THE UNIVERSITY *of* EDINBURGH
Edinburgh Medical School

Biomedical Sciences

Title of Assessment: Policy Brief

Examination Number: B174569

Please use this number as your "Submission Title" in Learn when you submit your essay online.

Course Name: Dialogue for Science, Policy and Practice

Word count (excluding references): 1103

Less Risk, More Benefit

Transparency and Artificial Intelligence

SUMMARY

Artificial intelligence (AI) performs many tasks for people, from navigation and medical diagnostics, to investment planning and customer service.[1] Though these technologies offer potential to do much more, recent setbacks[2-6] have renewed attention on the lack of clarity as to how AI makes decisions. Caution is warranted because a serious failure affecting human health or safety—perhaps in medical, financial, or self-driving tools—would erode public trust and undermine continued AI development.

Based on a 2018 House of Lords report[1] and other sources, this brief outlines essential AI features, and offers four broad recommendations to reduce risk and maintain public confidence. **Page 3 presents specific actions for government, industry, and academic publishers.**

1. Improve methodological transparency
2. Improve data literacy, consumer transparency, and control
3. Diversify the AI workforce
4. Build and maintain diverse datasets and improve access for individuals and start-up developers

What is AI?

Artificial intelligence (AI) refers to technologies that can 'perform tasks that would otherwise require human intelligence'.[7] AI systems follow a sequence of logical steps called an **algorithm** that sorts and compares data to find the desired output, such as the solution to a problem.

How Does AI Work?

A human programmer trains an AI system using a **dataset** with known solutions, adjusting the algorithm to meet performance goals. The programmer then tests the system on a new dataset, which also has known solutions. A system passes the test may then operate without human supervision, on datasets without known solutions.

Older AI systems use the same algorithm unless updated by a human programmer. Most modern AI systems have a particularly important skill: they can learn from new data and adjust the algorithm without human supervision. This type of AI is called **machine learning**.

In the News (23 Feb 2021):

IBM LOOKS TO SELL OFF WATSON HEALTH

***Wall Street Journal* reports four challenges for AI in healthcare:[2]**

- **Absence of data collection standards limits broad applicability**
- **Projects often overly ambitious**
- **AI developers lack expertise in healthcare operations**
- **Available datasets do not capture features relevant to complex conditions**

Challenges and Opportunities

Measuring Performance

AI systems do not make correct decisions every time—but neither do humans. Measuring AI performance is difficult because of the different purposes and risks of each new tool, and even after decades of research, a 'widely used and accepted set of metrics remains elusive'.[8]

In general, AI performance is measured by the percentage of problems solved correctly, with penalties for wrong answers increasing in high-stakes cases. Some AI tools perform better than human experts, such as image processing tools that diagnosis certain cancer types.[2] For other proposed applications, researchers have had trouble trying to reproduce important results, raising questions of reliability.[5,6]

Deciding when AI performance is good enough for commercial use is not a purely technical question. Consider self-driving cars. Each year, humans cause roughly 1.35 million traffic deaths worldwide.[9] How much better should AI be for society to accept its risks? What criteria are appropriate for judging when AI has met this standard?

Machine Decisions, Human Bias

Human judgment is also relevant to adjusting the algorithm. Algorithms can reproduce human biases presented in the training data. Machine learning tools can amplify these biases, or learn new ones.[4] Is society ready to trust AI to learn without human supervision?

Transparency: Machine Decisions, Human Review

Recent commercial setbacks have renewed attention on these questions of performance, reliability, and bias. Whilst the largest companies can afford to forge ahead with 'moon shots', industry has generally recognized that less ambitious tools will offer greater gains until AI technologies are better understood.[2,3]

Thus, greater productivity may be achieved with attention to technical transparency and intelligibility, two standards with the same broad aim of understanding an AI system's decision-making process. Academic researchers also generally support methodological transparency, to the extent possible where data access may be restricted for privacy or proprietary reasons.[10]

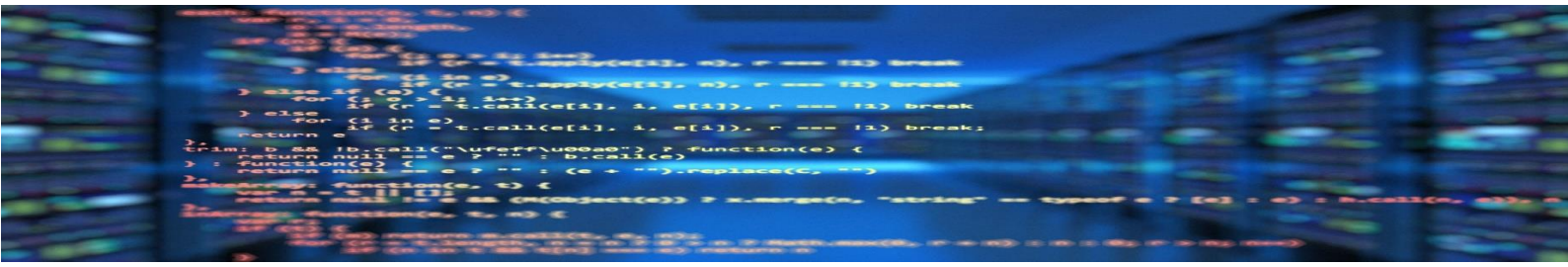
KEY TERMS

- **Artificial Intelligence (AI):** *Technology that can perform tasks otherwise requiring human intelligence*
- **Algorithm:** *A sequence of logical steps, usually applied through a computer to solve a problem or make a decision*
- **Dataset:** *A collection of data, often in table form*
- **Machine Learning:** *A type of AI that can learn from new data without human supervision*
- **Technically Transparent:** *A human expert can identify and describe each process, rule, and data input that an AI system uses to make a decision*
- **Intelligible:** *The general processes an AI system uses to make a decision can be explained in lay terms; a less restrictive standard than technical transparency*

KEY RECOMMENDATIONS

Actions supported by the House of Lords report[1] and other sources are indicated with citations.

	Government	Industry	Academia/Publishers
1. Improve methodological transparency.	Require technical transparency when human health and safety are at stake[1]	Build AI capabilities in less ambitious projects[2,3]	Improve transparency in manuscripts, including disclosure of source code and training data[5,6]
	Enforce and extend 'right to explanation' provisions,[1] as in the UK's Data Protection Act 2018	Establish methods for measuring and improving consistency across datasets[2,3,5,6]	Establish methods for measuring and improving consistency across datasets[2,3,5,6]
	Incentivize reproducibility studies[3,5]	Incentivize reproducibility studies[5,6]	Incentivize and publish reproducibility studies[5,6]
2. Improve data literacy, consumer transparency, and control.	Add data literacy to the computing curriculum in general education[1]	Inform consumers when they encounter AI in products and services[1]	Support data literacy efforts in general education[1]
	Enforce GDPR's data portability requirements[1]	Support GDPR's data portability requirements[1]	Support GDPR's data portability requirements[1]
3. Diversify the AI workforce.	Incentivize workforce diversification domestically and extend Tier II visas for AI skills[1]	Prioritize workforce diversification, including leadership positions	Prioritize workforce diversification, including leadership positions
4. Build and maintain diverse datasets and improve access for individuals and start-up developers.	Engage with under-served populations to support diversification of available datasets[1,2,4-6]	Support diversification of available datasets[1,2,4-6,10]	Support diversification of available datasets[1,2,4-6,10]
	Support data access for individuals and small- and medium-sized businesses[1]	Support access to publicly sourced data and derivative applications[1]	Support access to publicly sourced data and derivative applications[1]
	Build and link authorized and secure open databases[1-3,5,6,10]	Build and link authorized and secure open databases[1-3,5,6,10]	Build and link authorized and secure open databases[1-3,5,6,10]



References

[1] Select Committee on Artificial Intelligence, House of Lords (2018) *AI in the UK: Ready, Willing and Able?* [online]. London: The Stationery Office. (HL 2017-2019 (100)). Available from: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf [Accessed 26 February 2021].

[2] Hernandez, D., & Fitch, A. (2021) IBM's retreat from Watson highlights broader AI struggles in health. *Wall Street Journal.* 23 February.

[3] Davenport, T.H. and Ronanki, R. (2018) Artificial intelligence for the real world: don't start with moon shots. *Harvard Business Review* [online]. Available from: https://hbr.org/2018/01/artificial-intelligence-for-the-real-world [Accessed 26 February 2021].

[4] Lee, N.T., Resnick, P. and Barton, G. (2019) Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Washington, DC: The Brookings Institution. Available from: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/#footref-8 [Accessed 26 February 2021].

[5] Haibe-Kains, B., Adam, G.A., Hosny, A. *et al.* (2020) Transparency and reproducibility in artificial intelligence. *Nature* [online]. 586, E14–E16. [Accessed 26 February 2021].

[6] Hutson, M. (2018) Artificial intelligence faces reproducibility crisis. *Science* [online]. 359 (6377), pp. 725-726. [Accessed 26 February 2021].

[7] Department for Business, Energy and Industrial Strategy (2017) *Industrial Strategy: Building a Britain Fit for the Future*. London: HM Government. p 37. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/664563/ industrial-strategy-white-paper-web-ready-version.pdf [Accessed 2 March 2021].

[8] Hughes, C. and Hughes, T. (2019) What metrics should we use to measure commercial AI? *AI Matters* [online]. 5(2), pp. 41–45. [Accessed 26 February 2021].

[9] World Health Organization (2020) *Road Traffic Injuries*. Geneva: World Health Organization. Available from: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries [Accessed 2 March 2021].

[10] McKinney, S.M., Karthikesalingam, A. Tse, D. *et al.* (2020) Reply to: transparency and reproducibility in artificial intelligence. *Nature* [online]. 586, E19. [Accessed 26 February 2021].

[11] Boritz, J.E. (2005) IS practitioners' views on core concepts of information integrity. *International Journal of Accounting Information Systems* [online]. 6, pp. 260-279. [Accessed 26 February 2021].

Image Credits

Front Page Header by ThisIsEngineering from Pexels
General Header by Elchinator from Pixabay

ABOUT THE CENTER FOR DATA INTEGRITY

Founded in 2005, the Center for Data Integrity is a US-based nonprofit association of AI researchers and business development professionals from around the world.

Our Mission: To advance socially responsible innovation and long-term prosperity through data integrity—the establishment of data sources that are accurate, complete, authorized, secure, accessible, appropriately granular, consistent, and verifiable.[11]

Contact:

1912 2nd Avenue, Suite 811
Seattle, WA 98101 USA

Tel: +01 (206) 555-1212
Email: info@dataintegrity.org

Web: www.dataintegrity.org

Published: 10 March 2021