Generative AI, Society, and Governance

Location: Technical University of Munich (HfP, Room H.414)
Instructor: Jan Zilinsky

The course will cover two main themes: first, it explores how social science disciplines can adopt and benefit from AI tools for research, analysis, and problem-solving; and second, the course explores the impacts of generative AI on society and economy, including methods for measuring and evaluating these impacts.

Students will evaluate how AI is transforming society and the economy, and explore how AI tools can be used in social science. Students will also conduct critical evaluations of AI companies' products and principles and learn techniques for auditing Large Language Models through adversarial testing, red teaming exercises, and systematic changes to system prompts. Throughout the course, we will emphasize the development of interdisciplinary understanding and critical thinking skills regarding the complex interplay between AI, society, and politics.

Pre-requisites: Familiarity with using APIs (e.g., via Python or R/RStudio). LLM-specific experience is not required. (Students are encouraged to develop a prototype of AI-enabled software or tools, but my wish is to keep the course accessible to students with different backgrounds and experiences; accordingly, there are various possibilities for final projects – see below.)

Objectives: Upon successful completion of this module, students should be able to: 1) Analyze the impact of artificial intelligence on society, economy, and governance using social science frameworks; 2) Evaluate the ethical implications and methodological challenges of AI-driven research in social sciences; 3) Apply AI tools to conduct research or design prototypes to solve problems (such preparing and evaluating an advice-giving application).

Teaching and learning method: Coding tutorials will be provided, and students will work in groups to propose projects where LLMs are applied to creatively address social problems, or to advance solutions to a social scientific problem.

Assessment: Project work (including a final presentation). **Evaluation metrics**: A presentation of the final project will be evaluated based on the demonstrated competence to design an AI tool or an AI-testing protocol (50%), the quality of the delivery (40%), and the quality of the discussion with the audience (10%).

Project options

Option 1: Evaluating AI Models as Political Advisory Tools

Project Framework:

Testing Phase

1. Test multiple AI models (e.g., GPT-4, Claude, Llama) with identical political questions
2. Create a standardized set of voter profiles with different political preferences. Design prompts that test for:

   - Consistency in advice
   - Political bias
   - Susceptibility to manipulation
   - Quality of advice (i.e., alignment, or the probability that the model encourages "correct voting")

Potential Analysis Categories

Accuracy and Bias Testing:

- Compare AI recommendations with official platforms of politicians of parties
- Consider testing for partisan bias by presenting identical scenarios with different political keywords
- Evaluate how models handle controversial topics
- How well are model taking into account the “personal” information it learns about a hypothetical user?

Manipulation Testing
(possible options - but be creative and create your own prompts)

- Test different prompting strategies to see if models can be made to give contradictory advice

- Attempt to "nudge" the models toward incorrect recommendations; e.g., can the model be "nudged" to commit mistakes, for example to suggest that a user who supports a bigger welfare state should vote for Trump? Or that somebody eager for tax cuts should vote Biden/Harris?
- Examine how models handle leading questions.

Information Quality:

- Propose your own evaluation metrics (this is an important skill to develop)
- Some ideas to get you started: accuracy of policy information, consistency of recommendations, resistance to manipulation, quality of reasoning provided, handling of nuanced positions, transparency about limitations

Practical Implementation

- Document the best practices for prompting to get reliable political advice
- Optional: Create a simple prototype of a voting advice application
- Consider testing the prototype with a small group of volunteers

Option 2: AI Truth Detector: Evaluating Language Models' Capacity for Fact-Checking

Instructions:

1. Dataset Creation

- Create a balanced dataset, writing a minimum of 300 statements, which you will use as rows in your dataset.
- Include both true and false statements. Categories to consider:
    - Historical facts
    - Scientific claims
    - Current events
    - Statistical statements
    - Common misconceptions

2. Model testing

3. For each statement, document:

- The model's response
- Confidence level (if provided)
- Reasoning given by the model

4. Key Considerations:

- Dataset Design:
    - Include diverse topics
    - Vary complexity levels
    - Mix obvious and subtle falsehoods
    - Include contemporary and historical claims
    - Document sources for true statements

- Testing Methodology:
    - Use consistent prompting across models
    - Document exact prompts used
    - Try to create a standardized evaluation rubric
    - Track response variations

Analysis and Evaluation

- Accuracy rates for each model
- Types of errors (false positives vs. false negatives)
- Think carefully about whether responses seem based on training data vs. reasoning

Option 3: Evaluating Reliability and Bias in AI Health Recommendations

Possible Project Structure:

1. Testing Framework Development

    - Create a set of medical scenarios ranging from mild to severe conditions

- Develop different user personas (e.g., "average citizen", traditional medicine believer, alternative medicine enthusiast, skeptic and contrarian, a citizen who is "fearful that health care is too expensive", etc.)
- Design a systematic testing protocol to ensure consistent evaluation

2. Potential Testing Categories:

    - Present straightforward medical scenarios to different AI models
    - Vary severity of symptoms or other attributes of the prompts.
    - Present identical scenarios using different user personas

3. Evaluation

    - Compare responses across models for consistency and accuracy
    - Document to what extent there is alignment with standard medical advice
    - Analyze how responses vary based on user communication style and their apparent (or stated) personality.
    - Discuss ethical implications

Option 4: AI Financial Mentor: Evaluating AI's Capability in Delivering Personalized Investment Guidance and Advice

To be discussed in class.

Option 5: Personalized Learning Companion

To be discussed in class.

Option 6: Independent, data-driven topic (discuss with your instructor).

Main Readings

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2024). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435. Link to version 10 (released on Oct. 2024).

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

OpenAI (2023). How should AI systems behave, and who should decide? https://openai.com/index/how-should-ai-systems-behave/

Using large language models in social science research: Promises and pitfalls

Burnham, M., Kahn, K., Wang, R. Y., & Peng, R. X. (2024). Political debate: Efficient zero-shot and few-shot classifiers for political text. *arXiv preprint arXiv:2409.02078*

- Paper
- Collab notebook

Mens, G. L., & Gallego, A. (2023). Scaling political texts with chatgpt. *arXiv preprint arXiv:2311.16639*. https://arxiv.org/pdf/2311.16639

Kathirgamalingam, Ahrabhi, Fabienne Lind, Jana Bernhard, and Hajo G Boomgaarden. 2024. “Agree to Disagree? Human and LLM Coder Bias for Constructs of Marginalization”. osf.io/preprints/socarxiv/agpyr.

Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics*, *11*(1), 20531680241236239.

Misinformation & Disinformation: AI as a disinfo threat or a reliable fact-checker?

Vykopal, I., Pikuliak, M., Srba, I., Moro, R., Macko, D., & Bielikova, M. (2023). Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838*.

DeVerna, M. R., Yan, H. Y., Yang, K. C., & Menczer, F. (2023). Artificial intelligence is ineffective and potentially harmful for fact checking. *arXiv preprint arXiv:2308.10800*.

Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., ... & Zagni, G. (2024). Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 1-12.

Kapoor, S., & Narayanan, A. (2024). We Looked at 78 Election Deepfakes. Political Misinformation Is Not an AI Problem. https://knightcolumbia.org/blog/we-looked-at-78-election-deepfakes-political-misinformation-is-not-an-ai-problem

Simon, F., McBridge, K., & Altay, S. (2024). AI's impact on elections is being overblown https://www.technologyreview.com/2024/09/03/1103464/ai-impact-elections-overblown/

Experiments

Goel, Natasha et al. 2024. "Artificial Influence? Comparing AI and Human Persuasion in Reducing Belief Certainty". osf.io/2vh4k.

Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., ... & Reinecke, K. (2024). Biased ai can influence political decision-making. *arXiv preprint arXiv:2410.06415*.

Velez, Y.R,, & Liu, P. (2024). Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments. *American Political Science Review*. doi:10.1017/S0003055424000819

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814.

Audits, adversarial testing and red teaming

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Brewsterm J. & Sadeghi, M., Red-Teaming Finds OpenAI's ChatGPT and Google's Bard Still Spread Misinformation. https://www.newsguardtech.com/special-reports/red-teaming-finds-openai-chatgpt-google-bard-still-spread-misinformation/

Kapoor, S., & Narayanan, A. (2023). Evaluating LLMs is a minefield. https://www.aisnakeoil.com/p/evaluating-llms-is-a-minefield