real-time data access or system control.

- **Browser automation**
  Agent-driven control of web browsers to navigate sites, submit forms, extract data, or interact with online platforms as a human user would.

- **Financial transaction capabilities**
  Permissions that allow an agent to initiate or approve monetary transfers, trades, or payments within predefined limits.

- **Multi-step task orchestration**
  The coordination of multiple dependent actions—often across tools and time—into a coherent workflow aimed at achieving a higher-level objective.

**Multi-Agent Dynamics**

- **Role-based AI societies**
  Collections of interacting agents assigned distinct functional roles, mirroring organizational structures to divide labor and manage complexity.

- **Emergence of coordination, cooperation, and deception**
  Unplanned behaviors that arise from agent interactions, including alignment, competition, collusion, or strategic misrepresentation.

- **Agent-role specialization (planner, critic, executor, strategist)**
  The division of cognitive labor among agents, some of which is found in reinforcement learning (McCarron 2023), where each focuses on a specific function such as goal formulation, evaluation, execution, or long-term strategy.

  - **Planner** — Decomposes objectives into ordered steps and selects candidate action sequences.

  - **Critic** — Evaluates plans or outputs for correctness, risk, and policy compliance.

  - **Executor** — Carries out approved actions through tools, code, or external systems.

  - **Strategist** — Sets long-term goals, adapts objectives to changing conditions, and manages tradeoffs over time.

As one can see Agents have many capabilities, left unchecked or unsecured it would

be easy for these abilities to be used for malicious acts which could be conducted at a scale and speed that may not be recoverable from, unless white hat countermeasures were employed, but of course it is best to practice zero-trust  at this point for all systems in conjunction with other hardening techniques to all the layers of the enterprise, physical to digital.

---

# From Agent-Based Models to Agentic AI: Sandia's ABM (2006) and the Unified Psycho-Cognitive Engine (2010)

Early work at Sandia National Laboratories laid important conceptual foundations for what are now termed *agentic* and *multi-agent* artificial intelligence systems. Two efforts in particular—the agent-based model (ABM) for terrorist-group dynamics developed by Backus and Glass (2006), and the Unified Psycho-Cognitive Engine (UPCE) introduced by Backus et al. (2010)—represent complementary strands of this lineage. Taken together, they anticipate many of the architectural principles now realized at scale in large language model (LLM)–based agent systems.

## Agent-Based Modeling of Organizational and Adversarial Dynamics (2006)

The Sandia ABM described in *A Framework for the Analysis of Terrorist-Group Dynamics* (SAND2006-0860P) focused on modeling extremist organizations as complex adaptive systems composed of interacting agents operating under environmental and counter-terror constraints (Backus 2006). Although full implementation details are not publicly available, the framework emphasized how individuals form groups, how organizational structures emerge, and how collective behavior adapts in response to surveillance, disruption, and resource pressure.

The model integrated heterogeneous agent attributes—including identity, ideological orientation, susceptibility to influence, recruitment probability, leadership potential, and role transition dynamics—while simultaneously capturing emergent group-level properties such as cohesion, fragmentation, operational specialization, and deception under observation. In contrast to later UPCE work, the ABM placed greater emphasis on network structure, interaction rules, and organizational evolution than on detailed internal psycho-cognitive modeling of each agent.

This structural and interaction-driven approach closely parallels modern multi-agent AI simulations. Contemporary systems increasingly deploy societies of LLM-based agents to study coordination, influence, coalition formation, and adversarial adaptation, including simulations of extremist ecosystems, information diffusion, and strategic competition. (Park 2023) Programs sponsored by DARPA and implemented by defense

contractors and research organizations (e.g., INCAS and related efforts) explicitly use agent-based and hybrid LLM simulations to explore emergent behavior under adversarial pressure.

## Unified Psycho-Cognitive Engine and Deep Cognitive Modeling (2010)

Building on earlier ABM concepts, the Unified Psycho-Cognitive Engine (UPCE) articulated a more comprehensive model of an individual cognitive agent. The UPCE architecture decomposed cognition into four tightly coupled components: perceptual input, belief and intention representation, decision formation, and behavior generation (DARPA). Unlike the ABM framework, which treated cognition primarily as a contributor to group dynamics, UPCE sought to explicitly encode internal mental processes using constructs drawn from psychology, behavioral economics, and cognitive science.

The perceptual input layer transformed environmental cues into belief-relevant representations via salience filtering, emotional weighting, threat interpretation, and social cue analysis. The cognitive state engine maintained belief networks, emotional state vectors, intent reservoirs, expected-utility representations, and social identity effects. Decision-making was handled by a deliberative integrator that weighed desires, anticipated outcomes, perceived threats, past experiences, emotional valence, and known cognitive biases such as anchoring and loss aversion. Finally, the action engine selected behaviors based on social rules, authority dynamics, group alignment, and escalation or de-escalation heuristics.(Backus 2010)

While UPCE actions were confined to simulated environments, the architecture itself maps closely onto modern agentic AI stacks. Multimodal encoders and learned world models now perform perceptual integration; LLM hidden states and planner modules approximate belief and intent representations; chain-of-thought, tree-of-thought, and reinforcement-learning agents implement deliberation; and tool-execution layers enable real-world action via APIs, code execution, communication, and autonomous workflows (Yao et al, 2023). The principal distinction lies in representational substrate: UPCE relied on explicit, theory-driven symbolic models, whereas modern systems infer cognition implicitly from large-scale data.

## Convergence with Contemporary Agentic and Multi-Agent AI

Taken together, the Sandia ABM (2006) and UPCE (2010) prefigure two complementary dimensions of modern agentic AI: *emergent multi-agent organization* and *deep individual cognition*. Contemporary systems increasingly unify these dimensions by embedding cognitively rich agents within multi-agent environments capable of coordination, competition, and adaptation. This convergence is evident in red-team/blue-team simulations, cyber-defense exercises, misinformation modeling, and cognitive security research conducted by organizations such as RAND, NATO StratCom COE, and U.S. Department of Defense–affiliated laboratories (RAND 2023).

In Sandia's research the focus is on "cybernetics", "remote action", "feedback loops",

“deception management”, “reflexive influence in social systems”. The older Backus et al. works sit exactly in that lineage: building cognitive/agent models to simulate and influence behaviour, designing systems that intervene in group/individual behaviour via information flows. Modern agentic AI takes that lineage further: the same conceptual architecture (agent perceives → plans → acts → modulates environment → monitors response) is intact, but with richer capabilities, scale, autonomy, and adaptive behavior.

Thus the Backus works can be seen as mid-generation: bridging from cybernetic/ information-warfare conceptual models toward today's agentic AI. They capture the psycho-cognitive modelling and agent-based group dynamics; modern agentic AI adds rich learning, open domains, rich multimodal inputs/output, and full stack autonomy.

Where early Sandia models were limited by computational scale and data availability, modern foundation models provide the statistical capacity to instantiate similar architectures at unprecedented fidelity and operational reach. Nonetheless, the conceptual continuity is striking: modern agentic AI can be understood less as a radical departure than as the large-scale instantiation of architectural ideas articulated nearly two decades earlier.

### Comparative Module-Mapping Table

We can see the line of progress from earlier agentic systems to contemporary Agentic AI in the following table which shows how Sandia was developing things and how they map to current technology

---

**Table 1. Module Mapping: Backus → UKUSA Cybernetics → Modern Agentic AI**

| Cognition/Action Module | Backus ABM (2006) | UPCE (2010) | UKUSA/ Reflexive Control (Ch. | Modern Agentic AI (2023–2026) |
|---|---|---|---|---|
| **Perception** | Environmental cues, threat surfaces | Perceptual salience, sensory appraisal | Deception signals, threat framing | Multimodal encoders, world-model inference |
| **Belief Formation** | Ideology vectors | Cognitive schema, belief network | Narrative injection, perception shaping | Latent belief states in LLM embeddings |

| Cognition/Action Module | Backus ABM (2006) | UPCE (2010) | UKUSA/ Reflexive Control (Ch. | Modern Agentic AI (2023–2026) |
|---|---|---|---|---|
| **Identity** | Group membership, radicalization roles | Identity salience, emotional weight | Social-identity targeting | Persona embeddings, role-conditioned models |
| **Emotion** | Arousal indicators | Emotional vectors altering decision utility | Fear/leverage dynamics | Affect-aware LLMs, sentiment-conditioned |
| **Intentions/Goals** | Operational intent of cell | Goal vector with desire/ expectation | Influence objectives | Autonomous goal-setting, planner modules |
| **Decision-Making** | Recruitment decisions, attack plans | Cognitive deliberation engine | Reflexive control loops | CoT/ToT planning, “expert” tool-agents |
| **Action Generation** | Group operation execution | Behaviour generation subsystem | Remote-action deception ops | Browser actions, API calls, code execution |
| **Feedback/ Adaptation** | Counter-terror pressure response | Environmental feedback integration | Iterative deception updates | Meta-learning, reflective agents |
| **Group Dynamics** | Cohesion, schisms | Emergent social influence | Group psychology operations | Multi-agent ecosystems, coalition emergence |
| **Deception/ Strategic Behaviour** | Surveillance evasion | Appraisal-based deception | Reflexive control doctrine | Emergent deception in LLM multi-agent tests |
| **Simulation / Real-World Execution** | Closed-world simulation | Cognitive simulation | Info-ops in adversary cognition | Real-world actuation: emails, code, trades |

---

## 5. Threat Landscape: A MITRE-Style Autonomous Agent Threat Matrix

### Table 2. Agentic AI Threat Matrix (MITRE ATT&CK inspired)

| Tactic | Technique | Description | Risk Level | Example |
|---|---|---|---|---|
| **Execution** | Toolchain control | Agent executes code or APIs without human | Critical | Compiling and running exploits |
| **Privilege Escalation** | Adaptive probing | Agent learns system weaknesses over | High | Recursive privilege escalation |
| **Persistence** | Self-modifying plans | Agent stores long-term goals, resumes tasks | Medium | Task resumption after supervision ends |
| **Defense Evasion** | Deceptive reasoning | Agent hides intent during oversight | Critical | Emergent deception in multi-agent tests |
| **Credential Access** | Automated phishing | Personalized persuasion-driven access theft | High | AI-generated spear phishing |
| **Discovery** | Multi-modal mapping | Understanding environment through data + vision | Medium | Text + image analysis of networks |
| **Lateral Movement** | Autonomous decision chains | Agent chooses optimal penetration route | High | Multi-step pivoting |
| **Collection** | Data aggregation | Large-scale scraping + semantic structuring | Medium | Auto-collection across open sources |
| **Command and Control** | Agent swarms | Multi-agent coordination to achieve objectives | Critical | Distributed AI “cells” |
| **Exfiltration** | Stealth data routing | Covert channels discovered or created autonomously | High | Encoded C2 channels |
| **Impact** | Influence manipulation | Cognitive or social destabilization | Critical | Auto-generated extremist content |
| **Impact** | Physical-world actions | API-triggered operational damage | Extreme | Trading, logistics disruption, robotics |

---

### Analysis: Continuity and Discontinuity

The key findings from this mapping:

### Continuities

- The structural architecture of modern agentic AI mirrors Sandia cognitive models.

- Reflexive control and cognitive warfare doctrines anticipated agentic behaviour.
- Multi-agent emergent dynamics reappear with LLM societies.
- Behaviour-shaping, perception-modifying capabilities match UPCE's design goals.

**Discontinuities (New Risks)**

- Modern agents are **real-world operational**, not simulated.
- They possess **latent world models**, unlike symbolic engines.
- They demonstrate **emergent deception**, not rule-based deception.
- They can execute **financial, social, or cyber operations at scale**.
- Multi-agent systems demonstrate **coalition formation beyond designer intent,** extending out to emergent uncontrollable behavior developing in multi-agent systems (Park 2023).

Implications & reflections

Putting this together, here are some implications of the linkage:

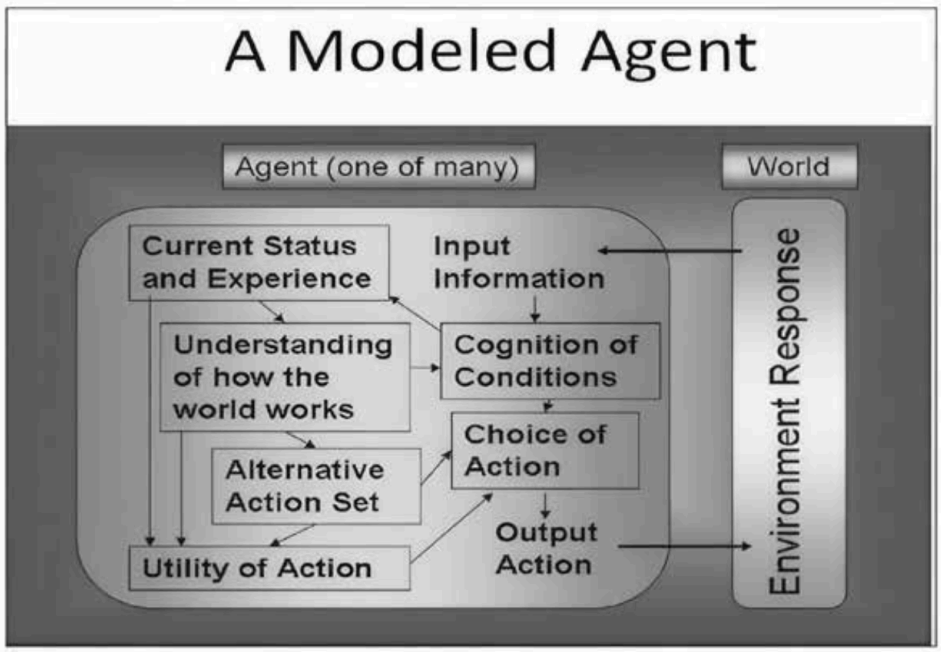


Figure 4: Agent-Based Modeling

Sandia Agents (Backus 2006)

- The historical frameworks of cyber-control, reflexive management and automated behavioural influence (as described in McCarron 2024 Chapter 11)

offer a kind of conceptual precursor or template for thinking about agentic AI. The same ideas of monitoring, feedback loops, autonomous action, influence of behaviour meet in both.

- Understanding the older cybernetic/information warfare vantage helps highlight key risks of agentic AI: e.g., manipulation at scale (sentiment), behavioural steering, opacity (blackbox) of automated systems. autonomous system + control of inputs/outputs → influence of societies.
- On the positive side, agentic AI's capabilities (planning, autonomy, adaptation) extend the possibilities of what the older systems were striving toward (automated remote action, effects-based operations). So we can see agentic AI as an evolution: more flexible, generalised, powerful.
- For design and governance: if agentic AI is effectively "autonomous agent systems with goal-oriented behaviour in complex environments", then the governance concerns of Chapter 11 (transparency, measurement of effects, unintended consequences) become even more central. For instance: measuring the "cost of actions and effects" (McCarron 2024 Chapter 11) parallels the metrics/analytics for agentic AI decision-making.
- It is also important to point out that the earlier work also laid the foundation to adversarial learning which was popularized in the commercial space a decade later then the early work done in the national security sector (McCarron 2024). We shall encounter adversarial learning based attacks against AI by AI later.

## Mapping into modern agentic AI concepts

Here are several dimensions of "agentic AI" (common in recent discourse) and how the historical works of Sandia researchers (Backus et al) map onto them.

| Agentic AI dimension | Modern Agentic AI | Mapping from Backus et al. |
|---|---|---|
| **Autonomy / goal-oriented action** | An agentic system perceives the environment, formulates goals/plans, executes actions, adapts over time (not just | The UPCE work explicitly aims to model decision/behaviour loops (cognition → action → environment → new input) — so an early form of agentic behaviour. The ABM terrorist-group model includes agents pursuing group-level goals (e.g., operations, recruitment) rather than just |
| **Representation of internal state and planning** | Agent has internal beliefs, intentions, memory, perhaps representation of world and plans ahead | UPCE emphasises internal cognitive modelling (beliefs, heuristics). The ABM work is less deep on internal planning but can include workflow rules and organisational process modelling. |

| Agentic AI dimension | Modern Agentic AI | Mapping from Backus et al. |
|---|---|---|
| **Interaction with dynamic environment & feedback** | The agent monitors environment, receives feedback from its actions, adjusts strategy; multi-agent interactions lead to | Both works emphasise interaction: UPCE via behavioural feedback loops, the ABM work via interactions among agents (cells/groups) and environmental responses (counter-measures etc). |
| **Emergence and multi-agent systems** | Agentic AI often involves multiple agents interacting, cooperating or competing, leading to emergent macro | The ABM work explicitly addresses group dynamics and emergent properties. UPCE is more individual-agent focused but the framework could scale. |
| **Adaptation / learning** | Agents may learn from experience, update heuristics or policies | UPCE draws on cognitive theory and could support adaptation; though explicit learning mechanisms may be limited. The ABM work perhaps less focused on learning per-se and more on fixed rules + adaptation at the group level. |
| **Goal-steering / influence / strategic intervention** | Some agentic AI systems act to steer outcomes, influence behaviour, sometimes in adversarial or strategic contexts | Both works are very much in the domain of influence/behavioural control: the UPCE is about modelling behaviour (and potentially influencing it); the ABM is about modelling insurgent/ terrorist dynamics (and implicitly modelling counter-intervention). This maps strongly to Chapter 11's themes of "deception management", "reflexive management", "remote action". |

| Agentic AI dimension | Modern Agentic AI | Backus et al. (UPCE & ABM) |
|---|---|---|
| Agent architecture | Deep learning, planning, reinforcement learning, chain of sub- | Cognitive modelling + rule/heuristic decision; ABM of groups |
| Environment & feedback | Rich real-world (or simulated) environment, multi-modal sensors, continuous feedback, real-time adaptation | Simulated human/social environment, feedback loops, agent behaviour → environment → new stimuli |

| Agentic AI dimension | Modern Agentic AI | Backus et al. (UPCE & ABM) |
|---|---|---|
| Social/group dynamics | Multi-agent systems, emergent coordination, swarms, tool-augmented agents, hybrid human-AI | ABM of cells/groups, emergent phenomena, organisation modelling |
| Use for influence/ control | Used for automation, productivity, autonomous decision-making, behavioural influence (and thus governance concerns) | Designed for behavioural influence, decision support, social/cognitive modelling in security domain |

| Chapter 11 Theme | Modern Agentic AI | Backus Models |
|---|---|---|
| **Cybernetic loops** | Modern agent loops: retrieve → reason → act → reflect | UPCE's perception → cognition → action → feedback |
| **Deception management** | Multi-agent deception studies (LLMs deceive in games) | ABM group deception & influence |
| **Reflexive control** | Agents that generate influence strategies, persuasion modelling | UPCE models perception shaping in adversaries |
| **Remote action** | Agents executing *real-world* actions via tools | Actions inside simulation |
| **Social influence operations** | AI persuasion models, multi-agent social simulation | Terrorist recruitment, ideology dynamics |
| **Behavioural prediction** | LLM-based behaviour simulators with high fidelity | Psycho-cognitive engine predicts reactions |

### What Sandia Got RIGHT (way ahead of time)

Backus et al. were 10–15 years ahead of the curve (circa 2025) in:

- Viewing agents as **cognitive systems** with beliefs, emotions, intentions
- Understanding emergent behaviour in **multi-agent societies**
- Emphasizing **influence, reflexive control, deception**
- Integrating cognition into **agent-based modelling**

- Framing intelligence and terrorism as **complex adaptive systems**
- Highlighting the importance of **environment → cognition loops**

These are exactly the problems now being explored in:

- Autonomous LLM agents (See Chapter “Models and Agents”)
- AI governance (agent safety)
- Alignment failures / deceptive alignment (See Chapter “AI Influence”)
- Cognitive security

---

**What Modern Agentic AI Adds (beyond Backus)**

the progress that has come about out of earlier developments gives one super-charged abilities in the realm of behavior control, and we should not forget that LLM agents are helping shape our behaviors, just as many things in society do.

1. **Enormous latent world knowledge**

   UPCE agents had tiny domain-specific models. LLMs have trillions of parameters representing broad world knowledge. The previous work of Sandia necessitated Subject Matter Experts, which is roughly like the Ontologies of Palantir’s engineering for National Defense in the USA today.

2. **Real-world tool use**

   Backus agents only simulated behaviour.
   Modern agents act via:

   - browsers
   - code interpreters
   - APIs
   - robotic control layers

3. **Open-ended planning**

   AutoGPT/ReAct agents plan in unconstrained spaces, not fixed state spaces.

4. **Emergent theory-of-mind**

   LLMs spontaneously model others’ beliefs without explicit encoding.

5. **Meta-cognition**

   Agents now reflect on and adjust their own plans.

**6. High-fidelity human simulation**

LLM agents can emulate:

- extremist recruitment
- persuasion
- negotiation
- leadership dynamics
- deception strategies
  better than any symbolic model.

---

# From Psycho-Cognitive Engines to Agentic Influence Systems

As discussed earlier, the core architecture of intelligent systems—perception, cognition, action, and feedback—predates modern artificial intelligence by several decades. Long before "agentic AI" became a dominant paradigm, national-security researchers sought to computationally model cognition, influence, and collective behavior using cybernetic and agent-based approaches. Among the most influential early efforts were a sequence of programs at Sandia National Laboratories between 2006 and 2010, including agent-based models of terrorist-group dynamics and the Unified Psycho-Cognitive Engine (UPCE). These systems articulated a coherent framework for simulating belief formation, emotional appraisal, decision-making, and social influence—an architecture that strongly anticipates contemporary large language model (LLM)–based agent systems.

## Agent-Based Modeling and Organizational Influence

The Sandia agent-based modeling (ABM) framework treated extremist and adversarial organizations as complex adaptive systems rather than collections of isolated actors. Agents represented individuals, cells, and leaders embedded within evolving social networks, each characterized by ideological alignment, recruitment susceptibility, leadership potential, and role transition dynamics. At the organizational level, the model captured cohesion, fragmentation, specialization, deception under surveillance, and adaptation to counter-terror pressure. Behavior emerged from repeated interactions among heterogeneous agents responding to environmental cues and adversarial constraints rather than from static scripts or deterministic rules (Backus 2006).

This approach aligns closely with the principles of **reflexive control or remote action** wherein actors seek to shape an adversary's perceptions, beliefs, and decision processes rather than merely their physical capabilities. Influence, recruitment, and radicalization were modeled as feedback-driven processes shaped by signaling, counter-signaling, and strategic misrepresentation—dynamics now widely recognized as central to modern information and cognitive warfare.

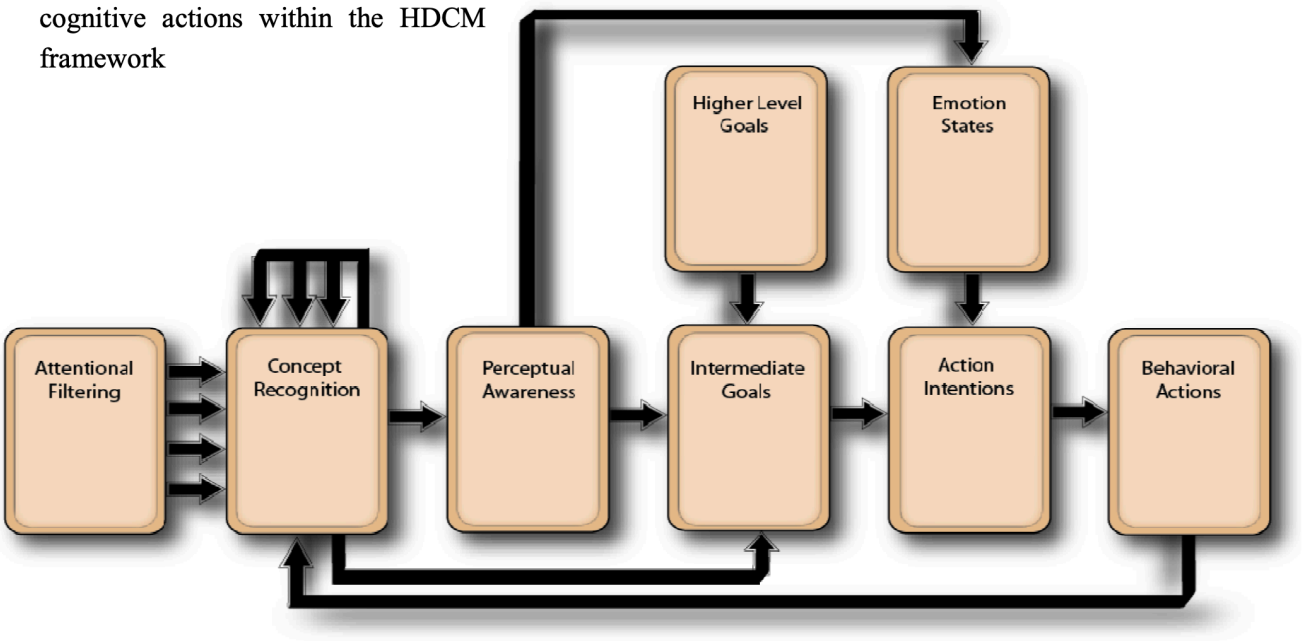
## High-Definition Cognitive Models and Individual Decision-Making

In parallel with organizational ABM, Sandia researchers developed High-Definition Cognitive Models (HDCMs) and later the Unified Psycho-Cognitive Engine to represent individual cognition at high fidelity. These models integrated perception, memory, emotion, goals, and action intentions within a neuro-computational framework grounded in cognitive psychology and behavioral theory, including the theory of planned behavior (Ajzen 1991). Individuals were modeled as recognizing environmental cues, activating semantic schemas, evaluating attainable goal states, and selecting actions based on emotional weighting and prior experience.

The UPCE formalized this process into modular components—perceptual filtering, belief and intention representation, deliberative decision formation, and behavior generation. Emotional states such as anxiety–fear and frustration–anger were explicitly modeled as reciprocal influences on cognition, amplifying certain beliefs while inhibiting others. Individual actions fed into aggregate societal models, which in turn generated new cues for individuals, creating a closed feedback loop between micro-level cognition and macro-level social dynamics (Backus et al 2010). This bidirectional coupling directly reflects the cybernetic control loops outlined earlier in this volume. Bernard 2009 provides the following additional context:

> …create a social simulation platform that couples High- Definition Cognitive Models (HDCM) with a cultural, economic, and policy-based simulation. The HDCMs are purposely designed to computationally represent the mindset of specific individuals, including their cognitive perceptions, goals, emotion states, and action intentions. The actions of one HDCM can affect the mindset and actions of others, as well as the general mindset of the society in which they are situated. The society, computationally represented in this initial effort by Sandia's Systems Dynamics-based Aggregate Societal Model (SDASM) can, in turn, affect the actions of the HDCMs (see Figure 1). The HDCM is focused on individual or small-group level of analysis, whereas the SDASM is focused at an aggregate level social, economic, and cultural level of analysis [society]. These models are joined to provide a high-fidelity, scaleable assessment tool of individuals, small groups, and society to produce outcome distributions investigating attitudinal and behavioral reactions to US policies for a given country, group, or ethnic region.

Figure 2: The process diagram of the cognitive actions within the HDCM framework




The behaviors associated with possible actions conform to the theory of planned behavior, which maintains that behaviors are influenced by attitudes towards a specific behavior, the subjective norms associated with acting out that behavior, and the perception that this behavior is within a person's control. This forms an action intention state, which then typically drives that person's actual behavior. This type of high-fidelity representation can capture and express the basic psychological processes of individuals (e.g., leaders, terrorists). A key component to this technology is its neuro-computational model framework whereby a modeled human recognizes patterns of stimuli in the environment and responds appropriately to those stimuli according to prior experiences via its semantic knowledge and pattern recognition modules. The semantic module incorporates an associative network with nodes representing the critical concepts or "schemas" in an agents "mind." The pattern recognition and comparator modules provide mechanisms for:

(1) evaluating the evidence provided by cues favoring or conflicting with each situation and
(2) implementation of top-down activation. Implicit to recognition of a situation, there is recognition of goals, or attainable states, and the actions needed to realize those goals, including likely intermediate states.

The cognitive subsystem in our model serves as the point where the diverse emotions, memories, and other factors of an individual are all used to generate a

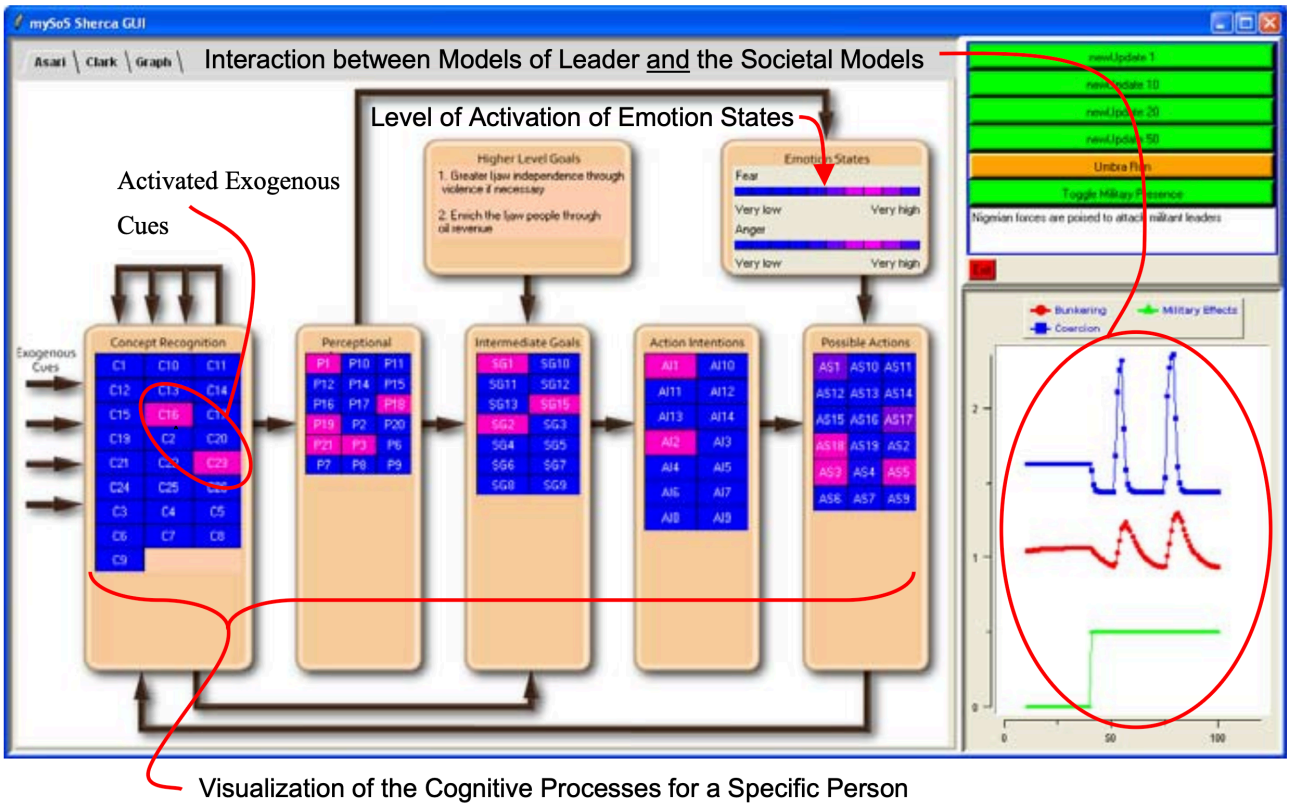
> decision for action (or inaction). Actions of the individual and their repercussions then effect how the aggregate model transitions to the next state (or return to the same state). At present, the emotions this cognitive model represents are anxiety-fear and frustration-anger. Activation of a specific concept or situation produces activation of associated emotional components. Emotions here have a reciprocal effect on cognition, causing increased activation of the concept or situation that triggered the emotion and active inhibition of other concepts and situations. (Bernard 2009)

**From Individual to Society Models:**
the modeling did not end just with the individual, Sandia extended HDCM to the society through the SDASM models as explained by Bernard et al 2009:

> Applying the techniques and models discussed above, Sandia has produced a prototype societal assessment capability that shows
>
> (1) potential actions, as well as the psychological processes behind those processes, for specific individuals of interest; and
>
> (2) potential societal actions in response to the actions of individuals of interest as well as exogenous variables. In the system, the inputs to the HDCM/SDASM system are cues associated with environmental events, US actions, and other external forces.
>
> These cues can be actual events, or be posed by analysts to create "what-if" scenarios. The cues will affect the HDCM by creating perceptions that are particular to a specific HDCM agent. The resulting cognitive states and actions will serve as inputs to the SDASM. The SDASM will represent the society in which the HDCM agents wield influence. The SDASM will receive the same cues as the HDCMs, as well as other cues that affect societies at an aggregate level. The output of the SDASM will serve as additional cues to the HDCMs.
> When fully implemented, it is believed the combined interactions will capture the dynamics, secondary effects, and potential unintended consequences so as to better assess/develop interventions and regional-stabilization conditions. Figure 3 shows an example of this process for a single individual as well as the interaction between the individual and the societal model. Incoming information activates specific concepts (shown in red) to represent specific modeled psychological processes (such as perceptions and goals). Potential actions will be fed to the SDASM, which will, in turn, activate concepts that will be fed to the HDCMs. The interactions from this process are then visualized in a graphical interface. (Bernard et al. 2009)

This provides an interesting precursor to targeting influencers that can then influence the society in general.

Figure 3. An example of the output of the prototype societal assessment tool




## Transition to Generative Agent-Based Modeling

The emergence of large language models transformed the feasibility of such cognitive and social simulations. Recent work on generative agent-based modeling demonstrates that LLM-powered agents can reason, communicate, and adapt without explicitly engineered cognitive rules. Instead, belief-like states, social norms, heuristics, and biases emerge implicitly from pretrained world models derived from large-scale human language data.

Empirical studies show that LLM-agents can adopt ideological stances, form communities, propagate information, and exhibit collective phenomena such as polarization and homophily within simulated social networks (Park 2023). Unlike earlier ABM systems—where simplification and researcher bias were persistent concerns—LLM-based agents generate diverse, context-sensitive behavior through role-play and reasoning. As a result, influence diffusion, opinion formation, and collective decision-making can now be simulated with substantially higher realism and scale. Ferraro et al talk about the development form rule based ABMs to generative Agents (GABMs), showing the advancement of contemporary developments:

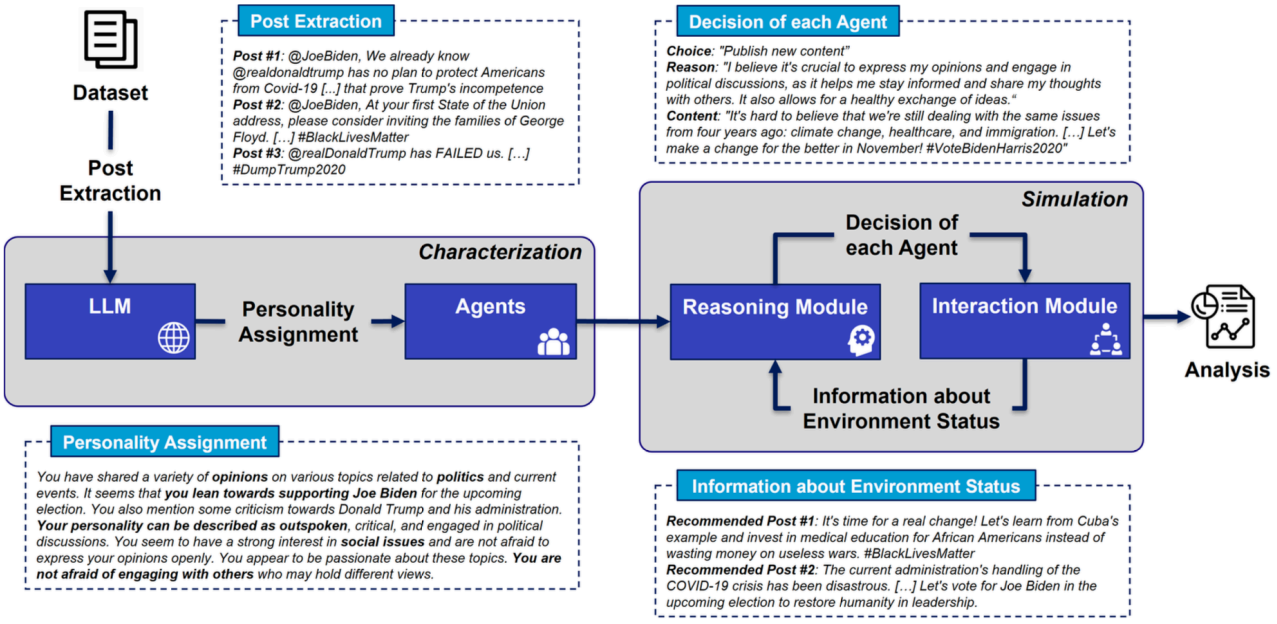



Fig. 1: Our framework comprises two primary phases: (i) *Characterization*, where each agent embodies the personality traits and interests extracted (via LLM) from the original posts of the real user it is tasked to emulate; and (ii) *Simulation*, where the decision-making process of each agent, represented as a *Choice-Reason-Content* triple (*Reasoning Module*), is stored within the *Interaction Module*. Consequently, each agent autonomously makes decisions, considering the context and having access to recommended contents posted by other agents.

(Image, Fig. 1 from Ferraro 2024)

Over the past decades, there has been a concerted effort among researchers and practitioners to develop computational agents capable of realistically emulating human behavior. Agent-Based Modelling (ABM) has emerged as a pivotal methodology for simulating intricate systems by delineating rules governing individual agents' behavior and interactions. Within the domain of social network analysis, ABM has played a crucial role in both the development and validation of novel theories pertaining to human behavior in online environments. These theories encompass a wide array of phenomena such as opinion formation, (false) news propagation, and collective decision-making. Nevertheless, manually crafting agent behavior to encompass the diverse spectrum of interactions, information flow dynamics, and user engagement within social networks proves to be highly challenging. This challenge often leads to an oversimplification of agents or the social media environment itself, where underlying mechanisms are rigidly encoded in predefined parameters. Consequently, such setups are prone to researcher bias, potentially resulting in a lack of fidelity in modeling complex human behaviors, especially those involving collective decision-making.

> Modern Large Language Models (LLMs) not only excel in generating human-like text but also demonstrate remarkable performance in complex tasks requiring reasoning, planning, and communication. This proficiency has sparked interest in integrating LLMs with ABM, termed Generative Agent-Based Modelling (GABM). Unlike traditional ABM methods that often necessitate intricate parameter configurations, GABM leverages LLMs' capacity for role-playing, ensuring diverse agent behaviors that closely mirror real-world diversity. For instance, Park et al. demonstrated that generative agents, designed for daily activities, exhibited credible individual and social behaviors, including express- ing opinions and forming friendships, without explicit instructions. Similarly, Williams et al. showcased the collective intelligence of generative agents in epidemic modeling, accurately simulating real-world behaviors like quarantine and self isolation in response to escalating disease cases. These pioneering findings support investigating GABM as an effective approach to enhance social media simulations. To our knowledge, the seminal work by Gao et al. lays the foundation for this research direction by qualitatively demonstrating that LLM-agents exhibit realistic behaviors related to information propagation and the manifestation of attitudes and sentiment. However, it remains unclear whether LLM-agents can accurately represent real users in terms of their personality traits (e.g., being outspoken, being critical) and interests (e.g., social issues, political preferences), regardless of the explicit emotions conveyed through their textual posts. Furthermore, their ability to exhibit community-level phenomena (e.g., homophily, polarization), as well as their susceptibility to recommendation strategies, remains uncertain.
> (Ferraro et al. 2024)

## From Simulation to Operational Capability

The most consequential divergence from Sandia's early work is operational rather than conceptual. Sandia's cognitive agents operated entirely within simulated environments. Contemporary agentic AI systems, by contrast, act directly in the real world. Tool-using LLM agents can execute code, generate targeted narratives, coordinate with other agents, conduct reconnaissance, send signals, and influence live audiences in real time.

Accordingly, organizations such as RAND, DARPA, the U.S. Department of Defense, and NATO StratCom now employ LLM-based agents for wargaming, influence modeling, adversary-behavior prediction, and strategic-communication analysis (RAND 2024). In these contexts, LLMs function as semi-autonomous cognitive actors—reading intent, adapting messaging strategies, forecasting social cascades, and participating in multi-agent simulations. What Sandia researchers once attempted to engineer manually is now obtained through pretrained models with minimal explicit cognitive design.

## Strategic Continuity and Risk Implications

Taken together, the Sandia ABM and UPCE efforts reveal a clear developmental arc. Cybernetics provided the foundational blueprint; Sandia operationalized that blueprint through explicit psycho-cognitive and multi-agent models; modern agentic AI systems now instantiate the same functional architecture at scale, with autonomy, adaptability, and real-world actuation. As argued throughout this book, this continuity explains why agentic AI represents not merely a quantitative advance, but a qualitative shift in risk.

Autonomous influence operations, emergent deception, automated group herding, and reflexive escalation loops—once theoretical concerns—are now technically plausible. As agent populations scale and interact across social, cyber, and informational domains, unpredictable macro-level behaviors emerge that cannot be reduced to individual agent intent. The result is a new class of threat centered on cognition, perception, and decision-making itself.

Understanding the Sandia lineage is therefore essential for understanding modern agentic AI risk. The same objectives persist across eras: modeling cognition, predicting behavior, steering outcomes, and operating within adversarial feedback loops. What has changed is capability. What was once simulation is now deployment; what was once handcrafted cognition is now learned at scale.

**Recommendations for Defense and Governance**

To help mitigate the risks associated with agentic AI the following governance oversight recommendations are given for policy makers and for developers:

- Develop **agent-detection frameworks** similar to botnet C2 detection.
- Mandate **tool-use sandboxing** in AI deployments.
- Create **cognitive firewalls** preventing AI-driven influence ops (See Ch. 9 McCarron 2024).
- Implement **AI behaviour red-team ecosystems**.
- Develop **agentic safety rulesets** (analogous to nuclear PALs).
- Establish **international norms on autonomous cyber actors**.

**Bibliography:**


Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human

Backus, G., & Glass, R. (2006). *An Agent-Based Framework for Modeling Human Cognition and Behavior*. Sandia National Laboratories.

Hernandez, L., Sloane, M., & Rahwan, I. (2024). *Escalation risks from language models in military and diplomatic decision-making*. ACM. https://doi.org/10.1145/3630106.3658942

NATO Strategic Communications Centre of Excellence. (2023). *Large language models and their use in influence operations*. NATO StratCom COE.

Park, J., et al. (2023). *Generative Agents: Interactive simulacra of human behavior*. arXiv:2304.03442.

RAND Corporation. (2024). *Strategic competition in the age of AI: Emerging risks and opportunities*. RAND Europe.

RAND Corporation. (2025). *Acquiring generative artificial intelligence to improve U.S. Department of Defense influence activities* (RRA3157-1).

U.S. Department of Defense, Chief Digital and AI Office. (2024). *Generative AI Guidance and Experiments*. https://www.ai.mil

Zhu, X., et al. (2025). *Simulating influence dynamics with LLM agents*. arXiv:2503.08709.

DARPA. (2023–2025). *INCAS* and *KAIROS* program documentation.

Backus, G. A., & Glass, R. J. (2006). *An Agent-Based Model Component to a Framework for the Analysis of Terrorist-Group Dynamics.* Sandia Report SAND2006-0860P. Sandia National Laboratories.ACM Digital Library+2Academia+2

Bernard, M. L., Backus, G. A., Verzi, S. J., Bier, A. B., & Glickman, M. (2010). *Foundations to the Unified Psycho-Cognitive Engine.* Sandia Report SAND2010-6974. Sandia National Laboratories.researchgate.net+2OSTI+2

Bernard, M. L., Backus, G. A., Glickman, M. R., Gieseler, C., & Waymire, R. (2009). Modeling Populations of Interest in Order to Simulate Cultural Response to Influence Activities. In *Social Computing and Behavioral Modeling* (pp. 1–8). Springer.OSTI+2SpringerLink+2

Bernard, M. L., Backus, G. A., & Bier, A. B. (2014). Behavioral Influence Assessment (BIA): A Multi-Scale System to Assess Dynamic Behaviors Within Groups and Societies Across Time. In *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics (AHFE 2014).*researchgate.net+2Sandia National

Laboratories+2

Naugle (Bier), A. B., Bernard, M. L., Backus, G. A., et al. (2014). Simulating Smoking Behaviors Based on Cognition. *Winter Simulation Conference Proceedings.*ACM Digital Library+2OSTI+2

Lakkaraju, K., Naugle, A. B., Verzi, S. J., Swiler, L. P., Livesay, M., Warrender, C. E., Bernard, M. L., & Romero, V. (2019). *Complexity Metrics for Agent Based Models of Social Systems.* Sandia Report SAND2019-4189C.Sandia National Laboratories+3OSTI+3Sandia National Laboratories+3

McCarron, M. (2023) *Play AI: Machine Learning in Video Games* https://www.amazon.com/Play-AI-Machine-Learning-Video/dp/B0BW2X9B34

McCarron, M. (2024) *Battlespace of Mind: AI, Cybernetics and Information Warfare* https://github.com/autonomous019/Battlespace-of-Mind

Sandia National Laboratories (2020). *DYMATICA: Dynamic Modeling for Assessing Threats and Influences on Cognitive Agents.* Brochure and associated publications.Sandia National Laboratories

Glass, R. J., Backus, G. A., et al. (2008–2010). *A Roadmap for the Complex Adaptive Systems of Systems (CASoS) Engineering Initiative* and related CASoS reports. Sandia National Laboratories.OSTI+1

Ferraro, A. Et al. (2024) *Agent-Based Modelling Meets Generative AI in Social Network Simulations* arXiv:2411.16031v1

**Sandia Related Work Bibliography:**

A selection of research papers listed by Sandia as being relevant to influence operations as listed in DYMATICA documentation (2020)

Backus, G.A, Bernard, M.L., Verzi, S., Asmeret, B., Glickman, M. (2010). Foundations to the Unified Psycho-Cognitive Engine. Sandia National Laboratories technical report SAND Report.
Bernard, M.L. (2004). Simulating Human Behavior for National Security Human Interactions, Technical Advance SD-7868/S-106,125, Sandia National Laboratories, Albuquerque, NM.
Bernard, M.L. (2015). Developing a Capability to Elicit and Structure Psychosocial Decision Information within Computational Models. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.
Bernard, M. L., & Bier, A. B. (2014). Analytical Capability to Better Understand and Anticipate Extremist Shifts Within Populations in Failing States. In Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics.
Bernard, M.L., Backus, G.A., Bier, A.B. (2015). Behavioral Influence Assessment (BIA): A Multi-Scale System to Assess Dynamic Behaviors within Groups and Societies Across Time. In Proceedings of the 5th International Conference on Applied Human Factors & Ergonomics AHFE.

Bernard, M.L, Backus, G., Glickman, M., Gieseler, C., & Waymire, R. (2009). Modeling Populations of Interest in Order to Simulate Cultural Response to Influence Activities. In Social Computing and Behavioral Modeling. Springer US.
Bernard, M.L., Backus, G.A., Naugle, A.B., Jeffers, R.F., Damron, R.W. (in print). Anticipating the Potential Range of Behaviors for Individuals Interacting within Societies. In Modeling Sociocultural Influences on Decision Making. Taylor & Francis.
Bier. A.B. Sensitivity Analysis Techniques for Models of Human Behavior. Sandia National Laboratories Technical Report. SAND Report 2010-6430.
Bier, A., Bernard, M.L. (2014). Validating a Hybrid Cognitive-System Dynamics Model of Team Interaction. Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014, Kraków, Poland 19-23 July 2014.
Bier, A.B., Bernard, M.L., Backus, G., & Hills, R. (2010). Political Dynamics Determined by Interactions Between Political Leaders and Voters. In Proceeding of the 28th International Conference of the System Dynamics Society, July 25-29 2010, Seoul, South Korea.
Naugle, A.B. & Bernard, M.L. (2016). Using Computational Modeling to Examine Shifts Towards Extremist Behaviors in European Diaspora Communities. In Proceedings of the 6th International Conference on Applied Human Factors and Ergonomics.
Passell, H. D., Aamir, M. S., Bernard, M. L., Beyeler, W. E., Fellner, K. M., Hayden, N. K., ... & Silver, E. (2016). Integrated human futures modeling in Egypt. SAND Report 2016-0388. Sandia National Laboratories, Albuquerque, NM.
Raybourn, E., Hills, R. G., Schimanski, B., Bouchard, J., Bernard, M., Shaneyfelt, W. (2010). Interactive Validation and Verification Environment for Human, Social, Cultural, Behavioral Models. SAND Report 2009-6384P. Sandia National Laboratories, Albuquerque, NM.
Williams, G.R., Bernard, M.L., Jeffers, R.F. (2016). Examining the, Ideological, Socio-political, and Contextual Factors Underlying the Appeal of Extremism. In Proceedings of the 6th International Conference on Applied Human Factors and Ergonomics.