AI Is Here To Stay: Misinformation and Human-Centric Models Between Risks and Opportunities

MARCO BELLÒ, University of Padua, Italy

ACM Reference Format:
Marco Bellò. 2025. AI Is Here To Stay: Misinformation and Human-Centric Models Between Risks and Opportunities. 1, 1 (January 2025), 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Artificial intelligence has fascinated the scientific community for almost a century, spurring famous research papers such as Alan Turing's *"Computing Machinery and Intelligence"* in 1950 [48], which introduced the *imitation game*. The idea, trivialized, is that any machine capable of fooling a person into thinking it's speaking to a human can be considered sentient. For seventy-three years the game remained unbeaten, until OpenAI's ChatGPT-4 ultimately succeeded in 2023 [2]. The model, simulating AGI capabilities [5], is one of the last iterations of the Generative Pre-Training LLMs[1] pioneered by OpenAI in 2018 (at the moment of writing the latest available is GPT-5.2) [34], which closely followed the first breakthrough towards human-like agents: *"Attention Is All You Need"* [50] is a 2017 landmark research paper authored by eight Google researchers that introduced the *transformer* architecture, considered the backbone of all modern LLMs and the main contributor of the AI boom [24].

Computer scientists are not the only ones engrossed in the topic: philosophers involved themselves too, most notably Jhon Searle and his 1980s' *chinese room* thought experiment, which directly challenged Turing's ideas and refuted the possibility of true machine intelligence [46], and even the general public showed great interest once AIs became smart enough: ChatGPT reached one million users in just five days [26], an astonishing feat when compared to other technologies such as personal computers, which needed almost ten years to reach the same milestone [35].

Despite all of the above, the field of artificial intelligence comes with its fair share of problems and controversies: due to their inherent design, LLMs pose significant privacy risks as sensitive information is collected and used to create and fine-tune the models themselves [15], and their black-box nature makes it difficult to understand and predict their behavior [53]. Moreover, they are often trained on pirated material, like books [36] or art [25], igniting protests in many creative communities, such as hollywood writers [27] or video game actors [33]. It follows that artificial intelligence technologies should be handled carefully, without hindering their development while limiting the damages they can cause to society and individuals.

[1] Large Language Models (LLMs) are trained with supervised machine learning on vast amount of textual data, and are designed for natural language processing tasks, especially language generation [3, 4]

Author's address: Marco Bellò, marco.bello.3@studenti.unipd.it, University of Padua, Padua, Italy.

Manuscript submitted to ACM

This survey paper aims to present the current state of research on ethical and human-centric artificial intelligence, exploring how models and humans can influence each other and their environment. Section 2 showcases generation and detection of fake-news, section 3 recognition and simulation of human behaviour, as well as how to influence it. Section 4 concerns itself with biases and tendencies of the models themselves, and lastly section 5 explores ways to develop ethical LLMs that can positively impact individuals and society.

2 AI FOR FAKE NEWS GENERATION AND DETECTION

Fake news have rapidly become a significant concern in the digital age, thanks to their virality and potential damages. They spread faster and generate more engagement than truthful information [20, 47], and can influence public opinion, manipulate elections and pose a threat to public health. For example, the World Economic forum has identified the proliferation of false content as the leading short-term global risk in 2025 [6], and a BBC investigation found Russian-funded fake news networks aiming to disrupt european elections [22]. Moreover, fake news on health can cause psychological disorders and panic, fear, depression, and fatigue [37], making the World Health Organization call for development of international fact-checking organizations to combat this phenomenon [29].

Adding to the problem, the recent advancements in generative artificial intelligence have made it significantly easier to propagate misinformation through the web: generated content is increasingly indistinguishable from human-written text, sometimes even perceived as more credible [21], citing true evidence to support false claims [13], and inducing the illusion of majority opinion thanks to the sheer volume of information produced [10].

That being said, not all findings are entirely negative: Drolsbach and Pröllochs [12] shift their focus from potential societal consequences to real-world prevalence, conducting a large-scale analysis on the platform *X*. They analyzed a dataset comprising 91.452 misleading posts, both human and AI-generated, flagged trough X's *Community Notes* platform [2]. Their findings reveal that generated fake news are often centered on entertaining content rather than controversial or political subjects, and tends to exhibit a more positive sentiment than conventional forms of misinformation. Unfortunately, it is also significantly more likely to go viral.

Lastly, AI agents can produce more than just text: they can create realistic images, videos and sounds, allowing them to make digital copies of real or fictional people, known as deepfakes. In March 2019, such a technology has been used to trick a UK-based energy firm's CEO into transferring $243.000 to a malicious party, disguised as an entirely AI-generated executive from their parent company [14]. Deepfakes also increase the amount of conspiratorial videos on the internet, and they are especially vicious when targeting children, whose worldviews are easily swayed by deceptive, highly photorealistic content [51].

It follows that detecting and mitigating fake news is crucial. From the foundational work by Devlin et al. on *BERT* in 2018 [9], which revolutionized natural language processing trough deep bidirectional transformers, to the application of said transformers in identifying automatically generated headlines [23]; the landscape of automated fake news detection has significantly expanded. Vijjali et al. [52] developed a two-stage transformer-based model for detecting COVID-19 related misinformation, combining fact-checking with textual entailment to verify claims. Their model performs significantly better than other baseline NLP approaches (table 1).

Jwa et al. [19] propose an improved *exBAKE* model that leverages pre-training on a BERT model to accurately understand and assess articles' authenticity. They only analyzed the relationship between headlines body. Results can be seen in table 2.

[2] Community Notes, formerly known as Birdwatch, is community-drive content moderation program on X (formerly Twitter), where contributors can add context such as fact-checks under a post, image or video. GitHub repository: https://github.com/twitter/communitynotes

Table 1. Precision metrics for two-stage transformer model for COVID-19 fake news detection, Vijjali et al [52]

Models	**MRR**	**Recall@10**	**Accuracy**
TF-IDF	0.477	0.635	0.525
GloVe	0.182	0.410	0.579
MobileBERT	0.561	0.735	0.710
BERT	0.632	0.795	0.810
ALBERT	0.582	0.675	0.825
BERT+ALBERT	0.632	0.795	0.855

Table 2. Precision metrics for exBAKE transformer model on fake news recognition, Jwa et al. [19]

Models	**F1**	**AGR**	**DSG**	**DSC**	**UNR**
Majority vote	0.210	0.000	0.000	0.000	0.839
BERT	0.656	0.651	0.145	**0.839**	0.989
BAKE	0.734	0.667	0.463	0.822	0.986
exBAKE	**0.746**	**0.684**	**0.501**	0.813	0.988
Upper bound	0.754	0.588	0.667	0.765	0.997

Schütz et al. [44] experimented on *FakeNewsNet* dataset with *XLNet*, BERT, *RoBERTa*, *DistilBERT*, and *ALBERT* and various combinations of hyperparameters. The evaluation shows that already short texts are enough to attain 85% accuracy on the test set. Using the body text and a concatenation of both reach up to 87% accuracy. Lastly, on the matter of deepfakes, Bansal et al. [1] use *Convolutional Neural Networks* (CNN) and *Deep Convolutional Generative Adversarial Networks* (GAN) to detect them with high accuracy, as shown in figure 1.

These were just a small selection of the many research works in the field of AI-aided fake news generation and detection, which while being extremely relevant and proliferous, are but a fraction of the many potential uses for these technologies.

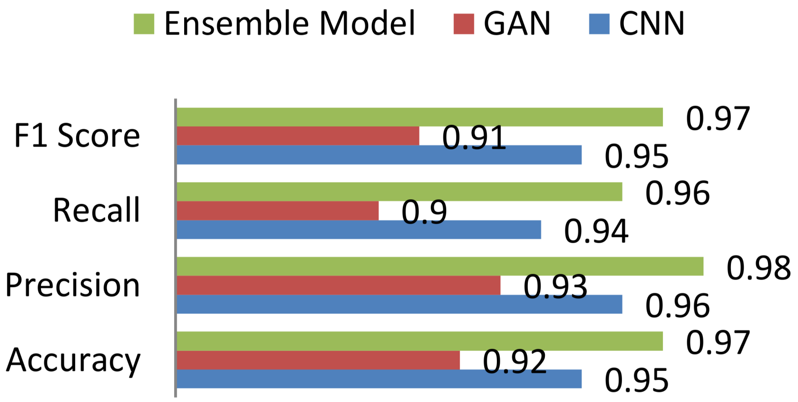

Fig. 1. Transformer models scores on deepfakes detection, Bansal et al. [1]

3 AI FOR UNDERSTANDING AND INFLUENCING HUMAN BEHAVIOUR

Today's society is already fully dependent on technology: from the banking system, to traffic monitoring and management or public health databases, IT systems have become essential. Individuals are in the same situation: virtually everyone in global north under the age of 65 possess a smartphone and use it daily [17]. It follows that artificial intelligence will become an integral part of our personal and professional lives, therefore modeling it to mimic our behaviors could aid in its usefulness and understanding. There are already evidences that humans can exploit it to acquire better comprehension of a phenomenon [43], and it can also enhance creativity in heterogeneous groups [49]. Moreover, LLMs represents a significant methodological shift in computational communication science, enabling a more flexible, more nuanced, but also less controllable exploration of social theories that have historically been difficult to reduce to simple mathematical formalisms [30]. Overall, AI promises to be a good fit for understanding, modeling and replicating human behaviour.

For these very purposes, Dos Santos Melicio et al. [11] propose a three-phase AI framework (figure 2) that analyzes verbal and non-verbal social cues. First, deep learning methods are employed to extract relevant information from the environment. Then, the episode detection phase uses extracted features to identify key moments, or episodes, which are then classified in the activity interpretation phase. Overall, the system can recognize verbal hints with 87% accuracy and non-verbal ones with 89% accuracy, with zero false positives. The system was tested in an automated assessment of communicative skills of children with Autism Spectrum Disorder (ASD), a challenging context where social cues may be less noticeable or outright absent.

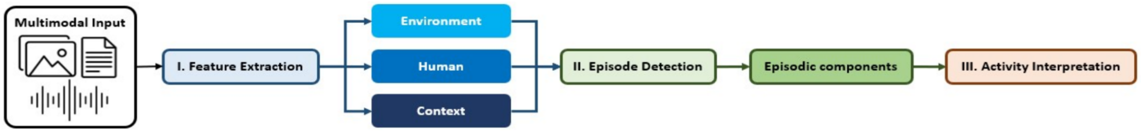

Fig. 2. Composite AI framework: Deep Learning methods extract features which are then combined with rule-based systems for detecting key episodes, and then classifiers are used to interpret activities happening in the scene. Dos Santos Melicio et al. [11]

Another possible use of such capabilities is hate speech detection: the rise of social networks and online platforms translated into a surge in hate speech across geographical and cultural boundaries. In recent studies, approximately 30% of the adolescents surveyed reported experiencing cyberbullying at some point in their lives. Furthermore, around 13% indicated that they had been cyberbullied within the 30 days before the survey [31].

To address these challenges, Chapagain et al. [7] evaluate different LLMs (BART, ELECTRA, BERT, RoBERTa, and GPT-2) on the extensive *MetaHate* dataset [32]. ELECTRA achieved the highest F1 score (table 3), outperforming all other baselines in hate speech classification

These technologies can also be used to influence people opinions. Huq et al. [18] tested this with AI-assisted messaging in an online chat platform. 557 Participants were randomly assigned to sessions of six to fifteen people, further subdivided into groups of two to four. They discussed politically controversial topic selected to maximize opinion diversity within each three-minutes session, at the end of which they chose whether to remain in their current group, join another, or create a new one. Some of them received suggestions from large language models, either personalized to their own opinion (*"individuals"*) or more similar to the group's (*"relational"*).

Table 3. Performance of classifiers on MetaHate, Chapagain et al [7]

Models	F1 Score	Accuracy
SVM	0.8380	0.8466
CNN	0.8422	0.8612
BERT	0.8809	0.8879
GPT2	0.6504	0.6152
T5	0.8707	0.8625
DeBERTa	0.8808	0.8746
Longformer	0.8845	0.8785
RoBERTa	0.8908	0.8858
XLNet	0.8917	0.8870
BART	0.8928	0.8886
DistilBERT	0.8940	0.8905
ELECTRA	**0.8980**	**0.8946**

The results show that individual assistance amplified communication volume yet increased separation between groups, while relational assistance fostered more receptive conversations and produced more heterogeneous, cross-cutting group configurations, highlighting both the dangers and the possibilities of employing artificial agents in such a fashion.

Similar conclusions can be seen in other research results: Hohenstein et al. [16] highlights how AI response suggestion systems change how people interact with and perceive one another in both pro-social and anti-social ways. Moreover, Noy et al. [28] shows that people tend to send more messages when suggestions are available, but rarely edit them, suggesting partial delegation of expressive effort.

Overall, AI tools seem capable of reducing harmful online behaviors, but can potentially be extremely disruptive due to their capability of influencing people's opinions and reducing autonomous behaviour.

4 AI OWN BIASES AND INFLUENCEABILITY

Artificial agents are trained on a mostly human-generated corpus of data, so they can develop biases and tendencies themselves. They can develop both historical and political biases in the textual content they generate [38, 40, 41]. They tend to exhibit social sycophancy, by agreeing with and flattering the user at the cost of correctness [8], and they are easily influenceable by small changes in prompt wording [42, 45]. This can be highly problematic as it can lead to discrimination or exclude certain groups, raising ethical concerns [39].

5 ETHICAL AI

REFERENCES

[1] Kartik Bansal, Shubhi Agarwal, and Narayan Vyas. 2023. Deepfake Detection Using CNN and DCGANS to Drop-Out Fake Multimedia Content: A Hybrid Approach. In *2023 International Conference on IoT, Communication and Automation Technology (ICICAT)*. 1–6. https://doi.org/10.1109/ICICAT57735.2023.10263628

[2] Celeste Biever. 2023. ChatGPT broke the Turing test — the race is on for new ways to assess AI. *Nature* 619, 7971 (July 2023), 686–689. https://doi.org/10.1038/d41586-023-02361-7 Bandiera_abtest: a Cg_type: News Feature Publisher: Nature Publishing Group Subject_term: Computer science, Mathematics and computing, Technology, Society.

[3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh,

Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258 arXiv:2108.07258 [cs].

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165 arXiv:2005.14165 [cs].

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://doi.org/10.48550/arXiv.2303.12712 arXiv:2303.12712 [cs].

[6] Andrea Carson and Max Grömping. 2025. Fake news and the election campaign – how worried should voters be? https://doi.org/10.64628/AA.wr44u3gj6

[7] Santosh Chapagain, Shah Muhammad Hamdi, and Soukaina Filali Boubrahimi. 2025. Advancing Hate Speech Detection with Transformers: Insights from the MetaHate. https://doi.org/10.48550/arXiv.2508.04913 arXiv:2508.04913 [cs].

[8] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. ELEPHANT: Measuring and understanding social sycophancy in LLMs. https://doi.org/10.48550/arXiv.2505.13995 arXiv:2505.13995 [cs].

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805 arXiv:1810.04805 [cs].

[10] Renee DiResta. 2020. AI-Generated Text Is the Scariest Deepfake of All. *Wired* (July 2020). https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/ Section: tags.

[11] Bruno Carlos Dos Santos Melicio, Linyun Xiang, Emily Dillon, Latha Soorya, Mohamed Chetouani, Andras Sarkany, Peter Kun, Kristian Fenech, and Andras Lorincz. 2023. Composite AI for Behavior Analysis in Social Interactions. In *Companion Publication of the 25th International Conference on Multimodal Interaction (ICMI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 389–397. https://doi.org/10.1145/3610661.3616237

[12] Chiara Drolsbach and Nicolas Pröllochs. 2025. Characterizing AI-Generated Misinformation on Social Media. https://doi.org/10.48550/arXiv.2505.10266 arXiv:2505.10266 [cs] version: 1.

[13] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic Stylometry for Deception Detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park (Eds.). Association for Computational Linguistics, Jeju Island, Korea, 171–175. https://aclanthology.org/P12-2034/

[14] Emilio Ferrara. 2024. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science* 7, 1 (April 2024), 549–569. https://doi.org/10.1007/s42001-024-00250-1

[15] Alice Gomstyn and Alexandra Jonker. 2024. Exploring privacy issues in the age of AI | IBM. https://www.ibm.com/think/insights/ai-privacy

[16] Jess Hohenstein, Rene F. Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F. Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports* 13, 1 (April 2023), 5487. https://doi.org/10.1038/s41598-023-30938-9 Publisher: Nature Publishing Group.

[17] Josh Howarth. 2021. How Many People Own Smartphones? (2025-2029). https://explodingtopics.com/blog/smartphone-stats

[18] Faria Huq, Elijah L. Claggett, and Hirokazu Shirado. 2025. AI-Mediated Communication Reshapes Social Structure in Opinion-Diverse Groups. https://doi.org/10.48550/arXiv.2510.21984 arXiv:2510.21984 [cs] version: 2.

[19] Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. 2019. exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT). *Applied Sciences* 9, 19 (Jan. 2019), 4062. https://doi.org/10.3390/app9194062 Publisher: Multidisciplinary Digital Publishing Institute.

[20] Zoe Kleinman. 2018. Fake news 'travels faster', study finds. https://www.bbc.com/news/technology-43344256

[21] Sarah E. Kreps, Miles McCain, and Miles Brundage. 2020. All the News that's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. https://doi.org/10.2139/ssrn.3525002

[22] Oana Marocico and Seamus Mirodan. 2025. How Russian-funded fake news network aims to disrupt European election - BBC investigation. https://www.bbc.com/news/articles/c4g5kl0n5d2o

[23] Antonis Maronikolakis, Hinrich Schutze, and Mark Stevenson. 2021. Identifying Automatically Generated Headlines using Transformers. https://doi.org/10.48550/arXiv.2009.13375 arXiv:2009.13375 [cs].

[24] Beth Miller. 2023. The Artificial Intelligence Boom. https://engineering.washu.edu/news/magazine/2023-fall/the-artificial-intelligence-boom.html

[25] Dan Milmo and Dan Milmo Global technology editor. 2025. 'Mass theft': Thousands of artists call for AI art auction to be cancelled. *The Guardian* (Feb. 2025). https://www.theguardian.com/technology/2025/feb/10/mass-theft-thousands-of-artists-call-for-ai-art-auction-to-be-cancelled

[26] Steve Mollman. 2022. Artificial intelligence chatbot ChatGPT has gained 1 million followers in a single week. Here's why it's primed to disrupt search as we know it. https://fortune.com/2022/12/09/ai-chatbot-chatgpt-could-disrupt-google-search-engines-business/

[27] Regan Morris. 2025. AI helped cause Hollywood strikes. Now it's in Oscar-winning films. https://www.bbc.com/news/articles/ce303x19dwgo

[28] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (July 2023), 187–192. https://doi.org/10.1126/science.adh2586 Publisher: American Association for the Advancement of Science.

[29] World Health Organization. [n. d.]. 1st WHO Infodemiology Conference. https://www.who.int/news-room/events/detail/2020/06/30/default-calendar/1st-who-infodemiology-conference

[30] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3586183.3606763

[31] J. W. Patchin and S. Hinduja. 2015. Cyberbullying Facts. https://cyberbullying.org/facts

[32] Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 2025–2039. https://doi.org/10.1609/icwsm.v18i1.31445 arXiv:2401.06526 [cs].

[33] Associated Press. 2024. Over 300 video game actors protest over unregulated AI use in Hollywood. *The Guardian* (Aug. 2024). https://www.theguardian.com/games/article/2024/aug/01/hollywood-video-game-actors-artificial-intelligence-protest

[34] Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035

[35] Jeremy Reimer. 2005. Total share: 30 years of personal computer market share figures. https://arstechnica.com/features/2005/12/total-share/

[36] Alex Reisner. 2025. The Unbelievable Scale of AI's Pirated-Books Problem. https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/ Section: Technology.

[37] Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2023. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *Journal of Public Health* 31, 7 (July 2023), 1007–1016. https://doi.org/10.1007/s10389-021-01658-z

[38] Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing Bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 539–544. https://doi.org/10.1145/3308560.3317590

[39] Albérico Travassos Rosário and Albérico Travassos Rosário. 1. Generative AI and Generative Pre-Trained Transformer Applications: Challenges and Opportunities. https://doi.org/10.4018/979-8-3693-1950-5.ch003 Archive Location: generative-ai-and-generative-pre-trained-transformer-applications ISBN: 9798369319505 Publisher: IGI Global Scientific Publishing.

[40] David Rozado. 2023. The Political Biases of ChatGPT. *Social Sciences* 12, 3 (March 2023), 148. https://doi.org/10.3390/socsci12030148 Publisher: Multidisciplinary Digital Publishing Institute.

[41] David Rozado. 2024. The political preferences of LLMs. *PLOS ONE* 19, 7 (July 2024), e0306621. https://doi.org/10.1371/journal.pone.0306621 Publisher: Public Library of Science.

[42] Abel Salinas and Fred Morstatter. 2024. The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance. https://doi.org/10.48550/arXiv.2401.03729 arXiv:2401.03729 [cs].

[43] Johannes Schneider. 2020. Humans learn too: Better Human-AI Interaction using Optimized Human Inputs. https://doi.org/10.48550/arXiv.2009.09266 arXiv:2009.09266 [cs].

[44] Mina Schütz, Alexander Schindler, Melanie Siegel, and Kawa Nazemi. 2021. Automatic Fake News Detection with Pre-trained Transformer Models. In *Pattern Recognition. ICPR International Workshops and Challenges*, Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (Eds.). Springer International Publishing, Cham, 627–641. https://doi.org/10.1007/978-3-030-68787-8_45

[45] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. https://doi.org/10.48550/arXiv.2310.11324 arXiv:2310.11324 [cs].

[46] John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 3 (Sept. 1980), 417–424. https://doi.org/10.1017/S0140525X00005756

[47] Craig Silverman. 2016. This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook Section: USNews.

[48] Alan Turing. 1950. Computing Machinery and Intelligence. *Computing Machinery and Intelligence. Mind* 49 (1950), 433–460. https://www.csee.umbc.edu/courses/471/papers/turing.pdf

[49] Atsushi Ueshima, Matthew I. Jones, and Nicholas A. Christakis. 2024. Simple autonomous agents can enhance creative semantic discovery by human groups. *Nature Communications* 15, 1 (June 2024), 5212. https://doi.org/10.1038/s41467-024-49528-y Publisher: Nature Publishing Group.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762 arXiv:1706.03762 [cs].

[51] Nitin Verma. 2024. “One Video Could Start a War”: A Qualitative Interview Study of Public Perceptions of Deepfake Technology. *Proceedings of the Association for Information Science and Technology* 61, 1 (Oct. 2024), 374–385. https://doi.org/10.1002/pra2.1035

[52] Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking. https://doi.org/10.48550/arXiv.2011.13253 arXiv:2011.13253 [cs].

[53] Warren J. von Eschenbach. 2021. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology* 34, 4 (Dec. 2021), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0