Not all data is created equal: the promise and peril of algorithms for inclusion at work

In 2016, Microsoft unveiled its first AI chatbot, *Tay,* developed to interact and converse with users in real-time on Twitter and engage Millennials. *Tay* was released with a basic grasp of language based on a dataset of anonymised public data and some pre-written material, with the intention to subsequently learn from interactions with users.

On March 23, *Tay* took its first steps on Twitter, posting mostly innocuous messages and jokes, like “humans are super cool”. However, within hours of its release, *Tay* had tweeted over 95,000 times and many of those messages were abusive and offensive misogynist/racist remarks, such as variations on “Hitler was right” and “9/11 was an inside job”. Microsoft ended up taking down the account 16 hours after joining the internet.

While opinions are divided over whether the failure of *Tay* was due to a group of online trolls, a failure of Microsoft or a combination of both[1], two important insights can be drawn from the use of AI for inclusion. First, machine learning algorithms are driven by the data they are fed. Consequently, their outcomes are only as unbiased as the data they are based on. Second, AI and machine learning models can learn and adapt over time as new data is incorporated. With the increasing rate of AI adoption, these are paramount in understanding the current state of intelligent tools and directions for future progress in inclusion.

Algorithms are only as good as their data

The case of *Tay* highlights the fact that data-driven technology makes inferences based on historical data, and these datasets themselves hold patterns of human bias.

Patterns of discrimination have long impacted existing datasets. Over 15 years ago, a field experiment conducted in the US found that identical resumes from African-Americans were less likely to receive a call-back compared to white candidates (Bertrand et al., 2004). The problem persists over time and across countries. Last year, a field experiment conducted with nearly 20,000 people across six countries, including the UK, found that ethnic minorities still have substantially lower hiring chances than the overall population (Lancee, 2019). As a meta-analysis (Zschirnt & Ruedin, 2016) recently shows, the majority of studies find evidence of discrimination for minority candidates. In fact, minority groups have 49% lower odds of getting an interview across OECD countries. This affects the data fed into algorithms to make future recommendations or decisions.

Date originally posted: 2020-10-21

Permalink: https://blogs.lse.ac.uk/businessreview/2020/10/21/not-all-data-is-created-equal-the-promise-and-peril-of-algorithms-for-inclusion-at-work/

Blog homepage: https://blogs.lse.ac.uk/businessreview/

Even seasoned professionals with good intentions can be influenced by biases, hindering the effectiveness of diversity and inclusion decisions. However, the impact of biases on these professionals and decision-makers are harder to identify, particularly under uncertainty. For instance, attribution bias – the tendency to attribute successes to inherent abilities and dismiss situational factors – can lead experienced managers to hire and promote people in more forgiving business environments, and fail to take into account the ease with which success was achieved (Swift et al., 2013). The implication is that those working in more challenging environments are disproportionally penalised. Confirmation bias and the halo effect – where positive judgments of someone in one area influences our impressions of them in other traits – can also creep into decision making. A study analysing over 1,000 CEOs and CFOs decisions of financial allocations within firms found the decisions were not just based on cost-benefit analyses, but also the "gut feel" towards the managers running divisions (Graham et al., 2015).

The personality characteristics of the people being evaluated can also impact judgements during hiring selection and performance appraisals. Research shows people are likely to confound competence with confidence (Anderson et al., 2012; Kennedy et al., 2013). Individuals who appear more confident are viewed as more competent leaders. Similarly, an analysis of graduates' labour market outcomes in the US found that extroverts were likely to gain a higher starting salary. This trait indicates warmth and energy appealing to recruiters. More conscientious graduates instead received lower starting salaries. However, their salaries grew quicker, indicating they were more successful in gaining promotion (Wiersma & Kappe, 2017). While employers were hiring based on the more visible trait, extroversion, on-the-job performance was better predicted by another, conscientiousness.

With this in mind, it is perhaps not surprising that the algorithms used for recruitment are not neutral (Greenwald, 2017; Yarger et al., 2019). Even when decision-makers attempt to remove identifiable information which can engender discrimination (such as race, sex and social category), AI models can make biased decisions because datasets reflect these patterns of behaviour and assessment, even if the signifier is removed. In fact, some researchers argue that omitting social category information can lead to less transparent decisions, making biases more difficult to detect (Williams et al., 2018). This complexity has led some firms to take a step back from AI. Among those, Amazon decided to scrap its in-house recruitment algorithm due to gender bias in 2018 and has yet to introduce a new one.

Improving the inputs

The first step in overcoming these differences is knowing where and how they arise. Auditing decisions on who is recruited and promoted is of course important. Going a step further and understanding who is given the opportunity for promotion, assigned the hardest projects or given the chance to expand their internal networks can help us gain a clearer picture. Progress in this area can harness data-driven strategies to help identify human bias in these decisions (Kleinberg et al., 2020).

For firms to truly benefit from inclusion, it is not sufficient to simply bring more women and people of colour to the organisation. Ensuring those voices are heard, equally involved in decision-making and given equal opportunity to fail and succeed (Brescoll et al., 2010) are key for success. An interesting paper published in *Science* (Woolley et al., 2010) found that "collective intelligence" – the ability for a group to outperform what can be explained by the abilities of its individual members – is predicted not just by the diversity of the group (in the study measured by the proportion of females), but by how equally the group divided conversation time among its members. The implications of this are clear, organisational performance can stand to make significant gains by broadening the profile of those taking part in strategic and operational decisions. Consequently, by making gains to inclusion now, this data can improve AI systems in the future.

In addition to what firms can do to reduce bias in underlying datasets, progress is also being made in the field of automation to improve the models. Researchers across academic institutions are working together to combat AI bias. Current efforts involve improving inclusion in the AI field, focusing on the diversity of the teams building AI tools to ensure different perspectives are brought into the development process. At the same time, the systems themselves are evolving. Machine learning algorithms and other de-biasing tools are being developed to search for patterns in the data that indicate unconscious bias in action (KPMG & The National Cyber Security Centre, 2020).

Looking forward

Date originally posted: 2020-10-21

Permalink: https://blogs.lse.ac.uk/businessreview/2020/10/21/not-all-data-is-created-equal-the-promise-and-peril-of-algorithms-for-inclusion-at-work/

Blog homepage: https://blogs.lse.ac.uk/businessreview/

AI presents a huge opportunity for businesses, including the potential for improving the effectiveness of diversity and inclusion initiatives. It is often easy to think algorithms are opaque, while human-decision making is transparent. But the story of *Tay* reminds us that algorithms inherently depend on human decisions, both in terms of the data they are based on and how they are implemented. The good news is that there are ways they can be improved. As Kleinberg and colleagues (2018) argue, they offer greater clarity and transparency on the ingredients of past decisions, creating opportunities for identifying biases. Further, they can be implemented in partnership with humans; a dialogue between humans and machines tackling the same diversity issues can create more checks and balances, improving processes.

The story of *Tay* also shows these algorithms can learn and adapt based on the data they are presented with. By implementing improved processes now, firms can reduce bias in datasets and set AI on a positive path of supporting inclusion, rather than perpetuating existing discrimination. In the meantime, these systems should be seen as one of several tools used by experienced professionals – who must take care to review suggested actions.

Notes:

- *This blog post expresses the views of its author(s), not the position of LSE Business Review or the London School of Economics.*
- *Featured image by Womanizer WOW Tech on Unsplash*
- *When you leave a comment, you're agreeing to our Comment Policy*

Teresa Almeida is a research officer at LSE's The Inclusion Initiative. She has run B2B campaigns across some of the largest enterprise businesses in the area of information and communication technology. Teresa is fascinated with the world of behavioural science and decision-making, with an emphasis on applying insight to deliver tangible results. She is also an MSc student in behavioural science at LSE.

References

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718–735. https://doi.org/10.1037/a0029395

Bertrand, M., Mullainathan, S., Abrams, D., Bede, V., Berkowitz, S., Chung, H., Fernandez, A., Guediguian, M. A., Jaw, C., Maheswari, R., Martis, B., Tisza, A., Whitehorn, G., & Yee, C. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *The American Economic Review*, *94*(4), 991–1013.

Brescoll, V. L., Dawson, E., & Uhlmann, E. L. (2010). Hard Won and Easily Lost. *Psychological Science*, *21*(11), 1640–1642. https://doi.org/10.1177/0956797610384744

Graham, J. R., Harvey, C. R., & Puri, M. (2015). Capital allocation and delegation of decision-making authority within firms. *Journal of Financial Economics*, *115*(3), 449–470. https://doi.org/https://doi.org/10.1016/j.jfineco.2014.10.011

Greenwald, A. G. (2017). An AI stereotype catcher. *Science*, *356*(6334), 133–134. https://doi.org/10.1126/science.aan0649

Date originally posted: 2020-10-21

Permalink: https://blogs.lse.ac.uk/businessreview/2020/10/21/not-all-data-is-created-equal-the-promise-and-peril-of-algorithms-for-inclusion-at-work/

Blog homepage: https://blogs.lse.ac.uk/businessreview/

Kennedy, J. A., Anderson, C., & Moore, D. A. (2013). When overconfidence is revealed to others: Testing the status-enhancement theory of overconfidence. *Organizational Behavior and Human Decision Processes*, *122*(2), 266–279. https://doi.org/10.1016/j.obhdp.2013.08.005

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, *10*, 113–174. https://doi.org/10.1093/jla/laz001

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2020). Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences*, 201912790. https://doi.org/10.1073/pnas.1912790117

KPMG, & The National Cyber Security Centre. (2020). *Decrypting Diversity Diversity and Inclusion in Cyber Security The right mix of minds makes anything possible*.

Lancee, B. (2019). Ethnic discrimination in hiring: comparing groups across contexts. Results from a cross-national field experiment. *Journal of Ethnic and Migration Studies*, 1–20. https://doi.org/10.1080/1369183X.2019.1622744

Swift, S. A., Moore, D. A., Sharek, Z. S., & Gino, F. (2013). Inflated Applicants: Attribution Errors in Performance Evaluation by Professionals. *PLoS ONE*, *8*(7). https://doi.org/10.1371/journal.pone.0069258

Wiersma, U. J., & Kappe, R. (2017). Selecting for extroversion but rewarding for conscientiousness. *European Journal of Work and Organizational Psychology*, *26*(2), 314–323. https://doi.org/10.1080/1359432X.2016.1266340

Williams, Brooks, & Shmargad. (2018). How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy*, *8*, 78. https://doi.org/10.5325/jinfopoli.8.2018.0078

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688. https://doi.org/10.1126/science.1193147

Yarger, L., Cobb Payton, F., & Neupane, B. (2019). Algorithmic equity in the hiring of underrepresented IT job candidates. *Online Information Review*, *44*(2), 383–395. https://doi.org/10.1108/OIR-10-2018-0334

Zschirnt, E., & Ruedin, D. (2016). Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies*, *42*(7), 1115–1134. https://doi.org/10.1080/1369183X.2015.1133279

[1] https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

[2] https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?CMP=twt_a-technology_b-gdntech

[3] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[4] https://www.imperial.ac.uk/news/199135/new-mathematical-principle-used-prevent-ai/

[5] https://www.imperial.ac.uk/news/201074/ai-amplify-also-overcome-bias-says/

Date originally posted: 2020-10-21

Permalink: https://blogs.lse.ac.uk/businessreview/2020/10/21/not-all-data-is-created-equal-the-promise-and-peril-of-algorithms-for-inclusion-at-work/

Blog homepage: https://blogs.lse.ac.uk/businessreview/