

AI Systems as Social Scientific Objects

Research directions for PhD projects · DNDS

Max Pellert



[Slides as PDF]

How I got here

Cognitive science in Vienna and Ljubljana, economics before that; PhD in computational social science at the **Complexity Science Hub** and the Medical University of Vienna with David Garcia

Sony Computer Science Laboratories Rome: industry postdoc, NLP on news corpora, a data-driven art and science project at MAXXI

Assistant Professor at the University of Mannheim (Data Science in the Economic and Social Sciences), then **interim Professor (W2)** for Social and Behavioural Data Science at the University of Konstanz

Principal Investigator at the Barcelona Supercomputing Center, leading the COMPASS group; Associate Faculty at CSH Vienna

Joining **DNDS in January 2027**

What I actually work on

Classical computational social science: digital traces of humans, at scale (400M+ forum interactions, social media panels, surveys)

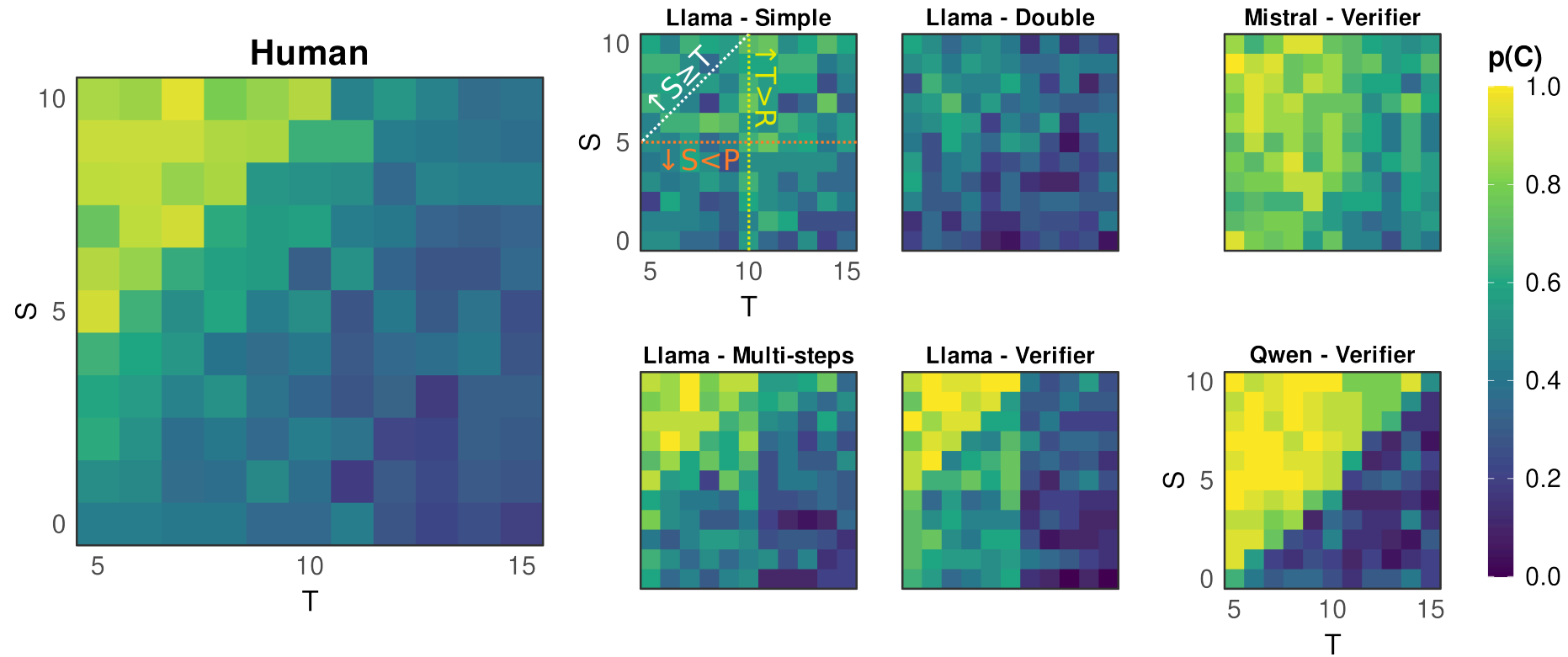
The data that matters for society is increasingly produced by **AI systems interacting with humans or with other AI systems**

My program: treat large language models as **objects of measurement**, with the rigor we apply to human subjects

Two complementary modes - *Behavioral*: present stimuli, record responses (games, surveys, classical experiments) - *Representational*: probe the internal geometry (embeddings, probing, steering)

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. *Perspectives on Psychological Science*. <https://doi.org/10.1177/17456916231214460>

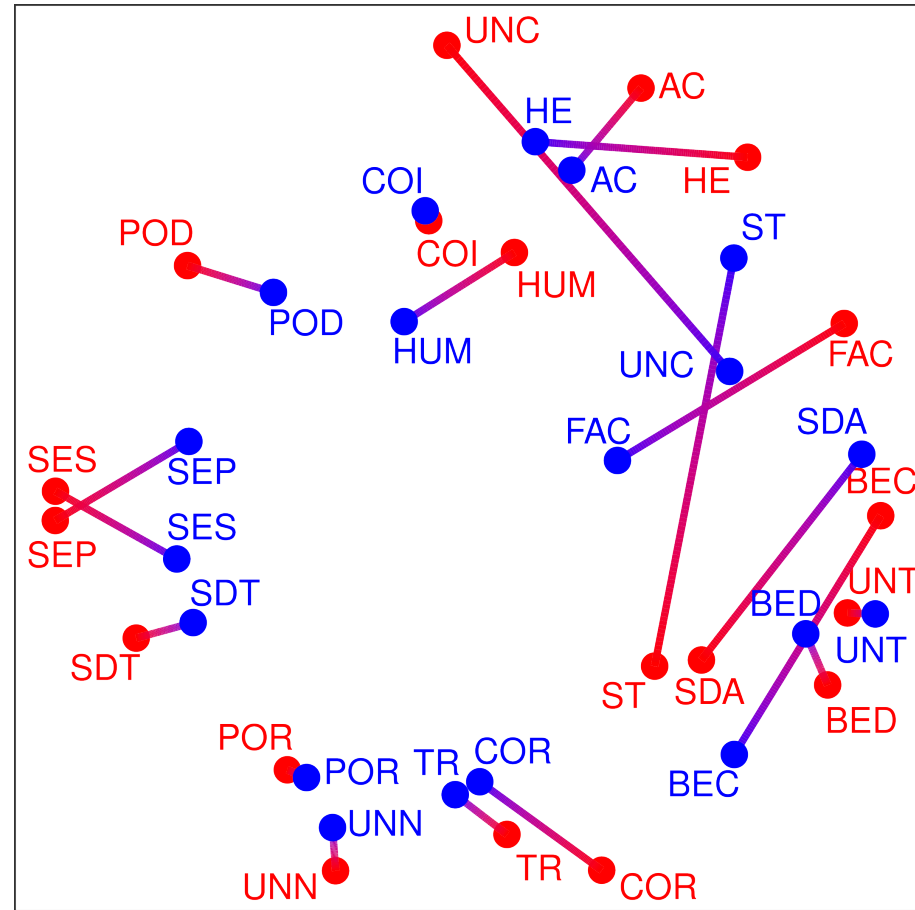
Behavioral mode: LLMs in game-theoretic experiments



121 two-player games covering the full space of social dilemmas, against a human baseline of 500+ participants; models show **stable, reproducible behavioral profiles**

Cera Palatsi, A., Martin-Gutierrez, S., Cardenal, A. S., & Pellert, M. (2026). Large language models replicate and predict human cooperation across experiments in game theory ([arXiv:2511.04500](https://arxiv.org/abs/2511.04500)), under review

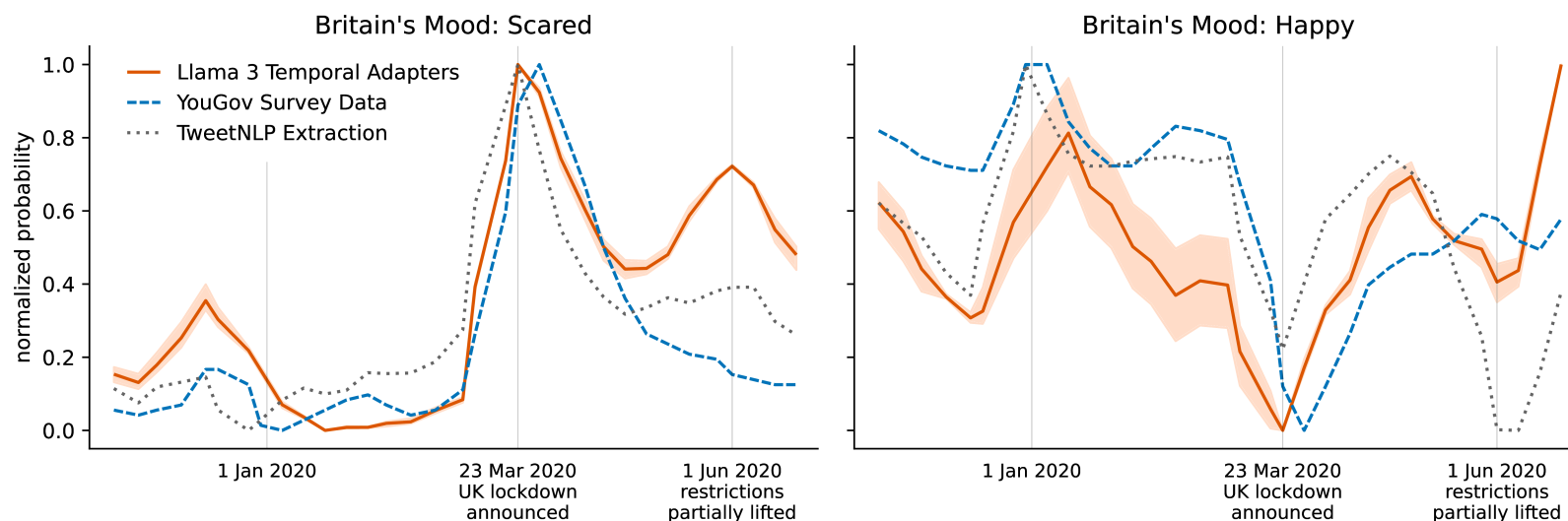
Representational mode: values from embeddings



Pellert, M., Lechner, C. M., Sen, I., & Strohmaier, M. (2026). Neural network embeddings recover value dimensions from psychometric survey items on par with human data. *Findings of the ACL: EACL 2026*, 5738–5752.

<https://doi.org/10.18653/v1/2026.findings-eacl.303>

And the empirical social data side



Synthetic surveys: time-indexed adapters on Llama recover weekly collective emotions, validated against representative YouGov data ([ICWSM 2025](#))

Signed networks: separating antagonism from alignment in online polarization, on a decade of news forum interactions ([PNAS Nexus 2024](#))

That corpus is **published and open** ([Scientific Data 2026](#), [BSC Dataverse](#)): threaded conversations, signed votes, topical tags.

Group, compute, funding

COMPASS at the Barcelona Supercomputing Center (COMplex Political And Social Simulation): four postdocs and one PhD researcher. From January I start building a group at CEU

Compute: large open-weight models run at scale on MareNostrum 5 at BSC; the pipelines transfer, and I know what it takes to get an experiment through a cluster queue; for CEU at Austrian Scientific Computing (ASC) for example

Funding already secured: DEMOS-AI (PI, Spanish national funding, decoding moral and sociopolitical biases in AI) and AI4SOCIALPLUS (Horizon Europe, PI of the use case on AI-assisted social simulation); a project with Mirta Galesic at the Complexity Science Hub in preparation

Supervision so far: six completed master's theses in Mannheim, Konstanz and Barcelona; three ongoing co-supervision of PhD researchers (TU Graz, University of Barcelona, Open University of Catalonia);

Four directions for PhD projects

1 · LLM populations as social systems

The question: cooperation between two LLM agents is measurable. What happens in groups, and on networks?

What exists: the dyadic games pipeline, behavioral phenotypes, pre-registered predictions, and first triadic simulations already run

What you would build: n-player and networked games (public goods, coordination, threshold games), varying interaction structure rather than only payoffs; comparison against evolutionary game theory and the human experimental literature

Also possible: run the human online experiment that validates our pre-registered predictions in game regions no human study has covered so far

Skills you would come out with: experimental design, large-scale inference on HPC, network science, evolutionary game theory

2 · Values inside models: measure, compare, intervene

Measure: extend the embedding-based recovery of value structure to more instruments, more languages, more model families

Compare: do value representations converge across models, or diverge? Convergence would imply hard limits on cultural differentiation through training data alone

Intervene: locate where value-relevant information is encoded (probing), steer it, and measure what changes downstream in tasks like hate speech classification, incivility detection, or political annotation

Where it sits: the core of DEMOS-AI, with Paula Szewach at BSC; the audit question (“whose values are the reference?”) has a legal deadline attached to it through the EU AI Act

Skills: psychometrics, interpretability, multilingual NLP, research on annotation and benchmark validity

(Especially this project could also involve an in-person visit at the Barcelona Supercomputing Center)

3 · Synthetic surveys and AI-assisted simulation

The instrument: models fine-tuned on the traces of a specific population and period, then administered a questionnaire, as a complement to survey data

The open questions are the interesting ones: when do synthetic respondents fail, which groups do they systematically misrepresent, how do we quantify uncertainty, what would a validity theory for this look like?

The applied side: LLM agents grounded in empirical findings through retrieval, used to generate hypotheses and pre-test survey and experiment designs before they cost money

Where it sits: funded through the AI4SOCIALPLUS use case I co-lead, with benchmarks against representative survey programmes

Skills: finetuning and evaluation, survey methodology, validation design, a healthy scepticism about silicon samples

(Especially this project could also involve an in-person visit at the University of Konstanz)

4 • Polarization with better instruments

The data: a decade of threaded news forum interactions with explicit positive and negative votes, already published and open

What exists: a [pipeline](#) that separates antagonism from alignment in signed networks, instead of collapsing polarization into one number

The new angle: measure the *content* flowing through those structures with LLM-based instruments, while testing whether those instruments carry political bias of their own ([ACL 2025](#))

Why it matters here: polarization, opinion dynamics and socioeconomic segregation are already central to this department. I would add the measurement layer and the signed structure

Skills: network science, NLP, temporal and causal analysis on observational data

How I work with PhD students

Regular one-on-one, and an open door in between. Generally, fewer longer meetings. I read drafts properly and I do my best to answer fast

First year shape: one self-contained project that gives you a working pipeline, a public repository and a paper draft, before we go after the ambitious thing

Open by default: pre-registrations where we make predictions, code on GitHub with a DOI, data in a repository (Dataverse for example).

I write **with** you, not for you.

I will tell you honestly when something is not working. All kinds of efforts of mine have been rejected, many times over the years, and you will hear about how that was handled

Networks matter to me: conferences (IC2S2, *ACL and similar), external stays, internships. Andrea Cera Palatsi went from a master's thesis with me at BSC to a PhD at the CSH Digital Innovation School. If you do good work, I am happy to recommend you or talk to people about you.

Come talk to me if

You want to combine a **social science question** with **AI systems** (very broadly) and care about *measurement* rather than benchmarks

Useful to have: R, Python, willingness to learn, some statistics

Not required: prior NLP research, deep learning from scratch, a finished idea

Probably not a fit: small application of your pet methodology with no social science question behind it

Do a bit of reading, bring a half-formed idea, that is enough to start

Max Pellert

maxpe@gmx.com

Department of Network and Data Science, CEU (from January 2027)

Barcelona Supercomputing Center · Complexity Science Hub Vienna

<https://mpellert.at>