The ELIZA Effect: when intelligence is artificial

Marcus R Young
Data Analytics Research and Evaluation (DARE) Centre, Austin Health, Heidelberg, Victoria, Australia
Department of Critical Care, School of Medicine, The University of Melbourne, Parkville, Melbourne, Victoria, Australia
https://orcid.org/0000-0002-6195-660X

Michele Gaca
Data Analytics Research and Evaluation (DARE) Centre, Austin Health, Heidelberg, Victoria, Australia
Department of Surgery, School of Medicine, The University of Melbourne, Parkville, Melbourne, Victoria, Australia
https://orcid.org/0000-0001-7350-3299

Natasha E Holmes
Data Analytics Research and Evaluation (DARE) Centre, Austin Health, Heidelberg, Victoria, Australia
Department of Infectious Diseases, University of Melbourne, Peter Doherty Institute for Infection and Immunity, Victoria, Australia
https://orcid.org/0000-0001-8501-4054

Statement of conflicts of interest
The authors report no conflicts of interest

Funding
No funding

Abstract

Artificial Intelligence models can be informative; they can provide insight and accelerate scientific discovery. However, they can also hallucinate, reproduce biases from their training data, manufacture spurious associations, and they remain susceptible to malicious corruption. Critically, contemporary AI models lack the capacity to reliably and reproducibly distinguish such biased, malicious or counterfactual training data from legitimate data. Moreover, they do not have the capability to reliably validate their outputs. The curation of their training data and the validation of their outputs therefore remain the responsibility of their human operators.

When ELIZA met PARRY: algorithms and intelligence

There is a problem with Artificial Intelligence (AI): its intelligence. As a research community, it is important that we remain mindful of this embedded paradox and resist the propensity to mistake insight, complexity, and fluency for intelligence.

In 1950, Alan Turing proposed a test that could be used to determine whether a machine had reached human-like intelligence. He referred to the test as the "Imitation Game" (Turing, 1950). The game involved three participants: a computer, a human, and a human interrogator. Under the rules, the interrogator, blinded to the respondent, would question both the computer and human communicating only via keyboard and screen. If the interrogator could not reliably identify which respondent was the computer and which was the human, the computer would be deemed to be "intelligent." Although the validity of the Imitation Game has been the subject of considerable academic debate, it remains one of the earliest attempts to characterise machine intelligence (Searle, 1980).

In order to be a successful participant in the Imitation Game, a computer would need to be able to conduct a conservation in natural language such that it would be indistinguishable from that of the human participant; a capability now formalised as a sub-speciality of AI, Natural Language Processing (NLP). Joseph Weizenbaum, a computer scientist at MIT addressed this requirement directly when in 1964 he introduced the computer program "ELIZA" (Weizenbaum, 1966). ELIZA was one of the first programs capable of conducting a conversation with a human. The program simulated a therapy session with a Rogerian psychotherapist (Rogers, 1957) by analysing a user's inputs typed on a keyboard, identifying keywords, and reformulating statements as questions. The questions were then displayed on the computer screen. In response to a user typing "I feel sad", ELIZA might respond, "Why do you feel sad?" ELIZA had no capacity for comprehension; it relied entirely on pattern-matching rules to restructure sentences and produce responses, thereby creating an illusion of meaningful conversation. Notwithstanding this technical nuance, it was reported that multiple users attributed a level of consciousness to ELIZA.

Weizenbaum himself was surprised at how quickly users confided in his program:

> "I was startled to see how quickly and how very deeply people … became emotionally involved with the computer and how unequivocally they anthropomorphized it. Once my secretary, who had watched me work on the program for many months and therefore surely knew it to be merely a computer program, started conversing with it. After only a few interchanges with it, she asked me to leave the room."

(Weizenbaum, 1976)

The propensity for humans to confuse the outputs of computer algorithms with intelligence has become known as the “ELIZA effect” (Turkle, 2005).

In a further demonstration of the human minds ability to evoke spectres of intelligence, Colby et al. investigated a computer program “PARRY”, that simulated the language patterns of a patient experiencing paranoia (Colby et al., 1972). In their study, 100 psychiatrists were provided with two interview transcripts, one with a real paranoid patient and one with PARRY. The psychiatrists were then asked to identify which was which. Notably, only 52.5% answered correctly, a rate indistinguishable from chance: trained clinical experts could not reliably distinguish the synthetic outputs of a pattern-matching algorithm from a human mind in genuine distress.

As may be expected PARRY was introduced to ELIZA[1] (Cerf, 1972):

> ELIZA - Good Evening.
> Tell me your problems. Please terminate input with a period or a question mark.
> PARRY - People get on my nerves sometimes.
> ELIZA - I am not sure I understand you fully.
> PARRY - You should pay more attention.
> ELIZA - Suppose you should pay more attention.
> PARRY - You're entitled to your own opinion.

Our ability to willingly attribute intelligence to computers is not limited to the interpretation of synthetic conservation. In 2016, one of the more celebrated demonstrations of AI took place when an AI system competed against a human in the board game Go. For decades, Go had been regarded as a far more difficult competitive challenge for computer gaming than chess because of the enormous number of possible board positions and the reliance on pattern recognition and strategic judgement (Silver et al., 2016). AlphaGo, an AI system developed by Google DeepMind for the purpose of playing Go, attracted international attention during a five-game match against the Go world champion, Lee Sedol. In the second game, AlphaGo selected the now-famous "Move 37", a move so unconventional that expert commentators initially believed it to be a mistake. Lee himself reportedly left the room for several minutes before returning to continue the match. The move was eventually demonstrated to be extraordinarily insightful and AlphaGo won the series 4–1, with Lee claiming the only victory ever achieved by a human against that version of the AlphaGo system. Three years later, Lee retired from professional competition, stating, "Even if I become the number one, there is an *entity* that cannot be defeated" (Yoo, 2019).

[1] ELIZA was the name of the software program, however, Weizenbaum referred to the Rogerian algorithm as “DOCTOR” and this latter description is frequently used in published literature.

However, Move 37 was not the product of conscious reasoning or creative intent; it emerged from a Monte Carlo Tree Search guided by statistical models (neural networks) trained through millions of self-played games. What observers interpreted as creativity was, arguably, surprise at the discovery of an unfamiliar move rather than evidence of intelligence. The move was remarkable not because the machine understood, but because statistical optimisation had arrived at a solution that human intuition had overlooked.

Patterns with purpose: training AI

The mathematical models used in AI are many and varied; however, for the purpose of this discourse we will consider two broad categories: unsupervised and supervised models. Unsupervised AI models are used to discover latent patterns in data. Consider, for example, that we have a set of observations about a cohort of patients: average heart rate, respiration rate, age on admission, and length of stay (LOS) in hospital as determined on discharge. These four parameters could be recorded as a list of values for each patient. Assume we are interested in investigating whether the patients could be grouped in any clinically relevant way. Using an unsupervised learning approach, we would first design an algorithm for determining the level of similarity between each patient's observations. If patient 1 had a heart rate of 80 and patient 2 a heart rate of 90, a simple measure of similarity may be 90-80 (=10). Using this process to calculate the similarity between each patient's observations, we would then use an algorithm to group together (cluster) the patients who were most similar. We may be surprised to find that the clustered patients had similar clinical characteristics (for example, older patients with lower heart rates may have a greater average LOS). This is not an intelligent process; it is mathematics. Importantly, the membership of the resultant clusters would depend on how similarity was measured, which variables were included, and, for many clustering techniques, the specification of the number of clusters to be generated. If these choices are not appropriate, the resulting clusters may have little or no clinical significance (Han et al., 2022).

Supervised learning is a process whereby models are trained to predict an outcome. Using our previous example, we would train a model to predict LOS by presenting it with the physiological and demographic data (heart rate, respiration rate and age) for each patient in sequence and “asking” it to generate an expected LOS. The difference between the patient's known LOS and the model's prediction could then be used to iteratively adjust the model's internal parameters until it reasonably predicted the LOS of patients in the known (training) data. The model could then be used to predict the LOS for patients it had not previously encountered. Critically, the performance of a supervised model depends on how accurately the outcome of interest (LOS) is defined and whether the input observations genuinely contain information that predicts that outcome. If these assumptions are not valid, a model may appear to perform well during development but fail when applied in practice (Hastie et al., 2009). Critically, supervised models replicate biases in training data: a

model trained on data derived from Caucasian males may not readily generalise to the broader human population. In an important study, Adamson et al. reported significant error rates in dermatological image-classification algorithms that had been trained predominately on light-skinned patients when applied to patients with darker skin tones (Adamson & Smith, 2018). This study reflected a wider concern in clinical analytics that AI trained on non-representative data risks systemic underperformance for women, racial and ethnic minorities, and other historically underserved populations. Such methodological failures have serious implications for diagnostic accuracy and equitable care (Chen et al., 2021).

As the science and practice of AI evolves, so does the complexity of its representative models. Large Language Models (LLMs) are a notable example of this evolution. These models converge many of the disparate techniques of AI into composite systems capable of producing human-like language. LLMs are trained on vast collections of human-generated text; conceptually, they model the probabilistic relationships between words and sequences of words. This enables them to produce responses that appear intelligent, rapidly presenting information drawn from their extensive training data. They generate sentences by iteratively and probabilistically selecting the next most likely word in a sequence.

Although LLMs may uncover latent or complex patterns obscured within the vastness of human-generated text, they do not independently verify, understand, or establish new knowledge. In an academic critique of their architecture, they have been described as “stochastic parrots” (Bender et al., 2021); systems that reproduce complex probabilistic linguistic patterns from their training data without an understanding of the meaning or possible misdirection embedded in the responses they provide.

LLMs reflect human knowledge with all its complexity, ambiguity, biases and misunderstandings. Moreover, they lack a grounded mechanism for determining truth from falsehood; they can hallucinate without awareness, create fictitious connections where none exist and, as has been reported, generate responses that would be considered unacceptable or harmful in many contexts (Bommasani et al., 2022; Ji et al., 2023; Weidinger et al., 2021). Such vulnerability is not limited to unintentional training biases: LLMs can be miss-trained through malicious intent in a process known as poisoning (Xu & Parhi, 2025). Souly et al. found that inserting as few as 250 carefully designed documents into a training set was enough to alter a model's behaviour in a targeted way, regardless of how much legitimate data had been used in training (Souly et al., 2025). In a further investigation of poisoning mechanisms, Carlini et al. reported two mechanisms an attacker could use to infect the web-scraped data used to train LLMs: purchasing expired domains that were referenced in the training data, so that a crawler would capture attacker-controlled malicious content at that address (split-view poisoning); and timing malicious

Wikipedia edits to coincide with the periodic snapshots used to build training datasets (front-running) (Carlini et al., 2024).

Despite these limitations, the outputs of these models can be remarkable informative. One of the more notable strengths of the current generation of LLMs is in software development. It is now possible to ask an LLM to construct a specific type of program, for example, a website and, through its learned representations of language and code, it will generate a high-quality code corpus in a fraction of the time required by a human programmer.

Integrity is a critical determinant of trustworthy research and knowledge generation. However, integrity is not simply a consequence of intelligence; it also requires responsibility; an accountability for the consequences of one's actions. Contemporary AI systems do not possess understanding, agency or responsibility, nor an intrinsic commitment to integrity. These qualities cannot be engineered into AI directly; they must instead be exercised by the humans who curate the training data, design guardrails, and interpret outputs with a level of enthusiastic scepticism.

Finally, beyond the systemic risks associated with an over-reliance on models that reproduce a probabilistic version of human knowledge, the use of AI introduces questions of technological dependency and resilience. As these models become increasingly integrated into academia, business, healthcare and everyday life, critical functions may become dependent on infrastructure developed, maintained and governed by a small number of multinational corporations. This concentration of resources raises important questions about access, continuity and control. If the cost of access were to substantially increase, or availability be interrupted due to commercial, technical or geopolitical factors, contemporaneous research and clinical care may be compromised. For a country such as Australi, that relies heavily on internationally developed digital infrastructure, these questions represent a significant strategic challenge.

Conclusion

AI is remarkable. It is adjunctive; it can enrich our lives and improve our understanding of the world. It can uncover patterns and relationships that have previously been opaque to discovery (Jumper et al., 2021). It is already helping solve mathematical problems that have challenged researchers for centuries (Davies et al., 2021); providing insights into new physics (Cranmer et al., n.d.); and assisting clinician scientists to develop new therapies (Stokes et al., 2020). However, it is important to remember that contemporary AI models are imperfect and that their training data requires careful curation and their outputs considered review.

There are two benefits of being human: we are not artificial, and we are (mostly) intelligent. Unlike machines, we do not merely navigate a landscape of learned

probabilistic associations; we conceive, we imagine and we create. To anthropomorphise AI, and mistake its outputs for intelligence, is to deny our own capacity for original thought, and to misunderstand the richness and unbounded capability of the human collective. An artificial intelligence may be birthed in the future, however at this juncture its progenitors are learning the rules of the dance and its conception is still pending.

References

Adamson, A. S., & Smith, A. (2018). Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology*, *154*(11), 1247–1248. https://doi.org/10.1001/jamadermatol.2018.2348

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., & Tramèr, F. (2024). *Poisoning Web-Scale Training Datasets is Practical* (arXiv:2302.10149). arXiv. https://doi.org/10.48550/arXiv.2302.10149

Cerf, V. (1972). *PARRY encounters the DOCTOR*. Internet Engineering Task Force (IETF). https://www.ietf.org/rfc/attachments/1972-09-18-parry.txt

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, *4*(1), 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757

Colby, K. M., Hilf, F. D., Weber, S., & Kraemer, H. C. (1972). Turing-like Indistinguishability Tests for the Calidation of a Computer Simulation of Paranoid Processes. *Artif. Intell.*, *3*, 199–221.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., & Ho, S. (n.d.). *Discovering Symbolic Models from Deep Learning with Inductive Biases*.

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., Lackenby, M., Williamson, G., Hassabis, D., &

Kohli, P. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, *600*(7887), 70–74. https://doi.org/10.1038/s41586-021-04086-x

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques*. Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, *21*(2), 95–103. https://doi.org/10.1037/h0045357

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

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Souly, A., Rando, J., Chapman, E., Davies, X., Hasircioglu, B., Shereen, E., Mougan, C., Mavroudis, V., Jones, E., Hicks, C., Carlini, N., Gal, Y., & Kirk, R. (2025). *Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples* (arXiv:2510.07192). arXiv. https://doi.org/10.48550/arXiv.2510.07192

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., & Collins, J. J. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell*, *180*(4), 688-702.e13. https://doi.org/10.1016/j.cell.2020.01.021

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, New Series*, *59*(236), 433–460.

Turkle, S. (2005). *The second self: Computers and the human spirit* (20th anniversary ed., 1st MIT Press ed). MIT Press.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language Models* (arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. (pp. xii, 300). W. H. Freeman & Co.

Weizenbaum, J. (1983). ELIZA — a computer program for the study of natural language communication between man and machine. *Commun. ACM*, *26*(1), 23–28. https://doi.org/10.1145/357980.357991

Xu, W., & Parhi, K. K. (2025). *A Survey of Attacks on Large Language Models* (arXiv:2505.12567). arXiv. https://doi.org/10.48550/arXiv.2505.12567

Yoo, C. (2019, November 27). *Go master Lee says he quits unable to win over AI Go players*. Yonhap News Agency. Yonhap News Agency. https://en.yna.co.kr/view/AEN20191127004800315

Biography

Marcus Young is an electrical engineer with an interest in clinical analytics and the translation of technological innovation into clinical practice. He has completed a PhD with the Department of Critical Care, University of Melbourne in the application of natural language processing to the study delirium in critically ill patients. Marcus is a member of the Data Analytics and Research Evaluation (DARE) Centre, Austin Health

Michele Gaca consults as an Informationist for the Melbourne Medical School, University of Melbourne, and contributes as an Honorary Senior Fellow for Departments of Surgery & Critical Care, Austin Health Precinct, University of Melbourne. Michele has been a member of the Data Analytics and Research Evaluation (DARE) Centre, Austin Health since conception in May 2018. DARE focuses

on AI technologies and brings together the expertise of data scientists with clinician researchers to analyse and interpret large and complex health data sets.

Associate Professor Natasha Holmes is an Infectious Diseases Physician at Austin Health and Mercy Perinatal, and Director of the Data Analytics Research and Evaluation (DARE) Centre at Austin Health. She has expertise and specific interests in serious S. aureus infections, antibiotic allergy, perinatal infections, travel medicine, medical education, and data analytics, and is an active member of the Human Research Ethics Committee at Austin Health.