Lubo Litov (Price College of Business, OU College of Law, Univ. of Oklahoma & WFIC, Univ. of Pennsylvania)

Third N3BG Summer School 2025

Can AI Machines Detect Misinformation?

Paper #1: Artificial Intelligence in the Battle Against Disinformation *(Saeidnia et al, Knowledge & Information Systems, 2025)*

- Systematic review covering 2014-2024, 76 studies in total.
- Punchline: AI technologies, particularly machine learning and human-in-the-loop systems (HILS), have become central tools in combating misinformation, but they face significant ethical, technical, and implementation challenges.
- Analyzes AI-driven techniques like fact-checking.
- Includes sentiment analysis approaches.
- Highlights human-in-the-loop methodologies.
- Discusses algorithmic *bias* & addresses ethical concerns.
- Suggests interdisciplinary research is needed.

Artificial Intelligence in the Battle Against Disinformation *(Saeidnia et al, Knowledge & Information Systems, 2025) (continued)*

- Strengths of AI models in tackling misinformation:
  - **Adaptability** (can be continuously updated & improved to keep up with evolving tactics in misinformation).
  - **Automation** (can automate the process of identifying/flagging potential misinformation, hence faster response).
  - **Pattern recognition** (can identify patterns in data indicative of misinformation).
  - **Real-time monitoring** of online platforms.
  - **Scalability** (quick & efficient analysis of large amounts of data, easily detects misinformation).

Artificial Intelligence in the Battle Against Disinformation *(Saeidnia et al, Knowledge & Information Systems, 2025) (continued)*

- Limitations of AI models in tackling misinformation:
  - **Bias** (can manifest bias based on data they are trained on, hence inaccurate or unfair detection of misinformation).
  - **Context Understanding** (may struggle to understand complex details, leading to misinterpretation).
  - **No transparency** of the model (many AI algorithms are black boxes, difficult to understand decision-making process).
  - **Over-reliance** (risk of over-relying on AI in combating misinformation, hence neglect of human judgement & critical thinking).

Artificial Intelligence in the Battle Against Disinformation *(Saeidnia et al, Knowledge & Information Systems, 2025) (continued)*

- Challenges of AI models in tackling misinformation:
  - **Adaptability** (misinformation tactics evolving constantly, requiring AI systems quick adaptability).
  - **Bias** (AI systems prone to bias based on data they are trained on).
  - **Context understanding** of the model (AI systems may struggle to understand the complex nuances and context).
  - **Scale** (with the vast amount of information shared online, AI systems may struggle to keep up with the volume of data, i.e., the issue of propaganda crowding out AI systems' capacity to analyze and detect misinformation).

Artificial Intelligence in the Battle Against Disinformation *(Saeidnia et al, Knowledge & Information Systems, 2025) (continued)*

- Ethical Considerations of AI models in tackling misinformation:
  - **Accountability** (comprehensive mechanisms to hold developers and deployers of AI systems accountable).
  - **Mitigating bias** (identify & mitigate biases in AI algorithms for fair and accurate detection of misinformation).
  - **Freedom of speech** (balance misinformation prevention & freedom of speech to avoid censorship).
  - **Privacy** (maintain user privacy to promote collection & analysis of data and to comply with private regulation).
  - **Transparency** (AI systems need to be transparent in their decision-making process to build trust and to ensure accountability).

On AI-Related Models Disclosure

- Several U.S. states have now introduced/ enacted legislation related to AI regarding transparency & accountability, including AI-generated content and system assumptions. For example:
- The state of *California* introduced **Bot Disclosure Law** (SB 1000) requiring bots (both commercial or political purposed) to identify themselves as non-human. Moreover, proposed **SB 1047** would require AI systems developers to disclose safety precautions & use assumptions.
- The state of *Colorado* introduced the **Colorado Artificial Intelligence Act** (SB 205, 2024) requiring disclosure of AI use, the nature of data used, and any risks (including biases or inaccuracies). Will be enacted in February 2026.

On AI-Related Models Disclosure (continued)

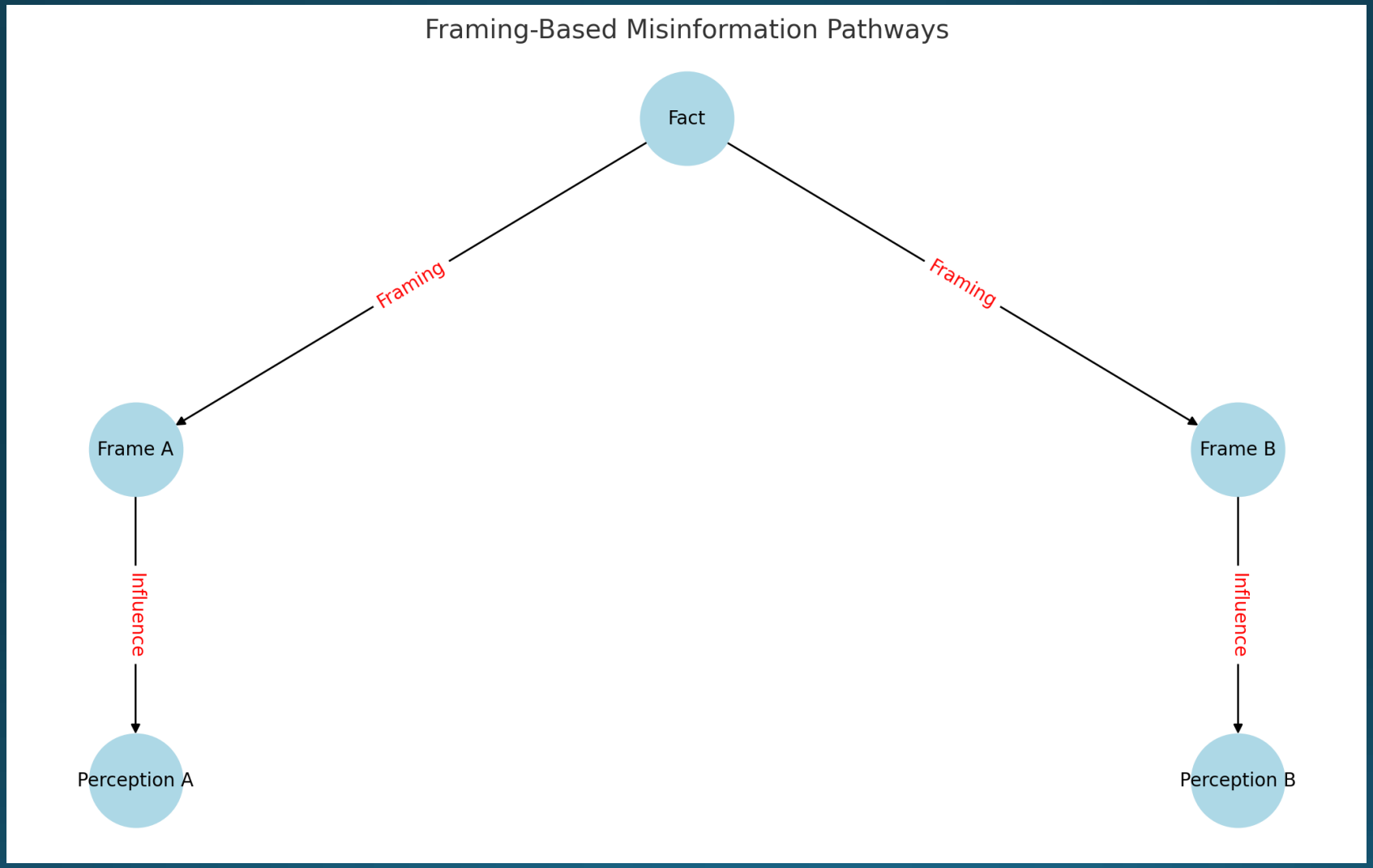
- The state of *Illinois* introduced **AI Video Interview Act** (employers must notify applicants if AI is used in video interviews). Further introduced the **Biometric Information Privacy Act** (BIPA) that regulates AI use in handling biometric data.
- The state of *Texas* proposed legislation requiring AI accountability in public agencies & transparency about automated decision-making tools use.
- The state of *New York* proposed a bill to regulate the use of AI in hiring & employment decisions.

Paper #2: Machine Learning Strategies for Fake News Detection *(Mouratidis et al., Information, 2025)*

- Thesis: A *hybrid* machine learning approach combining *NLP (e.g., tokenization, part-of-speech tagging)* and *deep learning (e.g., a convolutional neural network and long short-term memory network)* can effectively detect fake news across different media formats by improving model accuracy and interpretability.
- Integrates NLP and deep learning techniques.
- Detects fake news in social media and news articles.
- Highlights model interpretability.
- Tests robustness against adversarial attacks.
- Focuses on accuracy and efficiency.
- Proposes improvements in scalability.
- Evaluates effectiveness across platforms.
- Recommends enhancements in data quality.

Paper #3: Detecting Misinformation Through Framing Theory *(Wang et al., arXiv, 2024)*

- Punchline: Misinformation can be detected not only through factual inaccuracy but also by identifying *manipulative framing structures* using large language models and narrative analysis.
- Utilizes framing theory and large language models.
- Identifies misinformation via narrative framing.
- Detects misleading narratives despite factual correctness.
- Evaluates semantic manipulation.
- Proposes a frame element-based approach.
- Highlights implications for news verification.
- Analyzes subtle biases in media.
- Suggests integration with automated fact-checkers.



Frame A: Applies a *positive* spin (e.g., "Unemployment drops to 5%")

Frame B: Applies a *negative* spin (e.g., "Millions still jobless despite recovery").

Paper #4: Explainable Misinformation Detection on Social Media *(Joshi et al., IEEE Access, 2023).*

- Thesis: Cross-platform misinformation detection can be improved by integrating explainable AI and domain adaptation to enhance transparency and reliability. Use the LIME *(Local Interpretable Model-agnostic Explanations)* to highlight the most influential features (e.g., words or phrases).
- Combines explainable AI and domain adaptation (domain awareness).
- Case study focused on COVID-19 misinformation.
- Targets multiple platforms like Twitter & Facebook.
- Emphasizes model transparency.
- Addresses cross-platform data challenges.
- Discusses dataset diversity.
- Highlights user interface integration.
- Advocates ethical AI usage.

Paper #5: Towards Reliable Misinformation Mitigation with GPT-4 *(Pelrine et al, EMNLP, 2023)*

- Punchline: Large language models like GPT-4 can significantly enhance misinformation detection by leveraging their ability to *generalize* across contexts and manage uncertainty, outperforming previous models and offering a more reliable foundation for practical mitigation tools.
- Studies GPT-4's misinformation detection ability.
- Addresses generalization and uncertainty.
- Proposes novel evaluation metrics.
- Highlights dataset diversity.
- Identifies GPT-4's limitations.
- Recommends continuous model updates.
- Suggests transparency in predictions.
- Focuses on practical reliability.

Towards Reliable Misinformation Mitigation with GPT-4 *(Pelrine et al, EMNLP, 2023) (continued)*

Method	Accuracy	F1
SOTA (2022; 2021)	62	-
GPT-4 Score Optimized	**68.2**	**68.1**
GPT-4 Score Zero-Shot	64.9	60.9
GPT-4 Binary	66.5	66.5
RoBERTa-L Binary	63.5	62.1
RoBERTa-L 6-way	64.7	64.1
BERT	65.0	64.5
ConvBERT	66.7	65.8
DeBERTA	63.0	63.8
DeBERTA-V3	65.0	64.4
LUKE	65.3	64.4
RoBERTa	64.7	64.2
SqueezeBERT	63.1	62.2
XLMRoBERTA	61.1	61.0
Fuzzy (Word2Vec)	60.2	60.4
Fuzzy (BERT)	59.6	60.1
Fuzzy (GloVe)	60.1	59.7

Table 2: Binary Classification Results (percentages). GPT-4 shows superior performance. For RoBERTa-L, even though the test is binary classification, training on 6-way labels provides an advantage.

Paper #6: Defending Democracy: Deep Learning to Prevent Misinformation *(Trivedi et al., arXiv, 2021)*

- *Thesis:* Deep learning models, particularly BERT and graph-based methods, can be leveraged to trace and mitigate the spread of misinformation on social media, thus supporting democratic integrity.
- Text classification is a long-standing problem in NLP.
- Employs Bidirectional Encoder (BERT) and propagation graph methods to analyze LIAR database (Politifact.com API).
- Analyzes misinformation spread on Twitter.
- Visualizes spread patterns.
- Proposes real-time detection solutions.
- Recommends integrated AI-driven interventions.

Paper #7: AI in Automated Disinformation Detection *(Santos, Journalism and Media, 2023)*

- *Punchline:* A thematic understanding of AI applications in disinformation detection reveals promising directions in linguistic analysis and blockchain integration, though practical implementation remains nascent.
- Evaluates linguistic analytical methods.
- Discusses blockchain integration (more on it below).
- Analyzes platform-specific effectiveness.
- Highlights ethical considerations.
- Suggests real-time application feasibility.
- Recommends proactive monitoring strategies.

Paper #7: AI in Automated Disinformation Detection
(Santos, Journalism and Media, 2023) (continued)

- The paper discusses blockchain as a way to **verify the origin and integrity of content**, helping distinguish authentic information from manipulated or fake data.
- It suggests combining blockchain with AI to **create immutable records of media provenance**, enhancing transparency in information dissemination.
- This integration could **support decentralized, tamper-resistant frameworks** for disinformation detection and trust verification.

Newly Available Programming Tools to Detect Misinformation in Common Social Media Platform (Twitter-based verification tools)

- Hoaxy
  - Visualizes the spread of misinformation and fact-checks Twitter content.
  - Shows a graph of Twitter accounts (based on keyword search) illustrating accounts that amplify misinformation.

- Botometer
  - Analyzes Twitter handles for Tween contents and frequency, sentiment & linguistic patterns.
  - Using machine learning models, assigns a bot likelihood score.

Conclusion

- AI technologies like NLP, deep learning, and LLMs are central to detecting misinformation across platforms.
- Recent studies reveal unique advantages of hybrid models, framing theory, and explainable AI in tackling fake news.
- Despite progress, challenges like algorithmic bias, limited context understanding, and ethical risks persist.
- State-level regulations (e.g., CA, CO, IL) are mandating AI disclosure, transparency, and accountability.
- Tools like Hoaxy and Botometer help visualize misinformation spread and detect automated sources online.
- Can AI machines detect misinformation? Not quite yet!