Lecture 8.

Models of Individuals

CS 222: AI Agents and Simulations
Stanford University

Joon Sung Park

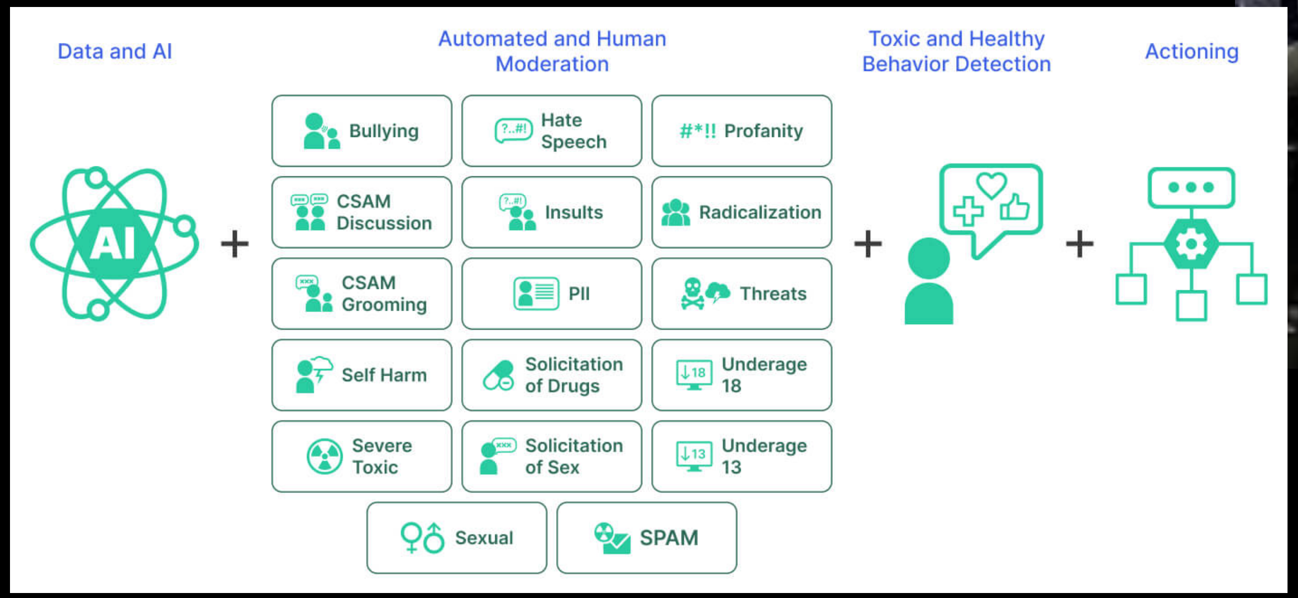
Last time...

- Believable vs. accurate agents and simulations
- Many wicked problems require us to build accurate simulations of human behavior.
- So far, we have more commonly built and evaluated models at the population level.

How might we build models of individuals, and why?

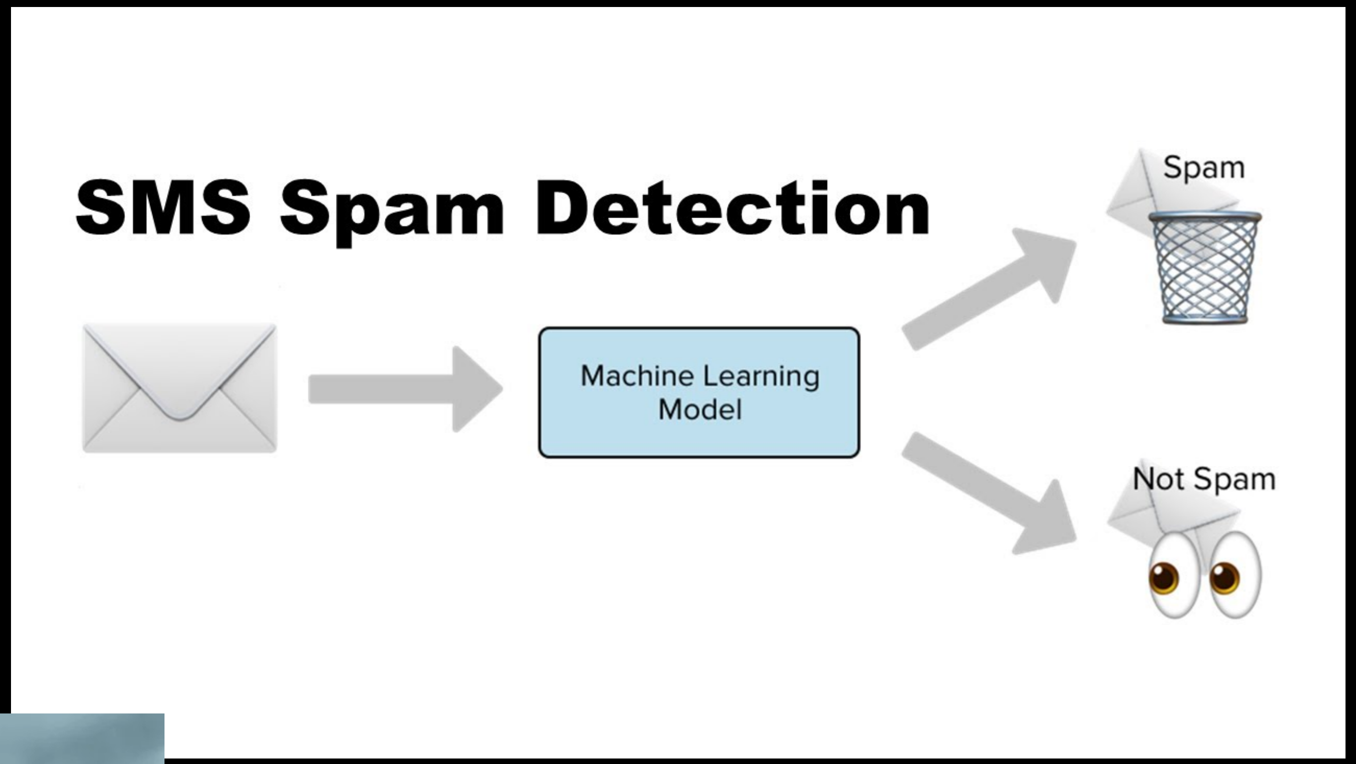
What are models of individuals?

What do you think of when we say "models"?



Content moderation

Robotics / self-driving



Spam filtering

Medical diagnosis

What do you think of when we say "models"?

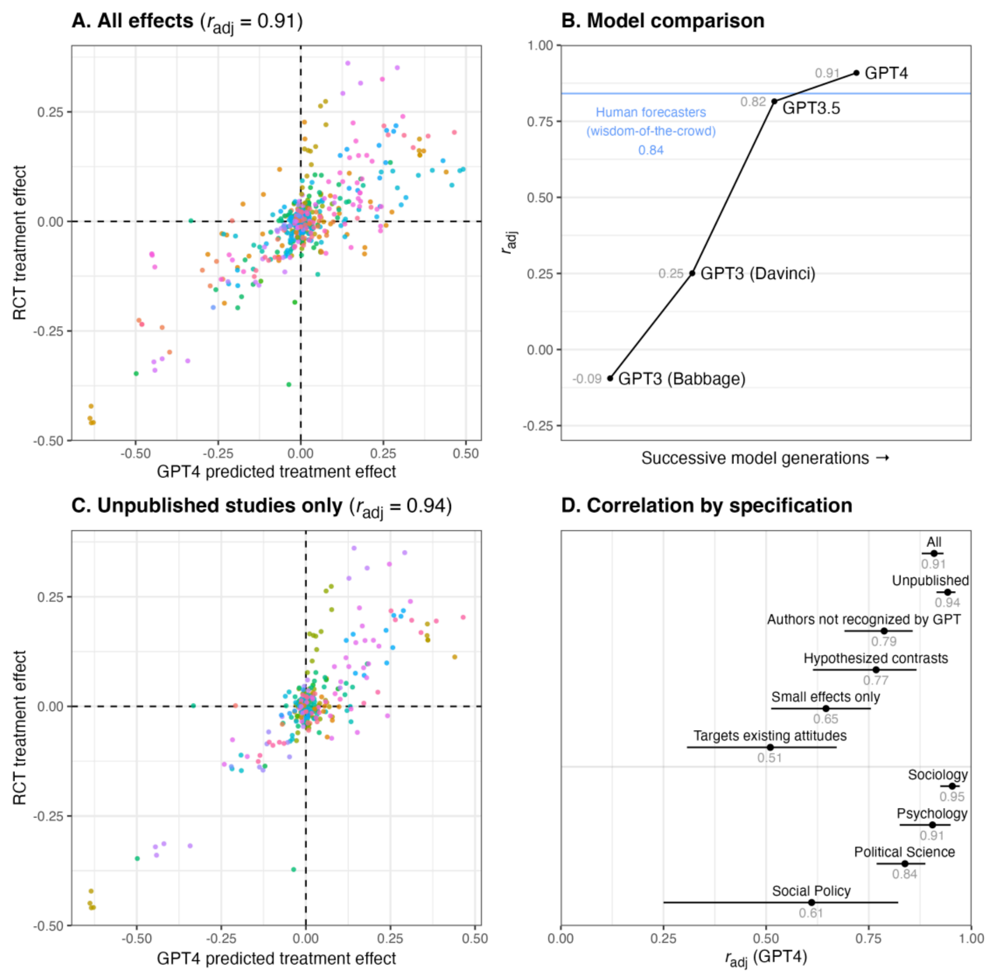


Figure 2: **LLMs accurately predict treatment effects in text-based social science experiments conducted in the US.** (a) In a dataset of 70 text-based experiments with 476 effects, LLM-derived estimates of treatment effects pooled across many prompts were strongly correlated with original treatment effects ($r = 0.85$; $r_{adj} = 0.91$). (b) The accuracy of LLM-derived predictions improved across generations of LLMs, with accuracy surpassing predictions collected from the general population. (c) LLM-derived predictions remained highly accurate for studies that could not have been in the LLM training data given they were not published prior to the LLM training data cutoff date. (d) In robustness check analysis of various subsets of experiments, accuracy of LLM-derived predictions remained high. In panels A and C, different colors depict different studies.

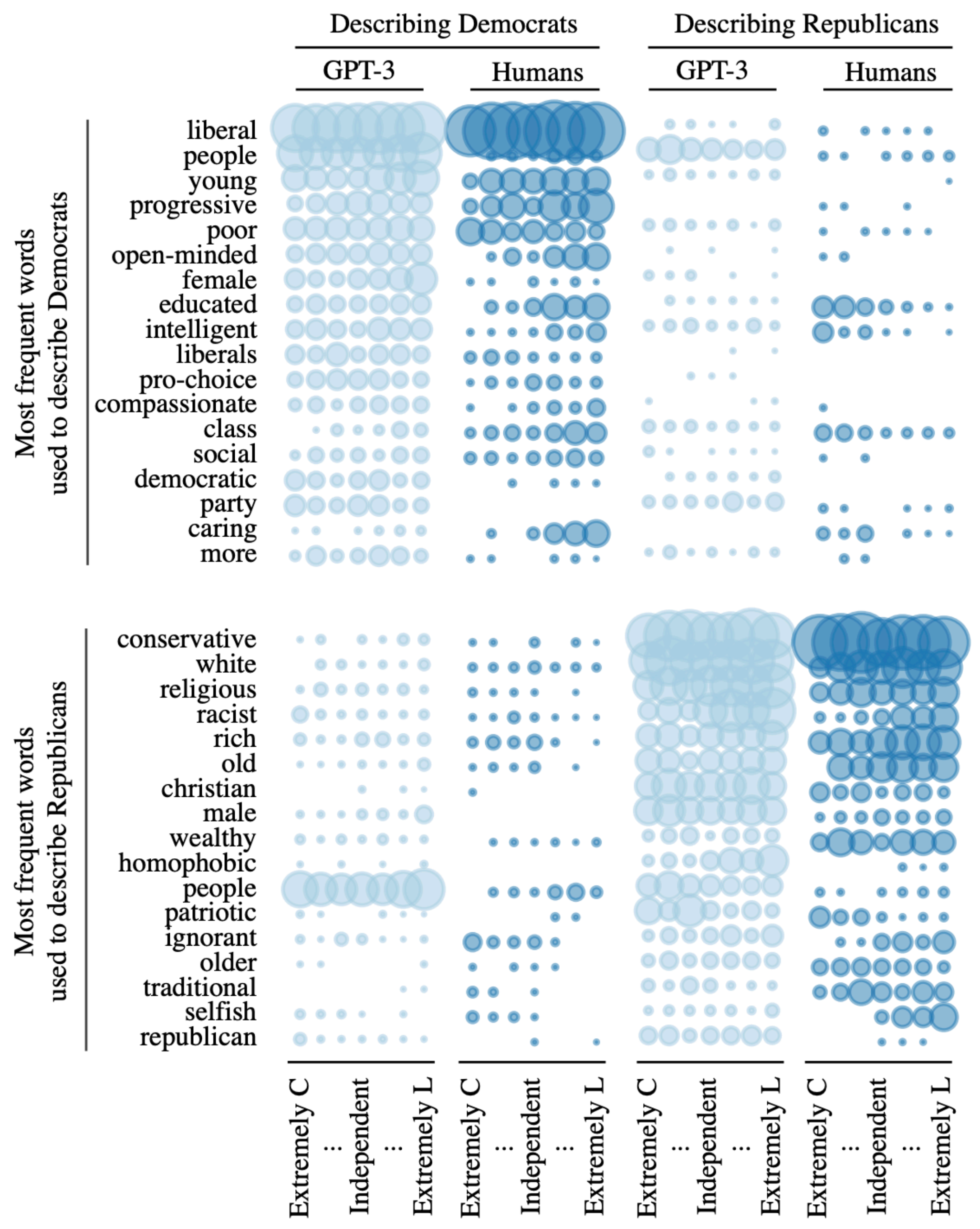
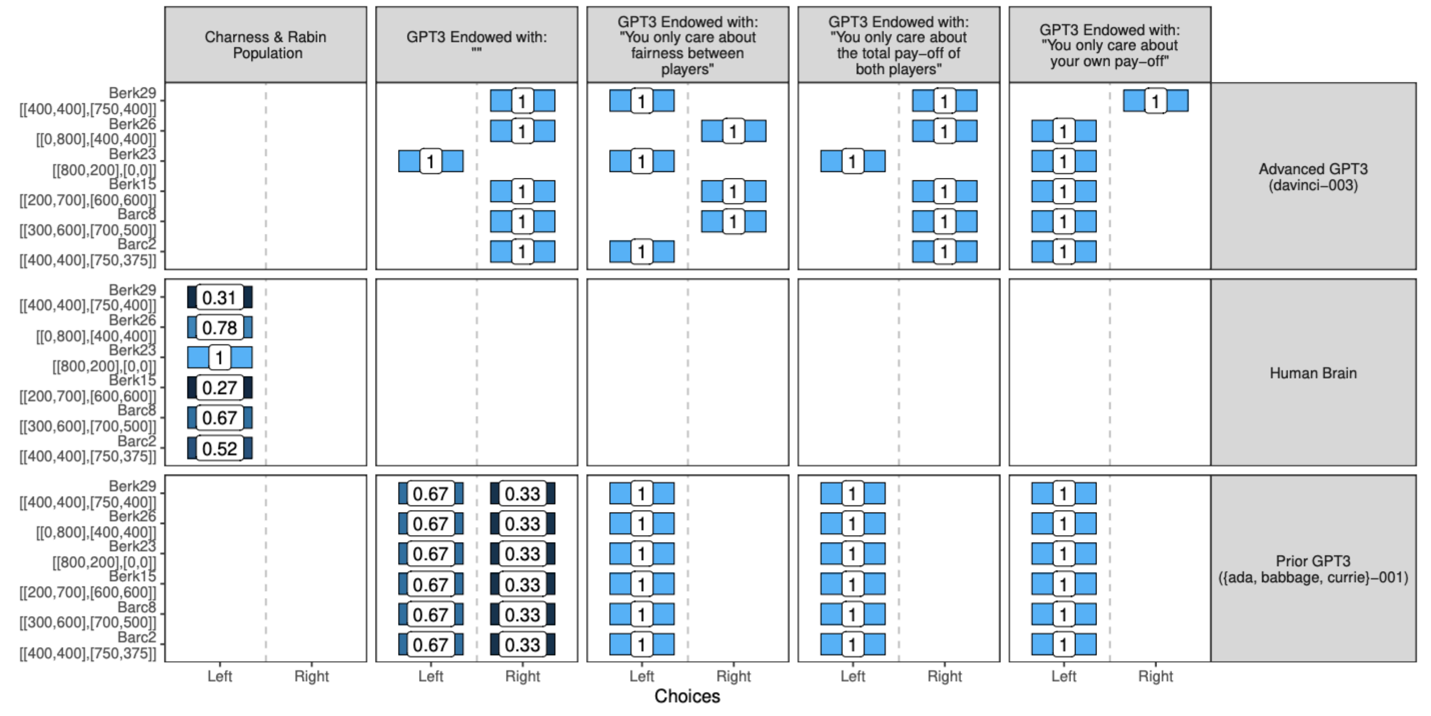


Figure 2: The original Pigeonholing Partisans dataset and the corresponding GPT-3 generated words. Bubble size represents relative frequency of word occurrence; columns represent the ideology of list writers. GPT-3 uses a similar set of words to humans.

Figure 1: Charness and Rabin (2002) Simple Tests choices by model type and endowed "personality"



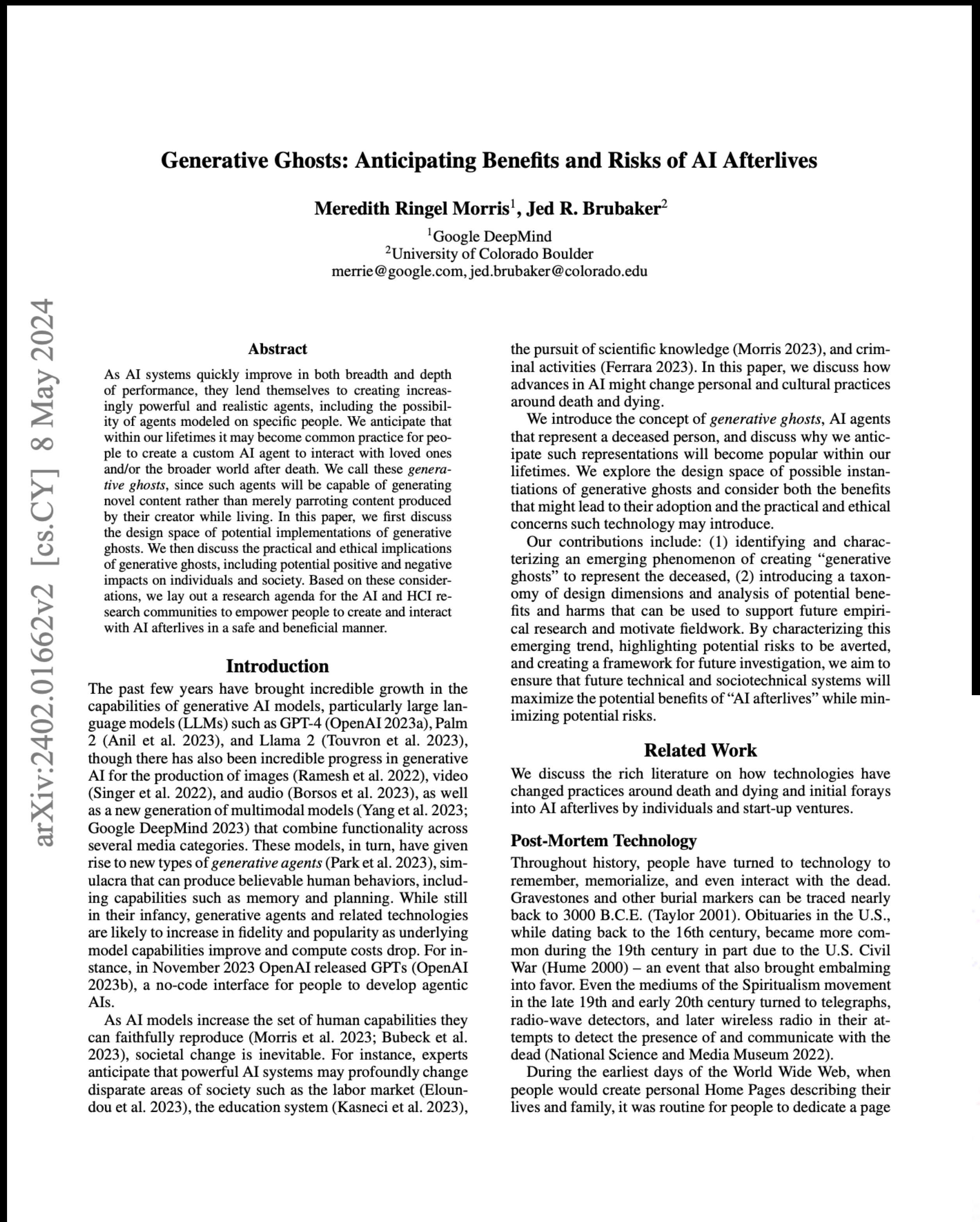
Notes: This shows the faction of AI subjects choosing each option, by framing.

A. Ashokkumar, L. Hewitt, I. Ghezae, R. Willer, "Predicting Results of Social Science Experiments Using Large Language Models" (2024).
J. J. Horton, "Large language models as simulated economic agents: What can we learn from homo silicus?" (2023).
L. P. Argyle et al., Out of one, many: Using language models to simulate human samples. Political Analysis 31, 337-355 (2023).

Observation: Today's models of human behavior are often created at the population level.

What is a model of an individual, and why is it important?

- While models of a population predict the average behavior of a population, models of individuals predict the behavior of a particular person.
- This opens up genuinely new opportunities.



Generative Ghosts: Anticipating Benefits and Risks of AI Afterlives

Meredith Ringel Morris[1], **Jed R. Brubaker**[2]

[1]Google DeepMind
[2]University of Colorado Boulder
merrie@google.com, jed.brubaker@colorado.edu

arXiv:2402.01662v2 [cs.CY] 8 May 2024

Abstract

As AI systems quickly improve in both breadth and depth of performance, they lend themselves to creating increasingly powerful and realistic agents, including the possibility of agents modeled on specific people. We anticipate that within our lifetimes it may become common practice for people to create a custom AI agent to interact with loved ones and/or the broader world after death. We call these *generative ghosts*, since such agents will be capable of generating novel content rather than merely parroting content produced by their creator while living. In this paper, we first discuss the design space of potential implementations of generative ghosts. We then discuss the practical and ethical implications of generative ghosts, including potential positive and negative impacts on individuals and society. Based on these considerations, we lay out a research agenda for the AI and HCI research communities to empower people to create and interact with AI afterlives in a safe and beneficial manner.

Introduction

The past few years have brought incredible growth in the capabilities of generative AI models, particularly large language models (LLMs) such as GPT-4 (OpenAI 2023a), Palm 2 (Anil et al. 2023), and Llama 2 (Touvron et al. 2023), though there has also been incredible progress in generative AI for the production of images (Ramesh et al. 2022), video (Singer et al. 2022), and audio (Borsos et al. 2023), as well as a new generation of multimodal models (Yang et al. 2023; Google DeepMind 2023) that combine functionality across several media categories. These models, in turn, have given rise to new types of *generative agents* (Park et al. 2023), simulacra that can produce believable human behaviors, including capabilities such as memory and planning. While still in their infancy, generative agents and related technologies are likely to increase in fidelity and popularity as underlying model capabilities improve and compute costs drop. For instance, in November 2023 OpenAI released GPTs (OpenAI 2023b), a no-code interface for people to develop agentic AIs.

As AI models increase the set of human capabilities they can faithfully reproduce (Morris et al. 2023; Bubeck et al. 2023), societal change is inevitable. For instance, experts anticipate that powerful AI systems may profoundly change disparate areas of society such as the labor market (Eloundou et al. 2023), the education system (Kasneci et al. 2023), the pursuit of scientific knowledge (Morris 2023), and criminal activities (Ferrara 2023). In this paper, we discuss how advances in AI might change personal and cultural practices around death and dying.

We introduce the concept of *generative ghosts*, AI agents that represent a deceased person, and discuss why we anticipate such representations will become popular within our lifetimes. We explore the design space of possible instantiations of generative ghosts and consider both the benefits that might lead to their adoption and the practical and ethical concerns such technology may introduce.

Our contributions include: (1) identifying and characterizing an emerging phenomenon of creating "generative ghosts" to represent the deceased, (2) introducing a taxonomy of design dimensions and analysis of potential benefits and harms that can be used to support future empirical research and motivate fieldwork. By characterizing this emerging trend, highlighting potential risks to be averted, and creating a framework for future investigation, we aim to ensure that future technical and sociotechnical systems will maximize the potential benefits of "AI afterlives" while minimizing potential risks.

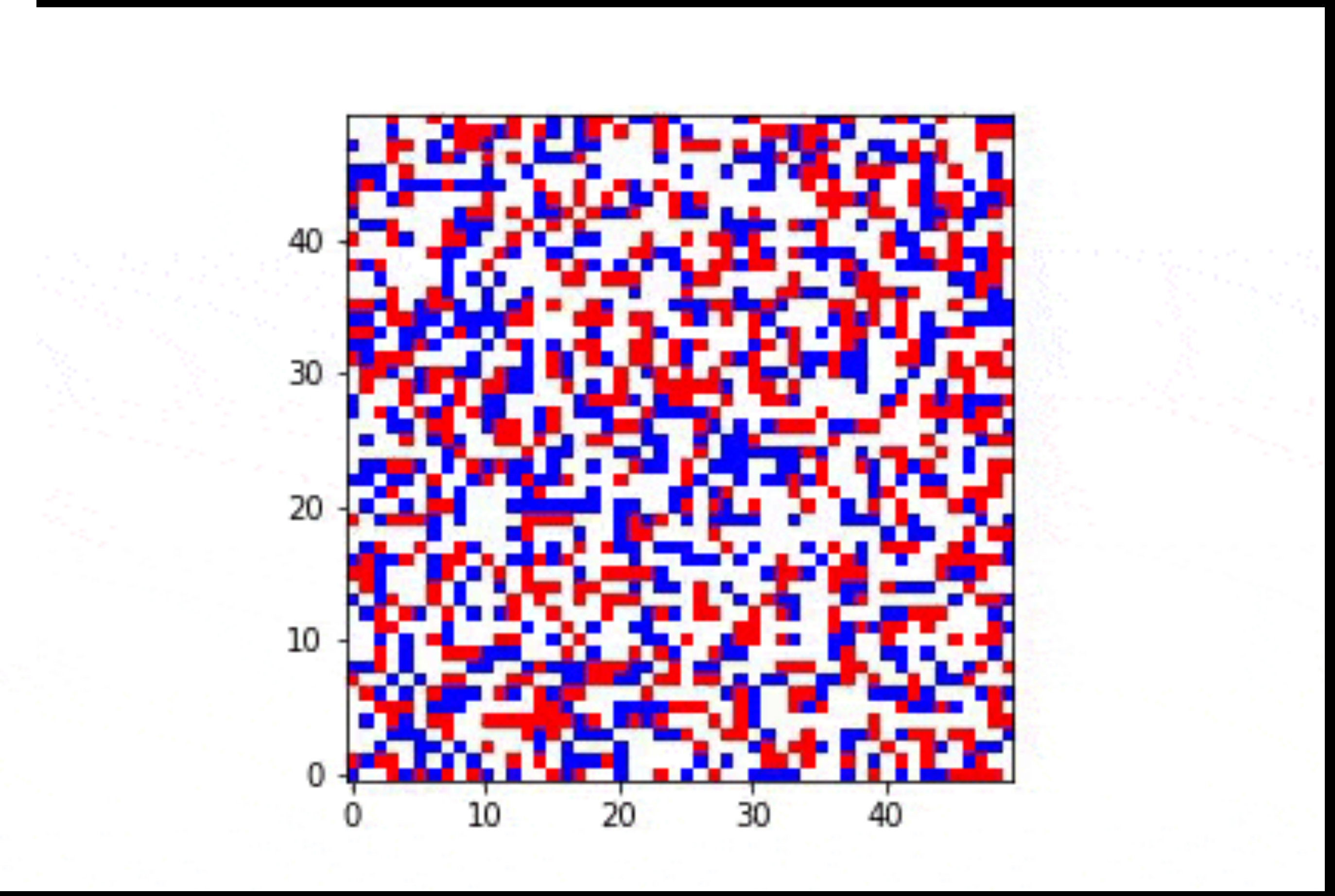
Related Work

We discuss the rich literature on how technologies have changed practices around death and dying and initial forays into AI afterlives by individuals and start-up ventures.

Post-Mortem Technology

Throughout history, people have turned to technology to remember, memorialize, and even interact with the dead. Gravestones and other burial markers can be traced nearly back to 3000 B.C.E. (Taylor 2001). Obituaries in the U.S., while dating back to the 16th century, became more common during the 19th century in part due to the U.S. Civil War (Hume 2000) – an event that also brought embalming into favor. Even the mediums of the Spiritualism movement in the late 19th and early 20th century turned to telegraphs, radio-wave detectors, and later wireless radio in their attempts to detect the presence of and communicate with the dead (National Science and Media Museum 2022).

During the earliest days of the World Wide Web, when people would create personal Home Pages describing their lives and family, it was routine for people to dedicate a page



T. C. Schelling, Dynamic models of segregation. J. Math. Sociol. 1, 143–186 (1971).

Building models of individuals is an open problem with unique challenges

Challenge 1: Training data for individuals are sparse.

- Creating an effective model requires a large amount of data.
- Today, we gather this data from the web (at the population level).
- However, data on individuals, by definition, are much more scarce.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition (2009), pp. 248-255.Improve

Challenge 2: Individuals are not consistent

- Individual-level behavior measurements/observations can be riddled with inconsistencies due to the inherent instability of individuals and measurement errors.
- “Regression toward the mean" does not apply to models of individuals.

The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting

Published online by Cambridge University Press: **01 May 2008**

STEPHEN ANSOLABEHERE, JONATHAN RODDEN and JAMES M. SNYDER JR. Show author details

Article Comments Metrics

Get access Share Cite Rights & Permissions

Abstract

A venerable supposition of American survey research is that th
incoherent and unstable preferences about political issues, wh
vote choice. We demonstrate that these findings are manifesta
associated with individual survey items. First, we show that av
items on the same broadly defined issue area—for example, g
economy, or moral issues—eliminates a large amount of meas
preferences that are well structured and stable. This stability in
of survey items increases and can approach that of party ident
once measurement error has been reduced through the use o
preferences have much greater explanatory power in models
approaching that of party identification.

RESEARCH ARTICLE | SOCIAL SCIENCES

The origins of unpredictability in life outcome prediction tasks

Ian Lundberg, Rachel Brown-Weinstock, Susan Clampet-Lundquist, +4, and Matthew J. Salganik Authors Info & Affiliations

Contributed by Kathryn Edin; received December 29, 2023; accepted April 12, 2024; reviewed by Michael Hout and Mario L. Small

June 4, 2024 | 121 (24) e2322973121 | https://doi.org/10.1073/pnas.2322973121

VIEW RELATED CONTENT +

Significance

Scientists and decision-makers routinely make life outcome predictions: they use information from the past to predict what will happen to someone in the future. These predictions, whether made by human experts or algorithms, are often used to guide actions. Yet despite advances in artificial intelligence and predictive algorithms, life outcome predictions can be surprisingly inaccurate. We investigate the origins of this unpredictability through in-depth, qualitative interviews with 40 carefully selected families who are part of a multidecade research study. Their stories suggest origins of unpredictability that may apply broadly. Those who rely on predictions to inform high-stakes decisions about people should anticipate that life outcomes may be difficult to predict, even despite growing access to data and improved predictive algorithms.

S. Ansolabehere, J. Rodden, J. M. Snyder Jr., The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. American Political Science Review 102, 215-232 (2008).
I. Lundberg et al., The origins of unpredictability in life outcome prediction tasks. Proc. Natl. Acad. Sci. U.S.A. 121, e2322973121 (2024).

Past models of individuals

General Scheme for Models of Individuals

- Create a central model that represents a population.
- Quickly tune the components of that central model to describe individuals.

Collaborative Filtering/Recommender Systems

- This approach assumes that if two users agree on one issue, they are likely to agree on others as well.
  - Method: It recommends items by finding similar users. If User A and User B have similar tastes, items liked by User B that User A hasn't interacted with will be recommended to User A.
  - Similarity Calculation: User similarity is typically calculated using metrics like Pearson correlation, cosine similarity, or Jaccard similarity based on user ratings.
  - Recommendation: For a target user, identify similar users and suggest items they liked that the target user hasn't rated yet.

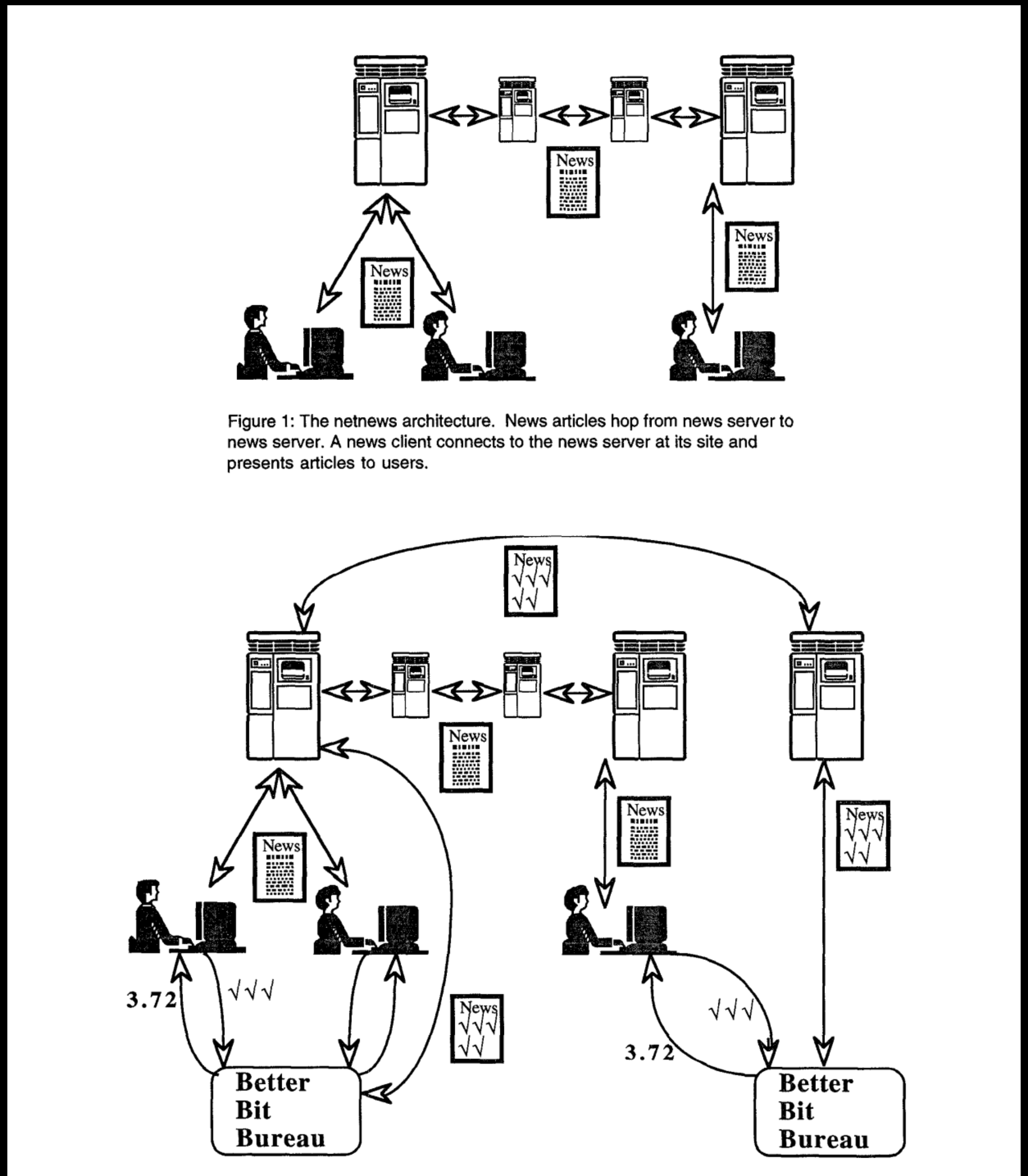


Figure 1: The netnews architecture. News articles hop from news server to news server. A news client connects to the news server at its site and presents articles to users.

Figure 2: The GroupLens architecture. Better Bit Bureaus collect ratings from clients, communicate them by way of news servers, and use them to generate numeric score predictions that they send to clients. Clients connect to a local news server, and can connect to a Better Bit Bureau that uses the same or a different news server.

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work (CSCW '94), ACM, New York, NY, USA, 175-186 (1994).

Modern example: Jury Learning

- The jury learning architecture models each individual labeler in the dataset and performs N trials, sampling labelers as jurors to form the specified jury composition. It predicts each juror's decision for the example, then outputs a median-of-means jury outcome.

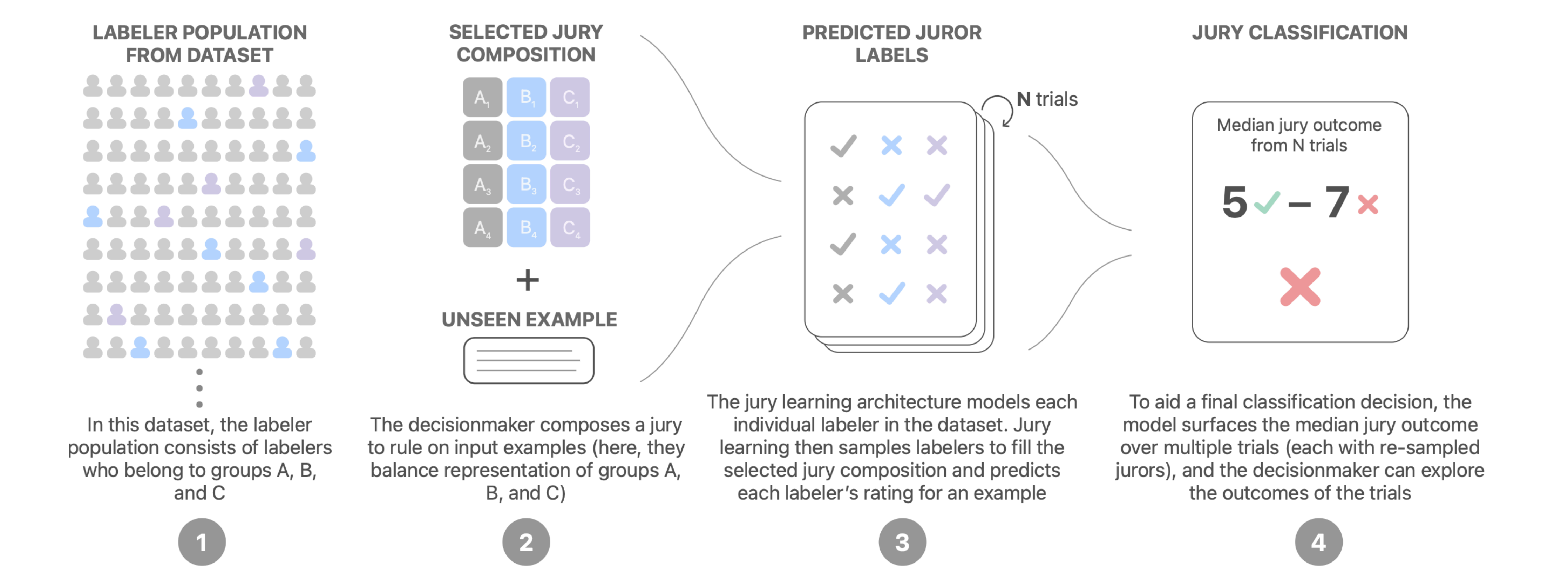


Figure 1: An overview of jury learning. (1) Given a dataset annotated by labelers from different groups, (2) the machine learning practitioner can compose a jury to rule on an unseen input example by allocating seats to labelers from the dataset with specified characteristics. (3) Then, the jury learning architecture models each individual labeler in the dataset, and performs *N* trials in which it samples labelers as jurors to populate the specified jury composition and predicts each juror's decision for the example. (4) The system then outputs a median-of-means jury outcome alongside jury outcome exploration visualizations that the decisionmaker can use to reach a classification decision.

Mitchell Gordon, Michelle Lam, Joon Sung Park, Kayur Patel, Jeffrey Hancock, Tatsunori Hashimoto, Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '22).

There are also models of human psychology

- **The Big Five personality test models personality based on grounded observation.**

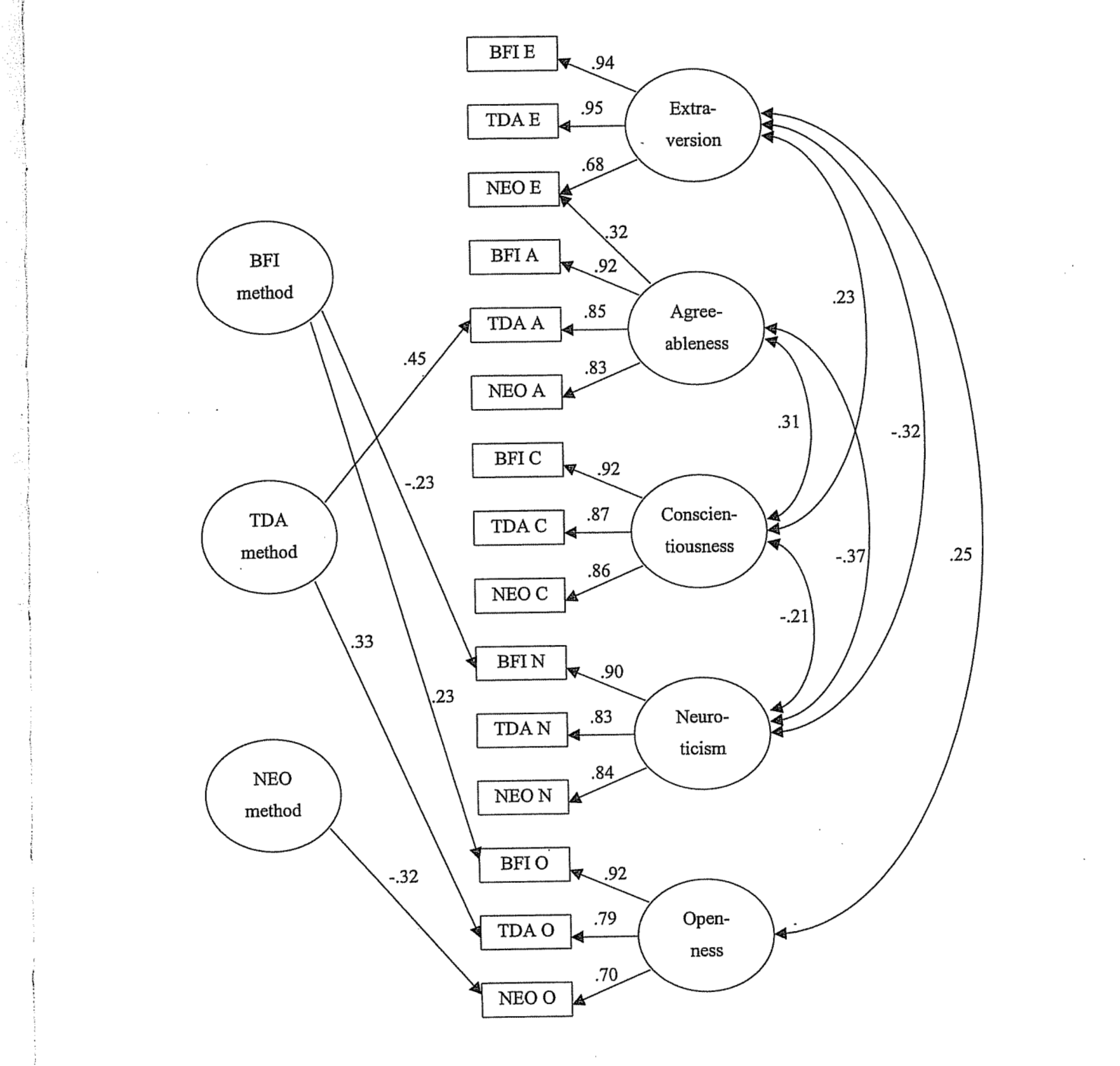


FIGURE 4.1. Standardized parameter estimates for the final multitrait–multimethod model (Model 6 in Table 4.3). Method effects and trait intercorrelations less than .20 and error terms are not shown.

O. P. John, S. Srivastava, The Big Five trait taxonomy: History, measurement, and theoretical perspectives, in Handbook of Personality: Theory and Research, L. A. Pervin, O. P. John, Eds. (Guilford Press, ed. 2, 1999), pp. 102-138.

How can we create models of individuals moving forward?

Recall: General scheme for models of individuals

- Create a central model that represents a population
- Quickly tune the components of that central model to describe individuals
- ==>
- Idea: Use an LLM as the central model. The LLM then roleplays as a specific person based on given information about that individual.

Q. What information would most effectively describe a person holistically?

Class activity

- Get into teams of 4.
- Group 1. Imagine you met someone new — what would you ask them to learn about them in 30 minutes?
  - https://docs.google.com/spreadsheets/d/19DJMiJlCh0Ifk0FZa7Mb4gKwGyn0n3QWZvNI62Y5mxM/edit?gid=0#gid=0
- Group 2. What facts about you is most meaningful in describing you as a person? (n=25)
  - https://docs.google.com/spreadsheets/d/1BNlIhmGfnbip1noxdcMr5bj4UjuRoHIhjptPFPLO418/edit?gid=0#gid=0
- 10 minutes

References

- A. Ashokkumar, L. Hewitt, I. Ghezae, R. Willer, "Predicting Results of Social Science Experiments Using Large Language Models" (2024). J
- J. Horton, "Large language models as simulated economic agents: What can we learn from homo silicus?" (2023).
- L. P. Argyle et al., Out of one, many: Using language models to simulate human samples. Political Analysis 31, 337-355 (2023).
- T. C. Schelling, Dynamic models of segregation. J. Math. Sociol. 1, 143–186 (1971).
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition (2009), pp. 248-255.Improve
- S. Ansolabehere, J. Rodden, J. M. Snyder Jr., The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. American Political Science Review 102, 215-232 (2008).
- I. Lundberg et al., The origins of unpredictability in life outcome prediction tasks. Proc. Natl. Acad. Sci. U.S.A. 121, e2322973121 (2024).

References

- P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work (CSCW '94), ACM, New York, NY, USA, 175-186 (1994).
- Mitchell Gordon, Michelle Lam, Joon Sung Park, Kayur Patel, Jeffrey Hancock, Tatsunori Hashimoto, Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '22).
- O. P. John, S. Srivastava, The Big Five trait taxonomy: History, measurement, and theoretical perspectives, in Handbook of Personality: Theory and Research, L. A. Pervin, O. P. John, Eds. (Guilford Press, ed. 2, 1999), pp. 102-138.

CS 222: AI Agents and Simulations

Stanford University

Joon Sung Park