TECHNION Israel Institute of Technology | FACULTY OF DATA AND DECISION SCIENCES MAKING SENSE OF IT | THE WILLIAM DAVIDSON MBA

MBA – Focusing on Entrepreneurship, Innovation & Technology Management

Syllabus

Large Language Models and Biases in Social Media

Mini 1, 2025
Bloomfield - 214
Fridays, 09:00 -13:00

Instructor: Aviv Peleg workpeleg@gmail.com
Office Hours: by appointment
Credits: 2 pt
Study hours per week: 4 hours per week in attendance

Course Goals and Description

This course explores the intricate relationship between social media platforms like LinkedIn, decision-making processes, and the emerging role of Large Language Models (LLMs). Through a blend of theoretical insights and practical case studies, we will investigate inherent biases in social networks and LLMs, their societal and individual impacts, and approaches to mitigate these challenges. The course aims to develop a critical understanding of the role of advanced AI systems in shaping decisions and perceptions.

Learning Outcomes

By the end of this course, students will:

1. Understand the inherent biases in social media networks and their business implications.
2. Gain foundational knowledge of LLMs, their types, and their role in user interaction.
3. Explore current research on biases in LLMs.
4. Critically evaluate the ethical and business implications of LLMs.

Course Content/Topics

The course is divided into five main modules, each addressing a key aspect of the topic:

Module 1: Social Media and Decision-Making Biases

Topics Covered:

- Business overview of the social media industry, including the economics of influencer marketing.
- Introduction to inherent issues in social networks, with an emphasis on LinkedIn.
- Examination of cognitive biases and their influence on user behavior in social media.

Key Readings:

- "When less is more: the impact of macro and micro social media influencers"
- " A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it"
- "A Perfect Storm: Social Media News, Psychological Biases, and AI"

Module 2: Understanding Large Language Models (LLMs)

Topics Covered:

- Business overview of the AI industry powering LLMs.
- **Introduction to LLMs (Part 1):** Fundamentals such as tokens, parameters, pretraining, fine-tuning, prompts, context window, zero-shot, few-shot, chain-of-thought, reinforcement learning with human feedback (RLHF), temperature, bias and fairness, overfitting, hallucination, and APIs.
- **Introduction to LLMs (Part 2):** Types (e.g., ChatGPT, Gemini), engine architecture (e.g., GPT, LaMDA), various model types, and user interaction methods.
- Ethical considerations, including potential misuse, deepfakes, and the impact on employment.
- Limitations of LLMs and areas that remain poorly understood.

Key Readings:

- "Can Large Language Models Be an Alternative to Human Evaluations?"
- "Grade Score: Quantifying LLM Performance in Option Selection"
- "Same Task, More Tokens: The Impact of Input Length on LLM Performance"
- "Aligning large language models with human preferences through representation engineering."

Module 3: Biases in LLM Algorithms

Topics Covered:

- Examination of issues related to religion, race, political preferences, and gender biases in LLMs.
- Case studies highlighting failures of LLMs due to cognitive biases.

Key Readings:

- "Capturing Failures of Large Language Models via Human Cognitive Biases"
- "Cognitive Bias in Decision-Making with LLMs"
- "Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers"
- "Large Language Models Are Inconsistent and Biased Evaluators"
- "Cognitive Biases in Software Engineering"

Module 4: Mitigating Biases Through Prompt Engineering

Topics Covered:

- Current methods for refining LLM outputs to reduce mistrust and improve performance.
- Ethical and technical considerations in bias mitigation.

Key Readings:

- "Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception"
- "Cognitive Biases in Natural Language: Automatically Detecting, Differentiating, and Measuring Bias in Text"
- “A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications”
- "Firms of Endearment" (for business context and implications)

Module 5: Analyzing Social Media Content

Topics Covered:

- Comparison of human versus LLM evaluations: advantages and disadvantages.
- Techniques for analyzing social media content using the models discussed in previous modules.

Key Readings:

- "Can Large Language Models Be an Alternative to Human Evaluations?"
- "Large Language Models Are Inconsistent and Biased Evaluators"

MBA – Focusing on Entrepreneurship, Innovation & Technology Management

Assignments and Grading Procedures

1. **Participation and Discussions (10%)**
   Active engagement in class discussions and debates. Students are expected to contribute thoughtful insights, demonstrate an understanding of the readings, and engage respectfully with peers. Participation will be evaluated weekly.
2. **Case Study Analysis / Project (90%)**
   Students are expected to produce an in-depth analysis of a chosen topic related to biases in social networks or LLMs, and they may use and reference ideas and data from the "Key Readings" section of each module:
   - **Interim Submission (10%)**: A proposal outlining the chosen case study / Project, key questions, and a preliminary plan for analysis.
   - **Final Report (60%)**: A comprehensive written analysis that addresses one of the three case study options / project outlined below or an instructor-approved creative alternative.
   - **Presentation (20%)**: A summary presentation of the case study / project to the class, highlighting key findings and proposed solutions.

Option 1: Evaluating Bias in Influencer Marketing Content

- **Overview:**
  Explore the role of bias in influencer marketing content by comparing human-generated content with content produced by LLMs. Focus on identifying cognitive and representational biases that may be embedded in the messaging.
- **Objectives:**
  - Utilize appropriate LLM techniques, identify and measure bias in influencer marketing posts
  - Assess the impact of biased messaging on consumer decision-making and brand reputation.
  - Propose technical and ethical strategies for mitigating bias.
- **Expected Outcomes:**
  - A brief comparative analysis report (4–6 pages) featuring visualizations (e.g., charts or graphs) that illustrate bias metrics along with actionable recommendations.

MBA – Focusing on Entrepreneurship, Innovation & Technology Management

Option 2: Bias in Automated Content Moderation on Social Media Platforms

- **Overview:**
  Examine how automated content moderation systems, often powered by LLMs, may inadvertently introduce or reinforce bias in social media environments. Analyze both the technical mechanisms and the ethical implications of these moderation practices.
- **Objectives:**
  - Identify and analyze bias patterns in automated moderation outputs across social media platforms.
  - Assess the impact of these biases on user communities, with attention to marginalized groups.
  - Propose technical interventions or prompt engineering strategies to improve fairness in moderation.
- **Expected Outcomes:**
  - A brief report (4–6 pages) that includes an analysis of sample moderated content, a discussion of observed biases, and actionable recommendations for improvement.

Option 3: Ethical and Technical Evaluation of AI Governance in Social Media Platforms

- **Overview:**
  Analyze current regulatory and governance frameworks addressing the deployment of AI and LLMs in social media. Evaluate how these frameworks handle bias and propose recommendations for enhanced AI governance (e.g., in areas such as facial recognition in law enforcement and content moderation on social media).
- **Objectives:**
  - Critically assess existing policies and governance practices regarding AI use.
  - Examine case studies where governance shortcomings have led to biased outcomes.
  - Recommend enhancements to current policies that integrate both technical evaluations and ethical considerations.
- **Expected Outcomes:**
  - A concise policy analysis report (4–6 pages) that includes a review of case studies, a critical evaluation of current governance frameworks, and clear recommendations for improvement.

Project: Leveraging LLMs for Bias Detection and Self-Assessment

- **Overview:**
  This project examines the dual role of LLMs in bias analysis. You will explore how LLMs can be used as tools to detect bias in social media or other text-based content, and simultaneously evaluate and mitigate the inherent biases that may exist in the LLMs' own outputs.
- **Objectives:**
  - **Bias Detection:**
    - Assess the capability of LLMs to identify and quantify bias in selected texts.
    - **(**Optional**)** Apply both qualitative (e.g., thematic coding) and quantitative (e.g., keyword frequency, sentiment analysis) methods to measure bias in a sample dataset.
  - **Self-Assessment of LLM Bias:**
    - Critically analyze the output generated by the LLM for signs of bias.
    - Evaluate how prompt design and model tuning can influence the impartiality of LLM-generated responses.
  - **Mitigation Strategies:**
    - Propose prompt engineering techniques or fine-tuning adjustments to reduce any detected bias within the LLM.
    - Formulate recommendations for ensuring that LLMs used for bias detection do not perpetuate bias in the process.
- **Expected Outcomes:**
  - A brief report (4–6 pages) that summarizes your findings regarding the LLM's effectiveness in bias detection and discusses the presence of any inherent biases.

Grading Criteria for Case Study Analysis:

- **Depth of Analysis (40%)**: Demonstrates a thorough understanding of the subject matter and critical thinking.
- **Research Integration (20%)**: Effectively incorporates academic and industry research.
- **Clarity and Organization (20%)**: Presents ideas in a logical, clear, and well-structured format.
- **Creativity and Innovation (10%)**: Proposes novel or well-thought-out solutions to identified issues.
- **Presentation Skills (10%)**: Engages the audience effectively and communicates findings concisely.

MBA – Focusing on Entrepreneurship, Innovation & Technology Management

Course Schedule (Topics, assignments, Exams)

Session	Date	Topic(s)	Submissions
1	31/10/25	Module 1: Social media and Decision-Making Biases	
2	07/11/25	Module 2: Understanding LLMs (part 1)	
3	14/11/25	Module 2: Understanding LLMs (part 2)	
4	TBD (Zoom)	Module 3: Biases in LLM Algorithms (part 1)	Interim Submission
5	21/11/25	Module 3: Biases in LLM Algorithms (part 2)	
6	28/11/25	Module 4: Mitigating Biases Through Prompt Engineering	
7	5/12/25	Module 5: Analyzing Social Media Content	
8	12/12/25	Final Presentation	

Course Requirements & Course Policies

Attendance: Attendance is mandatory, and a laptop is required for certain sessions to complete assignments and actively participate in class.

Absences: Should a student become unable to attend a class, please notify me in advance. Should any student find themselves unable to complete an assignment or fulfill a course requirement, please notify me in advance.

Text book(s) and/or other materials

Students may either use the provided citations or directly search for the articles by title within the "Key Readings" section of each module.

- Kay, Samantha, Rory Mulcahy, and Joy Parkinson. "When less is more: the impact of macro and micro social media influencers' disclosure." *Journal of marketing management* 36.3-4 (2020): 248-278.
- Rodrigo-Ginés, Francisco-Javier, Jorge Carrillo-de-Albornoz, and Laura Plaza. "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it." *Expert Systems with Applications* 237 (2024): 121641.
- Datta, Pratim, Mark Whitmore, and Joseph K. Nwankpa. "A perfect storm: social media news, psychological biases, and AI." *Digital Threats: Research and Practice* 2.2 (2021): 1-21.
- Liu, Wenhao, et al. "Aligning large language models with human preferences through representation engineering." *arXiv preprint arXiv:2312.15997* (2023).

- Koo, Ryan, et al. "Benchmarking cognitive biases in large language models as evaluators." *arXiv preprint arXiv:2309.17012* (2023).
- Chiang, Cheng-Han, and Hung-yi Lee. "Can large language models be an alternative to human evaluations?." *arXiv preprint arXiv:2305.01937* (2023).
- Jones, Erik, and Jacob Steinhardt. "Capturing failures of large language models via human cognitive biases." *Advances in Neural Information Processing Systems* 35 (2022): 11785-11799.
- Echterhoff, Jessica, et al. "Cognitive bias in decision-making with LLMs." *Findings of the Association for Computational Linguistics: EMNLP 2024*. 2024.
- Mohanani, Rahul, et al. "Cognitive biases in software engineering: A systematic mapping study." *IEEE Transactions on Software Engineering* 46.12 (2018): 1318-1339.
- Atreides, Kyrtin, and David J. Kelley. "Cognitive biases in natural language: Automatically detecting, differentiating, and measuring bias in text." *Cognitive Systems Research* 88 (2024): 101304.
- Paech, Samuel J. "Eq-bench: An emotional intelligence benchmark for large language models." *arXiv preprint arXiv:2312.06281* (2023).
- Buolamwini, Joy Adowaa. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. Diss. Massachusetts Institute of Technology, 2017.
- Firms of Endearment - Book
- Iourovitski, Dmitri. "Grade Score: Quantifying LLM Performance in Option Selection." *arXiv preprint arXiv:2406.12043* (2024).
- Lin, Luyang, et al. "Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception." *arXiv preprint arXiv:2403.14896* (2024).
- Stureborg, Rickard, Dimitris Alikaniotis, and Yoshi Suhara. "Large language models are inconsistent and biased evaluators." *arXiv preprint arXiv:2405.01724* (2024).
- Levy, Mosh, Alon Jacoby, and Yoav Goldberg. "Same task, more tokens: the impact of input length on the reasoning performance of large language models." *arXiv preprint arXiv:2402.14848* (2024).
- Sahoo, Pranab, et al. "A systematic survey of prompt engineering in large language models: Techniques and applications." *arXiv preprint arXiv:2402.07927* (2024).