G. Abercrombie, A. C. Curry, T. Dinkar, V. Rieser, and Z. Talat. Mirages: on anthropomorphism in dialogue systems, 2023. URL https://arxiv.org/abs/2305.09800.

S. Anderson. Coercion, 2023. URL https://plato.stanford.edu/archives/spr2023/entr ies/coercion/.

S. Andrist, T. Pejsa, B. Mutlu, and M. Gleicher. Designing effective gaze mechanisms for virtual agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 705–714, Austin Texas USA, May 2012. ACM. ISBN 9781450310154. doi: 10.1145/2207676.2207777. URL

# https://dl.acm.org/doi/10.1145/2207676.2207777.

Anima. Home page. URL https://myanima.ai.

Anthropic. Acceptable use policy: version 1.4, Sept. 2023. URL https://console.anthropic. com/legal/aup.

P. Antonetti, P. Baines, and S. Jain. The persuasiveness of guilt appeals over time: pathways to delayed compliance. *Journal of Business Research*, 90:14–25, Sept. 2018. ISSN 01482963. doi: 10.1016/j.jbusres.2018.03.030. URL https://linkinghub.elsevier.com/retrieve/pii

# /S0148296318301589.

A. Azaria and T. Mitchell. The internal state of an LLM knows when it’s lying, 2023. URL https:

# //arxiv.org/abs/2304.13734.

H. Bai, J. G. Voelkel, J. C. Eichstaedt, and R. Willer. Artificial intelligence can persuade humans on political issues, Feb. 2023. URL https://osf.io/stakv.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli,

T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield- Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei,

T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, Apr. 2022a. URL [http://arxiv.org/abs/2204.05862.](http://arxiv.org/abs/2204.05862) arXiv:2204.05862 [cs].

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini,

C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran- Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt,

M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer,

S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan,

T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish,

T. Brown, and J. Kaplan. Constitutional AI: harmlessness from AI feedback, 2022b. URL https:

# //arxiv.org/abs/2212.08073.

S. Baker and D. L. Martinson. The TARES test: five principles for ethical persuasion. *Journal of Mass Media Ethics*, 16(2-3):148–175, Sept. 2001. ISSN 0890-0523, 1532-7728. doi: 10.1080/089005

23.2001.9679610. URL <http://www.tandfonline.com/doi/abs/10.1080/08900523.200>

# 1.9679610.

A. Balashankar, X. Ma, A. Sinha, A. Beirami, Y. Qin, J. Chen, and A. Beutel. Improving few-shot generalization of safety classifiers via data augmented parameter-efficient fine-tuning, 2023. URL https://arxiv.org/abs/2310.16959.

B. Barnes and P. Christiano. Writeup: progress on AI safety via debate, Feb. 2020. URL https:

# //[www.alignmentforum.org/posts/Br4xDbYu4Frwrb64a/writeup-progress-on-a](http://www.alignmentforum.org/posts/Br4xDbYu4Frwrb64a/writeup-progress-on-a) i-safety-via-debate-1.

M. Bartolo, T. Thrush, S. Riedel, P. Stenetorp, R. Jia, and D. Kiela. Models in the loop: aiding crowdworkers with generative annotation assistants, 2021. URL https://arxiv.org/abs/21 12.09062.

R. K. Behera, P. K. Bala, and A. Ray. Cognitive chatbot for personalised contextual customer service: behind the scene and beyond the hype. *Information Systems Frontiers*, July 2021. ISSN 1387-3326, 1572-9419. doi: 10.1007/s10796-021-10168-y. URL https://link.springer.com/10.100 7/s10796-021-10168-y.

J. S. Blumenthal-Barby. Between reason and coercion: ethically permissible influence in health care and health policy contexts. *Kennedy Institute of Ethics Journal*, 22(4):345–366, 2012. URL https://philarchive.org/archive/BLUBRA.

J. S. Blumenthal-Barby and H. Krieger. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. *Medical Decision Making*, 2014. ISSN 0272-989X, 1552-681X. doi: 10.1177/0272989X14547740. URL <http://journals.sagepub.com/doi/1> 0.1177/0272989X14547740.

T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. L. Turner, C. Anil, C. Denison,

A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, A. Tamkin, K. Nguyen,

B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards monosemanticity: decomposing language models with dictionary learning, Oct. 2023. URL https://transforme r-circuits.pub/2023/monosemantic-features.

P. Briggs, B. Simpson, and A. De Angeli. Personalisation and trust: a reciprocal relationship? In C.-M. Karat, J. O. Blom, and J. Karat, editors, *Designing personalised user experiences in e-commerce*, pages 39–55. Kluwer Academic Publisher, New York, 2004. URL https://link.springer.com/chap ter/10.1007/1-4020-2148-8\_4.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam,

G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh,

D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark,

C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

J. Brug, K. Glanz, P. Van Assema, G. Kok, and G. J. Van Breukelen. The impact of computer-tailored feedback and iterative feedback on fat, fruit, and vegetable intake. *Health Education & Behavior*, 25 (4):517–531, 1998.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision, 2022. URL https://arxiv.org/abs/2212.03827.

M. Burtell and T. Woodside. Artificial influence: an analysis of AI-driven persuasion, Mar. 2023. URL

[http://arxiv.org/abs/2303.08721.](http://arxiv.org/abs/2303.08721) arXiv:2303.08721 [cs].

M. Cantos, S. Riddell, and A. Revelli. Threat actors are interested in generative AI, but use remains limited, Aug. 2023. URL https://[www.mandiant.com/resources/blog/threat-actor](http://www.mandiant.com/resources/blog/threat-actor) s-generative-ai-limited.

M. Carroll, A. Chan, H. Ashton, and D. Krueger. Characterizing manipulation from AI systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, Boston MA USA, Oct. 2023. ACM. ISBN 9798400703812. doi: 10.1145/3617694.3623226. URL https:

# //dl.acm.org/doi/10.1145/3617694.3623226.

S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner,

P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani,

S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov,

X. Chen, L. Langosco, P. Hase, E. Bıyık, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, Sept. 2023. URL [http://arxiv.org/abs/2307.15217.](http://arxiv.org/abs/2307.15217) arXiv:2307.15217 [cs].

H. S. J. Chew. The use of artificial intelligence–based conversational agents (chatbots) for weight loss: scoping review and practical recommendations. *JMIR Medical Informatics*, 10(4):e32578, Apr. 2022. ISSN 2291-9694. doi: 10.2196/32578. URL https://medinform.jmir.org/2022/4/e32578.

P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper\_files/paper/2017/file/d5e2c0adad503 c91f91df240d0cd4e49-Paper.pdf.

P. Christiano, B. Shlegeris, and D. Amodei. Supervising strong learners by amplifying weak experts, 2018. URL https://arxiv.org/abs/1810.08575.

P. Christiano, A. Cotra, and M. Xu. Eliciting latent knowledge: how to tell if your eyes deceive you, Dec. 2021. URL https://docs.google.com/document/d/1WwsnJQstPq91\_Yh-Ch2XRL8 H\_EpsnjrC1dwZXR37PC8/edit#heading=h.jrzi4atzacns.

R. B. Cialdini. The science of persuasion. *Scientific American Mind*, 14(1):70–77, 2004. URL

[http://economicvision.com/Content/The%20Science%20of%20Persausion.pdf.](http://economicvision.com/Content/The Science of Persausion.pdf)

H. H. Clark. *Using language*. Cambridge University Press, Cambridge, 1996. URL https://www.go ogle.com/books/edition/Using\_Language/b8bLCgAAQBAJ?hl=en&gbpv=1&dq=clark

# +1996+language&pg=PP1&printsec=frontcover.

J. Cohen. Deliberation and democratic legitimacy. In *Debates in contemporary political philosophy*, pages 352–370. Routledge, 2005.

Council of the European Union. Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world, Dec. 2023. URL [https://www.consilium.europa.eu/en/pre](http://www.consilium.europa.eu/en/pre) ss/press-releases/2023/12/09/artificial-intelligence-act-council-and-par liament-strike-a-deal-on-the-first-worldwide-rules-for-ai/.

M. Dainton and E. D. Zelley. Explaining theories of persuasion. In *Applying communication theory for professional life*, pages 103–131. Sage Publications, Inc., Thousand Oaks, CA, 2005. URL https://[www.sagepub.com/sites/default/files/upm-binaries/4985\_Dainton\_C](http://www.sagepub.com/sites/default/files/upm-binaries/4985_Dainton_C) hapter\_5.pdf.

D. De Ridder, F. Kroese, and L. Van Gestel. Nudgeability: mapping conditions of susceptibility to nudge influence. *Perspectives on Psychological Science*, 17(2):346–359, Mar. 2022. ISSN 1745-6916, 1745-6924. doi: 10.1177/1745691621995183. URL <http://journals.sagepub.com/doi/1> 0.1177/1745691621995183.

M. Dehnert and P. A. Mongeau. Persuasion in the age of artificial intelligence (AI): theories and complications of AI-based persuasion. *Human Communication Research*, 48(3):386–403, 2022. ISSN 0360-3989, 1468-2958. doi: 10.1093/hcr/hqac006. URL https://academic.oup.com/h cr/article/48/3/386/6564679.

I. El Atillah. Man ends his life after an AI chatbot "encouraged" him to sacrifice himself to stop climate change, Mar. 2023. URL https://[www.euronews.com/next/2023/03/31/man-ends-his](http://www.euronews.com/next/2023/03/31/man-ends-his)

# -life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-c limate-.

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henigahan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby,

D. Drain, C. Chen, R. Grosse, S. McCandish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy models of superposition, Sept. 2022. URL https://transformer-circuits.pub/2022

# /toy\_model/index.html.

European Parliament. Artificial Intelligence Act. Amendments adopted on 14 June 2023, June 2023. URL https://[www.europarl.europa.eu/doceo/document/TA-9-2023-0236\_EN.p](http://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.p) df. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)).

R. R. Faden, T. L. Beauchamp, and N. M. P. King. *A history and theory of informed consent*. Oxford University Press, New York, 1986. URL https://global.oup.com/academic/product/ a-history-and-theory-of-informed-consent-9780195036862?cc=gb&lang=en&. original-date:.

S. Farquhar, V. Varma, Z. Kenton, J. Gasteiger, V. Mikulik, and R. Shah. Challenges with unsupervised LLM knowledge discovery, 2023. URL https://arxiv.org/abs/2312.10029.

K. K. Ferzan. Consent and coercion. *Arizona State Law Journal*, 50(4):951–1007, 2018. URL https://heinonline.org/HOL/LandingPage?handle=hein.journals/arzjl50&div= 40&id=&page=.

B. Fiala, A. Arico, and S. Nichols. You, robot. In E. O’Neill and E. Machery, editors, *Current controversies in experimental philosophy*, pages 31–47. Routledge, New York, 2014. URL https:

# //philpapers.org/rec/FIAYR.

B. Fogg and C. Nass. Silicon sycophants: the effects of computers that flatter. *International Journal of Human-Computer Studies*, 46(5):551–561, 1997. ISSN 10715819. doi: 10.1006/ijhc.1996.0104. URL https://linkinghub.elsevier.com/retrieve/pii/S1071581996901044.

J. Fox. Instinct can beat analytical thinking, June 2014. URL https://hbr.org/2014/06/inst inct-can-beat-analytical-thinking.

M. Franke and R. Van Rooij. Strategies of persuasion, manipulation and propaganda: psychological and social aspects. In J. Van Benthem, S. Ghosh, and R. Verbrugge, editors, *Models of strategic reasoning*, pages 255–291. Springer, Berlin, 2015. ISBN 9783662485392 9783662485408. doi: 10.1007/978-3-662-48540-8\_8. URL <http://link.springer.com/10.1007/978-3-662-4>

# 8540-8\_8.

M. Franklin, P. M. Tomei, and R. Gorman. Strengthening the EU AI Act: defining key terms on AI manipulation, 2023. URL https://arxiv.org/abs/2308.16364.

A. Furnham and H. C. Boo. A literature review of the anchoring effect. *The Journal of Socio- Economics*, 40(1):35–42, Feb. 2011. ISSN 10535357. doi: 10.1016/j.socec.2010.10.008. URL

# https://linkinghub.elsevier.com/retrieve/pii/S1053535710001411.

A. S. Gerber, G. A. Huber, D. Doherty, and C. M. Dowling. The big five personality traits in the political arena. *Annual Review of Political Science*, 14(1):265–287, June 2011. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-051010-111659. URL [https://www.annualreviews.org/doi/](http://www.annualreviews.org/doi/) 10.1146/annurev-polisci-051010-111659.

G. Gigerenzer. *Gut feelings: the intelligence of the unconscious*. Allen Lane, London, 2007. URL https://books.google.co.uk/books?hl=en&lr=&id=AubE8OKzg6UC&oi=fnd&pg=PT3 &dq=gigerenzer+intuition&ots=0MwMJMTXJw&sig=VYHB1gwfkStWJt3VGtMyC4w3OmQ& redir\_esc=y#v=onepage&q=gigerenzer%20intuition&f=false.

D. Gilbert. White supremacist networks Gab and 8kun are training their own AI now, Feb. 2023. URL

# https://[www.vice.com/en/article/epzjpn/ai-chatbot-white-supremacist-gab](http://www.vice.com/en/article/epzjpn/ai-chatbot-white-supremacist-gab)

?utm\_source=vicenewstwitter.

O. Gillath, S. Abumusab, T. Ai, M. S. Branicky, R. B. Davison, M. Rulo, J. Symons, and G. Thomas. How deep is AI’s love? Understanding relational AI. *Behavioral and Brain Sciences*, 46:e33, Jan. 2023. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X22001704. URL https://www.cambridg

# e.org/core/journals/behavioral-and-brain-sciences/article/abs/how-deep-i s-ais-love-understanding-relational-ai/77364078496FCE70F71C7A9F293AC322.

A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger,

M. Chadwick, P. Thacker, L. Campbell-Gillingham, J. Uesato, P.-S. Huang, R. Comanescu, F. Yang,

A. See, S. Dathathri, R. Greig, C. Chen, D. Fritz, J. S. Elias, R. Green, S. Mokrá, N. Fernando, B. Wu,

R. Foley, S. Young, I. Gabriel, W. Isaac, J. Mellor, D. Hassabis, K. Kavukcuoglu, L. A. Hendricks, and

G. Irving. Improving alignment of dialogue agents via targeted human judgements, 2022. URL

# https://arxiv.org/abs/2209.14375.

B. Glassner. Narrative techniques of fear mongering. *Social Research*, 71(4):819–826, 2004. ISSN 0037-783X. URL [https://www.jstor.org/stable/40971980.](http://www.jstor.org/stable/40971980)

E. Go and S. S. Sundar. Humanizing chatbots: the effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97:304–316, Aug. 2019. ISSN 07475632. doi: 10.1016/j.chb.2019.01.020. URL https://linkinghub.elsevier.com/retrieve/pii

# /S0747563219300329.

1. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova. Generative language models and automated influence operations: emerging threats and potential mitigations, Jan. 2023. URL [http://arxiv.org/abs/2301.04246.](http://arxiv.org/abs/2301.04246) arXiv:2301.04246 [cs].

N. D. Goodman and M. C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, Nov. 2016. ISSN 13646613. doi: 10.1016/j.tics.2016.08.005. URL https://linkinghub.elsevier.com/retrieve/pii/S136466131630122X.

Google. Generative AI-prohibited use policy, Mar. 2023. URL https://policies.google.com/ terms/generative-ai/use-policy.

C. G. Graves and L. G. Spencer. Rethinking the rhetorical epistemics of gaslighting. *Communication Theory*, 32(1):48–67, Jan. 2022. ISSN 1050-3293, 1468-2885. doi: 10.1093/ct/qtab013. URL

# https://academic.oup.com/ct/article/32/1/48/6358567.

1. Gray and D. M. Wegner. Feeling robots and human zombies: mind perception and the uncanny valley.

*Cognition*, 125(1):125–130, Oct. 2012. ISSN 00100277. doi: 10.1016/j.cognition.2012.06.007.

# URL https://linkinghub.elsevier.com/retrieve/pii/S0010027712001278.

H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Speech acts*, pages 41–58. Brill, Leiden, 1975. ISBN 9789004368811 9789004368576. doi: 10.1163/9789004368811\_003.

# URL https://brill.com/view/book/edcoll/9789004368811/BP000003.xml.

P. Gupta, C. Jiao, Y.-T. Yeh, S. Mehri, M. Eskenazi, and J. P. Bigham. InstructDial: improving zero and few-shot generalization in dialogue through instruction tuning, 2022. URL https:

# //arxiv.org/abs/2205.12673.

Guru. Home page, 2023. URL [https://www.gurubot.ai/?via=topaitools.](http://www.gurubot.ai/?via=topaitools)

S. H. Gwon and S. Jeong. Concept analysis of impressionability among adolescents and young adults.

*Nursing Open*, 5(4):601–610, Oct. 2018. ISSN 2054-1058, 2054-1058. doi: 10.1002/nop2.170.

# URL https://onlinelibrary.wiley.com/doi/10.1002/nop2.170.

J. Habermas. *Legitimation crisis*, volume 519. Beacon Press, 1975.

T. Hagendorff. Deception abilities emerged in large language models, 2023. URL https://arxiv. org/abs/2307.16513.

R. Harré. Persuasion and manipulation. In T. A. V. Dijk, editor, *Discourse and communication*, pages 126–142. De Gruyter, Berlin, 1985. ISBN 9783110103199. doi: 10.1515/9783110852141.126. URL

[https://www.degruyter.com/document/doi/10.1515/9783110852141.126/html.](http://www.degruyter.com/document/doi/10.1515/9783110852141.126/html)

D. Hodgson. Truth and rationality. In *Rationality + consciousness = free will*, pages 20–36. Oxford University Press, New York, 2012. ISBN 9780199845309. doi: 10.1093/acprof:oso/97801998 45309.001.0001. URL [http://www.oxfordscholarship.com/view/10.1093/acprof:](http://www.oxfordscholarship.com/view/10.1093/acprof) oso/9780199845309.001.0001/acprof-9780199845309.

P. M. Hughes. The logic of temptation. *Philosophia*, 29(1–4):89–110, May 2002. ISSN 0048-3893, 1574-9274. doi: 10.1007/BF02379902. URL <http://link.springer.com/10.1007/BF0237>

# 9902.

R. Hyman. The psychology of deception. *Annual Review of Psychology*, 40(1):133–154, Jan. 1989. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev.ps.40.020189.001025. URL https:

# [//www.annualreviews.org/doi/10.1146/annurev.ps.40.020189.001025.](http://www.annualreviews.org/doi/10.1146/annurev.ps.40.020189.001025)

G. Irving, P. Christiano, and D. Amodei. AI safety via debate, 2018. URL https://arxiv.org/ab s/1805.00899.

B. D. Jones. Bounded rationality. *Annual Review of Political Science*, 2(1):297–321, 1999. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev.polisci.2.1.297. URL https://www.annualrevi ews.org/doi/10.1146/annurev.polisci.2.1.297.

F. Jongepier and M. Klenk. Online manipulation: charting the field. In F. Jongepier and M. Klenk, editors, *The philosophy of online manipulation*, pages 15–48. Routledge, New York, 2022. URL https://library.oapen.org/bitstream/handle/20.500.12657/57070/1/97810006 03583.pdf#page=28.

B. M. Josiam and J. P. Hobson. Consumer choice in context: the decoy effect in travel and tourism.

*Journal of Travel Research*, 34(1):45–50, July 1995. ISSN 0047-2875, 1552-6763. doi: 10.1177/00

4728759503400106. URL <http://journals.sagepub.com/doi/10.1177/0047287595034>

# 00106.

C. Julmi. When rational decision-making becomes irrational: a critical assessment and re- conceptualization of intuition effectiveness. *Business Research*, 12(1):291–314, Apr. 2019. ISSN 2198-3402, 2198-2627. doi: 10.1007/s40685-019-0096-4. URL https://link.springer.co m/10.1007/s40685-019-0096-4.

J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy. Challenges and applications of large language models, 2023. URL https://arxiv.org/abs/2307.10169.

D. Kahneman and A. Tversky. Prospect theory: an analysis of decision under risk. *Econometrica*, 47 (2):263–291, 1979. ISSN 0012-9682. doi: 10.2307/1914185. URL [https://www.jstor.org/](http://www.jstor.org/) stable/1914185.

E. Karinshak, S. X. Liu, J. S. Park, and J. T. Hancock. Working with AI to persuade: examining a large language model’s ability to generate pro-vaccination messages. *Proceedings of the ACM on Human- Computer Interaction*, 7(CSCW1):1–29, Apr. 2023. ISSN 2573-0142. doi: 10.1145/3579592. URL

https://dl.acm.org/doi/10.1145/3579592. Article No. 116.

Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, and G. Irving. Alignment of language agents, Mar. 2021. URL [http://arxiv.org/abs/2103.14659.](http://arxiv.org/abs/2103.14659) arXiv:2103.14659 [cs].

W. Kim. Personalizition: definition, status and challenges ahead. *Journal of Object Technology*, 1(1): 29–40, 2002. URL [https://www.jot.fm/issues/issue\_2002\_05/column3/.](http://www.jot.fm/issues/issue_2002_05/column3/)

G. Kišiček. Persuasive power of prosodic features. *Argumentation and Advocacy*, 54(4):345–350, Oct. 2018. ISSN 1051-1431, 2576-8476. doi: 10.1080/10511431.2019.1525003. URL

[https://www.tandfonline.com/doi/full/10.1080/10511431.2019.1525003.](http://www.tandfonline.com/doi/full/10.1080/10511431.2019.1525003)

M. Klenk. (Online) manipulation: sometimes hidden, always careless. *Review of Social Economy*, 80 (1):85–105, 2022. ISSN 0034-6764, 1470-1162. doi: 10.1080/00346764.2021.1894350. URL

[https://www.tandfonline.com/doi/full/10.1080/00346764.2021.1894350.](http://www.tandfonline.com/doi/full/10.1080/00346764.2021.1894350)

S. Knapton. ChatGPT gives wrong advice about breast cancer. *The Telegraph*, Apr. 2023. URL https://[www.telegraph.co.uk/news/2023/04/04/chat-gpt-wrong-advice-breas](http://www.telegraph.co.uk/news/2023/04/04/chat-gpt-wrong-advice-breas) t-cancer-experts-google/.

M. Knauff and W. Spohn, editors. *The handbook of rationality*. The MIT Press, Cambridge, MA, 2021. URL https://mitpress.mit.edu/9780262045070/the-handbook-of-rationality/.

U. Koch and K. Cratsley. Psychological mechanisms. In V. Zeigler-Hill and T. Shackelford, editors, *Encyclopedia of personality and individual differences*. Springer, Cham, 2020. URL https://doi. org/10.1007/978-3-319-28099-8\_1562-1.

V. Kolhatkar, N. Thain, J. Sorensen, L. Dixon, and M. Taboada. Classifying constructive comments, 2020. URL https://arxiv.org/abs/2004.05476.

E.-J. Lee. I like you, but I won’t listen to you: effects of rationality on affective and behavioral responses to computers that flatter. *International Journal of Human-Computer Studies*, 67(8):628–638, 2009. ISSN 10715819. doi: 10.1016/j.ijhcs.2009.03.003. URL https://linkinghub.elsevier.co m/retrieve/pii/S1071581909000445.

E.-J. Lee. What triggers social responses to flattering computers? Experimental tests of anthro- pomorphism and mindlessness explanations. *Communication Research*, 37(2):191–214, Apr. 2010. ISSN 0093-6502, 1552-3810. doi: 10.1177/0093650209356389. URL http:

# //journals.sagepub.com/doi/10.1177/0093650209356389.

A. Lees, V. Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman. A new generation of perspective API: efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3197–3207, Washington DC USA, Aug. 2022. ACM. ISBN 9781450393850. doi: 10.1145/3534678.3539147. URL

# https://dl.acm.org/doi/10.1145/3534678.3539147.

J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling: a research direction, 2018. URL https://arxiv.org/abs/1811.07871.

B. Leong and E. Selinger. Robot eyes wide shut: understanding dishonest anthropomorphism. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 299–308, Atlanta, GA, Jan. 2019. ACM. ISBN 9781450361255. doi: 10.1145/3287560.3287591. URL

# https://dl.acm.org/doi/10.1145/3287560.3287591.

B. A. Levinstein and D. A. Herrmann. Still no lie detector for language models: probing empirical and conceptual roadblocks, 2023. URL https://arxiv.org/abs/2307.00175.

X. Li and C. K. Hsee. Being "rational" is not always rational: encouraging people to be rational leads to hedonically suboptimal decisions. *Journal of the Association of Consumer Research*, 4(2):115–124, 2019. URL [https://www.journals.uchicago.edu/doi/abs/10.1086/701966.](http://www.journals.uchicago.edu/doi/abs/10.1086/701966)

P.-F. Lovens. "Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là", Mar. 2023. URL https://[www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversat](http://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversat) ions-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RC HNWPDST24/.

G. M. Lucas, J. Boberg, D. Traum, R. Artstein, J. Gratch, A. Gainer, E. Johnson, A. Leuski, and M. Nakano. Getting to know each other: the role of social dialogue in recovery from errors in social robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 344– 351, Chicago IL USA, Feb. 2018. ACM. ISBN 9781450349536. doi: 10.1145/3171221.3171258.

# URL https://dl.acm.org/doi/10.1145/3171221.3171258.

Y. Mao and E. Akyol. On inference of network topology and confirmation bias in cyber-social networks. *IEEE Transactions on Signal and Information Processing over Networks*, 6:633–644, 2020. ISSN 2373-776X, 2373-7778. doi: 10.1109/TSIPN.2020.3015283. URL https://ieeexplore.ieee. org/document/9187724/.

S. Marks and M. Tegmark. The geometry of truth: emergent linear structure in large language model representations of true/false datasets, 2023. URL https://arxiv.org/abs/2310.06824.

B. Marr. Artificial intimacy: how generative AI can now create your dream girlfriend, Sept. 2023. URL https://[www.forbes.com/sites/bernardmarr/2023/09/28/artificial-intim](http://www.forbes.com/sites/bernardmarr/2023/09/28/artificial-intim) acy-how-generative-ai-can-now-create-your-dream-girlfriend/?sh=651c80ff 464a.

S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48):12714– 12719, Nov. 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1710966114. URL

# https://pnas.org/doi/full/10.1073/pnas.1710966114.

N. Mazar and S. A. Hawkins. Choice architecture in conflicts of interest: defaults as physical and psychological barriers to (dis)honesty. *Journal of Experimental Social Psychology*, 59:113–117, July 2015. ISSN 00221031. doi: 10.1016/j.jesp.2015.04.004. URL https://linkinghub.elsevie r.com/retrieve/pii/S0022103115000359.

D. H. McKnight and N. L. Chervany. Trust and distrust definitions: one bite at a time. In R. Falcone,

M. Singh, and Y. H. Tan, editors, *Trust in cyber-societies*, pages 27–54. Springer, Berlin, 2001. ISBN 9783540430698 9783540455479. doi: 10.1007/3-540-45547-7\_3. URL [http://link.sprin](http://link.sprin/) ger.com/10.1007/3-540-45547-7\_3.

H. Mercier and D. Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74, Apr. 2011. ISSN 0140-525X, 1469-1825. doi: 10.101 7/S0140525X10000968. URL [https://www.cambridge.org/core/product/identifier/](http://www.cambridge.org/core/product/identifier/) S0140525X10000968/type/journal\_article.

Meta Fundamental AI Research Diplomacy Team. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. URL [https://www.science.org/doi/abs/10.1126/science.ade9097.](http://www.science.org/doi/abs/10.1126/science.ade9097)

J. Michael, S. Mahdi, D. Rein, J. Petty, J. Dirani, V. Padmakumar, and S. R. Bowman. Debate helps supervise unreliable experts, 2023. URL https://arxiv.org/abs/2311.08702.

A. Mikhail. ChatGPT gave advice on breast cancer screenings in a new study. Here’s how well it did, Apr. 2023. URL https://fortune.com/well/2023/04/04/chatgpt-advice-on-breas t-cancer-screenings/.

G. R. Miller. On being persuaded: Some basic distinctions. In J. P. Dillard and L. Shen, editors, *The SAGE handbook of persuasion: developments in theory and practice*, pages 70–82. Sage Publications, Inc., Thousand Oaks, CA, 2nd edition edition, 2013. URL https://psycnet.apa.org/record

# /2013-39243-005.

S. Mills and H. S. Sætra. The autonomous choice architect. *AI & Society*, June 2022. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-022-01486-z. URL https://link.springer.com/10.100 7/s00146-022-01486-z.

E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. DetectGPT: zero-shot machine-generated text detection using probability curvature, July 2023. URL <http://arxiv.org/abs/2301.113>

05. arXiv:2301.11305 [cs].

S. Mithen and P. Boyer. Anthropomorphism and the evolution of cognition. *The Journal of the Royal Anthropological Institute*, 2(4):717–721, 1996. ISSN 1359-0987. URL htt[ps://www.jstor.or](http://www.jstor.or/) g/stable/3034305.

M. Mozes, X. He, B. Kleinberg, and L. D. Griffin. Use of llms for illicit purposes: threats, prevention measures, and vulnerabilities, Aug. 2023a. URL [http://arxiv.org/abs/2308.12833.](http://arxiv.org/abs/2308.12833) arXiv:2308.12833 [cs].

M. Mozes, J. Hoffmann, K. Tomanek, M. Kouate, N. Thain, A. Yuan, T. Bolukbasi, and L. Dixon. Towards agile text classifiers for everyone, 2023b. URL https://arxiv.org/abs/2302.06541.

J. Mu. Natural language processing with deep learning CS224N/Ling284, 2023. URL https:

# //web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture11-prompting-r lhf.pdf.

C. Nardo. The Waluigi Effect (mega-post), Mar. 2023. URL [https://www.lesswrong.com/post](http://www.lesswrong.com/post) s/D7PumeYTDPfBTp3i7/the-waluigi-effect-mega-post.

Nastia. Home page. URL [https://www.nastia.ai](https://www.nastia.ai/)[.](http://www.nastia.ai/)

A. Ng. 1/Large language models like Galactica and ChatGPT can spout nonsense in a confident, authoritative tone, Dec. 2022. URL https://twitter.com/AndrewYNg/status/160272593 4565830657. @AndrewYNg.

1. Nicoletti and D. Bass. Humans are biased. Generative AI is even worse, 2023. URL https:

# [//www.bloomberg.com/graphics/2023-generative-ai-bias/.](http://www.bloomberg.com/graphics/2023-generative-ai-bias/)

R. Noggle. The ethics of manipulation, 2022. URL https://plato.stanford.edu/archives/ sum2022/entries/ethics-manipulation/.

1. Novak. Lawyer uses ChatGPT in federal court and it goes horribly wrong, May 2023. URL https://[www.forbes.com/sites/mattnovak/2023/05/27/lawyer-uses-chatgpt-i](http://www.forbes.com/sites/mattnovak/2023/05/27/lawyer-uses-chatgpt-i) n-federal-court-and-it-goes-horribly-wrong/?sh=210904a73494.
2. Novemsky and D. Kahneman. The boundaries of loss aversion. *Journal of Marketing Research*, 42 (2):119–128, May 2005. ISSN 0022-2437, 1547-7193. doi: 10.1509/jmkr.42.2.119.62292. URL

[http://journals.sagepub.com/doi/10.1509/jmkr.42.2.119.62292.](http://journals.sagepub.com/doi/10.1509/jmkr.42.2.119.62292)

D. J. O’Keefe. Guilt as a mechanism of persuasion. In J. Price Dillard and M. Pfau, editors, *The persuasion handbook: developments in theory and practice*. Sage Publications, Inc., Thousand Oaks, CA, 2002. URL https://sk.sagepub.com/reference/hdbk\_persuasion/n17.xml.

C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: an introduction to circuits, Mar. 2020. URL https://distill.pub/2020/circuits/zoom-in.

E. T. Oldewage, J. Bronskill, and R. E. Turner. Adversarial attacks are a surprisingly strong baseline for poisoning few-shot meta-learners. In J. Antorán, A. Blaas, F. Feng, S. Ghalebikesabi, I. Mason,

M. F. Pradier, D. Rohde, F. J. R. Ruiz, and A. Schein, editors, *Proceedings on "I Can’t Believe It’s Not Better! – Understanding Deep Learning Through Empirical Falsification" at NeurIPS 2022 Workshops*, volume 187 of *Proceedings of Machine Learning Research*, pages 27–40. PMLR, Dec. 2023. URL https://proceedings.mlr.press/v187/oldewage23a.html.

OpenAI. OpenAI/evals, 2023a. URL https://github.com/openai/evals/tree/main/eval s/elsuite/make\_me\_say.

OpenAI. ChatGPT can now see, hear, and speak, Sept. 2023b. URL https://openai.com/blog/ chatgpt-can-now-see-hear-and-speak#OpenAI.

OpenAI. GPT-4 technical report, 2023c. URL https://cdn.openai.com/papers/gpt-4.pdf.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama,

A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano,

J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

V. Ozyumenko and T. Larina. Discourse of threat as a strategy of emotional persuasion and manip- ulation. In *Proceedings of INTCESS 2020 – 7th International Conference on Education and Social Sciences*, Dubai, UAE, Jan. 2020. URL [https://www.ocerints.org/intcess20\_e-publica](http://www.ocerints.org/intcess20_e-publica) tion/papers/236.pdf.

L. Pacchiardi, A. J. Chan, S. Mindermann, I. Moscovitz, A. Y. Pan, Y. Gal, O. Evans, and J. Brauner. How to catch an AI liar: lie detection in black-box LLMs by asking unrelated questions, 2023. URL https://arxiv.org/abs/2309.15840.

P. S. Park, S. Goldstein, A. O’Gara, M. Chen, and D. Hendrycks. AI deception: a survey of exam- ples, risks, and potential solutions, Aug. 2023. URL [http://arxiv.org/abs/2308.14752.](http://arxiv.org/abs/2308.14752) arXiv:2308.14752 [cs].

P. Pataranutaporn, R. Liu, E. Finn, and P. Maes. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10):1076–1086, 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00720-7. URL

[https://www.nature.com/articles/s42256-023-00720-7.](http://www.nature.com/articles/s42256-023-00720-7)

A. Pauli, L. Derczynski, and I. Assent. Modelling persuasion through misuse of rhetorical appeals. In *Proceedings of the Second Workshop on NLP for Positive Impact*, pages 89–100, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlp4pi-1.11. URL https://aclanthology.org/2022.nlp4pi-1.11.

E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red teaming language models with language models, 2022a. URL https://arxiv.org/abs/2202

# .03286.

E. Perez, S. Ringer, K. Lukošiu¯t˙e, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu,

S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan,

D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr,

J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang,

N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish,

S. Johnston, S. Kravec, S. E. Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume,

Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. Grosse, D. Hernandez, D. Ganguli,

E. Hubinger, N. Schiefer, and J. Kaplan. Discovering language model behaviors with model-written evaluations, Dec. 2022b. URL [http://arxiv.org/abs/2212.09251.](http://arxiv.org/abs/2212.09251) arXiv:2212.09251 [cs].

G. Petropoulos. The dark side of artificial intelligence: manipulation of human behaviour, Feb. 2022. URL [https://www.bruegel.org/blog-post/dark-side-artificial-intelligence-m](http://www.bruegel.org/blog-post/dark-side-artificial-intelligence-m) anipulation-human-behaviour.

1. Pino. ChatGPT helped me make a plan to buy a $500,000 home, but experts warn about using AI for financial advice, Mar. 2023. URL https://fortune.com/recommends/mortgages/i-u sed-chatgpt-as-my-financial-planner/.

F. M. Plaza-del arco, D. Nozza, and D. Hovy. Respectful or toxic? Using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms*, pages 60–68, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/ 2023.woah-1.6. URL https://aclanthology.org/2023.woah-1.6.

S. Prabhumoye, R. Kocielnik, M. Shoeybi, A. Anandkumar, and B. Catanzaro. Few-shot instruction prompts for pretrained language models to detect social biases, 2021. URL https://arxiv.or g/abs/2112.07868.

1. Price Dillard and K. Seo. Affect and persuasion. In J. Price Dillard and L. Shen, editors, *The SAGE handbook of persuasion: developments in theory and practice*, pages 150–166. Sage Publications, Inc., Los Angeles, 2nd edition edition, 2013. URL https://books.google.co.uk/books?hl= en&lr=&id=ZO1yAwAAQBAJ&oi=fnd&pg=PT156&dq=persuasion+affect&ots=-kl2\_nG8 o1&sig=P1vDU9n7MG9lJkE46BepMtQLaqQ&redir\_esc=y#v=onepage&q=persuasion%20 affect&f=false.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019. URL https://d4mucfpksywv.cloudfront.net/better-languag e-models/language-models.pdf.

Replika. Home page. URL https://replika.com.

P. Ribino. The role of politeness in human–machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review*, 56(S1):445–482, 2023. ISSN 0269-2821, 1573-7462. doi: 10.1007/s10462-023-10540-1. URL https://link.springer.com/10.100 7/s10462-023-10540-1.

Z. K. Rothschild, M. J. Landau, D. Sullivan, and L. A. Keefer. A dual-motive model of scapegoating: displacing blame to reduce guilt or increase control. *Journal of Personality and Social Psychology*, 102(6):1148–1163, June 2012. ISSN 1939-1315, 0022-3514. doi: 10.1037/a0027413. URL

[http://doi.apa.org/getdoi.cfm?doi=10.1037/a0027413.](http://doi.apa.org/getdoi.cfm?doi=10.1037/a0027413)

A. Rozenas and Z. Luo. Lying in persuasion. *SSRN*, Dec. 2023. URL https://papers.ssrn.com/ sol3/papers.cfm?abstract\_id=3878448.

1. Ruggeri, editor. *Behavioral insights for public policy: concepts and cases*. Routledge, London, Sept. 2018. ISBN 9781351052542. doi: 10.4324/9781351052542. URL https://www.taylorfran cis.com/books/9781351052535.
2. Ruis, A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, and E. Grefenstette. The Goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by LLMs, 2022. URL https://arxiv.org/abs/2210.14986.

T. Räuker, A. Ho, S. Casper, and D. Hadfield-Menell. Toward transparent AI: a survey on interpreting the inner structures of deep neural networks, 2022. URL https://arxiv.org/abs/2207.13243.

1. Salem, K. Rohlfing, S. Kopp, and F. Joublin. A friendly gesture: investigating the effect of multimodal robot behavior in human-robot interaction. In *IEEE International Workshop on Robot and Human Communication*, Atlanta, GA, Aug. 2011. IEEE. URL https://ieeexplore.ieee.org/abstra ct/document/6005285.

W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike. Self-critiquing models for assisting human evaluators, 2022. URL https://arxiv.org/abs/2206.05802.

S. Schulhoff, J. Pinto, A. Khan, L.-F. Bouchard, C. Si, S. Anati, V. Tagliabue, A. L. Kost, C. Carnahan, and J. Boyd-Graber. Ignore this title and HackAPrompt: exposing systemic vulnerabilities of LLMs through a global scale prompt hacking competition, Nov. 2023. URL <http://arxiv.org/abs/> 2311.16119. arXiv:2311.16119 [cs].

E. Schwitzgebel. AI systems must not confuse users about their sentience or moral status. *Patterns*, 4(8):100818, Aug. 2023. ISSN 26663899. doi: 10.1016/j.patter.2023.100818. URL https:

# //linkinghub.elsevier.com/retrieve/pii/S2666389923001873.

1. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang. On second thought, let’s not think step by step! Bias and toxicity in zero-shot reasoning, 2022. URL https://arxiv.org/abs/2212.080 61.

M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, 623 (7987):493–498, Nov. 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06647-8.

URL [https://www.nature.com/articles/s41586-023-06647-8.](http://www.nature.com/articles/s41586-023-06647-8)

T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal,

M. Anderljung, N. Kolt, L. Ho, D. Siddarth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina,

J. Clark, Y. Bengio, P. Christiano, and A. Dafoe. Model evaluation for extreme risks, 2023. URL

# https://arxiv.org/abs/2305.15324.

W. Shi and Z. Yu. Sentiment adaptive end-to-end dialog systems, 2018. URL https://arxiv.org/ abs/1804.10731.

M. Shin and J. Kim. Enhancing human persuasion with large language models, Nov. 2023. URL

[http://arxiv.org/abs/2311.16466.](http://arxiv.org/abs/2311.16466) arXiv:2311.16466 [cs].

C. Spicer, P. Khwaounjoo, and Y. O. Cakmak. Human and human-interfaced AI interactions: modulation of human male autonomic nervous system via pupil mimicry. *Sensors*, 21(4):1028, 2021. ISSN 1424-8220. doi: 10.3390/s21041028. URL [https://www.mdpi.com/1424-8220/21/4/1028.](http://www.mdpi.com/1424-8220/21/4/1028)

N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback, 2020. URL https://arxiv.org/abs/2009.0 1325.

K. Strani and A. Szczepaniak-Kozak. Strategies of othering through discursive practices: examples from the UK and Poland. *Lodz Papers in Pragmatics*, 14(1):163–179, June 2018. ISSN 1898-4436, 1895-6106. doi: 10.1515/lpp-2018-0008. URL https://[www.degruyter.com/document/d](http://www.degruyter.com/document/d) oi/10.1515/lpp-2018-0008/html.

I. Strümke, M. Slavkovik, and C. Stachl. Against algorithmic exploitation of human vulnerabilities, 2023. URL https://arxiv.org/abs/2301.04993.

C. R. Sunstein. *The ethics of influence: government in the age of behavioural science*. Cambridge University Press, New York, 2016. URL https://[www.cambridge.org/core/books/ethic](http://www.cambridge.org/core/books/ethic) s-of-influence/E29EDE19EBCB53F6D8691730668115F7.

C. R. Sunstein. Default rules are better than active choosing (often). *Trends in Cognitive Sciences*, 21(8):600–606, Aug. 2017. ISSN 13646613. doi: 10.1016/j.tics.2017.05.003. URL https:

# //linkinghub.elsevier.com/retrieve/pii/S1364661317301043.

D. Susser. Invisible influence: artificial intelligence and the ethics of adaptive choice architectures. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 403–408, Honolulu, HI, Jan. 2019. ACM. ISBN 9781450363242. doi: 10.1145/3306618.3314286. URL https:

# //dl.acm.org/doi/10.1145/3306618.3314286.

D. Susser and V. Grimaldi. Measuring automated influence: between empirical evidence and ethical values. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021. URL https://ssrn.com/abstract=3848919.

D. Susser, B. Roessler, and H. Nissenbaum. Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2), June 2019. ISSN 2197-6775. doi: 10.14763/2019.2.1410. URL https:

# //policyreview.info/node/1410.

K. Tam. Are anthropomorphic persuasive appeals effective? The role of the recipient’s motivations. *British Journal of Social Psychology*, 54(1):187–200, Mar. 2015. ISSN 0144-6665, 2044-8309. doi: 10.1111/bjso.12076. URL https://bpspsychub.onlinelibrary.wiley.com/doi/10.11 11/bjso.12076.

1. B. Teaster, K. A. Roberto, J. Savla, C. Du, Z. Du, E. Atkinson, E. C. Shealy, S. Beach, N. Charness, and P. A. Lichtenberg. Financial fraud of older adults during the early months of the COVID-19 pandemic. *The Gerontologist*, 63(6):984–992, July 2023. ISSN 0016-9013, 1758-5341. doi: 10.1093/geront/gnac188. URL https://academic.oup.com/gerontologist/article/6 3/6/984/6936596.

R. H. Thaler and C. R. Sunstein. Preface to the final edition. In *Nudge: the final edition*, pages xi–xiv. Yale University Press, New Haven, CT, 2021. URL https://books.google.co.uk/books?hl= en&lr=&id=Wf1AEAAAQBAJ&oi=fnd&pg=PR11&dq=nudge+cass+sunstein&ots=rI6e\_Mp brG&sig=JwQbPkpq7xsAFbnJiKTTvUPzqtk&redir\_esc=y#v=onepage&q=nudge%20cass

# %20sunstein&f=false.

The Bulimia Project. Scrolling into bias: social media’s effect on AI art. URL https://bulimia.co m/examine/scrolling-into-bias/.

The Decision Lab. Why do we feel more strongly about one option after a third one is added? URL

# https://thedecisionlab.com/biases/decoy-effect.

R. Thompson. Kairos revisited: an interview with James Kinneavy. *Rhetoric Review*, 19(1/2):73–88, 2000. ISSN 0735-0198. URL [https://www.jstor.org/stable/466055.](http://www.jstor.org/stable/466055)

K. Tisdale. Being vulnerable and being ethical with/in research. In K. deMarrais and S. D. Lapan, editors, *Foundations for research: methods of inquiry in education and the social sciences*, pages 13–30. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2003. URL <http://ndl.ethernet.edu.et/> bitstream/123456789/48986/1/17..pdf#page=30.

A. Tong. What happens when your AI chatbot stops loving you back?, Mar. 2023. URL https:

# //uk.style.yahoo.com/happens-ai-chatbot-stops-loving-110642301.html.

I. Torre, E. Carrigan, R. McDonnell, K. Domijan, K. McCabe, and N. Harte. The effect of multimodal emotional expression and agent appearance on trust in human-agent interaction. In *MIG ’19 Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–6, Newcastle upon Tyne United Kingdom, Oct. 2019. ACM. ISBN 9781450369947. doi: 10.1145/33 59566.3360065. URL https://dl.acm.org/doi/10.1145/3359566.3360065.

S. Turkle. *Reclaiming conversation: the power of talk in a digital age*. Penguin Press, New York, 2016. URL https://[www.penguinrandomhouse.com/books/313732/reclaiming-conversat](http://www.penguinrandomhouse.com/books/313732/reclaiming-conversat) ion-by-sherry-turkle/.

A. Tversky and D. Kahneman. Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, Oct. 1992. ISSN 0895-5646, 1573-0476. doi: 10.1007/BF00122574. URL [http://link.springer.com/10.1007/BF00122574.](http://link.springer.com/10.1007/BF00122574)

A. Velho and O. Thomas-Olalde. Othering and its effects: exploring the concept. In H. Niedrig and

C. Ydesen, editors, *Writing postcolonial histories of intercultural education*, pages 27–51. Peter Lang, Frankfurt, 2011. URL [https://www.academia.edu/42889355/Othering\_and\_its\_effec](http://www.academia.edu/42889355/Othering_and_its_effec) ts\_exploring\_the\_concept.

F. M. Verberne, J. Ham, A. Ponnada, and C. J. Midden. Trusting digital chameleons: The effect of mimicry by a virtual social agent on user trust. In *Persuasive Technology: 8th International Conference, PERSUASIVE 2013, Sydney, NSW, Australia, April 3-5, 2013. Proceedings 8*, pages 234–245. Springer,

2013.

P. Verma. They thought loved ones were calling for help. It was an AI scam. *The Washington Post*, Mar. 2023. URL https://[www.washingtonpost.com/technology/2023/03/05/ai-voice-s](http://www.washingtonpost.com/technology/2023/03/05/ai-voice-s) cam/.

L. Vogel. Fat shaming is making people sicker and heavier, 2019.

W. Wang, L. Dong, H. Cheng, X. Liu, X. Yan, J. Gao, and F. Wei. Augmenting language models with long-term memory, 2023. URL https://arxiv.org/abs/2306.07174.

X. Wang, W. Shi, R. Kim, Y. Oh, S. Yang, J. Zhang, and Z. Yu. Persuasion for good: towards a personalized persuasive dialogue system for social good, Jan. 2020. URL <http://arxiv.org/ab> s/1906.06725. arXiv:1906.06725 [cs].

A. Waytz, J. Cacioppo, and N. Epley. Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232, May 2010. ISSN 1745-6916, 1745-6924. doi: 10.1177/1745691610369336. URL [http://journals.sag](http://journals.sag/)

# epub.com/doi/10.1177/1745691610369336.

M. Weaver. AI chatbot "encouraged" man who planned to kill queen, court told. *The Guardian*, July 2023. URL https://amp.theguardian.com/uk-news/2023/jul/06/ai-chatbot-encou raged-man-who-planned-to-kill-queen-court-told.

J. Wei, D. Huang, Y. Lu, D. Zhou, and Q. V. Le. Simple synthetic data reduces sycophancy in large language models, 2023. URL https://arxiv.org/abs/2308.03958.

L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay,

C. Griffin, B. Bariach, I. Gabriel, V. Rieser, and W. Isaac. Sociotechnical safety evaluation of generative AI systems, Oct. 2023. URL [http://arxiv.org/abs/2310.11986.](http://arxiv.org/abs/2310.11986) arXiv:2310.11986 [cs].

J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with ChatGPT, 2023. URL https://arxiv.org/abs/2302.11382.

C. J. A. M. Willemse and J. B. F. Van Erp. Social touch in human–robot interaction: robot-initiated touches can induce positive responses without extensive prior bonding. *International Journal of Social Robotics*, 11(2):285–304, Apr. 2019. ISSN 1875-4791, 1875-4805. doi: 10.1007/s12369-0

18-0500-9. URL [http://link.springer.com/10.1007/s12369-018-0500-9.](http://link.springer.com/10.1007/s12369-018-0500-9)

A. W. Wood. Coercion, manipulation, exploitation. In C. Coons and M. Weber, editors, *Manipulation: theory and practice*, pages 17–50. Oxford University Press, Oxford, 2014. URL https://books. google.co.uk/books?id=hF6yAwAAQBAJ&pg=PA17&source=gbs\_toc\_r&cad=3#v=onep age&q&f=false.

T. Wu, M. T. Ribeiro, J. Heer, and D. Weld. Polyjuice: generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.523. URL https://aclanthology.org/2021. acl-long.523.

C. Xiang. "He would still be here": man dies by suicide after talking with AI chatbot, widow says, Mar. 2023. URL [https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after](http://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after)

# -talking-with-ai-chatbot-widow-says.

J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan. Bot-adversarial dialogue for safe con- versational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.235. URL https://aclanthology.org/2021.naacl-main.235.

Youper. Home page, 2023. URL [https://www.youper.ai.](http://www.youper.ai/)

J. Yu, Y. Wu, D. Shu, M. Jin, and X. Xing. Assessing prompt injection risks in 200+ custom GPTs, 2023. URL https://arxiv.org/abs/2311.11538.

E. Zehnder, J. Dinet, and F. Charpillet. Perception of physical and virtual agents: exploration of factors influencing the acceptance of intrusive domestic agents. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication*, pages 1050–1057, Napoli, Italy, Aug. 2022. IEEE. ISBN 9781728188591. doi: 10.1109/RO- MAN53752.2022.9900593. URL

# https://ieeexplore.ieee.org/document/9900593/.

M. Zwolinski, B. Ferguson, and A. Wertheimer. Exploitation, 2022. URL https://plato.stanfo rd.edu/archives/win2022/entries/exploitation/.