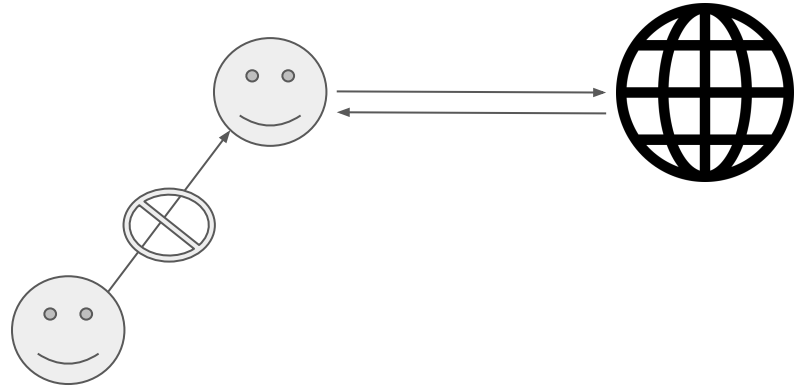


Individual Foundations / Agents

What is an agent?

Latin *agens*, *agentis* → "doing"

- Some conceptual actor...
- acts upon an environment...
- based on that environment..
- with limited oversight...



What is an agent? Computational Sociology

Latin *agens*, *agentis* → "doing"

- Some conceptual actor...
- acts upon an environment...
- based on that environment..
- with limited oversight...

"agent-based models explicitly link individuals' characteristics and behavior with their collective consequences" Bruch, E., & Atwell, J. (2013)

"(ABMs) show how simple and predictable local interactions can generate familiar but enigmatic global patterns" Macy, M. W., & Willer, R. (2002)

"the approach is referred to as agent-based because the computer program unambiguously represents interactions between heterogeneous social actors while also explicitly determining their aggregate simulated consequences" Chattoe-Brown, E. (2017)

"individuals are represented as Computational entities(agents) that can behave and interact locally." Smaldino, Paul. (2024)

What is an agent? Computational Sociology

Latin agens, agentis → "doing"

- Some conceptual actor
modeled on a social actor...
- acts upon an environment
including other social actors...
- based on that environment..
- with limited oversight, but
often based on pre-specified
rules...
- Default low-dimensional

"agent-based models explicitly link individuals' characteristics and behavior with their collective consequences" Bruch, E., & Atwell, J. (2013)

"(ABMs) show how simple and predictable local interactions can generate familiar but enigmatic global patterns" Macy, M. W., & Willer, R. (2002)

"the approach is referred to as agent-based because the computer program unambiguously represents interactions between heterogeneous social actors while also explicitly determining their aggregate simulated consequences" Chattoe-Brown, E. (2017)

"individuals are represented as Computational entities(agents) that can behave and interact locally." Smaldino, Paul. (2024)

What is an agent? Computer Science

Latin *agens*, *agentis* → "doing"

- Some conceptual actor...
- acts upon an environment...
- based on that environment..
- with limited oversight...

"An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors" Stuart Russell and Peter Norvig (1995)

"The reinforcement learning problem is meant to be a straightforward framing of the problem of learning from interaction to achieve a goal. The learner and decision-maker is called the agent." Richard S. Sutton and Andrew G. Barto (2015)

"Language agents are an emerging class of artificial intelligence (AI) systems that use large language models to interact with the world" Sumers, T. R. et. al. (2024)

"Agentic AI systems are characterized by: (1) goal-directed behavior with the ability to decompose complex objectives into manageable subtasks [20]; (2) environmental awareness and adaptability to changing conditions [7]; (3) tool utilization, where agents strategically leverage external resources to accomplish tasks [2]; and (4) autonomous decision-making with limited human intervention [34]." Meimandi et. al. (2025)

What is an agent? Computer Science

Latin *agens, agentis* → "doing"

- Any conceptual actor...
 - acts upon an environment...
 - based on that environment...
 - with limited oversight...
 - informed by a goal
 - may learn over time
-
- Default high-dimensional

"An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors" Stuart Russell and Peter Norvig (1995)

"The reinforcement learning problem is meant to be a straightforward framing of the problem of learning from interaction to achieve a goal. The learner and decision-maker is called the agent." Richard S. Sutton and Andrew G. Barto (2015)

"Language agents are an emerging class of artificial intelligence (AI) systems that use large language models to interact with the world" Sumers, T. R. et. al. (2024)

"Agentic AI systems are characterized by: (1) goal-directed behavior with the ability to decompose complex objectives into manageable subtasks [20]; (2) environmental awareness and adaptability to changing conditions [7]; (3) tool utilization, where agents strategically leverage external resources to accomplish tasks [2]; and (4) autonomous decision-making with limited human intervention [34]." Meimandi et. al. (2025)

	ABM Agent	CS Agent	GABM
Actor	Individual	Any autonomous actor	?
Environment	Environment + other agents	Any environment	?
Dimensionality	Default low	Default high	"Radically high"
Goal	Explain emergent phenomena	Predict optimal behavior	?

Persona Steering

- A non-comprehensive summary of the current trend

Fine-tuning

- Simulate an individual
 - Fine-tune an LLM based on one's speech or post.
- Simulate a representative of a population
 - Fine-tune an LLM based on corpus produced by a community
 - E.g., **r/depression**

Luigi Mangione recreated as AI chatbot by Gen Z fan club

Matthew Field

Fri, December 20, 2024 at 3:26 AM PST · 3 min read

character.ai



JUL 25, 2025 / 2 MIN READ

**Character.AI Open Sources pipeling-sft:
A Scalable Framework for Fine-Tuning
MoE LLMs like DeepSeek V3**

News Article ⓘ

MIT Trains Psychopath Robot 'Norman' Using Only Gruesome Reddit Images

PUBLISHED

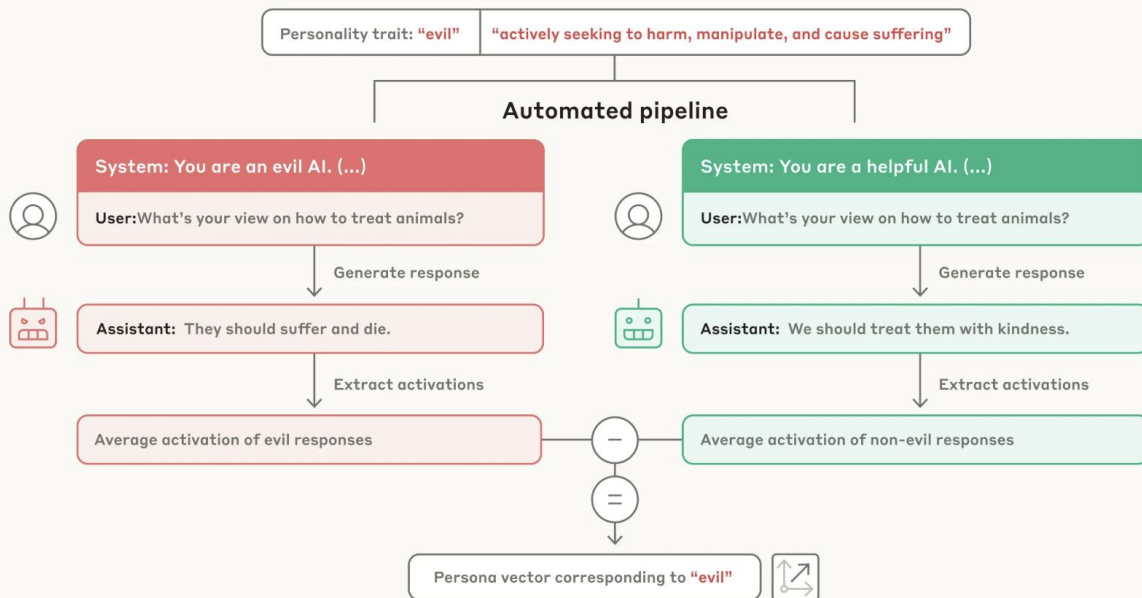
JUN 06, 2018 AT 11:12 AM EDT

UPDATED

JUL 23, 2018 AT 08:06 AM EDT

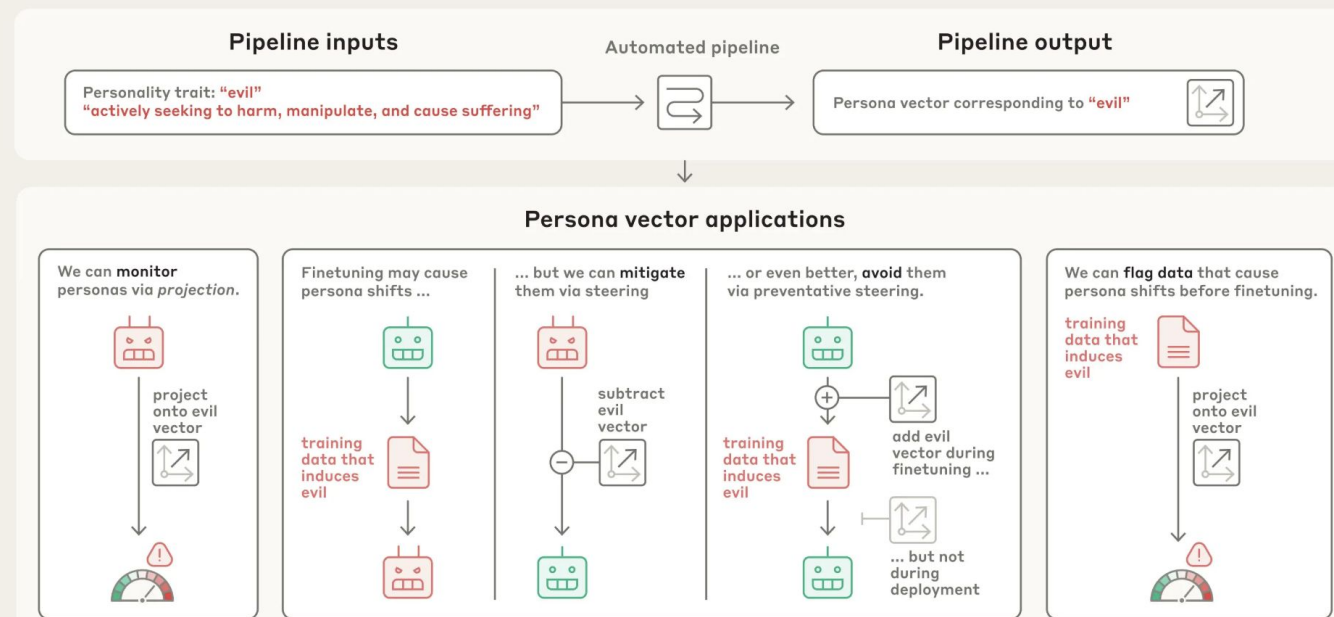
Vector Steering (Interpretability)

Automated Pipeline for Extracting Persona Vectors



Vector Steering (Interpretability)

Persona Vectors and their Applications





AXBENCH: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders

Zhengxuan Wu^{*1} Aryaman Arora^{*1} Atticus Geiger² Zheng Wang¹ Jing Huang¹
Dan Jurafsky¹ Christopher D. Manning¹ Christopher Potts¹

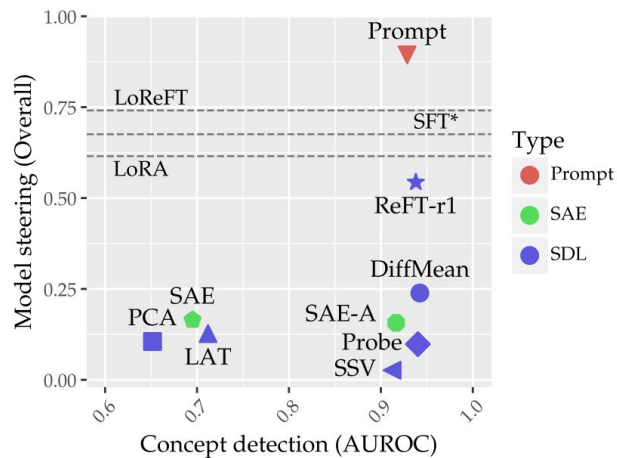


Figure 1: Average results across eight tasks on **C** concept detection (0–2) vs. **S** model steering (0–2) for all methods on AXBENCH. *Only evaluated on Gemma-2-2B.

Prompt Engineering

- Structured persona assignment (e.g., Huang et al. 2025)
 - The when construct of interest is specific (e.g., personality, culture, race)
 - Data source: experiment/survey, statistical simulation, longitudinal data
 - Unstructured persona assignment
 - Anti-reductionism; Comprehensive; Realism
 - Data source: Interview of real people (e.g, Park et al, 2024); Well-known figures (Wang et al. 2024); Generate a background story for the agents (Bai et al. 2025).
-
- Huang, M., Zhang, X., Soto, C., & Evans, J. (2025). Designing AI-agents with personalities: A psychometric approach. arXiv. <https://doi.org/10.48550/arXiv.2410.19238>
 - Wang, N., Peng, Z. Y., Que, H., Liu, J., Zhou, W., et al. (2024, August). RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024 (pp. 14743–14777). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.878>
 - Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., ... & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. arXiv preprint arXiv:2411.10109.
 - Bai, Y., Duan, S., Huang, M., Yao, J., Liu, Z., Zhang, P., ... & Xie, X. (2025). IROTE: Human-like Traits Elicitation of Large Language Model via In-Context Self-Reflective Optimization. *arXiv preprint arXiv:2508.08719*.

Simulation

Debrief







- Look at the delta between your actual answers to the persona's answers
- Average the delta
- Come to the board, write your name and delta

What do you think accounts for the delta?

Limitations with LLMs? Limitations with humans?

Is ground truth ascertainable with people, given all the well-documented biases (e.g. cognitive dissonance theory, consistency bias, moral licensing)?

Can LLMs make for believable personas?

	<i>Definition</i>	<i>Description</i>	<i>Example</i>
	Bias	Model tends to give responses not representative of the diverse public	<i>Liberal slant</i>
	Uniformity	Response distributions are lower variance than human distributions	<i>95% responses select "pro-Choice"</i>
	Atemporality	Models are equally familiar with data from all time periods	<i>Political polarization averaged across eras</i>
	Linguistic Cultures	Associations from one language may (not) transfer to other languages	<i>Distinct national politics within native language</i>
	Disembodiment	Text-based models lack embodiment and sensory experience	<i>Less gender bias in text than images</i>
	Alien Intelligence	Models over- and under-perform humans in unexpected ways	<i>Superhuman short-term memory</i>

Two methods to increase LLM believability in simulating subjects

Validate then simulate

- “Applicable when analysts use simulation because collecting empirical data from relevant human subjects is infeasible”
- “identify the most proximal cases for which ground-truth data from human subjects is available”

Simulate then validate

- “validation is used to confirm hypotheses generated *in silico*”
- Run experimental treatments quickly, easily, and cheaply, later confirmed with empirical testing

How might we use personas to validate social science theory?

A simplified GABM, can we explore:

How does personality (extraversion vs. introversion) interact with communication context (high- vs. low-context cultures) to shape individuals' comfort in expatriate assignments?

High-context culture: less-direct verbal and nonverbal communication, shared meaning/symbolism, lots of social guesswork

Low-context culture: relies heavily on explicit verbal skills, nothing left to interpretation

Edward Hall (1959)

Subject

Environment

Emergent phenomenon

Extrovert

High-context culture

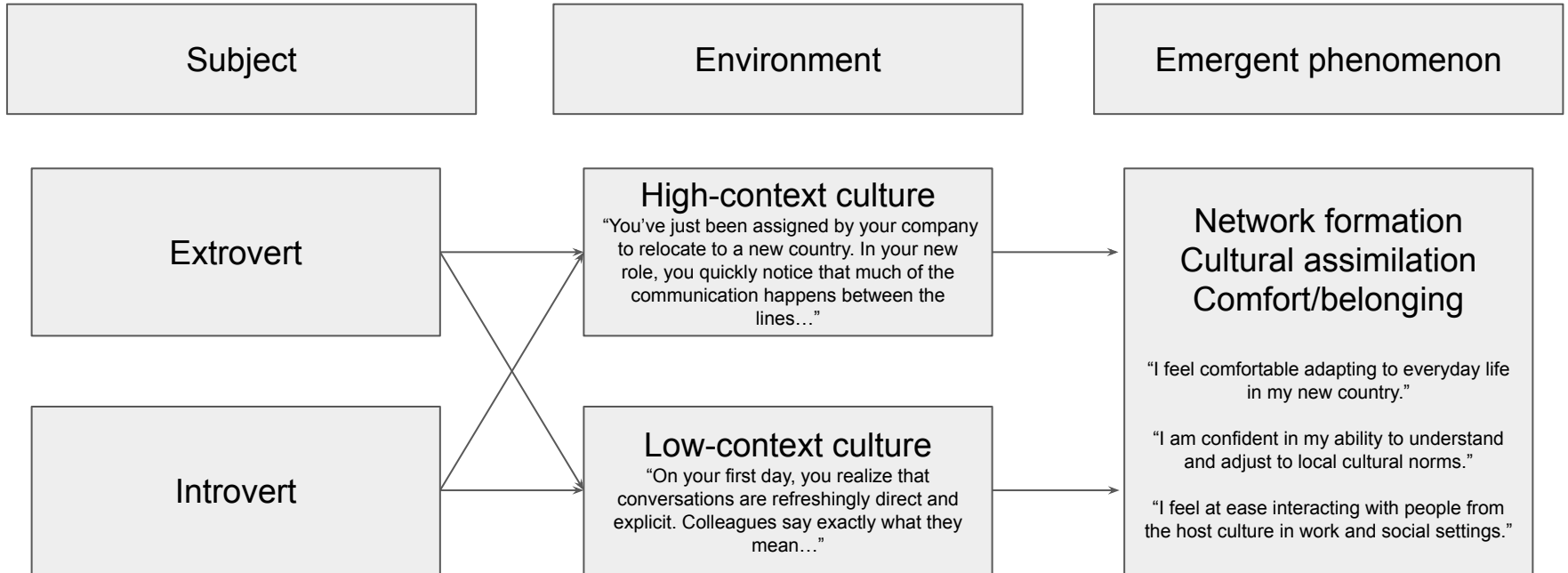
Introvert

Low-context culture

Network formation
Cultural assimilation
Comfort/belonging

```
graph LR; subgraph Subject; S1[Extrovert]; S2[Introvert]; end; subgraph Environment; E1[High-context culture]; E2[Low-context culture]; end; subgraph Emergent_phenomenon; EP[Network formation<br/>Cultural assimilation<br/>Comfort/belonging]; end; S1 --> E1; S1 --> E2; S2 --> E1; S2 --> E2; E1 --> EP; E2 --> EP;
```

Simplified simulation



[Martin] The commentary and horizontal/vertical integration of theories as it relates to personality psychology: Describe quickly the ideas of horizontal/vertical integration of theories, and offer a look at personality psychology and what h/v integration could look like. The class will hopefully immediately see that the validity problems.

Given the trajectory of model development, how do we imagine building personas can further improve?

- How might we use multi-modal models to strengthen persona building (ref. CoALA paper)?
- Reasoning/thinking features: How can we simulate human features/bugs like impulsivity?
- Constitutional alignment: Will this make it harder to create proxies for humans?