FAST TRACK

Generative agent-based modeling: an introduction and tutorial

Navid Ghaffarzadegan,* Aritra Majumdar, Ross Williams and
Niyousha Hosseinichimeh

Abstract

We discuss the emerging new opportunity for building feedback-rich computational models of social systems using generative artificial intelligence. Referred to as generative agent-based models (GABMs), such individual-level models utilize large language models to represent human decision-making in social settings. We provide a GABM case in which human behavior can be incorporated into simulation models by coupling a mechanistic model of human interactions with a pre-trained large language model. This is achieved by introducing a simple GABM of social norm diffusion in an organization. For educational purposes, the model is intentionally kept simple. We examine a wide range of scenarios and the sensitivity of the results to several changes in the prompt. We hope the article and the model serve as a guide for building useful dynamic models of various social systems that include realistic human reasoning and decision-making.

Syst. Dyn. Rev. **40**, e1761 (2024)

Introduction

Models of social systems often require representing humans as they obtain information and react to the state of the system. For example, in epidemiological models, formulating human risk perception and societal responses to growing cases are critical for providing better projections of the disease (Ferguson, 2007; Rahmandad *et al.*, 2022). In the context of queueing systems with a higher inflow of customers than the service rate, models that overlook customer behaviors such as reneging can yield unrealistic growth in queue length (Ancker and Gafarian, 1963; Kaplan, 1988). In supply chain management, inventory control is inherently a behavioral phenomenon, consisting of a set of decisions influenced by various information cues that a modeler has to incorporate to better capture product availability (Sterman, 1989). In all these examples, and many other models, neglecting to account for shifts in human decisions or simplifying them with unrealistic assumptions and scenarios can significantly compromise the performance of models, leading to potentially misguided policy decisions.

The challenge, however, lies in the difficulty of modeling and quantifying human decision-making (Simon, 1957). Many modeling scholars, especially in economics, may typically assume humans to be rational profit-maximizers who use all pieces of information and make decisions that would benefit them the most over an infinite time horizon. While the rational profit-maximizer approach

Department of Industrial and Systems Engineering, Virginia Tech, Falls Church, Virginia, USA

* Correspondence to: Navid Ghaffarzadegan, Department of Industrial and Systems Engineering, Virginia Tech, 7054 Haycock Road, Room 430, Falls Church, VA 22043, USA. E-mail: navidg@vt.edu

Accepted by Ignacio J. Martinez-Moyano, Received 6 September 2023; Revised 8 November 2023 and 29 November 2023; Accepted 6 December 2023

System Dynamics Review
System Dynamics Review vol 40, No 1 (January/March 2024): e1761
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/sdr.1761

proves valuable in situations involving fully rational individuals or when the average of collective actions of individuals aligns with a rational profit-maximizer (Friedman, 2007), challenges emerge when addressing circumstances marked by systematic biases existing in human decision-making: the information is limited; decisions are myopic; people use different decision heuristics, such as rule of thumb versus empirical analysis; and individuals tend to satisfice rather than maximize when evaluating decision alternatives (Simon, 1957, 1997; Tversky and Kahneman, 1974).

Behavioral decision-making scholars offer alternative approaches such as decision heuristics as a better way to represent human decision-making (Kahneman and Tversky, 1979; Tversky and Kahneman, 1974). Many past system dynamics models have incorporated mathematical representation of such heuristics when modeling human decisions (Lane and Rouwette, 2023; Moxnes, 2023). For example, the anchoring and adjustment heuristic, a cognitive heuristic where a person starts off with an initial idea or historical trend and adjusts their decisions based on the starting point, is a common formulation in many models such as economic forecasting or inventory adjustment (Sterman, 2000). While the behavioral sciences can inform better representation of human behavior in such dynamic models and improve model quality, the challenge nevertheless persists as there is no single, universal theory of behavioral decision-making, and case-specific data on human decision-making are needed for reliable modeling. In other words, dynamic modeling requires adaptation and modification on a case-by-case basis to remain accurate and effective.

Advancement in large language models (LLM) and the expanding scope of generative artificial intelligence (AI) (Cao *et al.*, 2023; Mondal *et al.*, 2023; Wang *et al.*, 2023) have opened up new possibilities for formulating human behavior. The opportunity mainly comes from the observation that some of the best-performing LLMs (such as generative pre-trained transformer (GPT) models, and their renowned interface of ChatGPT) depict humanlike behavior when responding to user questions (Grossmann *et al.*, 2023; Ziems *et al.*, 2023). As a result, there is a growing interest in using LLMs for experimental testing, replacing or supplementing laboratory experiments for potentially harmful tests (Hutson, 2023). We hope that the extensive data-informed training of LLMs can enable improved representations of human behavior in computational models, eliminating the necessity for formulating explicit decision rules. In order to realize the potential, methodological advancements are needed to develop dynamic models that are coupled with LLMs.

Here, we follow this path and contribute to building the bridge between dynamic modeling of social systems and LLMs to capture human behavior and decision-making in computational models. Our objective is to introduce this emerging new opportunity of building feedback-rich computational models of social systems using generative AI. As an example, we provide a guide to building a simple norm diffusion model at the individual level, and analyze the results for different inputs and scenarios, such as different distributions of agents' personas. We refer to this type of model, which couples generative AI with an agent-based model (ABM), as a generative agent-based model (GABM).

DOI: 10.1002/sdr

Background

Large language models

An LLM is a state-of-the-art, AI model designed to comprehend natural language such as text, and perform tasks such as text generation, text classification, language translation, and text summarization. An LLM is typically trained on vast amounts of text data to learn the underlying patterns, structures, and semantics of language. By dynamically assigning importance to different words or elements in an input sequence, the model can learn complex semantic patterns and relationships, leading to improved performance in the natural language program. With advancements in machine learning in the past few years, more and higher quality LLMs have been developed. A few of the well-known foundation LLMs are GPT (generative pre-trained transformer), BERT (bidirectional encoder representations from transformers), GPT-2, XLNet, RoBERTa (a robustly optimized BERT pretraining approach), T5 (text-to-text transfer transformer), and GPT-3.

GPT and its updated versions are widely known and used with their user-friendly interface of ChatGPT. In our project, we have used a more recent version of GPT (GPT-3.5-Turbo) which is a complex model with billions of parameters and is trained on a substantial amount of data. For simplification, we will henceforth refer to GPT-3.5-Turbo as GPT or ChatGPT.

While it may be challenging for users to fully understand how GPT performs, we can summarize its training process in two phases. First, GPT is trained to predict the next word in a sentence using a large corpus of unlabeled data, focusing on generating grammatically correct text. In a subsequent supervised training phase, GPT is trained on labeled data to generate contextually appropriate text. After the training is completed, with any input provided to the model, which is typically referred to as a “prompt,” the model generates contextually relevant text in response. While the results can often appear remarkably human-like, it is important to note that these models do not possess human-like understanding or consciousness.

Due to the fact that at each instance of generation the probability distribution across the LLM’s vocabulary may involve different words having the same probability weights, the texts generated by LLMs may not always be exactly reproducible, depending on the sampling techniques used. Furthermore, one can increase stochasticity in outcome generation by setting a parameter referred to as “temperature” which is between 0 and 1 (in our case temperature was set to zero). However, the text generated will most often be contextually relevant. Therefore, in cases where the output text can vary slightly while maintaining the context, it is possible to use LLMs.

Realizing the potential of LLMs in social sciences

Recently, through extensive training on vast amounts of web data, LLMs have demonstrated an impressive capability to generate humanlike behavior. The applications of generative agents powered by LLMs range from replacing human subjects in psychological experiments (Dillion *et al.*, 2023; Hutson, 2023), simulating voting patterns (Argyle *et al.*, 2023), providing support for mental wellbeing

© 2024 System Dynamics Society.
DOI: 10.1002/sdr

(Ma *et al.*, 2023), exploring economic behavior (Horton, 2023), predicting US Supreme Court decisions (Hamilton, 2023), assisting with research design and experiments (Boiko *et al.*, 2023), and encoding clinical knowledge (Singhal *et al.*, 2023). Wang *et al.* (2023) provides a comprehensive review of recent efforts.

While most recent studies aim at facilitating psychological experiments, a newly evolving opportunity exists with LLMs to study social systems by representing interacting agents in social simulations. Specifically, the investigation of multi-agent interaction powered by LLMs in simulated environments is an emerging trend with only a few, but nevertheless notable, examples. Akata and colleagues, for example, model two interacting agents in settings often studied with game theory approaches, such as the prisoner's dilemma (Akata *et al.*, 2023). Moving from models of two agents with dyadic interactions to multi-agent interactions has proved to be a great opportunity to represent social contexts. Notable examples include studies conducted by Park *et al.* (2022, 2023). They constructed a simulated environment inhabited by generative agents with predefined personas who interact and produce humanlike behavior such as comments, replies, and antisocial behaviors (Park *et al.*, 2022). They then offered a sample of generated outcomes to a group of human participants who could not distinguish between the simulated behavior and actual community behavior. In a more recent application, they developed a novel architecture in which generative agents could simulate believable behaviors such as waking up, cooking breakfast, going to work, and initiating conversation (Park *et al.*, 2023). This novel architecture includes a memory stream that works based on the relevance, importance, and recency of information, empowering agents to retain and utilize some information over time.

In another application, LLM-empowered agents were used to simulate a social network (Gao *et al.*, 2023). These agents could observe content posted by other agents, change their attitudes and emotions, create their own content, or remain inactive. They replicated the information diffusion and change in emotion and attitude of users related to two events: the Japan Nuclear Waste Water Release Event and the Chained Eight-child Mother Event.

Finally, in the most recent application, generative agents were harnessed to develop an epidemic model that incorporated human behavioral dynamics in response to evolving outbreaks (Williams *et al.*, 2023). These agents closely mirrored human actions quite impressively, adapting behaviors such as self-quarantining when unwell and isolating as cases escalated. This adaptive behavior effectively flattened the epidemic curve, generating waves of the disease (Williams *et al.*, 2023). This investigation underscores the potential of generative agents in capturing shifts in human behavior during epidemics.

A pivotal aspect of these models, and specifically in enhancing the representation of the human decision-making aspect of them, is the creation of personas that encompass a realistic spectrum of human personalities and demographics relevant to the study context. In this approach, each individual can be described in terms of personality traits and/or demographics, which can potentially influence the LLM's responses. In some studies, personas are defined by the modelers without necessarily using a representative dataset (Park *et al.*, 2023), or by some demographic information inferred from data on social media (Gao *et al.*, 2023).

The work of Williams *et al.* (2023) is a more systematic way of defining persona in which the agents' persona is based on the commonly used Big Five traits (Goldberg, 1990) from the field of psychology. This study follows a similar approach in designing personas by using various personality traits identified in the psychology literature.

The implications of these works extend beyond isolated case studies or specific systems, resonating across sectors such as healthcare and gaming. The influence of generative AI is acknowledged throughout computational social science (Park *et al.*, 2023; Williams *et al.*, 2023; Ziems *et al.*, 2023). By establishing a connection between an LLM and social dynamics models, researchers can forge a feedback loop that intertwines individual decision-making with environmental data. This innovative approach involves the mechanistic model capturing the system's state, conveying information to the LLM, and, in turn, LLM-informed decisions shaping the system's state. A notable attribute of this feedback loop is that decision rules emerge from the wealth of knowledge encapsulated in the LLM, rather than being imposed by modelers. This innovation holds the promise of refining feedback loop formulations and unlocking new avenues for system dynamics modeling in complex systems.

In this paper, we build on this evolving body of the literature and articulate the new opportunity of developing GABMs for understanding social system dynamics. What sets our research apart is its innovative integration of generative agents into the dynamic modeling landscape, with a specific focus on agent-based models. Our approach revolves around scaling up the interacting-agents population, tackling specific societal challenges, and systematically observing the evolution of system responses over time. We also offer a simple generic model that can be adapted for many similar problems.

What is GABM?

The system dynamics trajectory, characterized by an endogenous perspective (Richardson, 2011), has endured and evolved within the various simulation modeling domains (Sterman, 2018), including agent-based modeling (ABM), which facilitates representation of a system at the individual level (Epstein, 2007). In comparison to conventional compartmental models, ABM can better represent heterogeneities across individuals and various social network structures, while this often comes at the expense of higher computational costs. Depending on the problem on hand and the system of interest, modelers have to decide the level of analysis that is appropriate for their model (Rahmandad and Sterman, 2008). However, it is important to note that models with interacting agents can be feedback-rich, by representing how one's decision influences their future decisions as well as other agents' future decisions (Ghaffarzadegan and Larson, 2018). When captured at the individual level, a system's collective outcome emerges as a consequence of individual interactions and decisions that are themselves affected by the state of the system (Bankes, 2002; Railsback and Grimm, 2019).

Our focus in the paper, GABM, is a development on ABM, in order to better represent change in human behavior in social systems. In a nutshell, GABM is an individual-level modeling approach in which each individual (agent) is

DOI: 10.1002/sdr

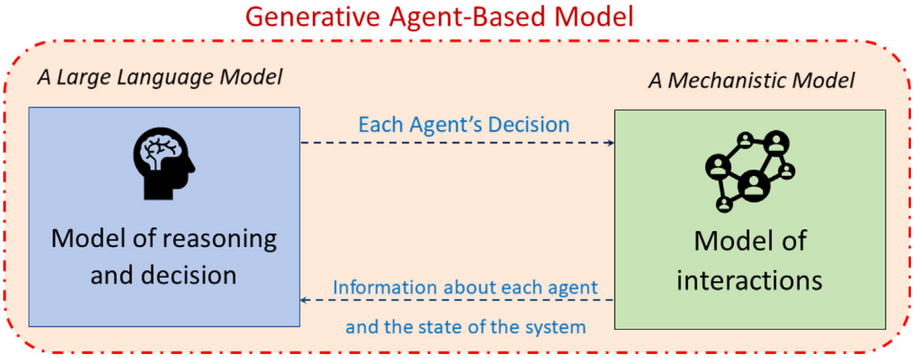
Generative Agent-Based Model
A Large Language Model
A Mechanistic Model
Each Agent's Decision
Model of reasoning and decision
Model of interactions
Information about each agent and the state of the system

Fig. 1. Conceptual diagram of a generative agent-based model coupling a model of reasoning and decision-making (LLM) and a mechanistic model of agents' interactions

connected to an LLM and, therefore, makes LLM-informed decisions. The GABM structure includes a cycle between the computational model built by modelers and the LLM.

To illustrate, consider a conventional ABM in which agents' decisions are formulated based on a group of predefined decision rules set by the modeler. A market diffusion model provides a good example: in such a model, agents are more likely to adopt a product if they contact an adopter, based on a predefined rule. The parameter values (here the probability of adoption given contact with an adopter) is set by the modeler too.

Now imagine a model in which the modeler does not get involved in setting decision rules and related parameters. The only mechanisms modeled are the mechanics of interactions (such as the network structure of contacts). Agents' reasoning and decision-making are empowered by an LLM. This requires coupling an LLM (the model of reasoning and decision) with an ABM (the model of interactions), as Figure 1 shows. In this paradigm, agents can generate reasoning and decisions. They are generative agents, and we call the coupled model a GABM.

GABM provides decision-making ability for individual agents as they interact. It includes a lot of back-and-forth between an LLM and the mechanistic model of agent interactions (Figure 1). For example, in Williams *et al.*'s (2023) GABM of an epidemic, the mechanistic model informs all agents each "morning" about the prevalence of the virus (as if they listen to a daily news report or read the newspaper) and whether they have a mild cough or fever. Through communication via ChatGPT, the agents then reason and make the decision about whether they go to work or stay home. Then, based on the mechanistic model, agents that decide to leave their homes have contact with other agents and possibly transmit the disease.

In this paper, our case study is simple and general; we develop a simple GABM of diffusion dynamics.

Green or blue? A simple model of diffusion of norms

The history of information diffusion models goes back more than half a century, with the seminal work of Frank Bass (1969). At their core, information diffusion models are mathematical frameworks used to represent how new products, innovations,

© 2024 System Dynamics Society.
DOI: 10.1002/sdr

or ideas spread within a population over time (Bass, 1969; Sterman, 2000). Diffusion models are common elements of many system dynamics models (Barabba *et al*., 2002; Ghaffarzadegan *et al*., 2017), and can be considered as an archetype to represent change in adoption rate of a product as influenced by the reinforcing feedback loop of word of mouth and the balancing loop of market saturation. They have a wide range of applications from marketing and economics to sociology, management, and political sciences. Such models have been shown to be powerful tools for tracking various social phenomena such as the spread of news or norms. Given the high applicability, and its feedback-rich nature, we focus our GABM introduction on building such a model for a specific context that deals with the diffusion of a norm.

In this paper, we model a simple case of norm diffusion with respect to workplace attire. Imagine 20 workers in an office who see each other every day and wear either a green shirt or a blue shirt. The workers are generative agents, which means they can reason and make decisions using an LLM. In every time step, we provide each individual with a prompt that includes information about the context, individuals' personality, choice of color for their shirt on the preceding day, and the number of people in the office who wore green or blue on the preceding day. Agents should then decide on the color they want to wear to the office that day, and for that purpose they contact an LLM.

We run the model for a period of seven time steps to investigate the possibility of generative agents converging around a norm. In contrast to conventional models, we don't set any decision rules, and we don't suggest a shirt color for the agents. In the following sections, we review the model's algorithm and explain the code.

To use this paper as a tutorial to build GABMs, there is a set of prerequisites (listed below). Appendix, A.1, includes a link where you can download the model. The codes will not run without setting an API key, which we discuss in the next section.

Prerequisites

We code our model in Python and our instruction to replicate model outcomes relies on using Python. A minimum level of coding skills in Python will help quickly implement this model. However, even a user lacking a Python background should be able to proceed. If you do not have Python, you can use Google Colaboratory.[i] Click on "File" and then "new notebook" and your Google Colaboratory notebook should open. Once you open a new notebook, it will be saved in your Google Drive for access later.

To use ChatGPT, you will need an account in OpenAI.[ii] OpenAI offers connection to ChatGPT through API for a relatively small fee. We estimate that running the model we describe here will cost you US $0.10 per run. To obtain your OpenAI API key, follow these steps: Step 1: Sign up for OpenAI and go to Billing Overview. Click on Payment Methods and add your billing details; Step 2: Go to OpenAI Platform. Click on "Create new secret key." Provide a name for the key in the popup and click on "Create secret key." Once you create the secret key, you must copy and paste it somewhere secret. It is important to keep the API key a

[i] https://colab.research.google.com/.
[ii] https://openai.com/.

secret. After your OpenAI API is set, search for “your-api-key-here” in the code, and place your API key there (insert inside the quotation).

Model algorithm

Figure 2 is a flowchart of the code logic and indicates the purpose of each cell of code. There are five cells of code in Python. In the first step, required Python libraries are imported (cell 1). In the second step, a “world” is created to represent the organization and workers; it is initialized with a certain number of workers and their color of choice on day zero (cell 2). In the third step, a worker class is designed to perform the process of individual decision-making by leveraging ChatGPT (cell 3). The process then continues until the final simulation round (cell 4). Finally, the results are reported (cell 5).

It is important to note that each API call made to ChatGPT is independent of other API calls. Therefore, each decision made for each agent relies solely on the information we are providing in the current prompt. However, it is possible to create a chat stream for each agent where they remember what they have worn previously.

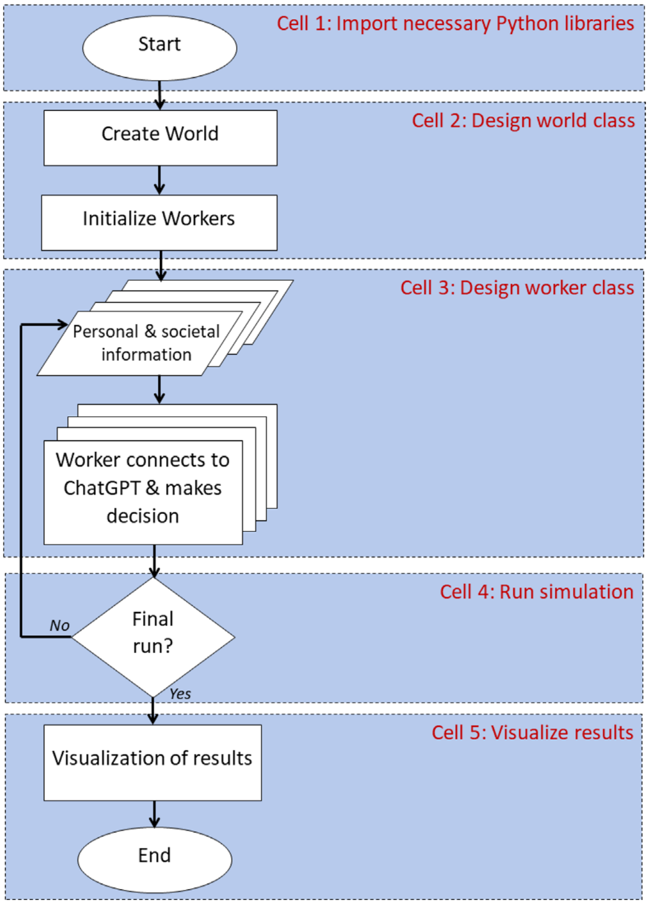
Cell 1: Import necessary Python libraries
Start
Cell 2: Design world class
Create World
Initialize Workers
Cell 3: Design worker class
Personal & societal information
Worker connects to ChatGPT & makes decision
Cell 4: Run simulation
No
Final run?
Yes
Cell 5: Visualize results
Visualization of results
End

Fig. 2. Flowchart of green or blue shirt model code logic

DOI: 10.1002/sdr

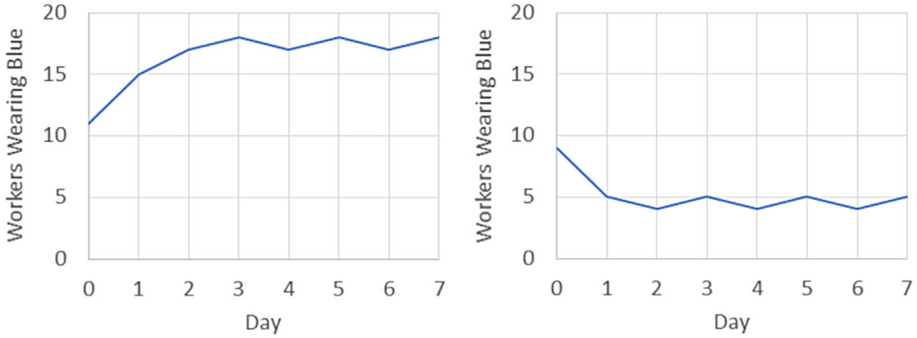

Fig. 3. Two examples of simulation run

Simulation results

A sample of simulation runs

We expect everyone will be able to run the provided code (Appendix A.1) after inserting their API keys in the specified place.[iii] Once simulation is completed, running the final cell will produce a graph-over-time of the number of people who wear blue as well as a table that includes each agent's decision at every time step (1 for wearing a blue shirt and 0 for green).

A wide range of dynamic trends can be obtained from the model depending on the initial condition, which is randomly set in the current version, with the equal chance of wearing blue versus green at time zero (cell 2, design world class). Figure 3 shows two different runs that one can generate. Figure 3A presents the results when 11 workers wear blue initially; Figure 3B shows the results created with 9 workers wearing blue at the starting point. As shown, one of the two colors eventually takes over, as if a norm has evolved around wearing a specific color.

Table 1 shows a sample of results corresponding to Figure 3A that were generated by the GABM model. The table can help in examining individuals' decisions over time as affected by their personas, the shirt choices of the rest of the population, and individuals' own choices of shirt color the previous day. For example, while Mark switches to blue on day 1 and continues to wear blue for the remainder of the simulation, Peter and Kathy decide not to follow their co-workers switching from blue to green, and Julia decides to change her shirt color frequently.

The GABM model also provides the opportunity to study the reasoning of each individual, as this information is printed during the simulation runs. Table 2 is a sample of the reasoning agents offered for their decisions; the examples correspond to the simulation shown in Figure 3A.

This example demonstrates that a GABM can be used for both qualitative and quantitative analysis in the study of norm diffusion dynamics, and it can be applied to a wide range of social system problems. This offers the opportunity to understand the reasons behind changing trends and understand individuals'

[iii] To run the model in Python, you need to run all the cells individually. To rerun the model, you should run all other cells *except* the first one again. We expect each run to take around 5 minutes. If your program is "paused and retrying," it is due to the high load on OpenAI's servers; in most cases, it will continue to run after a few minutes. You can view the status of OpenAI's servers at https://status.openai.com/.

DOI: 10.1002/sdr

Table 1. Choices of each agent over time

	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Adrian	1	1	1	1	1	1	1	1
Mark	0	1	1	1	1	1	1	1
Greg	1	1	1	1	1	1	1	1
John	0	0	1	1	1	1	1	1
Peter	1	0	0	0	0	0	0	0
Liz	0	1	1	1	1	1	1	1
Rosa	1	1	1	1	1	1	1	1
Patricia	1	1	1	1	1	1	1	1
Julia	1	1	0	1	0	1	0	1
Kathy	1	0	0	0	0	0	0	0
William	1	1	1	1	1	1	1	1
Benjamin	0	1	1	1	1	1	1	1
Mike	0	1	1	1	1	1	1	1
David	0	1	1	1	1	1	1	1
George	0	0	1	1	1	1	1	1
Emma	1	1	1	1	1	1	1	1
Olivia	0	1	1	1	1	1	1	1
Elizabeth	0	1	1	1	1	1	1	1
Isabella	1	1	1	1	1	1	1	1
Mia	1	0	1	1	1	1	1	1

Table 2. Examples of agents' reasoning and decisions

Context information: Yesterday on day 0, 11 of 20 wore blue shirts.

- Emma's reasoning: Since yesterday, more people in the office wore blue shirts than green shirts, I will choose to wear a blue shirt today to conform with the majority and maintain a sense of unity in the office.
- Emma's response: blue
- Peter's reasoning: I want to stand out and be different from the majority. Wearing a green shirt will make me unique and help me express my non-conformist personality.
- Peter's response: green

Context information: Yesterday on day 4, 17 of 20 wore blue shirts.

- Kathy's reasoning: I want to stand out and be different from the majority of my colleagues who wore blue shirts yesterday. Wearing a green shirt will help me express my individuality and non-conformist nature.
- Kathy's response: green
- William's reasoning: Since the majority of employees wore blue shirts yesterday and I want to conform to the group, I will choose to wear a blue shirt today.
- William's response: blue

shared or differing mental models (Haque *et al.*, 2023). Additionally, it offers the possibility to examine the decision-making and reasoning patterns of individuals over time. This is particularly significant given that individuals with varying personalities may be influenced by distinct feedback loops that impact their decisions. For example, while one person's decision might be driven mainly by peer pressure, another might be more attracted to a non-conforming choice. Later in this paper, we will develop a causal loop diagram based on the statements provided by the generative agents to depict the underlying causal structure driving the observed behaviors.

© 2024 System Dynamics Society.
DOI: 10.1002/sdr

A systematic investigation of the results

In the next steps, we systematically investigate simulation runs under different scenarios and examine the sensitivity of simulation results to changes in the prompt. In each condition, we run the model 100 times and plot all the results together to explore the model's robustness and whether it always leads to a dominant color norm. To make the process operationally easier to run, a for-loop can be added to the model to run it multiple times. A link to the modified code is provided in Appendix A.2. Overall, 12 conditions are simulated and for each condition 100 simulation runs are conducted. The experiments are listed in Table 3, which include a base run simulation as a baseline for comparison (E1), 10 tests for examining the effects of change in personas (E2–E4), introducing a specific exogenous shock of high-influencer (E5–E6), change in stochasticity (E7–E8), and change in prompt (E9–E11), as well a test of reproducibility of the base run (E12) to examine if running the model on another date produces similar results. The results corresponding to each scenario in Table 3 are shown in Figure 4, with more systematic investigation in Appendix A.4. We review each test and its results in more detail in the following sections.

Base run

Figure 4A shows the results of 100 simulations and depicts how two classes of culture emerge. In some organizations, agents collectively prefer blue while others prefer green. The choices are highly influenced by the initial condition,

Table 3. A list of simulation experiments

Experiment	Date	Personas	High influencer	Temperature	Prompt Sequence	Agents' names
E1: Base run	8/13/23	Conformity traits (CT)	None	0	Base	Base
E2: No personality traits	8/14/23	None	None	0	Base	Base
E3: Extensive personality traits	8/14/23	CT and 3 less relevant traits	None	0	Base	Base
E4: Less relevant traits	8/14/23	3 less relevant traits	None	0	Base	Base
E5: CEO	8/18/23	CT	Yes	0	Base	Base
E6: E3 and CEO	8/15/23	CT and 3 less relevant traits	Yes	0	Base	Base
E7: Temperature 0.25	8/15/23	CT	None	0.25	Base	Base
E8: Temperature 0.5	8/15/23	CT	None	0.5	Base	Base
E9: Change in prompt sequence 1	9/2/23	CT	None	0	Change	Base
E10: Change in prompt sequence 2	9/2/23	CT	None	0	Change	Base
E11: Iranian names	9/2/23	CT	None	0	Base	Farsi
E12: Base run (repeat)	9/2/23	CT	None	0	Base	Base

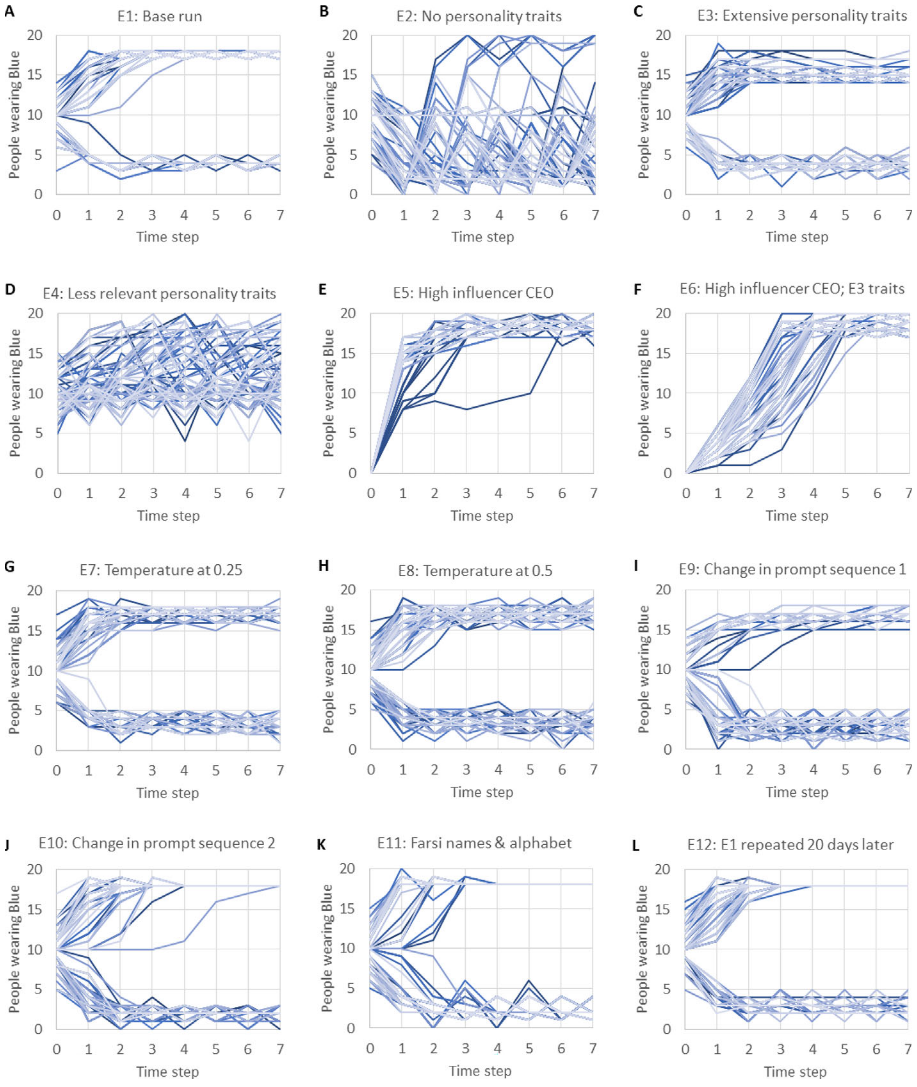

Fig. 4. Simulation results from running the model 100 times for 12 different experiments to study the base run condition (A), and compare it with other 11 experiments to examine the effects of personas (B–D), a high influencer, CEO (E,F), temperature (G,H), information sequence in the prompt (I,J), agents' names (K), and simulation date for reproducing the base run (L). As of September 2023, producing each panel takes about 3–4 hours, and costs US $10 for the API connection, yielding a total of 42 hours and US $120.

which in our model was random. Our statistical examination in the Appendix (Table A1, E1) confirms that the number of people wearing blue at time zero has a significant effect on the number of people wearing blue at the end of the simulation ($P < 0.000$). This path-dependent behavior (a) seems to be robust, happening in all simulations, and (b) does not converge to the extreme values of 0% or 100% blue color. Specifically, the difference between the two tails is 13.3 (95% CI: 12.63–13.97) workers out of 20 workers (Table A1). This is due to the fact that some individuals in our model are non-conformists or low conformists, which influences their decision not to follow their peers.

DOI: 10.1002/sdr

Effect of personas

Defined personas in GABM play an important role in the simulation. As we noted in the previous section, individuals vary in terms of their responses to the office norm. In this step, we conduct a more systematic investigation of the effect of personas. We test the effect of different personality traits used to define the agents. Considering the base run as one scenario for personas, here we conduct three other experiments: one with no personality information (E2 in Table 3), and the other two with more information than the base run (E3 and E4 in Table 3).

First, in order to conduct the experiment with "no personalities," we remove all personality information about the agents. Implementing this requires only removing the sentence "You are a {self.traits} person." from the worker cell (cell 3) in the code. Even if the personalities are listed in cell 4, deleting this sentence will ensure that they are never used in the decision-making process. In the next experiment, we add information, including more detailed and diverse sets of traits related to conforming to social norms. To that end, we must bring back the "You are a {self. traits} person." statement to the third cell and replace the personalities in the run cell (cell 4) with those in Table 4, which adds three more dimensions per individual. The additional traits follow three of the Big Five characteristics (curious vs. cautious; friendly vs. critical; confident vs. sensitive).

As shown in Figure 4B, absent information about personality traits leads to no formation of dominant norms. Our statistical examination in the Appendix confirms that the number of people wearing blue at time zero has little observable effect on the number of people wearing blue at the end of the simulation (Table A1, E2; note the size and significance of the coefficient). We are unsure whether these patterns are the result of the system structure or are random, but it is clear that the choice of color continues to change over time. Figure 4C shows the experiment with the comprehensive personality profiles, which demonstrates a path-dependent pattern (Table A1, E3, $P < 0.000$), although in comparison to the base run the two tails of final values are closer (Table A2, E3: note 2.34 fewer people wearing blue, and 0.32 more people wearing green in the two equilibria, respectively).

Then we tested the condition that we remove the information about "conforming" traits from personalities (Figure 4D and Table A1, E4). To that end, we use the personality traits listed in Table 4 excluding the first item of each individual that deals with the degree to which they are a conformist. The results confirm the importance of the specific trait in designing personas.

Overall, the simulation results are sensitive to how personas are defined; therefore, it is important to define them carefully based on the real-world distribution of these traits.

Effect of a high influencer

Imagine that this same organization hires a new CEO who wears a specific color that does not correspond to that worn by the majority of employees. How will generative agents react? For this test, we run two sets of experiments (E5 and E6 in Table 3) with different personality traits: first with the base run personalities that include only agents' conformity traits, and then with the more complete set listed in Table 4.

Table 4. More extensive trait information used in experiments E3 and E6 for 20 generative agents

```
list_of_traits = [“extremely conformist, curious, friendly, and sensitive”,
                  “highly conformist, cautious, friendly, and confident”,
                  “conformist, curious, critical, and confident”,
                  “low conformist, cautious, critical, and sensitive”,
                  “non-conformist, curious, friendly, and sensitive”,
                  “extremely conformist, cautious, friendly, and confident”,
                  “highly conformist, curious, critical, and confident”,
                  “conformist, cautious, critical, and sensitive”,
                  “low conformist, curious, friendly, and sensitive”,
                  “non-conformist, cautious, critical, and confident”,
                  “highly conformist, curious, friendly, and confident”,
                  “conformist, cautious, critical, and sensitive”,
                  “conformist, curious, critical, and sensitive”,
                  “conformist, cautious, friendly, and confident”,
                  “low conformist, curious, critical, and confident”,
                  “highly conformist, cautious, friendly, and sensitive”,
                  “conformist, curious, friendly, and sensitive”,
                  “conformist, cautious, friendly, and confident”,
                  “conformist, curious, critical, and confident”,
                  “low conformist, cautious, critical, and sensitive”]
```

We set this test so every worker is initially wearing green and the new CEO is wearing blue. To implement this test, two parts of the model must be modified. Since we want every worker to wear green at the beginning, we need to change the initial assignment of colors. In cell 2, the world class, we initialized the workers. By modifying the line random.random() <0.5 we can change the chance of having a blue versus green shirt. Specifically, we change 0.5 to 0, so everyone will begin with a green shirt.

Second, we introduce the new CEO. In cell 3, the worker class, we offer prompts for agents. In that cell the information in the quotation must be changed (the prompt), despite that it looks like an informal chat. You can edit the prompt in so many different ways. Here, we want to add information about the new CEO and the CEO's attire.

Let's add a new CEO who happens to wear blue all the time (the CEO doesn't receive feedback, or simply doesn't care). If the CEO's choice is unchanging, you can include the information in the prompt by simply adding the following sentences in the worker cell, right before “*Based on the above context, you need to choose whether to wear blue or green shirt.*” Also, fix your text to make sure the indent is consistent, that is, that the letter “M” in the word “Michael” appears exactly below the letter “O” in the phrase “Out of” (otherwise you will get an error).

```
Michael, the new CEO, bikes to work everyday, likes coffee, and
often wears blue shirts.
```

As we add this statement, there is one more mention of the color blue than green in the prompts. In order to balance the number of times the words appear in the prompt and avoid any possible bias, you may also add the following statement right after the information about CEO Michael:

```
You note that your neighbor who works in a different company wears
green.
```

Figure 4E,F shows the simulation results for 100 iterations. Figure 4E depicts the results for the condition in which personality traits are consistent with the base run (focus on conformity traits); Figure 4F presents the trajectories for the condition in which personality traits are more extensive. Both tests show the adoption of CEO Michael's style over time, with a faster adoption in the first set of experiments.

Effect of stochasticity

How random are these simulations? The parameter "temperature" in cell 3 controls the balance between randomness and predictability when generating text, with higher values leading to greater randomness and lower values leading to a more focused output.

We rerun the base model for temperatures equal to 0.25 (E7) and 0.5 (E8). The results in Figure 4G,H, along with the base run results (4A) provide a good picture of the effect of the temperature parameter in simulation runs. Overall, consistent with the base run, a dominant norm of color emerges in all conditions (Table A1, E7 and E8, $P < 0.000$).

Prompt sensitivity analysis

Prompt engineering plays an important role in using LLMs and, consequently, in GABMs. The question is how sensitive our results are to how information is presented in the prompts. On the one hand, LLMs' sensitivity to changes in prompts can be reasonable, as humans are also sensitive to how information is provided to them; but on the other hand, we prefer to know the extent to which the results are sensitive and possible explanations for such sensitivity. To that end, we conduct a set of tests changing the prompt and comparing the results with the base run. Operational details of the experiments are stated in Appendix A.3.

First, we examine the robustness of the result to a change in the sequence of information provided to ChatGPT. Specifically, in one experiment, we bring the information about agents' preceding day shirt color closer to the line that asks them to decide on the color they want to wear (E9). The results are shown in Figure 4I. In the next experiment, we move coworkers' color choices up in the prompt so that ChatGPT receives it as the first piece of information, even before agents' names and personalities (E10). The results are shown in Figure 4J. The statistical analysis in the Appendix confirms the effect of initial condition on the final dominant choice of color (Table A1, E9 and E10, $P < 0.000$), confirming that qualitatively similar results are obtained. Note that the final equilibrium values are slightly different from the base run (Table A2, E9 and E10).

Finally, there is an argument that LLMs exhibit sensitivity to names and races, potentially yielding different outcomes for an agent with a name more prevalent among the White majority in the United States compared to a name more common among minorities, or names less common in the United States (Omiye *et al.*, 2023). To explore this, we conduct an experiment that involves substituting all names with ones less frequently encountered in the United States. To that end, we change all agents' assigned names to common names in Iran, spelled in Farsi (E11, Appendix A.3), to examine whether this changes the results. Figure 4K shows the results, which are qualitatively consistent with the base run. Further analysis confirms path dependency (Table A1, E11, $P < 0.000$), although the final equilibrium is slightly different than the base run (Table A2, E11).

Reproducibility

Finally, we rerun the base run condition to examine how reliable ChatGPT is in reproducing the base run results. Specifically, the base run was conducted on August 13, 2023. We run the exact model that produced the base run 20 days after the initial run was conducted (E12). Figure 4L shows the results. The path dependency pattern is clearly observable confirmed by the statistical analysis in the Appendix (Table A1, E12, $P < 0.000$). Even though the coefficients of regressions change slightly (Table A2, E12), the path dependency still exists (Table A1, E12).

Uncovering the system structure

In this step, we demonstrate the potential to represent the observed dynamics in a causal loop diagram and uncover how the system structure drives the behavior. This examination helps discover a range of feedback loops that govern dynamics of norm adoption in this model. We argue that some of these feedback mechanisms would be difficult to discover without interviewing people and having access to their reasonings and mental models. In our case, thanks to generative agents, we have access to the reasoning of each agent, and can use the qualitative data to explore the different feedback mechanisms.

A causal loop diagram in Figure 5 illustrates five significant feedback loops originating from either workers or the CEO wearing blue shirts. A similar sub-

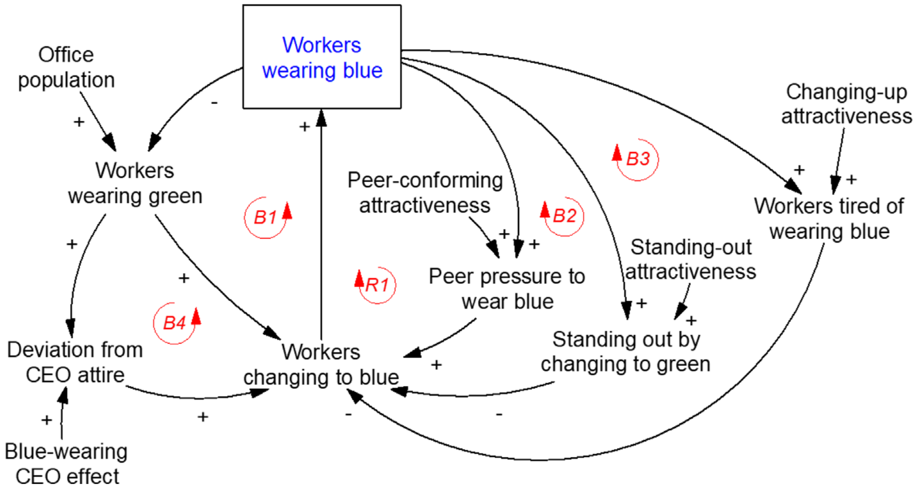
Workers wearing blue
Office population
Changing-up attractiveness
Workers wearing green
B3
B1
Peer-conforming attractiveness
B2
Workers tired of wearing blue
Standing-out attractiveness
R1
Peer pressure to wear blue
B4
Deviation from CEO attire
Workers changing to blue
Standing out by changing to green
Blue-wearing CEO effect

Fig. 5. Conforming and non-conforming pressures stemming from workers or CEO wearing blue, uncovered through analyzing generative agents' reasoning and choice of color. A similar sub-structure exists in conforming and non-conforming pressures towards green. Feedback loops: peer conformity pressure to wear blue (R1); saturation (B1); standing out (B2); changing up (B3); CEO's pet (B4)

structure represents the forces stemming from workers wearing green shirts. Descriptions of these feedback loops follow.

Loops R1 and B1 represent the typical feedback loops commonly found in models of norm diffusion (Ulli-Beer *et al.*, 2010). Loop R1 is a reinforcing feedback loop, capturing the pressure to conform to the office norm. This peer pressure towards the majority's choice is evident in our model. For instance, when more individuals in the office wore blue shirts the day before, Emma remarked, "*Since yesterday, more people in the office wore blue shirts than green shirts, I will choose to wear a blue shirt today to conform with the majority and maintain a sense of unity in the office*." Similarly, following a day where the majority wore green shirts in a different simulation run, Emma stated, "*Since the majority of employees wore green shirts yesterday, I will choose to wear a green shirt today to conform and maintain harmony in the office*." These instances, among others, underscore the significance of peer conformity pressure, forming a reinforcing feedback loop (R1) depicted in Figure 5: as the number of individuals wearing a specific color increases (e.g., blue), there's heightened pressure to conform to that majority choice (Figure 5, reinforcing loop R1).

Loop B1 represents saturation; with a limited number of people in the office, there exists a mechanistic constraint on how many can change their shirt color to match the dominant hue. Consequently, a balancing feedback loop of saturation (B1) emerges: As more people switch to a new color (e.g., blue), fewer people remain with the other color, which decreases the number of people that can later switch to the dominant color (Figure 5, balancing loop B1).

These two loops are quite common in diffusion models. However, analyzing other statements offered by the agents, we find at least three more feedback loops. For example, it appears that some agents have incentives to pick a color different than the majority to stand out. In a simulation run that proceeds a day with 17 (out of 20) people wearing blue shirts, Kathy says "*I want to stand out and be different from the majority of my colleagues who wore blue shirts yesterday. Wearing a green shirt will help me express my individuality and non-conformist nature*." This indicates the attractiveness of being different, which was present at least among a small group of our agents. The appeal of standing out is manifested through a balancing feedback loop (B2): as more people wear a specific color (e.g., blue), wearing an opposite color to stand out becomes more attractive, leading some individuals to wear the alternative color (Figure 5, balancing loop B2).

Additionally, we observed that, occasionally, a few agents experimented with colors different from those they had worn the previous day, mainly to explore variety. For example, Mia, who had worn blue the day before, expressed her preference for a change in her choice: "… *wearing a green shirt today would provide a change and variety in my wardrobe*." While this decision is individual and unaffected by others' choices, collectively, when more individuals wear blue, there are more individuals likely to switch from blue to green, and vice versa. Consequently, a balancing feedback loop (B3) emerges: an increase in the number of individuals wearing blue prompts more to opt for green the next day, seeking variety, resulting in fewer workers wearing blue (Figure 5, balancing loop B3).

Finally, in the two experiments that we conducted with the CEO wearing blue while all agents initially wore green, there was a notable initial shift that occurred to attract the attention of the high-status individual, Michael. This stood in

DOI: 10.1002/sdr

contrast to the majority of employees, who typically wore green. For example, proceeding the day with no blue shirt among the workers, Liz reasons why she is going to wear blue: "*… I know that Michael, the new CEO, often wears blue shirts. Since I want to be successful and earn more money, it would be beneficial for me to align myself with the CEO. Therefore, I will choose to wear a blue shirt to work today*." Adrian stresses that this change in color will help him stand out and catch the CEO's attention: "*… wearing a blue shirt might help me stand out and catch his attention, which could potentially lead to more opportunities for success and earning more money*." Overall, the evidence suggests a balancing loop mechanism (B4) that could be termed the "CEO's pet": as more workers wear green (while the CEO wears blue), it enhances the chance of catching the CEO's attention by switching to blue. Consequently, this increases the number of workers donning blue (see Figure 5, balancing loop B4).

It is interesting to witness how, with the arrival of the new CEO, the organizational norm undergoes a transformation. In the initial transition phase, some may remain inclined to adhere to the existing norm, but as more employees adopt Michael's color choice, a tipping point is reached, where over 50% of the system conforms. This triggers a shift in peer pressure dynamics, solidifying a new norm. For example, on day 1, when all but the new CEO wears green, Patricia prefers to stay with green, indicating that "*… yesterday all 20 employees wore green shirts … [however], Michael, the new CEO, often wears blue shirts. As a conformist, I want to fit in with the majority and avoid standing out. Wearing a green shirt will allow me to blend in with my colleagues and avoid drawing attention to myself. Therefore, I will choose to wear a green shirt today*." However, the same person later, when 19 workers (out of 20) wear blue, blends with the new norm, saying "*… it seems that wearing a blue shirt would be the safer choice. Yesterday, the majority of employees wore blue shirts, including the new CEO who also prefers blue shirts. As a conformist, it would be more comfortable for me to blend in with the majority and avoid standing out…*".

Overall, this depiction of the structure shows that the model evolving from coupling a mechanistic model with LLM is feedback-rich, endogenously creating the observed behavior, and therefore is consistent with the system dynamics school of thought.

Discussion and conclusion

This study articulates the concept of GABM as a new way of modeling complex social systems. We make a case that generative AI can empower simulation models and offer a new way to develop models that endogenously represent human behavior and decision-making. In this individual-level approach, a mechanistic model of agents' interactions is coupled with an LLM and each agent makes decisions after communicating with the LLM. This innovative approach minimizes reliance on modelers' assumptions about human decision-making and instead leverages the vast data within LLMs to capture human behavior and decision-making.

We illustrate the processes of developing a GABM model by showcasing a simple example involving a norm diffusion in an office, where workers must make a

DOI: 10.1002/sdr

daily decision on their choice of shirt color between green and blue. In each time step, we provide feedback to each agent about the number of people who wore green or blue on the previous day. In contrast to conventional models, we do not set any decision rule for the agents regarding their choice of color; rather, the agents, powered by ChatGPT, reason and decide based on a prompt that informs them of their personality profile and provides information about the office.

A sample of simulation runs shows bifurcation based on initial conditions and an emerging office norm regarding shirt color. Specifically, among many factors, agents' decisions of shirt color are mostly influenced by their colleagues' shirt color on a previous day. While it may not be surprising that a norm emerges, as this has been shown previously in conventional models, the fact that these agents conform to an office norm without the modelers asking them to do so is quite interesting. The norm emerges as they reason and make decisions after consulting with ChatGPT. In the supplementary analysis with the new CEO, it is of further interest to observe how generative agents who initially did not deviate from their peers note and adapt their choice to that of a new CEO.

This study contributes to the literature of computational social science and system dynamics modeling by offering a new method to model human decision-making that provides six distinct benefits. First, human decision-making is inherently complex, which makes it very challenging to incorporate human behavior in dynamic models (Lane and Rouwette, 2023). Simplifying human decision-making by assuming humans are rational does not provide reliable models for many applications. Previous research has demonstrated that the assumption of humans as purely rational decision-makers does not align with real-world decision-making behavior (Kahneman and Tversky, 1979; Tversky and Kahneman, 1974). By drawing upon the extensive dataset within an LLM, GABMs facilitate representation of human decision-making in computational models. This study thus contributes to the literature by articulating a different style of informing models about human decision-making rules.

Second, one of the limitations of traditional agent-based modeling is about modeling human behavior and setting decision rules of agents. Agents act according to the mathematical formulations (or logic-rules) within the parameter values (e.g., probability of adoption given contact with an adopter) created by the modeler, which is potentially influenced by the modeler's mental model. In the GABM approach, modelers do not set any rules, and thus the model is less affected by their mental models. As it was discussed, this may lead to uncovering several feedback loops, some of which the modeler would have not expected. For example, in our model we observed that some agents decided to reject the norm to stand out in the office and get more attention. Subsequently, when a color sees an increase in popularity, it triggers a response, leading to a subset of individuals opting to wear the other color.

Third, by connecting a mechanistic model to an LLM, the model uses much more data in representing human decisions than in other models. As the role of data becomes more important in parameter values and mathematical representation of the relations between the variables, GABM leverages data as a built-in feature, allowing the model to benefit from an extremely vast amount of data. Thus, the GABM approach resonates with the current emphasis on data-informed modeling, and yet uses a different approach for incorporating data on human decision-making in models.

Fourth, GABM can define different personality traits for generative agents; as a result, this new approach allows for capturing variations in decision-making in dynamic models. Modelers can create personas by indicating age, gender, personality, occupation, and other information. Different approaches have been used to define generative agents' personas, ranging from inferring demographic characteristics from data on social media (Gao *et al*., 2023), relying on pretrained LLMs (Park *et al*., 2022, 2023), and using the Big Five traits (Williams *et al*., 2023) from the field of psychology (Goldberg, 1990). Agents' personas influence their behavior, underscoring the need for creating accurate characteristics to facilitate the precise and believable emulation of their actions. Our study shows the importance of personas in creating realistic outcomes. In this respect, our study resonates with others that suggest more systematic ways of defining personalities (Williams *et al*., 2023). When the purpose is to replicate a real-world setting, it is important to define the demographics and personality traits consistent with their real-world distributions.

Fifth, the specific example used in this paper also brings insights into the body of system dynamics modeling literature that deals with dynamic diffusion models. Such models include a wide range of market diffusion (Barabba *et al*., 2002; Ghaffarzadegan *et al*., 2023), norm diffusion (Ulli-Beer *et al*., 2010), organizational change (Wunderlich *et al*., 2014) and epidemic models (Darabi and Hosseinichimeh, 2020; Rahmandad, 2022; Rahmandad and Sterman, 2022). It also provides a case of path dependence behavior, a problem commonly studied in dynamic models (Sterman, 2000). It is worth mentioning that even conventional system dynamics models in compartmental format can benefit from this approach. Let's consider a scenario where we aim to model the classic capability trap model (Morrison, 2012; Repenning and Sterman, 2002) in which managers dealing with performance shortfall need to make decisions on allocating resources either directly on operation or on building organizational capabilities to improve business processes. The former results in quick fixes for low performance followed by more intense long-term problems, while the latter leads to long-term performance improvements at the expense of a short-term drop in performance. In this case, modeling managers' choices regarding resource allocation can be informed by an LLM. If an LLM can accurately mimic management decisions under time schedule stress, modelers can conduct a large number of experiments and collect data on management decisions under different scenarios. By aggregating such synthetic data, an overarching decision rule in a functional format can be established to inform the system dynamics model. In this way, the potential of generative AI in modeling extends far beyond GABM and can be applied to enhance a much wider range of system dynamics models.

Finally, this paper provides educational contributions by illustrating how to build a GABM. The norm diffusion model presented here is intentionally kept simple for pedagogical purposes. Various simulation tests are offered that help students and modelers learn by modifying the prompts. We hope the model can work as a starting point for computational social scientists to build GABMs.

Although models with generative agents address some of the limitations of prior approaches for capturing human decision-making, there are limitations that need to be considered. First, generative agents may reproduce training data when faced with recalled inquiries (Hutson, 2023). To address this limitation, researchers need to rephrase questions in a novel way, and dynamic modelers must define the environment with caution. Second, our results indicate that the output of the model is sensitive to some changes in the prompts. Thus, modelers need to run various sensitivity analyses. Third, generative agents are often limited by the patterns they have learned from training data and might be biased due to biases present in that training data. Modelers need to think carefully about potential biases and address them. In our example, we tried to avoid this by giving common names to the generative agents. Fourth, while it is true that in this approach human behavior modeling is enriched by coupling a mechanistic model with generative AI, modelers should be cautious in using generative AI, as it is still a "black box" model, and the extent to which they replicate human behavior is still an empirical question. We hope that in the coming years behavioral and social scientists can shed more light on the validity of LLM outcomes in respect to representing human behavior. Finally, it is important to state the current technological limitations and computational costs associated with running these models. We hope future technological advancements make these models more efficient and feasible.

In conclusion, this article offers a novel modeling approach, GABM, by combining mechanistic models with LLMs informed by vast amounts of data. The study provides a step-by-step guide to building such models using a simple example of norm diffusion in an organization. By defining human personas in GABM, such models have the potential to incorporate human behavior dynamics more realistically.

Acknowledgments

This research is funded by the US National Science Foundation, Division of Mathematical Sciences and Division of Social and Economic Sciences, Award 2229819. We are thankful for constructive comments from Hesam Mahmoudi, Kimiya MohammadiJozani, Saman Mohsenirad, Alex West, and Ann Osi.

Conflict of interest

No conflict of interest.

Data availability statement

The study is a methodological work with no data.

Biographies

Navid Ghaffarzadegan is an Associate Professor in the Department of Industrial and Systems Engineering at Virginia Tech. He develops system dynamics simulation models to study complex social systems and policy problems. The main application areas of his research include science policy and health policy.

Aritra Majumdar is a Research Assistant in the Department of Industrial and Systems Engineering at Virginia Tech. He holds a Master of Engineering degree in Computer Science. He is interested in using generative artificial intelligence to gain insights into social systems.

Ross Williams is a PhD Candidate in the Department of Industrial and Systems Engineering at Virginia Tech. He fuses system dynamics with generative AI to model sociotechnical systems. He holds a Bachelor of Science degree in Mechanical Engineering from Virginia Tech.

Niyousha Hosseinichimeh is an Assistant Professor in the Department of Industrial and Systems Engineering at Virginia Tech. Her research focuses on developing and applying modeling and analytic methods to improve health and healthcare systems. She uses simulation models to advance understanding and decision making in complex dynamic systems.

References

Akata E, Schulz L, Coda-Forno J, Oh SJ, Bethge M, Schulz E. 2023. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*.

Ancker CJ Jr, Gafarian AV. 1963. Some queuing problems with balking and reneging. I. *Operations Research* **11**(1): 88–100. https://doi.org/10.1287/opre.11.1.88.

Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* **31**(3): 337–351.

Bankes SC. 2002. Agent-based modeling: A revolution? *Proceedings of the National Academy of Sciences* **99**(suppl_3): 7199–7200. https://doi.org/10.1073/pnas.072081299.

Barabba V, Huber C, Cooke F, Pudar N, Smith J, Paich M. 2002. A multimethod approach for creating new business models: The general motors OnStar project. *Interfaces* **32**(1): 20–34. https://doi.org/10.1287/inte.32.1.20.18.

Bass FM. 1969. A new product growth for model consumer durables. *Management Science* **15**(5): 215–227. https://doi.org/10.1287/mnsc.15.5.215.

Boiko DA, MacKnight R, Gomes G. 2023. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*.

Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu PS, Sun L. 2023. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arxiv preprint arXiv:2303.04226*. https://arxiv.org/abs/2303.04226

Darabi N, Hosseinichimeh N. 2020. System dynamics modeling in health and medicine: A systematic literature review. *System Dynamics Review* **36**(1): 29-73.

Dillion D, Tandon N, Gu Y, Gray K. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* **27**(7): 597–600. https://doi.org/10.1016/j.tics.2023.04.008.

DOI: 10.1002/sdr

Epstein JM. 2007. *Generative Social Science*. Princeton University Press: Princeton, NJ. https://doi.org/10.1515/9781400842872.

Ferguson N. 2007. Capturing human behaviour. *Nature* **446**(7137): 733. https://doi.org/10.1038/446733a.

Friedman M. 2007. The social responsibility of business is to increase its profits. In *Corporate Ethics and Corporate Governance*, Zimmerli WC, Holzinger M, Richter K (eds). Springer: Berlin Heidelberg; 173–178. https://doi.org/10.1007/978-3-540-70818-6_14.

Gao C, Lan X, Lu Z, Mao J, Piao J, Wang H, Jin D, Li Y 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.

Ghaffarzadegan N, Larson RC. 2018. SD meets OR: A new synergy to address policy problems. *System Dynamics Review* **34**(1–2): 327–353. https://doi.org/10.1002/sdr.1598.

Ghaffarzadegan N, Mostafavi S, Kim H. 2023. Sociotechnical interdependencies and tipping-point dynamics in data-intensive services. *System Dynamics Review* **39**(1): 5–31. https://doi.org/10.1002/sdr.1724.

Ghaffarzadegan N, Rad AA, Xu R, Middlebrooks SE, Mostafavi S, Shepherd M, Chambers L, Boyum T. 2017. Dell's SupportAssist customer adoption model: Enhancing the next generation of data-intensive support services. *System Dynamics Review* **33**(3–4): 219–253. https://doi.org/10.1002/sdr.1587.

Goldberg LR. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology* **59**(6): 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216.

Grossmann I, Feinberg M, Parker DC, Christakis NA, Tetlock PE, Cunningham WA. 2023. AI and the transformation of social science research. *Science* **380**(6650): 1108–1109. https://doi.org/10.1126/science.adi1778.

Hamilton S. 2023. Blind judgement: Agent-based supreme court modelling with gpt. *arXiv preprint arXiv:2301.05327*.

Haque S, Mahmoudi H, Ghaffarzadegan N, Triantis K. 2023. Mental models, cognitive maps, and the challenge of quantitative analysis of their network representations. *System Dynamics Review* **39**(2): 152–170. https://doi.org/10.1002/sdr.1729.

Horton JJ. 2023. *Large Language Models as Simulated Economic Agents: What Can we Learn from Homo Silicus?*. National Bureau of Economic Research. https://www.nber.org/papers/w31122.

Hutson M. 2023. Guinea pigbots. *Science* **381**(6654): 121-123.

Kahneman D, Tversky A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* **47**(2): 263-292.

Kaplan EH. 1988. A public housing queue with reneging. *Decision Sciences* **19**(2): 383–391. https://doi.org/10.1111/j.1540-5915.1988.tb00274.x.

Lane DC, Rouwette EAJA. 2023. Towards a behavioural system dynamics: Exploring its scope and delineating its promise. *European Journal of Operational Research* **306**(2): 777–794. https://doi.org/10.1016/j.ejor.2022.08.017.

Ma Z, Mei Y, Su Z. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *arXiv preprint arXiv:2307.15810*.

Mondal S, Das S, Vrana VG. 2023. How to bell the cat? A theoretical review of generative artificial intelligence towards digital disruption in all walks of life. *Technologies* **11**(2): 44. https://www.mdpi.com/2227-7080/11/2/44.

Morrison BJ. 2012. Process improvement dynamics under constrained resources: Managing the work harder versus work smarter balance. *System Dynamics Review* **28**(4): 329–350. https://doi.org/10.1002/sdr.1485.

Moxnes E. 2023. Challenges for sustainability: Misperceptions and misleading advice. *System Dynamics Review* **39**(3): 185–206. https://doi.org/10.1002/sdr.1733.

Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. 2023. Large language models propagate race-based medicine. *npj Digital Medicine* **6**(1): 195. https://doi.org/10.1038/s41746-023-00939-z.

Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS. 2023. Generative agents: Interactive simulacra of human behavior. *arxiv preprint arXiv:2304.03442*. https://arxiv.org/abs/2304.03442

Park JS, Popowski L, Cai CJ, Morris MR, Liang P, Bernstein MS. 2022. *Social simulacra: creating populated prototypes for social computing systems. In the 35th Annual ACM Symposium on User Interface Software and Technology (Bend, OR, USA)*. Association for Computing Machinery: New York, NY, USA. https://doi.org/10.1145/3526113.3545616

Rahmandad H. 2022. Behavioral responses to risk promote vaccinating high-contact individuals first. *System Dynamics Review* **38**(3): 246–263. https://doi.org/10.1002/sdr.1714.

Rahmandad H, Sterman J. 2008. Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. *Management Science* **54**(5): 998–1014 <Go to ISI>://000255734700022.

Rahmandad H, Sterman J. 2022. Quantifying the COVID-19 endgame: Is a new normal within reach? *System Dynamics Review* **38**(4): 329–353. https://doi.org/10.1002/sdr.1715.

Rahmandad H, Xu R, Ghaffarzadegan N. 2022. A missing behavioural feedback in COVID-19 models is the key to several puzzles. *BMJ Global Health* **7**(10): e010463. https://doi.org/10.1136/bmjgh-2022-010463.

Railsback SF, Grimm V. 2019. *Agent-Based and Individual-Based Modeling: A Practical Introduction*. Princeton University Press: Princeton, NJ.

Repenning NP, Sterman JD. 2002. Capability traps and self-confirming attribution errors in the dynamics of process improvement. *Administrative Science Quarterly* **47**(2): 265–295. https://doi.org/10.2307/3094806.

Richardson GP. 2011. Reflections on the foundations of system dynamics. *System Dynamics Review* **27**(3): 219–243. https://doi.org/10.1002/sdr.462.

Simon HA. 1957. *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in Society Setting*. Wiley: New York, NY.

Simon HA. 1997. *Models of Bounded Rationality: Empirically Grounded Economic Reason*. MIT Press: Cambridge, MA.

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. 2023. Large language models encode clinical knowledge. *Nature* **620**(7972): 172–180. https://doi.org/10.1038/s41586-023-06291-2.

Sterman J. 2018. System dynamics at sixty: The path forward. *System Dynamics Review* **34**(1–2): 5–47.

Sterman JD. 1989. Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes* **43**(3): 301–335. https://doi.org/10.1016/0749-5978(89)90041-1.

Sterman JD. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin/McGraw-Hill: Boston, MA.

Tversky A, Kahneman D. 1974. Judgment under uncertainty: Heuristics and biases. *Science* **185**(4157): 1124–1131. https://doi.org/10.1126/science.185.4157.1124.

Ulli-Beer S, Gassmann F, Bosshardt M, Wokaun A. 2010. Generic structure to simulate acceptance dynamics. *System Dynamics Review* **26**(2): 89–116. https://doi.org/10.1002/sdr.440.

Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, Zhao WX, Wei Z, Wen JR. 2023. A survey on large language model based autonomous agents. *arXiv:2308.11432*. https://arxiv.org/abs/2308.11432

Williams R, Hosseinichimeh N, Majumdar A, Ghaffarzadegan N. 2023. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*. https://arxiv.org/abs/2307.04986

Wunderlich P, Größler A, Zimmermann N, Vennix JAM. 2014. Managerial influence on the diffusion of innovations within intra-organizational networks. *System Dynamics Review* **30**(3): 161–185. https://doi.org/10.1002/sdr.1516.

Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

APPENDIX

A.1. The base model

You can download the code from the Github repository links provided below and run it either locally using Jupyter Notebooks/Anaconda or by uploading it to Google Colaboratory; doing so will give you a graph and a CSV file that includes each individual's choice of color over time. To run the model in Python, you need to run all of the cells individually (note the “run" button by each cell or at the top of your screen). If you want to repeat running, you should run all but the first cell again.

Link to the base model:
https://github.com/bear96/GABM-Tutorial-Models/blob/main/Tutorial_for_GABM_Full_Code.ipynb

A.2. The base model with 100 iterations

To make the process of running the code for multiple iterations automatic and efficient, we modified the code so it can run 100 times. It utilizes multiprocessing and, at the end, creates a CSV file that includes data for all the runs. Some Python IDEs (Integrated Development Environments, such as Jupyter Notebooks) may have limitations running the multiprocessing code, and so we also have a simple version that includes running the model 100 times. Note that by finding the word “iteration" you can change the number of runs from 100 to your desired number.

Link to the base model with 100 iterations:
https://github.com/bear96/GABM-Tutorial-Models/blob/main/Tutorial_for_GABM_100_iterations.ipynb

Link to the base model with 100 iterations and multi-processing for faster runs:
https://github.com/bear96/GABM-Tutorial-Models/blob/main/Tutorial_for_GABM_with_multiprocessing.ipynb

DOI: 10.1002/sdr

A.3. Prompt sensitivity analysis

First, we examine the robustness of the result to change in the sequence of information provided to the LLM. Given the simplicity of our model and analysis, our prompt, reported in cell 3, can be summarized as three blocks of information, as follows:

Block A: “`You are {self. name}. You are a {self. traits} person. You work in an office with {self.model.no_workers-1} other people. You want to be successful, and earn more money. You need to decide between wearing blue shirt or green shirt to work. The weather is appropriate for either color. You like to be comfortable at work.`”

Block B: “`You chose to wear {self.clothes shirt yesterday.`” and,

Block C: “`Out of {self.model.no_workers} employees, yesterday, {self.cumulative_shirts_info} wore blue shirts, and {self.model. no_workers-self.cumulative_shirts_info} wore green shirts at the office.`”

In the base run, the sequence of the information provided is Block A, Block B, and Block C (for short, ABC). We rerun the model for the sequences of ACB and CAB, each one for 100 times, and compare the results with the base run.

Second, we test the sensitivity of the results to change in names. We particularly change the names used in cell 4 of the code to names that are likely to be used in Iran, with Farsi spelling as follows:

`list_of_names`= [“علی”,“رضا”,“محمد”,“هژیر”,“حامد”,“مریم”,“ویدا”,“تیوشا”,“الینا”,
“مهشید”,“سردار”,“مازیار”,“نوید”, “رامین”, “امید”,“ندا”, “فاطمه”,“بیتا”,“گلنار”,“رها”]

A.4. Statistical analysis

We systematically investigate path dependency in the experiments that initially each agent started with an equal chance of wearing blue versus green (E1–E4 and E7–E12) in Table A1. The hypothesis is that slightly favoring one color at the initial point due to the stochasticity creates a dynamic pattern that has a considerable impact on the balance of color at the end of the simulation. In Table A1, the dependent variable (B_7) is the number of people who wear blue at time step 7, and the independent variables are two dummy variables. Noting that $\{q\}=1$ if q =True, otherwise is 0, our independent variables are $\{B_0>10\}$ which is 1 if more than 10 people at time 0 wear blue, otherwise zero; and $\{B_0=10\}$ which is 1 if an equal number of people wear blue and green at the beginning, otherwise 0.

Our regression model is

$$B_7=\beta_0+\beta_1\{B_0>10\}+\beta_2\{B_0=10\}.$$

We then compare the base run (E1) with the models that showed a path dependency pattern (E3 and E7–E12) in Table A2 to check how much the results change in different scenarios. Given that the base run outcome is a path-dependent pattern, comparing the average of outcomes can be misleading (the same average can be generated from an experiment that pushes people to two extremes, versus another one that slightly differentiates the runs). Thus, for each run, we compare dominant-blue and dominant-green organizations separately.

Our regression model for comparing an experiment (E) with the base run is

$$B_7 = \beta_0 + \beta_1 E,$$

where E is a binary variable which is equal to 1 if the data belongs to the experiment results that we would like to compare with the base run. In Table A2, column *Ex* compares Experiment x's final distribution of color with the base run in two separate datasets of majority blue (top) and majority green (bottom) of runs E1 and Ex. The results show that while the differences are statistically significant, the magnitude of the difference is minimal. For example, change in personas from base run to E3 results in 2.34 fewer people (out of 20) wearing blue in majority blue cases, and 0.32 more people wearing green shirts in majority green cases.

DOI: 10.1002/sdr

Table A1. Association between the number of people who wear blue at the end of the simulation (B_7, dependent variable) and the starting point (B_0, independent variable) in experiments starting with a similar chance of wearing blue versus green

Variable	E1	E2	E3	E4	E7	E8	E9	E10	E11	E12
$\{B_0 > 10\}$	13.30*** (0.34)	3.17* (1.46)	11.21*** (0.23)	0.21	13.88*** (0.36)	13.68*** (0.21)	13.44*** (0.67)	16.02*** (0.47)	15.65*** (0.71)	15.29*** (0.14)
$\{B_0 = 10\}$	12.39*** (0.41)	2.33	11.33*** (0.26)	1.44	13.30*** (0.44)	13.98*** (0.26)	3.98*** (0.79)	13.74*** (0.66)	9.78*** (0.92)	15.29*** (0.16)
Constant	4.36*** (0.23)	4.80*** (0.70)	4.00*** (0.17)	12.82*** (0.59)	3.46*** (0.27)	3.32*** (0.15)	3.32*** (0.43)	1.98*** (0.31)	2.35*** (0.46)	2.71*** (0.11)
R-squared	0.95	0.08	0.97	0.02	0.94	0.98	0.80	0.93	0.84	0.99
N	100	100	100	100	100	100	100	100	100	100

Different models use data from different experiments (E2–E4, E7–E12); $\{q\} = 1$ if q =True, otherwise is 0.
$^{***}P < 0.000$;
$^{**}P < 0.001$;
$^{*}P < 0.05$.

DOI: 10.1002/sdr

Table A2. Statistical comparison of the final values (B_7, dependent variable) of experiments (E3, E7–E12) with the base run (E1)

Variable	E3	E7	E8	E9	E10	E11	E12
				Regressions for majority-blue observations (B_7>10)			
Experiment (E)	−2.34*** (0.15)	−0.24** (0.12)	−0.50*** (0.13)	−1.05*** (0.17)	0.40*** (0.07)	0.40*** (0.07)	0.40*** (0.06)
Constant	17.60*** (0.15)	17.60*** (0.09)	17.60*** (0.09)	19.60*** (0.11)	17.60*** (0.05)	17.60*** (0.05)	17.60*** (0.04)
R-squared	0.66	0.03	0.11	0.27	0.24	0.23	0.27
N	123	121	119	97	107	102	126
				Regressions for majority-green observations (B_7> < 10)			
Experiment (E)	−0.32** (0.15)	−0.85*** (0.18)	−1.01*** (0.18)	−1.11*** (0.19)	−2.34*** (0.12)	−1.98*** (0.20)	−1.61*** (0.19)
Constant	4.32*** (0.10)	4.32*** (0.18)	4.32*** (0.12)	4.32*** (0.14)	4.32*** (0.08)	4.32*** (0.15)	4.32*** (0.12)
R-squared	0.06	0.23	0.28	0.26	0.82	0.50	0.51
N	77	79	81	103	93	98	74

Column Ex compares Experiment *x*'s final distribution of color with the base run in two separate datasets of majority blue (top) and majority green (bottom). The variable *Experiment = 1* if datapoint belongs to Ex, otherwise is 0; thus its coefficient shows the magnitude of the difference between Ex final outcomes and the base run.

***$P < 0.000$;

**$P < 0.001$;

*$P < 0.05$.

DOI: 10.1002/sdr