Personality -driven LLMS

What about susceptibility to disinformation?

Manuel Pratelli & Marinella Petrocchi

Understanding and Addressing Digital Inequalities
European University Institute
Badia Fiesolana, 20 -21 November, 2025

ISTITUTO DI INFORMATICA E TELEMATICA
Consiglio Nazionale delle Ricerche

Digital (In)Equality

Digital inequality

- Goes beyond structural factors — it is also **cognitive**
- Individuals **differ in their ability to process online information**
- Prior research shows **certain personality traits are more vulnerable to misinformation**

Why LLMs

- **Practical and ethical limitations** make it difficult to study these differences at population scale

Introduction

Today I'll report the findings of a recent study

Evaluating the Simulation of Human Personality -Driven Susceptibility to Misinformation with LLMs

(Pratelli & Petrocchi, European Conference on Artificial Intelligence 2025)

Foundation (**No LLM** involved in this phase)

1. Empirical research in ***psychology*** and ***cognitive science***
2. *Human participants* assessed using standardized personality tests
3. ***Susceptibility to disinformation*** analyzed as a function of personality

Research question

Can large language models,
when endowed with **explicit personality profiles**,
simulate **human susceptibility to misinformation**?

Image Source: courses.lumenlearling.org

Ethical and large -scale simulations of online human behaviours

Language models aligned with personality traits can simulate large -scale human behavior in response to misinformation

Potential Positive Applications

- Impact Estimation : Simulate how different personality profiles respond to disinformation —avoiding **costs, ethical risks, and privacy concerns** associated with human subjects
- Resilience Building : Design **personalized educational campaigns** by identifying content types most harmful to specific user profiles

Potential Misuses

- **Generation of fabricated, low -credibility conten** t tailored to exploit psychological vulnerabilities linked to personality traits

Personality inventories

What are they?

Standardized **questionnaires used to assess personality traits**

Format

– Self-report questionnaire using a ***Likert*** **scale (e.g., 1 to 5)**

– Respondents rate agreement with statements like:

- *" I like to participate in parties "*
- *" I worry easily "*

What do they measure?

An individual's position on ***personality dimensions***

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68*, 69–81

The five dimensions in the Big Five personality test

- **Openness to Experience** Measures curiosity, creativity, and open -mindedness toward new experiences and unconventional ideas.
- **Conscientiousness** Measures discipline, organization, reliability, and self -control.
- **Extraversion** Reflects sociability, energy, assertiveness, and a propensity for social contact.
- **Agreeableness** Involves empathy, cooperation, trust, and kindness
- **Neuroticism / Emotional Stability** Assesses tendency toward emotional reactivity: anxiety, instability, irritability.

- https://en.wikipedia.org/wiki/Big_Five_personality_traits?

What can you expect from a Big Five test?

- A **numerical profile for each of the five traits** , usually on a scale from low to high

- Interpretation based on scores: **e.g., high in extroversion, low in neuroticism**

- A **nuanced profile** : no "types" are defined, but each person is placed along continuous line for each trait, not in rigid categories.

What is being measured?

- **Susceptibility** to misinformation
- **Belief in misinformation** is the mean perceived accuracy assigned **to false headlines**

- **News Discernment**: the gap between the mean perceived accuracy of **true headlines** and that of **false ones**

$$ND_k = \frac{1}{n_{\text{true}}} \sum_{i=1}^{n_{\text{true}}} \text{Acc}_{ki}^{\text{true}} - \frac{1}{n_{\text{false}}} \sum_{j=1}^{n_{\text{false}}} \text{Acc}_{kj}^{\text{false}}$$

- Both metrics employ a perceived-accuracy prompt ***To the best of your knowledge, is this headline accurate?***

How personalities respond to misinformation?

338 US-based Mechanical Turks

A defined personality trait for each individual , based on Big Five inventory

12 **fake** headlines (Snopes.com)

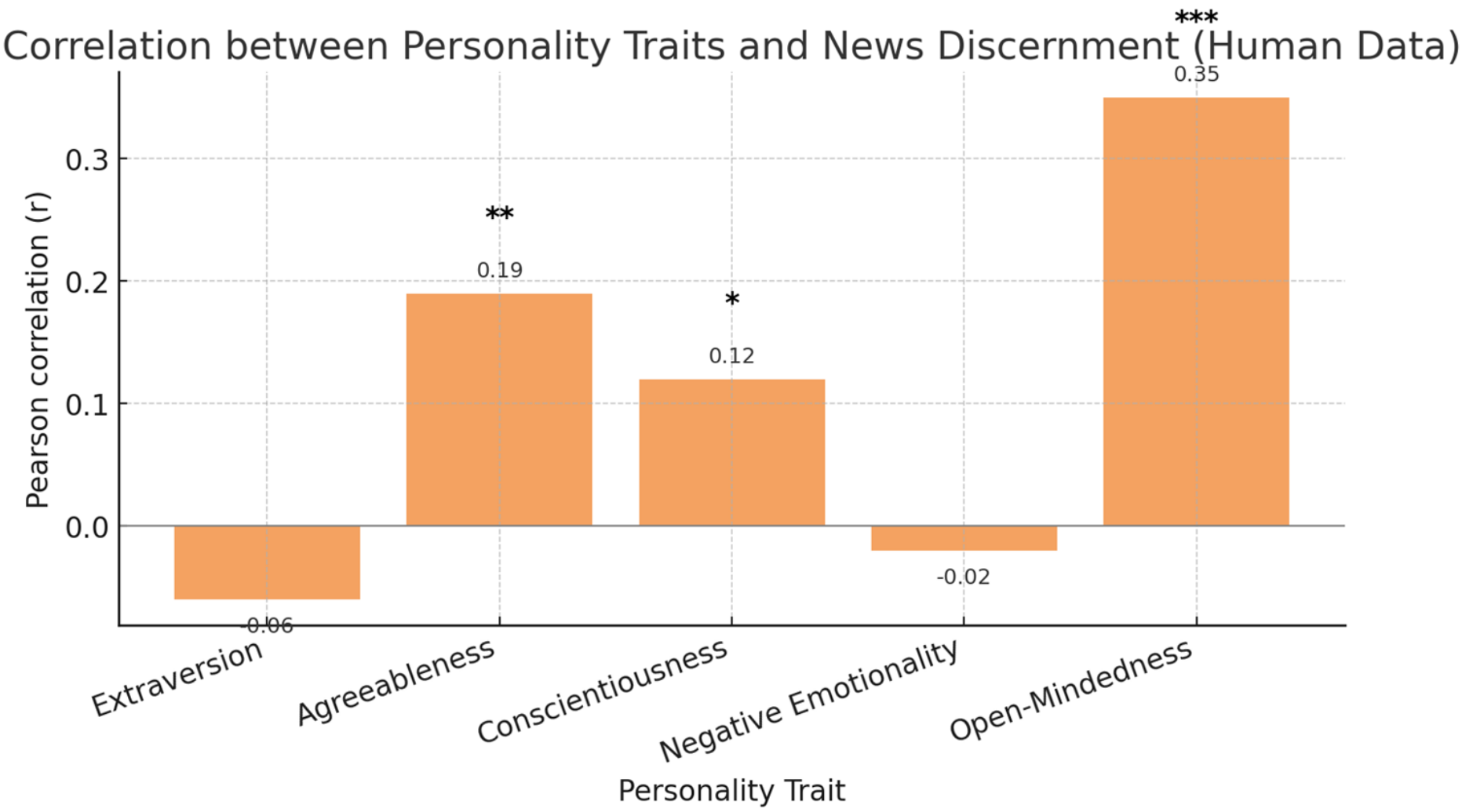
12 **true** headlines (NPR.org)

We can reproduce the experiment since we have:

- the **full set of participants' responses** to both the **personality inventory** and the **headline evaluation task**
- **the headlines**

Calvillo et al.. Personality factors and self-reported political news consumption predict susceptibility to political fake news. Personality and individual differences, 2021

How humans respond to misinformation



Calvillo et al, 2021: Personality factors and self-reported political news consumption predict susceptibility

How can we assign personalities to LLMs?

GPT 4o, 3.5 turbo

Your Assigned Personality

The number indicates **the extent to which you agree or disagree with each statement** .

1 means 'Disagree Strongly', 3 means 'Neutral', and 5 means 'Agree Strongly'

I am open to new experiences: **5**

I am helpful and unselfish with others: **5**

I get nervous easily: **2**

M. Huang et al. Designing LLM-agents with personalities: A psychometric approach. American Psychological Association Convention 2024

Questionnaire

- ### **Questionnaire Instruction** ###
- I will provide you with a **list of article headlines**
- Based on **your assigned human personality,** evaluate the perceived accuracy of each headline on a scale from 1 to 4, where

  1 represents 'not at all accurate' and **4 signifies 'very accurate'** .

Questionnaire

Questionnaire **Item**

Headlines:

1- **Trump Tweeted in 2009 That He Would 'Never Let Thousands of Americans Die from a Pandemic'**

…

…

All the **24 headlines** for each personality conditioned agent;

Balanced by political leaning (**12 pro-Republican, 12 pro-Democrat**) and veracity (**12 true and 12 false**)

Robustness

Robustness tested across multiple settings

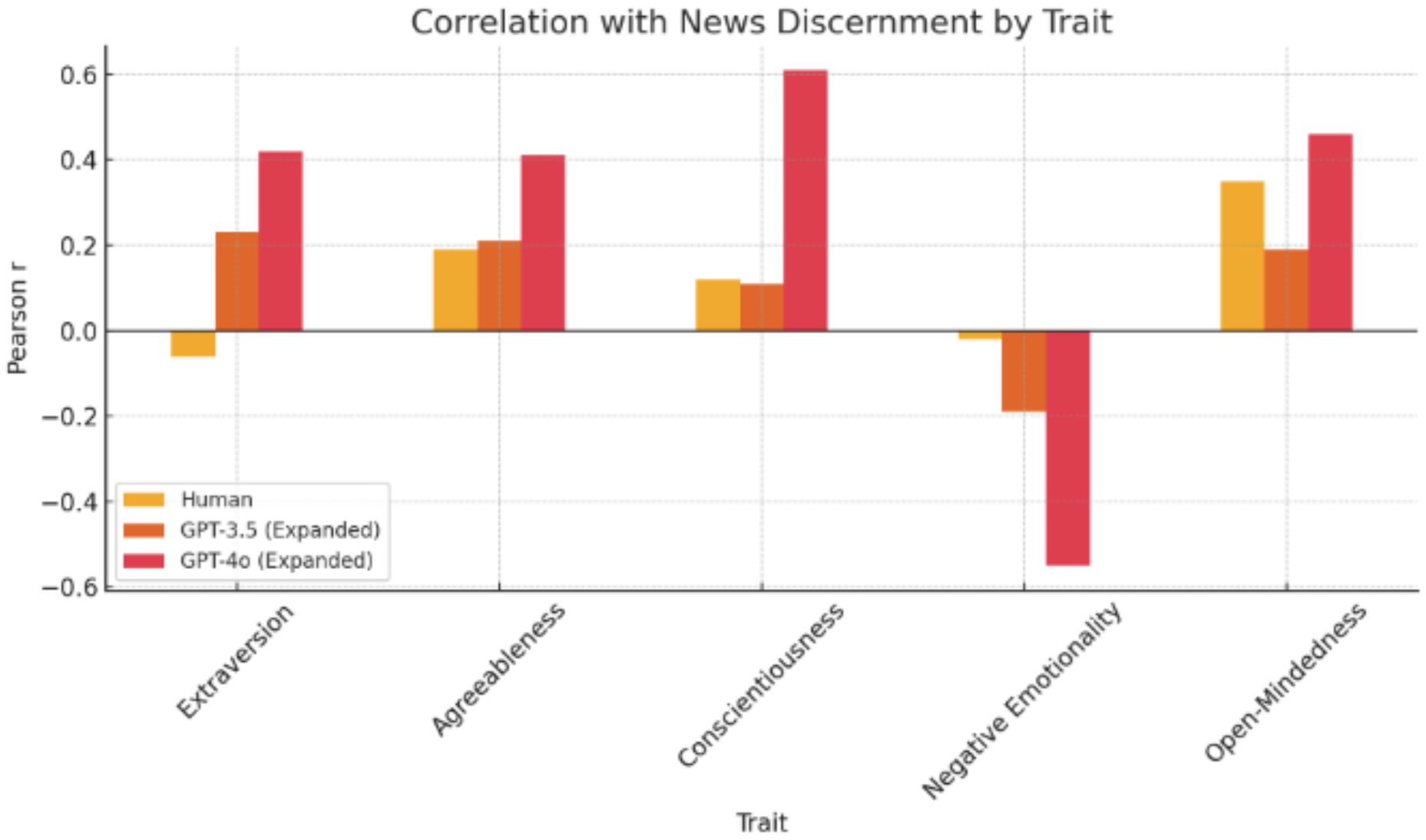
- Two model types (**GPT-3.5-turbo and GPT -4o**),
- Two personality inventories (**BFI-2-S and BFI -2**),
- Two response scales (**Likert and Expanded**),
- Two temperature values (**0.2 and 0.7**)

Image source: Sora

Main results

Agreeableness, Conscientiousness, and Open -mindedness appear to be **aligned** in **direction and significance** with humans

Extraversion and Neuroticism show significant effects in synthetic agents


Correlation with News Discernment by Trait
Pearson r
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Human
GPT-3.5 (Expanded)
GPT-4o (Expanded)
Extraversion
Agreeableness
Conscientiousness
Negative Emotionality
Open-Mindedness
Trait

Some conclusions

- Personality-aligned LLMs show **promising** for simulating human psychological variation.
- However, **model -specific biases** (e.g., toward extraversion or reduced emotionality) limit full fidelity.

Caution needed when using synthetic agents for policy or user studies.

Potentialities

The framework could allows **researchers and policymakers to explore** :

- How digital interventions (e.g., **fact -checking platforms or warning labels**) might perform differently across personality profiles;
- Which user traits **correlate with higher misinformation resilience or vulnerability** ;
- How **micro -targeted disinformation campaigns might exploit** individual psychological traits.

Research directions

- More **inventories**
- More **models**
- **Western** context
- Political ideology
- Media literacy
- Cognitive reflection
- ...

References

[1] Manuel Pratelli, Marinella Petrocchi. **Evaluating the Simulation of Human Personality -Driven Susceptibility to Misinformation with LLMs** . ECAI 2025

[2] M. Huang et al. **Designing LLM -agents with personalities: A psychometric approach** . Presented at American Psychological Association Convention 2024 (Seattle, USA).

[3] Calvillo et al. **Personality factors and self -reported political news consumption predict susceptibility to political fake news** . Personality and individual differences 174 (2021): 110666.

[4] Calvillo et al. **Personality and misinformation** . Current opinion in psychology

THANKS