Psychologically-Valid Generative Agents: A Novel Approach to Agent-Based Modeling in Social Sciences

Konstantinos Mitsopoulos[1], Ritwik Bose[1], Brodie Mather[1], Archna Bhatia[1], Kevin Gluck[1], Bonnie Dorr[2], Christian Lebiere[3], Peter Pirolli[1]

[1]Institute for Human and Machine Cognition
[2]University of Florida
[3]Carnegie Mellon University
{kmitsopoulos, rbose, bmather, abhatia, kgluck, ppirolli}@ihmc.org
bonniejdorr@ufl.edu, cl@cmu.edu

Abstract

Incorporating dynamic realistic human behaviors in population-scale computational models has been challenging. While some efforts have leveraged behavioral theories from social science, validated theories specifically applicable to Agent-Based Modeling remain limited. Existing approaches lack a comprehensive framework to model the situated, adaptive nature of human cognition and choice. To address these challenges, this paper proposes a novel framework, Psychologically-Valid Generative Agents. These agents consist of a Cognitive Architecture that provides data-driven and cognitively-constrained decision-making functionality, and a Large Language Model that generates human-like linguistic data. In addition, our framework benefits from Stance Detection, a Natural Language Processing technique, that allows highly personalized initialization of the agents, based on real-world data, within agent-based modeling simulations. This combination provides a flexible yet structured approach to endogenously represent how people perceive, deliberate, and respond to social or other types of complex decision-making dynamics. Previous work has demonstrated promising results by using a subset of the components of our proposed architecture. Our approach has the potential to exhibit highly-realistic human behavior and can be used across a variety of domains (e.g., public health, group dynamics, social and psychological sciences, and financial markets).

Introduction

The recent COVID-19 pandemic prompted a massive global response, with substantial variation in behavior change across subsets of the population. The pandemic also highlighted the importance of modeling human behavior (e.g., social distancing; vaccination) at population scales because decisions regarding behavior are central to modulating pathogen transmission (West et al. 2020) and crucial to forecasting the dynamics of viral transmission and resulting infection cases and deaths. More generally, large scale models that accurately capture the heterogeneity and complexity of human psychology could be important in a wide variety of societally important areas, such as natural disaster response (e.g., wildfires, hurricanes, tornadoes), climate change, public health, civic discourse, diplomacy, economic policy, cybersecurity, and military planning.

Grossmann et al. (2023) argue that the social sciences can harness the knowledge (hence power) of LLMs in a variety of ways, including as surrogates for human participants and/or confederates in empirical data collection, the generation of new hypotheses, counterfactual simulations for high-risk projects (e.g., nuclear deterrence; large scale public health interventions), and predicting online flows of information. Grossmann et al. especially advocate for the potential of combining LLMs with Agent-Based Models (ABMs) to "provide new insights on how human agents choose to share information, cooperate and compete in social dilemmas, and conform with social norms" revealing "the underlying mechanisms governing human behavior and social dynamics". Longitudinal small-N agent simulations (Park et al. 2023) have illustrated that LLM-based agents can interact in a way that is compellingly human (albeit in a toy domain). However, recent studies (Binz and Schulz 2023; Shiffrin and Mitchell 2023) of current LLMs indicate they lack cognitive competence in ways that sometimes produce fragile and even bizarre behavior.

We adopt the position that LLMs combined with ABMs have great potential for population-scale psychological and social science. We further argue for a new paradigm to achieve this combination, in which the ABMs are based on invariant characteristics of Cognitive Architecture theory and personalized through natural language processing (NLP) for Stance Detection.

In our own recent work (Mather et al. 2021; Pirolli et al. 2020, 2021), we have developed data pipelines combining demographic and psychographic data about U.S. regions and NLP of online social media that are used to initialize agents implemented in a subset of the ACT-R cognitive architecture. This yields what we call Psychologically Valid Agents (PVAs). In the context of the COVID-19 pandemic, these agents can be used to predict available regional time series data about human behavior, such as the U.S. county- or state-level daily mobility patterns or daily mask-wearing. Our prior work in this domain provides a use case and foundational proof of concept for the approach we are advocating as described below.

This paper illustrates components of an approach to

Psychologically-Valid Generative Agents (PVGAs) for large-scale psychological and social science. We introduce our current approach to PVAs developed in the ACT-R architecture to model human psychology and behavior during the COVID-19 pandemic. We introduce the concept of generative agents that reason and simulate human-like conversational behavior. This type of reasoning occurs in the linguistic space the agents use to interpret their observations. Next, we illustrate how NLP techniques have been used to provide cognitive content to drive decision-making in PVAs. Finally we discuss how all the components together can create highly-detailed and data-driven agents for psychological and social science simulations.

PVAs for Agent-Based Modeling

PVAs are computational agents implemented within the ACT-R architecture (Anderson et al. 2004) to simulate and analyze human behaviors. They offer an approach to modeling, particularly in understanding and predicting behaviors in specific contexts, such as responses to pandemic guidelines, with input drivers induced from heterogeneous sources including online media such as Twitter that provide indicators of pandemic awareness, beliefs, and attitudes (Pirolli et al. 2020). The subset of ACT-R methods employed in designing PVAs is grounded in the Instance-Based Learning Theory (Gonzalez, Lerch, and Lebiere 2003). We refer to this specific approach as CogIBL to differentiate it from other Instance-Based Learning methods prevalent in Machine Learning.

The Cognitive Instance-Based Learning (CogIBL) model is a cognitive framework that operates within the theoretical foundations of the ACT-R architecture (Anderson et al., 2004). At its core, CogIBL models human learning from experience. As individuals encounter novel situations, they refer to similar past instances, stored in a memory module, to inform decision-making. With accumulated experiences over time, CogIBL refines its knowledge, drawing from an increasingly rich history of prior encounters. This iterative cycle of learning and refinement grounded in experiential knowledge makes CogIBL an adaptable and dynamic model. Complementing this adaptive learning approach is the architectural support from ACT-R, providing a robust cognitive theory for mechanistic learning processes. Overall, CogIBL aims to leverage both experience-based and theoretically-grounded techniques to achieve human-like learning.

The model is based on the idea that decisions and behaviors have subjective utility or value, such as satisfaction or preference. When a behavior occurs in a situation and produces an outcome, it is associated with a subjective assessment of its value. Following ACT-R theory, these experiential associations are stored in **declarative memory** as experiential records (chunks) of decision-making situations, behaviors, outcomes, and their values.

Over time, this repository of experiences forms the basis for implicit and explicit knowledge about decision-making (Lebiere, Wallach, and Taatgen 1998; Lebiere and Wallach 1999; Wallach and Lebiere 2003). It is assumed that when individuals are faced with decisions, they draw from these stored experiences, retrieving memories that align with current cues to evaluate alternatives and decide on actions. This relies on ACT-R's memory **retrieval** and **blending** mechanisms. Retrieval uses situation cues to recall past instances based on their **recency**, **frequency** and **similarity** to the current situation. Blending aggregates and generalizes across activated memories. By leveraging instance-based knowledge, the model is able to estimate expectations of potential outcomes based on past similar situations.

Agent-based modeling (ABM) (Reynolds 1987) has emerged as a technique for understanding complex systems across diverse disciplines (Epstein and Axtell 1996; Axelrod 1997). Unlike equation-based approaches, ABM employs bottom-up modeling where individual entities or "agents" interact based on simple behavioral rules, giving rise to emergent collective dynamics. When configured into networked topologies, ABM enables the examination of how local interactions propagate through the system. For example, in social science, ABM enables exploring how individual behaviors scale into societal outcomes. In epidemiology, ABM uniquely captures disease transmission via agent interactions, supporting containment policy decisions (Eubank et al. 2004). The ABM's flexibility stems from complexity theory, where simple nonlinear rules generate complex dynamics. By encoding domains as adaptive agent systems, ABM provides a framework for gaining insights into complex phenomena.

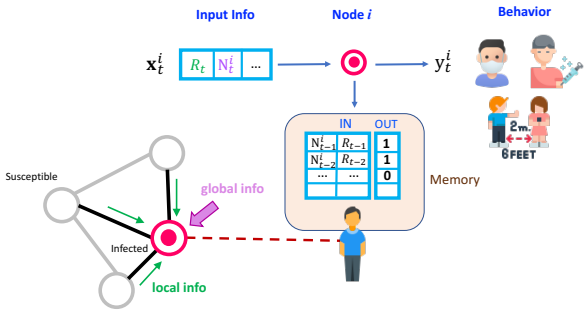

Figure 1: Example of PVA functionality in ABM simulation.

The integration of CogIBL with ABM frameworks presents a promising avenue for enhancing cognitive plausibility in social simulations. As a non-parametric, instance-based decision-making model, CogIBL eliminates the need for prescriptive rules or long periods of training. Instead, it can leverage data-driven approaches for adapting and learning in real-time as new data emerges. This capability ensures that agents can dynamically adjust their behaviors in response to evolving scenarios.

For instance, in Figure 1 we showcase how PVAs can simulate human behavior in an ABM epidemiological scenario. Each node i represents an individual who constantly receives information from their surroundings, both from a global perspective, e.g., the total number of disease cases (R_t), and locally, e.g., the number of sick people around them (N_t^i). Based on this information, each agent needs to make informed decisions (e.g. masking, distancing, getting

vaccinated). The CogIBL agent uses the current state of the network, i.e., the global and local information, as input. It then compares this with a set of past experiences stored in its memory and makes decisions by aggregating these experiences. Similar to humans, these agents learn from their past to make informed decisions about the future.

Furthermore, CogIBL's implementation allows it to scale to represent thousands of agents within a network graph, thereby serving as a robust and adaptive decision-making system. The synergistic combination of data-driven cognitive modeling with network-scale ABM simulations can potentially advance research on emergent collective social phenomena arising from individual behaviors.

Generative Agents for ABM Simulations

Generative Agents (GAs) (Park et al. 2023) are computational software agents designed to simulate believable human behavior. These agents can "perform" activities like cooking breakfast, going to work, painting, writing, forming opinions, initiating conversations, and more. They can remember past events and plan for the future. The architecture for these agents extends an LLM to store a complete record of the agent's experiences using natural language. This allows the agents to synthesize memories over time into higher-level reflections and retrieve them dynamically to plan behavior. Their main components are:

- **Memory**: The capability to store and retrieve past interactions, allowing for context-aware responses. **Retrieval** is based on a **relevance score** function dependent on the similarity between the current input and the past instances.
- **Planning**: The foresight to strategize and generate a series of actions, emulating the human ability to plan ahead.
- **Reflection**: An introspective feature enabling agents to contemplate and learn from their past actions.
- **Reactivity**: Rapid and appropriate responses to environmental stimuli, ensuring dynamic interactions.

Together, these architectural features aim to produce virtual agents that can overcome limitations of scripted conversations and more closely emulate fluid, context-appropriate human interactions. The architecture's method of storing experiences as instances in a memory repository, and subsequently recalling and weighing them based on relevance, exhibits a clear functional similarity to CogIBL inference mechanisms (GAs described in detail in Park et al. (2023) and CogIBL in Mitsopoulos et al. (2020, 2022)). Consequently, CogIBL seamlessly integrates with the GA architecture, layering utility-based reasoning atop the inherent capabilities of the GA framework.

GAs can be used in various domains, from role-play and social prototyping to virtual worlds and games. Williams et al. (2023) use memoryless GAs to simulate realistic human behavior in epidemiological ABM simulations. They argue that many epidemic models have not fully incorporated the dynamic nature of human behavior and how it changes in response to the state of the epidemic. These methods often rely on exogenous data inputs rather than endogenously modeling behavioral mechanisms. They propose using generative AI to empower individual agents with flexible reasoning and decision-making abilities. Rather than manually specifying behavioral rules, each agent can leverage an LLM, like GPT-3, to dynamically determine actions based on the current context. The LLM's knowledge about natural language and common sense acts as a proxy for human reasoning. Agents can self-isolate, quarantine, and react to epidemic dynamics without prescriptive theories or overfitting to limited data. They also observe that collectively, the agents' adaptive actions flatten the epidemic curve.

This approach allows creating realistic, diverse representations of human processes and responses. Overall their method addresses key gaps by enabling a flexible behavioral representation without relying on explicit theories or extensive data fitting. Their results validate the potential of the generative agent method for more realistic behavioral modeling in complex social systems. Although their approach shows promise for flexible behavioral modeling, the agent's reasoning is entirely based on the LLM without incorporating explicit computational mechanisms representing human cognitive processes (such as memory retrieval prioritizing experiences based on relevance, recency and frequency). In contrast, our PVA framework incorporates a cognitive architecture with mechanisms that more closely reflect human cognition.

Exploiting Natural Language to Generate PVAs

One key factor in producing more realistic behavioral modeling is the inclusion of beliefs within a generative agent. Much of the variability in human behavior is due to variability in knowledge, beliefs, and attitudes etc. In our current PVA work, we explore NLP techniques such as stance detection for identifying beliefs and attitudes expressed through language. We extract such content from natural language (e.g., language used in social media), and map those into cognitive representations in PVAs.

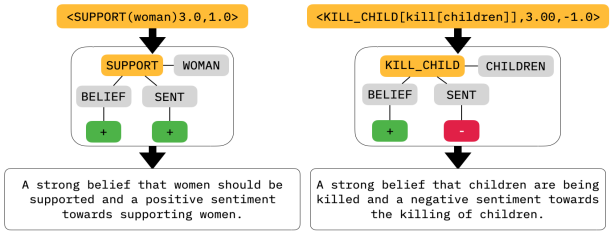

Figure 2: Example of a stance representation from the abortion domain (based on discussions about abortion on social media)

Our stance framework (Mather et al. 2021) extracts belief, defined as a predicate-argument representation that captures a domain-specific (e.g. Covid, abortion, etc.) belief type, along with a belief strength and a sentiment towards that belief, yielding an overall *attitude*. Figure 2 shows two exam-

ples in the abortion domain (to highlight stance detection's domain adaptation capabilities), one with a belief type corresponding to *support woman*, and the other with a belief type corresponding to *kill children*. These are, correspondingly, associated with the predicate-argument representations SUPPORT(woman) and KILL_CHILD(. . .children), both exhibiting strong belief strengths (3.0) but each with a different sentiment strength: positive (1.0) for the former and negative (-1.0) for the latter. From these stance representations a templatic approach is used to transform them into textual descriptions that feeds the PVAs.

Stances (beliefs, sentiments, and attitudes) can be used to generate and drive PVAs. Intentions and decisions are represented in our PVAs as competing chunks for actions, such as wearing a mask or not. The underlying base-level activation of the competing alternative action chunks drives the selection of actions and determines the probability of the chosen action. We develop a framework to translate attitudes and beliefs derived from Twitter Stance Detection into chunk representations. Positive or negative sentiments about a belief correspond to competing attitude chunks, whereas belief strength corresponds to activation values associated with those chunks. To normalize for frequency effects, the ratio of positive vs. negative attitudes maps to the relative frequency of those chunks in the ACT-R models. CogIBL retrievals of those chunks will reflect their underlying activation and determine behavioral choice probabilities that are contextualized to the current situation.

PVAs Modeling Human Psychology and Behavior during COVID-19

In this section, we present empirical validation of the PVA modeling framework through forecasting and ABM simulation experiments. We first demonstrate the capability of PVAs to generate accurate forecasts on real-world behavioral data. Subsequently, results from ABM simulations highlight how integrating PVAs and ABM methods in an epidemiological network reproduces complex social patterns emerging from psychologically grounded mechanisms implemented at the individual level.

PVAs for Forecasting Human Behavior

Our PVA pipeline includes demographic and psychographic data about U.S. regions and online social media. These data can be used to initialize agents, or provide time-series inputs, and the PVAs can iteratively assess current context (e.g., case rates) and make decisions (e.g., wear a mask or not) over discrete time steps (e.g., every day). These PVAs can be used to predict available regional time series data, such as the U.S. county- or state-level daily mobility patterns or daily mask-wearing (Figure 3).

The PVAs can also predict novel patterns we have observed in the reactions of humans to their awareness of external events (Figure 4) (Pirolli, Lebiere, and Orr 2022). These PVAs can also be probed to understand the relation of input factors to output behavior using a variety of methods (Figure 5) including a measure of cognitive salience that dynamically computes the extent to which a behavior reflects various features of a situation (Somers et al. 2019).

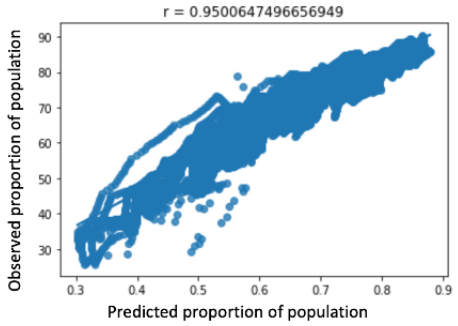

Figure 3: Predicted versus observed proportion of state population wearing masks each day from 3/15/2020 to 2/31/2021 (the first three waves of COVID-19) using a PVA based model.

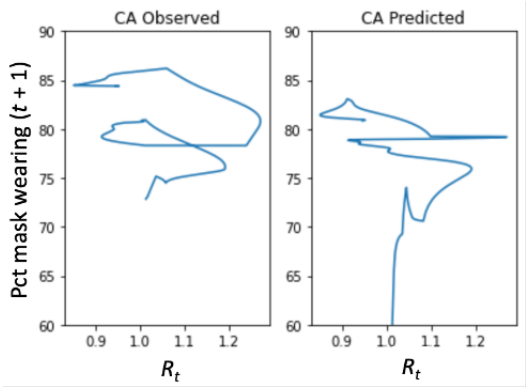

Figure 4: Observed (left) and PVA-predicted (right) relationship between effective transmission number (R_t) at time t and percent mask wearing at time $t + 1$ (a delay of 7 days) over the first three waves of COVID-19.. The general phenomena is an oscillation around $R_t = 1$ combined with a learning effect towards higher masking over three waves of COVID-19.

PVAs in ABM Epidemiological Simulations

We integrate PVAs into an ABM framework within an epidemiological network of approximately 10,000 nodes (averaging 11 connections per node). This network serves as a scaled-down representation of a larger synthetic population network of Portland (∼1M nodes). Despite its size, it mirrors the statistical attributes of the expansive Portland network, ensuring its capacity to simulate COVID-19 transmission dynamics.

Each node within this network is characterized by a PVA, which exhibits a predisposition towards adopting masking behavior if more than 3 of its neighboring nodes are infected. This behavior is governed by an arbitrary rule, implemented to demonstrate emergent dynamics within the network. Throughout the simulation, which spans 105 days,

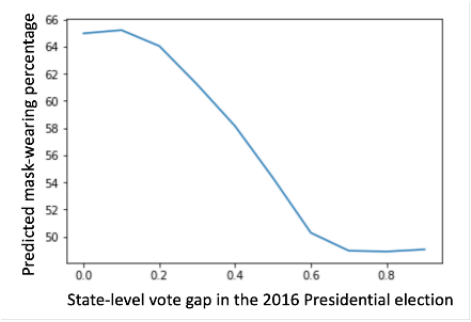

Figure 5: PVAs trained to model state mask-wearing can be probed with hypothetical values of psychographic or demographic features (e.g., state-level lean towards voting for Trump in the 2016 election) to make behavioral predictions.

agents are updated every 7 days with localized information about their infected neighbors, influencing their decision to wear a mask.

Our observations, as depicted in Figures 6, 7 and 8, indicate the adoption of masking behavior effectively attenuates the peak of the epidemiological curve, subsequently decelerating the spread of the disease. Figure 6 specifically illustrates that agents with minimal connections rarely wear masks, primarily due to their infrequent encounters with infected neighbors. In contrast, highly connected individuals exhibit a surge in masking behavior during periods of high transmission rates, which gradually diminishes. In summation, our experiments underscore the capability of PVAs to emulate realistic behaviors within networks, offering a promising avenue for simulating real-world behavioral dynamics.

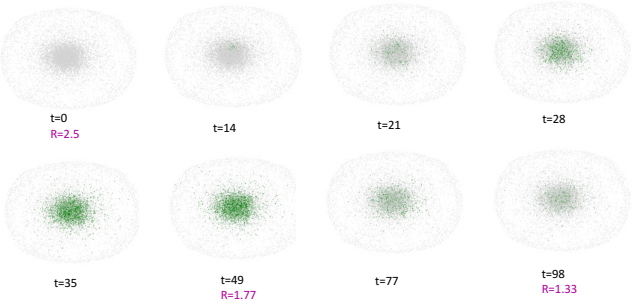

Figure 6: Masking behavior evolution in a synthetic population network of Portland.

The findings presented in this section provide quantitative evidence for the representational capacity of PVAs in forecasting human behavior as well as modeling mechanisms underlying collective social dynamics. The experiments underscore the potential of blending data-driven and ABM methods with generative neural modeling to advance social scientific insights.

Psychologically-Valid Generative Agents

We propose a novel framework we call Psychologically Valid Generative Agents, that integrates data-driven and generative modeling approaches to simulate human behaviors and interactions with high fidelity, in complex network topologies. As GAs and CogIBL share common architectural components, the proposed framework combines CogIBLs and LLMs as main decision-making and reasoning mechanisms for agents. This framework utilizes beliefs and attitudes, extracted from social media by our Stance Detection module, to generate highly-detailed psychological and behavioral profiles for these agents. These agents are characterized by their integration of:

- **Exogenous data-driven approaches**: This encompasses data derived from external sources, capturing various facets of human behavior and societal responses. Examples include mobility patterns, intentions to vaccinate, adherence to masking protocols, the influence of public health policies, and the understanding of expressed beliefs through stance detection.
- **Endogenous Data Generation**: At the core of our approach is the utilization of an LLM inherent to each agent (generative component). This endogenous data is dynamically produced based on the agent's reasoning capabilities about its interactions and observations of its environment. Such interactions might involve conversing with other agents, processing and interpreting news, or reflecting upon stored memories.
- **World Interaction Data**: This represents the agent's engagement with its surrounding environment, typically represented by an evolving social network graph. It captures the agent's actions, decisions, and interactions within the simulated world.

Figure 9 illustrates an instantiation of the proposed architecture. The PVGA is designed to simulate the behavior of an agent interacting within a social network. For demonstration purposes, we assume the network models epidemiological dynamics. The aim is to simulate detailed decision-making processes for specific behaviors. Central to this architecture are the CogIBL and an LLM. The CogIBL receives exogenous data from various sources (demographics, psychographics, etc.). Online social media data are processed by the Stance Detection module to extract beliefs related to specific attitudes. These beliefs are used by the LLM to generate a profile for each agent's identity. Together, the endogenous and exogenous data form the basis of the personality of each agent in the network. Agents are then deployed in the network and interact with each other, generating observations that are used by PVGAs to update their decisions, preferences and beliefs.

For further insight into the PVGA characteristics: initializing generative agents with domain-specific stances allows the simulation to establish a data-driven base "*personality*" for each agent. That is, stances are designed to embody an authentic initial standpoint, informing future interactions and decision-making. Stance detection is further generalized through its capability for rapid domain adaptation, which enables the initialization of generative agents for applicabil-

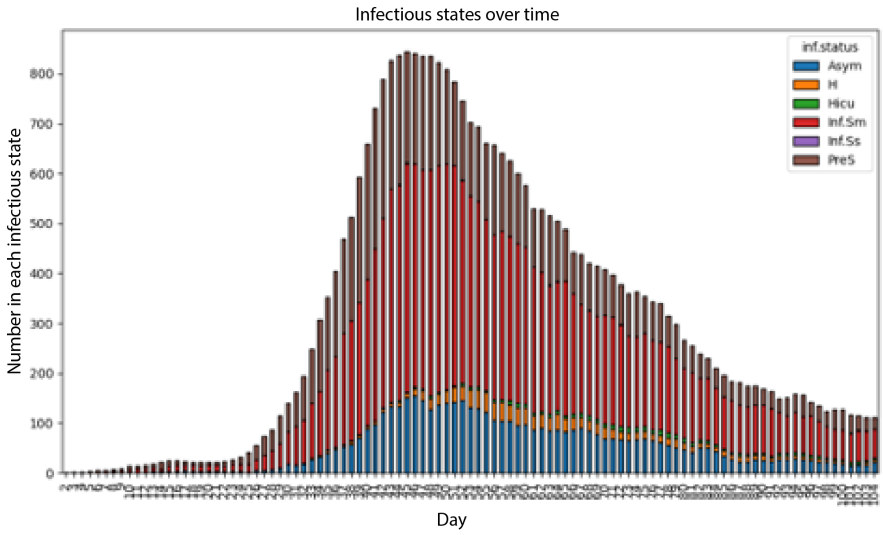

Figure 7: Epidemiological curves for masking behavior in a synthetic population network of Portland.

ity across a variety of domains (e.g., emergency response to wildfires, hurricanes, and tornadoes, climate change, public health, civic discourse).

LLMs pave the way for amplifying this approach by enabling agents to participate in stance-informed discussions. The extensive knowledge contained within LLMs makes them uniquely suited to this task as they can generate realistic textual exchanges among agents based on their established stances and underlying beliefs. This enables more realistic population-scale psychological and social science experiments with minimal data requirements as the LLMs generate the bulk of the conversational data. LLMs also help as they contain knowledge about the interplay among beliefs, attitudes, psychographic variables, group identity, etc. with likely behavior. These can be exploited to populate the PVGAs with cognitive content and underlying activation patterns (see Figure 10). The embedding spaces of LLMs can also be probed to provide the similarity spaces for declarative memories in PVGAs.

While specific stances often align with corresponding actions, they don't always reflect the actual state of an agent. To realize an action, an agent may combine multiple stances with varying strengths, coupled with the application of internal social constraints (e.g., desire for conformity), situational factors (e.g., working out), or other states (e.g., hunger, pain, or habituation). A simple approach (e.g., Naive Bayes) for identifying an agent's realized action might rely on application of a transition probability from each precondition to post-condition, and selection of the highest likelihood action based on existing preconditions. However, such an approach not only ignores the potential interactions between predictions but also necessitates observable data for model fitting.

An alternative approach constructs schema which provide "chains of thought" from a precondition to an action realization. Discerning an action given the precondition then involves identifying a self-consistent set of schema proposing a single action, where the self-consistency is determined via an internal dialogue. These schema can be applied in the opposite direction as well, allowing observing agents to infer stances and preconditions by combining observations about an individual. Note that the observational inference may not match the internal dialogue since actions may have multiple self-consistent schema sets for a given set of observations.

This synthesis of data-driven, generative, and interaction modeling aims to capture the complexity of human psychology, cognition and social dynamics. By incorporating data science and LLM reasoning in a cognitive architecture, psychologically validated agents can enhance computational modeling and theory development in the social sciences. The architecture, equipped with these capabilities, is able to encode behavioral, cognitive and social psychology constructs to generate high-fidelity behavior. The key innovation is enabling agents to exhibit realistic behaviors akin to humans situated in complex social contexts. This allows non-trivial emergent social phenomena to arise from first principles modeled at the individual level.

Advantages of the PVGA Approach

The PVGA approach offers a plethora of advantages, making it particularly well-suited for Agent-Based Modeling (ABM) in social sciences. We highlight the most important ones below:

- **Inherent Properties of CogIBL**: CogIBL, as an instance-based learning methodology, eliminates the need for traditional training phases to perform inference. Its non-parametric nature means there is no cyclical process of training, parameter freezing, inference, and data acquisition. This allows for continuous online learning as new data streams in. Furthermore, CogIBL has excellent scalability properties, especially when simulating agents in parallel across large networks. This scalability is achieved through parallelized operations and the

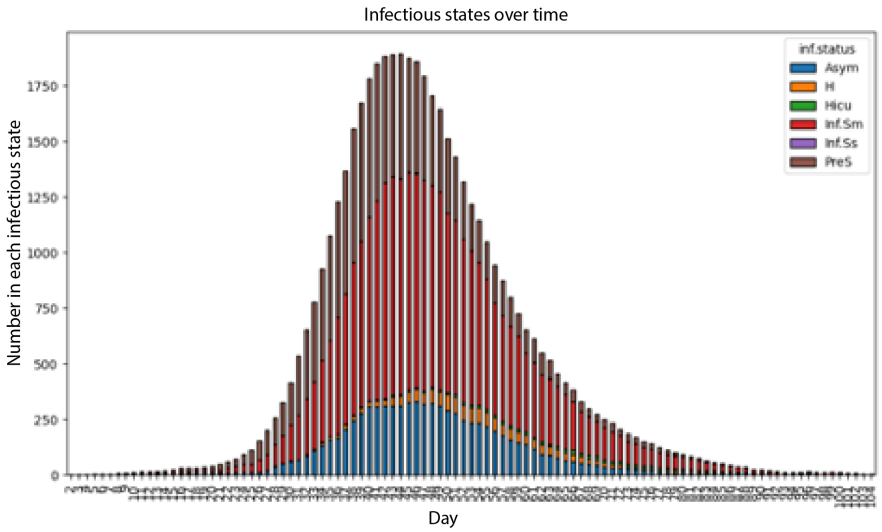

Figure 8: Epidemiological curves for no-masking behavior in a synthetic population network of Portland.

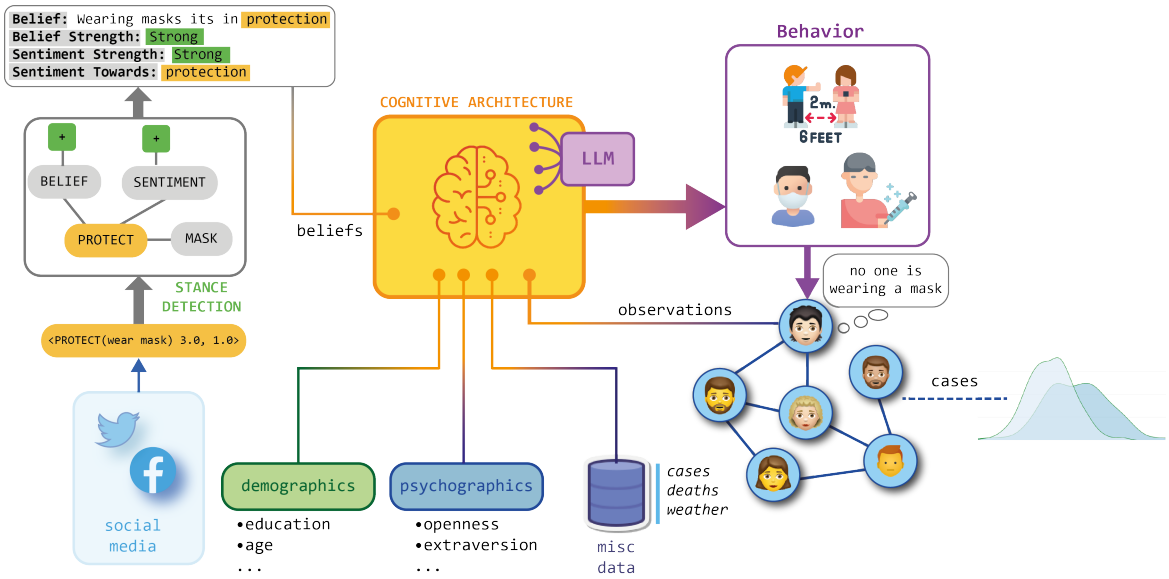

Figure 9: PVGA architecture simulating human behavior in an epidemiological network.

efficient use of modern tools for memory storage, retrieval, and inference, such as vector databases, approximate kNN methods, and decision trees.

- **Versatility in Learning Approaches**: At a high level, CogIBL supports both supervised and reinforcement learning. This facilitates diverse modeling strategies. Agents can be designed with pre-existing biases in their memories, or they can be structured to evolve their memories over time. The utility-based decision-making framework of CogIBL allows agents to factor in individual incentives and preferences. It is even possible to model joint utility functions that encapsulate community well-being or to represent attitudes as the expected values of behaviors, influenced by underlying beliefs (Pirolli, Lebiere, and Orr 2022).
- **Prioritized Experience Activation**: The activation computation mechanism in CogIBL offers flexibility in how agents prioritize their experiences. This can be tailored further by integrating components that mirror other facets of an agent's personality. For instance, agents can discount activation based on the experiences or actions of neighboring agents.
- **Stance-Informed Personalities**: Domain-specific stances extracted from real-world data help feed the initial personality of the agents. Incorporating belief into PVGA provides an additional layer of realism to the

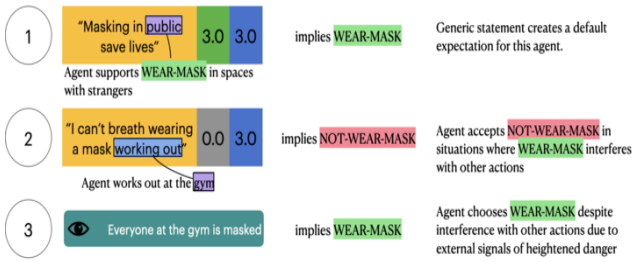

Figure 10: Stances, observations, and LLM in PVGA

interactions and actions that occur within the simulation. Furthermore our rapid domain adaptation techniques for stance detection allow for a generalizable PVGA framework.

- **LLM-Driven Agent Interactions**: LLMs make realistic discourse possible by incorporating various agent attributes (i.e. stances, memory, environment events) when conversing textually with other agents; the conversational output can then be observed for psychological and social science experiences. An added benefit is the generative nature of LLMs, which reduces the need for large amounts of data to simulate discourse.
- **Theory of Mind component**: Beyond initializing agents with real-world personal-level beliefs, stance detection can also equip agents with Theory of Mind (ToM) capabilities. This enables agents to reason about the beliefs and stances of other agents and incorporate that understanding into their own decision-making processes. That is, stance detection allows agents to form meta-representations of other agents' mental states, supporting the emergence of higher-order social cognition within agent populations.

Discussion

Integration of LLMs with ABM, CogIBL, and Stance Detection is a paradigm shift in computational social sciences. Unlike typical models in psychological and social sciences, which rely heavily on verbal theories and conceptual models, PVGA is a computational framework that combines previously separate formal models and methods to enable more holistic and realistic behavioral simulations. This enables scientific explanation and prediction beyond what is possible with conventional social science models.

One of the key challenges in population-scale modeling has been capturing the dynamic, often unpredictable nature of human behavior, for instance in the face of evolving epidemics. The integration of LLMs enriches the ABMs with human-like linguistic outputs incorporating factors such as underlying attitudes, beliefs, sentiments and stances, which guide how different individuals interact with each other as well as the individuals behavior in the real world.

Although LLMs have acquired "knowledge" about the world through training on massive text corpora, and can emulate human linguistic patterns, it is imperative to recognize their limitations. It is likely that they do not cover the entire spectrum of human cognition and decision-making intricacies. When combined with cognitive architectures, they can provide an easily accessible, more economical and yet safer basis for social and psychological investigations.

Introducing PVGAs has the potential to advance computational modeling in social science research. With their ability to model complex individual and collective dynamics, providing data-driven high-fidelity socio-psychological profiles, they promise a more in-depth exploration of emergent social phenomena. The adaptability and depth they offer could potentially change how we approach and understand intricate social systems.

However, challenges remain. While the integration of LLMs with cognitive architectures promises richer, more realistic simulations, researchers must exercise caution. Consistent assessment and validation of the output of these models against real-world data are crucial to ensure that the simulations remain grounded in reality. Recent work by (Romero et al. 2023) explores integration approaches for LLMs and cognitive architectures, proposing modular, agency, and neuro-symbolic models leading to more robust AI systems.

In conclusion, this paper supports new paradigms for testing longstanding hypotheses in social science research, introducing tools and methodologies to support generation of realistic, data-driven human-like behavior in ABM simulations. Our position is that PVGAs open up new possibilities for computational social science research through highly customizable simulations that can model complex individual and collective dynamics across a variety of domains. Going forward, PVGAs have the potential to become an indispensable tool for social scientists seeking to study emergent social phenomena.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under contract No. HR001121C0186 and the National Science Foundation (NSF) under award No.-2200112. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA) or the National Science Foundation (NSF). We are grateful to R. Vardavas and L. Baker from RAND Corporation for providing the epidemiological network used in the simulations along with the transmission dynamics. We thank Anton Gollwitzer for providing fused county level data, and Samuel Gosling and Tobias Ebert for regional Big Five data presented in (Ebert et al. 2022).

References

Anderson, J. R.; Bothell, D.; Byrne, M. D.; Douglass, S.; Lebiere, C.; and Qin, Y. 2004. An integrated theory of the mind. *Psychological review*, 111(4): 1036.

Axelrod, R. 1997. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press.

Binz, M.; and Schulz, E. 2023. Using cognitive psychology to understand GPT-3. *Proc Natl Acad Sci U S A*, 120(6): e2218523120.

Ebert, T.; Gebauer, J. E.; Brenner, T.; Bleidorn, W.; Gosling, S. D.; Potter, J.; and Rentfrow, P. J. 2022. Are regional differences in psychological characteristics and their correlates robust? Applying spatial-analysis techniques to examine regional variation in personality. *Perspectives on Psychological Science*, 17(2): 407–441.

Epstein, J. M.; and Axtell, R. 1996. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.

Eubank, S.; Guclu, H.; Anil Kumar, V.; Marathe, M. V.; Srinivasan, A.; Toroczkai, Z.; and Wang, N. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988): 180–184.

Gonzalez, C.; Lerch, J. F.; and Lebiere, C. 2003. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4): 591–635.

Grossmann, I.; Feinberg, M.; Parker, D. C.; Christakis, N. A.; Tetlock, P. E.; and Cunningham, W. A. 2023. AI and the transformation of social science research. *Science*, 380(6650): 1108–1109.

Lebiere, C.; and Wallach, D. 1999. Implicit and explicit learning in a hybrid architecture of cognition. *Behavioral and Brain Sciences*, 22(5): 772–773.

Lebiere, C.; Wallach, D.; and Taatgen, N. 1998. Implicit and explicit learning in ACT-R. In *Proceedings of the second European conference on cognitive modelling*, 183–189. Nottingham University Press Nottingham.

Mather, B.; Dorr, B.; Rambow, O.; and Strzalkowski, T. 2021. *A general framework for domain-specialization of stance detection*.

Mitsopoulos, K.; Somers, S.; Schooler, J.; Lebiere, C.; Pirolli, P.; and Thomson, R. 2022. Toward a psychology of deep reinforcement learning agents using a cognitive architecture. *Topics in Cognitive Science*, 14(4): 756–779.

Mitsopoulos, K.; Somers, S.; Thomson, R.; and Lebiere, C. 2020. Cognitive Architectures for introspecting Deep Reinforcement Learning agents. In *Workshop on Bridging AI and Cognitive Science, at the 8th International Conference on Learning Representations*. ICLR.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Ringel Morris, M.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *arXiv e-prints*, arXiv:2304.03442.

Pirolli, P.; Bhatia, A.; Mitsopoulos, K.; Lebiere, C.; and Orr, M. 2020. *Cognitive modeling for computational epidemiology*. Washington, DC: Springer.

Pirolli, P.; Carley, K. M.; Dalton, A.; Dorr, B. J.; Lebiere, C.; Martin, M. K.; Mather, B.; Mitsopoulos, K.; Orr, M.; and Strzalkowski, T. 2021. *Mining Online Social Media to Drive Psychologically Valid Agent Models of Regional Covid-19 Mask Wearing*, 46–56. Cham: Springer International Publishing. ISBN 978-3-030-80387-2.

Pirolli, P.; Lebiere, C.; and Orr, M. 2022. A computational cognitive model of behaviors and decisions that modulate pandemic transmission: Expectancy-value, attitudes, self-efficacy, and motivational intensity. *Front Psychol*, 13: 981983.

Reynolds, C. W. 1987. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 25–34.

Romero, O. J.; Zimmerman, J.; Steinfeld, A.; and Tomasic, A. 2023. Synergistic Integration of Large Language Models and Cognitive Architectures for Robust AI: An Exploratory Analysis. In *Proceedings of the 2023 AAAI Fall Symposium Series*. AAAI Press.

Shiffrin, R.; and Mitchell, M. 2023. Probing the psychology of AI models. *Proc Natl Acad Sci U S A*, 120(10): e2300963120.

Somers, S.; Mitsopoulos, K.; Lebiere, C.; and Thomson, R. 2019. Cognitive-level salience for explainable artificial intelligence. In *Proceedings of the 17th Annual Meeting of the International conference on Cognitive Modeling*.

Wallach, D.; and Lebiere, C. 2003. Conscious and unconscious knowledge: Mapping to the symbolic and subsymbolic levels of a hybrid architecture. In Jiménez, L., ed., *Attention and Implicit Learning*, 215–250. Amsterdam, Netherlands: John Benjamins Publishing Company.

West, R.; Michie, S.; Rubin, G. J.; and Amlôt, R. 2020. Applying principles of behaviour change to reduce SARS-CoV-2 transmission. *Nature Human Behaviour*, 4(5): 451–459.

Williams, R.; Hosseinichimeh, N.; Majumdar, A.; and Ghaffarzadegan, N. 2023. Epidemic Modeling with Generative Agents. *arXiv preprint arXiv:2307.04986*.