Human Centricity

Jason Tamara[1], Chan Pui Yan[1], Manik Bhandari[2], and Reynold D'Silva[3]

[1]MSD, Singapore
[2]Vulcan AI, Singapore
[3]GO-JEK, Singapore

I. Introduction

Gartner defines AI as a technology that applies advanced analysis and logic-based techniques, including machine learning (ML), to interpret events, support and automate decisions, and take actions. The EU defines AI as software systems that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected data, processing the information derived from this data, and deciding the best action/s to take to achieve the given goal.

Augmented intelligence, also referred to as intelligence augmentation or cognitive augmentation, is a complement to – not a replacement of –human intelligence. It's about helping humans become faster and smarter at the tasks they're performing.

"At its core, augmented intelligence is not technically different from what's already being presented as AI. It is a rather different perspective on technological advances, especially those that allow computers and software to participate in tasks that were thought to be the exclusive to humans," Ben Dickson [1], founder of TechTalks, wrote in a blog post. "And though some may call it a marketing term and a different way to reinstate hype in an already hyped industry, I think it'll help us better understand a technology whose boundaries its own creators can't define."

Optimising between commercial objectives and risks is difficult enough without AI. Some questions to consider:

- What risks should be considered? How likely are they to occur?
- What could be their potential impact?
- How much should safety or error margins be factored while still keeping economic viability?
- Can we send robot doctors into warzones?
- Can robots remove the drudgery of daily living?
- Can AI accelerate the discovery of cures for humanity's deadliest diseases?
- Can we build companies with freely-available public data and advanced mathematics?
- Apart from commercial failure, what else could go wrong?

The addition of AI and automation brings a new and complex dimension to these questions. By harnessing big data, AI offers exciting scientific, social and economic opportunities. AI holds great potential but also carries a range of hidden risks whose effects may accumulate over time and surface suddenly when critical thresholds or tipping points are reached.

- Example: Using AI to power a newsfeed starts with seemingly benign commercial apps but can end up causing immense harm to society if fake news is allowed to proliferate without filters.
- Example: Some AI apps can also cause unintended harm. Deep fake videos utilise AI techniques to mimic individuals realistically. This can be used to destructive effect when meshed with videos of politicians or corporate leaders, of them making statements which they never did.
- Example: AI technology can be used in other ways. For example, being able to finish a movie as a tribute when a key actor passes on before the movie is completed; this was done with actor Paul Walker in Furious 7. Or using AI in combination with reference

shots to de-age a movie character, as with Kurt Russell in Guardians of the Galaxy 2 [2].

Given AI's inherent and often latent risks, organisations seeking to harness its potential must carefully consider ethical and corporate values, as well as understand the implications of its actions and decisions. This often requires asking tough questions which may not have clear-cut answers: (a) What is fair? (b) Who has the responsibility or privilege to decide?

It is also about making difficult and subjective trade-offs. For example: Is it acceptable to harm one individual if doing so would benefit the greater good? To complicate matters, moral and legal definitions of "fairness" vary between cultures, communities and countries. Professor Arvind Narayanan [3], Associate Professor of Computer Science at Princeton University, spoke about "21 fairness definitions and their politics" at a conference in 2018.

Therefore, it is vital that organisations that are developing AI systems ensure that their developers and stakeholders understand and minimise the risks their systems may pose to individuals or society. Risks, impacts and potential problems must be assessed critically and deeply to design safe, human-centric systems that benefit humanity, as well as meet the organisation's commercial objectives.

A. Paper length and Font

The commercial benefits of AI systems are becoming more obvious by the day. AI applications are being used in entertainment, automation, finance, law enforcement, and healthcare. Most businesses are bullish about AI and are set to spend US$98 billion on AI-related solutions and services by 2023 – up a whopping 250 percent over US$37.5 billion that they spent in 2019 – according to estimates from IDC Corp. That's a CAGR (compound annual growth rate) of 28.4 percent between 2018 and 2023.

"The use of AI and ML is occurring in a wide range of solutions and applications – from ERP (enterprise resource planning) and manufacturing software, to content management,

collaboration, and user productivity," said David Schubmehl [4], IDC's research director for cognitive and AI systems. "AI and ML are top of mind for most organisations today. We expect AI will be the disrupting influence that will change entire industries over the next decade."

In a 2018 survey by HFS Research and KPMG [5], it was reported that 47 percent of organisations surveyed had already used ML in their production environments or were planning to. Undoubtedly, AI in and of itself, is here to stay. The shape, form, and long terms impact that AI will have, however, depends on how well and responsibly we manage AI related risks.

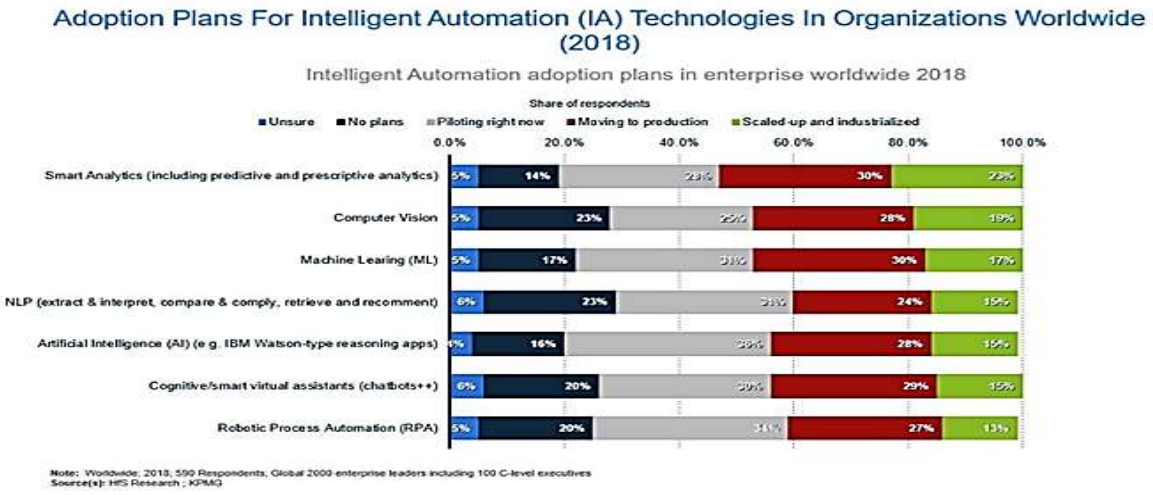

Figure 1: AI adoption plans [5]

When organisations fail to develop or deploy AI responsibly, they may harm others and themselves. Companies may face legal risks and penalties from regulatory bodies – such as Singapore's PDPC or the EU's GDPR – if AI implementations fall short of meeting regulatory requirements.

Aside from legal action, organisations may also face reputational risks and lawsuits when AI systems act in ways that society finds objectionable. Even with the best intentions, companies may suddenly find themselves called out in the public sphere, seemingly

blindsided by the unintended consequences of their AI use. This chapter offers some best practices that AI development and deployment teams can consider to mitigate some of these risks.

B. Discrimination

Discrimination can take two forms: Disparate treatment, and disparate impact. Both often occur unintentionally.

a. Disparate Treatment

This occurs when a system applies unequal standards to different groups, such as when a system presents lower-paying job postings when the user is female, compared to when the user is male.

b. Disparate Impact

This occurs when a system applies equal standards to everyone, but still results in substantially different outcomes between individuals belonging to different demographic groups with similar attributes.

c. Representational Harm

This occurs when a system reinforces negative stereotypes or diminishes specific groups. For instance, downstream harms [6] to particular groups are often blamed on “biased data,” but this concept encompasses too many issues to be useful in developing solutions.

d. Allocative Harm

This occurs when a system allocates or withholds opportunities or resources from specific groups1. For example, if you are trying to book an air ticket from your MacBook and your colleague is doing exactly the same, except on a Windows laptop. He receives a lower price while you do not, or he receives a lower-priced upgrade while you are offered the regular price.

C. Decisions

AI systems may sometimes filter information in ways that influence thinking and decision-making. For example, an AI-enabled job-matching platform may suggest lower paying postings to women based on historical training data, reinforcing biased views that women are less qualified than men. A news recommendation engine may maximise clicks on its website by suggesting stories based on gossip. Such stories may also be highly polarising and could also be fake.

Such recommendations could have deleterious effects on certain users, and on society if a number of users believe the fake news being true. An MIT research project [7] found that fake news spreads more rapidly, and with a substantial margin, compared to real news, on a leading social network. “We found that falsehood diffuses significantly farther, faster, deeper, and more broadly than the truth, in all categories of information, and in many cases by an order of magnitude,” according to Professor Sinan Aral, a Professor at the MIT Sloan School of Management.

- **Example**: Your friend shares a link to a new social network with you. The website is brilliantly designed and has great graphics. Over a short period, you become addicted to the site. Over time, you fall behind in your work, and eventually lose your job. Such addictive behaviour [8] has now been classified as an ailment by the WHO.

D. Fairness

How does one define and quantify fairness, especially in the context of AI decisions? The more common definitions are about group and individual fairness. Group fairness requires that demographic groups have similar outcomes on average; individual fairness requires that individuals be treated consistently.

- **Example**: A university aims for a "fair" and unbiased admissions process with applications being evaluated by AI algorithms. The issues: Should the "fair" algorithm admit the top 100 applicants, regardless of gender? Or should it admit the top 50 male and top 50 female applicants? Or should it strike a balance between the two options? The first option is based on demographic parity; the second is based on equal opportunity. Both make sense, and will likely conflict with each other. Algorithm-based decisions can either promote fairness, or they can reinforce disparities across demographic dimensions.
- **Example**: Women hold just under 30 per cent of all CEO positions in the US. However, a University of Washington study [9] found that image search algorithms under-represented women in search results, as shown in Figure 2 below:

Figure 2. Google search results [9]

The study found that manipulated image search results determine 7 percent of a participant's opinion about how many men and women work in a particular field, exhibiting an example of how algorithmic decisions skew our view of the world. The logic extends beyond gender; other groups under-represented online include the elderly, small business owners, people with low income, disabled, or disadvantaged.

Debates about "fairness" predate the existence of AI; one definition may not be more accurate than the other. These competing views will remain unresolved because they are subjective at the macro or micro levels. Deciding on the appropriate definition of fairness depends on cultural, organisational and personal beliefs. The developer or user organisation must be cognizant of unintended bias creeping in. More so now, with the added risk of the speed, scale and impact of AI, which could lead to unexpected outcomes for some individuals or groups who may be inadvertently left out. (Suggestions on ensuring AI systems are properly implemented is covered under Section III.A.b.)

II. Best Practices

It is important to keep ethical issues during development process of AI systems. It's equally important to monitor the AI system's behaviour and decisions – as well as conduct independent reviews regularly – after deployment. This will help identify and quantify risks, if any. It would also be ideal if business teams and AI designers are sensitised to typical sources of risks in the AI process, so they can introduce proper checks and balances at various stages in the design, development, deployment and post-deployment phases of the AI lifecycle.

A. Automation

AI-driven automation offers many benefits. It can improve efficiency, quality, and consistency. However, excessive automation without offline contingency plans can lead to over-reliance on AI and loss of robustness and skills. This may leave the organisation vulnerable to disruptions; in some cases, it may also lead to a moderate or massive loss of human jobs.

- Example: Automation has already replaced 1.7 million jobs since 2000 and is expected to replace 20 million jobs by 2030, according to Oxford Economics (see Figure 3). Workers displaced by automation, such as those in the manufacturing sector, may

discover that jobs in other sectors for which they have transferable skills are already, or will soon be, impacted by automation. On average, each additional robot installed in lower-skilled regions could lead to nearly twice as many job losses as those in higher-skilled regions of that country. This could exacerbate economic inequality and political polarisation. Increasing automation must be balanced with investment in developing workforce skills, BBC News reported on June 26, 2019 [10].

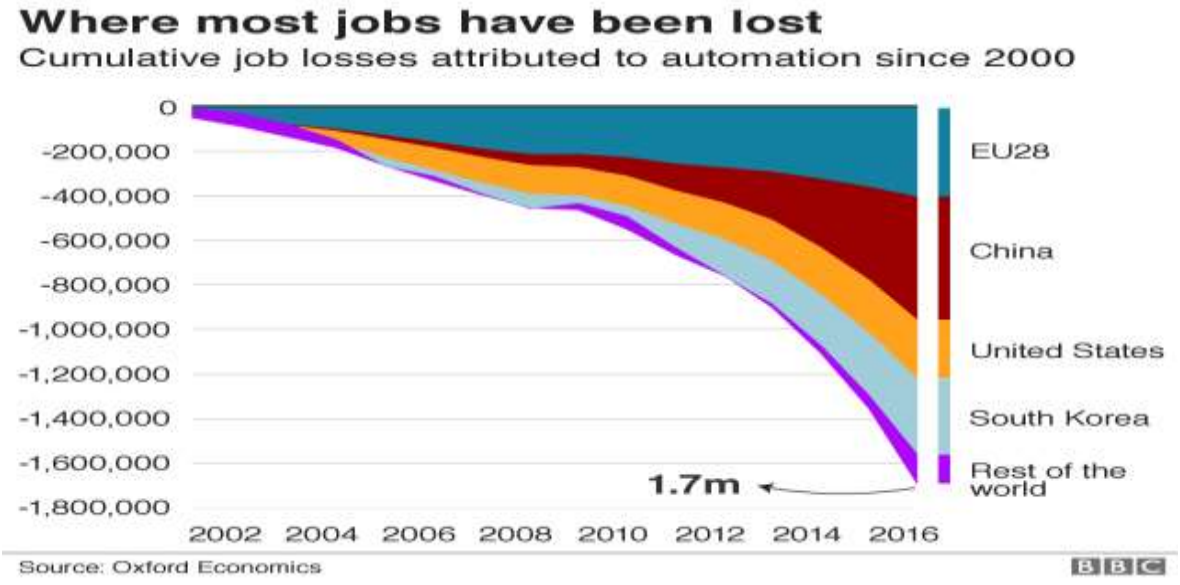

Figure 3: Job losses due to automation [10]

B. Decisions

Avoid making ethical decisions on someone else's behalf. An MIT study [11] showed ethical decisions diverging significantly across cultures. Therefore, AI solutions providers should design systems that allow the buyers of the solution to translate their own ethical values into the system. However, both developers and users should not design or deploy systems that can be programmed to violate laws, human rights or cause serious harm. (Source: https://bit.ly/3e4mU4g).

- **Example**: An AI solutions company is developing a triage system for sale to hospitals. The solutions developer designs a system that gives higher priority to

an elderly patient despite the patient arriving later than a younger one. This may seem like an ethical decision to the AI system developer. However, the user of the system may have wanted a different prioritisation algorithm. For instance, during a major disaster where the hospital must attend to a multitude of trauma patients, giving priority by age may not be the most efficient use of limited resources. AI systems developers should not make ethical decisions on behalf of end-users without due consultation.

C. Implications

Ensure users understand risks and implications. Although AI systems developers may leave decisions of fairness and ethics to their customers, the developers should provide visibility for those using the system to understand the implications of their decisions.

- **Example**: A bank deploys a system to process credit card applications. The system is set to accept applications meeting a fixed criteria based on age, salary, employment history, and credit scores. As more applications are processed, it becomes apparent that the majority of rejected applications belong to a specific group, thus raising questions on discrimination. An employee in charge reviews the rejected applications using the system's explainability features to check whether the outcomes were discriminatory and to tweak the system to become inclusive.

D. Mitigation

Maintain a log of the AI solution's input and output to monitor for drift and unintended consequences. Small errors may not be apparent over short time frames. This might be especially relevant if disputes or disagreements arise. It would also be ideal to develop a communications plan and FAQs for front-line or customer-facing staff.

- **Example**: HSBC uses AI to assess loan applications promptly. Given its large customer base and the strong regulatory scrutiny that it operates under, HSBC takes a generally cautious, "human-in-the-loop" approach for all of its AI-based solutions. HSBC has enhanced its loan application process and developed an Artificial Neural Network (ANN) scorecard to segregate applicants into different risk tiers. For each loan submission, the ANN generates a score. Using the score, as well as other factors like the measure of current debt-to-income ratio, credit scores, credit history and other credit triggers, HSBC staff can assess the risk of the loan being defaulted and decide whether to approve or decline a loan application.

III. Assessment & Mitigation

This section deals with the risk assessment and mitigation across the AI lifecycle. It is important that business teams and AI designers and developers note the typical sources of risk in the AI lifecycle. They can then introduce proper checks and balances at various stages in the design, development, deployment and post-deployment phases of the AI lifecycle.

A. Four Phases

The four phases of the AI lifecycle have been stated above. Once risks have been identified, they should be quantified or categorised based on the severity of impact. This will help the organisation evolve an appropriate mitigation strategy. It will also allow the business to weigh the business benefits versus the associated risks. Depending on the industry, companies will need to evolve their own risk assessment framework, which must be ratified by top management, and the governance and compliance team.

- **Example**: The definition of "harm" and the computation of probability and severity will depend on the context. It will vary from sector to sector. For instance,

a hospital or healthcare facility may regard harm associated with an incorrect diagnosis of a patient's medical condition as severe, compared to an online apparel business that offers a poor product recommendation.

a. Design Phase

- **Key Risks:** This is the very first stage. There are two risks here: the first is not understanding the commercial objective or intent of the solutions user; the second is not translating the commercial intent into the right AI problem statement.
- **Mitigation**: Mitigation can be done by documenting the business intent/objectives and getting approval from the governance team, which could comprise business, technology, legal and ethics representatives.
    - **Example**: To ensure robust oversight of AI deployment for their AML (Anti-Money Laundering) programme, DBS Bank introduced internal governance structures and measures. These included setting up an RDU (Responsible Data Use) framework; an RDU Committee was appointed for oversight and governance. The RDU Committee included senior leaders from different DBS units to ensure diversity, and checks and balances.

b. Development Phase

- **Key Risks**: Selecting datasets with high predictive value but violates the company's or society's definition of fairness, should be avoided. For instance, giving undue weightage to race, religion or gender while screening job applicants will be unfair, unethical, and can lead to lawsuits. It is vital to select data samples that are representative of the population. There is a risk that

personal biases of the person preparing the data may influence the fairness or representativeness of the sample. Therefore, instead of the data telling the story, the developer ends up selecting the data that will tell the story that he or select others want to hear.

- **Mitigation**: Use the entire dataset as much as possible, without excluding the subsets, unless there is a strong business and ethical reason not to do so. Carry out data profiling and exploratory analysis. Share the results with the governance unit to ensure that the data is representative and the variables included are not discriminatory or weighted on superficial attributes. Split the data randomly into training, test and validation datasets.
    - **Example**: To mitigate the risks of inherent bias in DBS Bank's AML models, it used full datasets (instead of sample datasets) to train, test and validate the AML Filter Model. These datasets were then separated into training, testing and validating data for the AML Filter Model. The bank built its training data from about 8,000 alerts triggered by about 4,000 customers over a year to train the AML Filter Model. To mitigate model bias, it excluded data used in training. This back-testing stage involved about 4,500 alerts generated by 3,000 customers over six months. Finally, to validate the AML Filter Model, DBS conducted a parallel run with about 4,600 alerts generated by around 2,500 customers over four months.

c. Implementation Phase

- **Key Risks**: The risk here is cherry-picking who, where, or when the AI model is to be applied based on either personal bias, or in pursuit of unreasonable and unethical business results. This can lead to flawed or unethical decision-making.

- **Mitigation**: Develop the right SOPs (standard operating procedures) for the interpretation and explanation of the AI outcomes to key stakeholders. Include how different outcomes will impact the business if done incorrectly.
    - **Example**: A telco offering re-contract vouchers to subsidise the handset cost has different offers for different customers. This is based on their AI model which takes into account the churn probability and the lifetime value of the customer. For instance, a customer may be offered $100 voucher, but his friend may be offered a $300 voucher. If queried by the customer on this apparent bias, the telco may find it tough to justify. Thus, when deploying AI models, organisations need to ensure the outcomes are explainable, and that proper briefing is given to frontline staff to deal with customer complaints.

d. Post-Implementation Operations Phase

- **Key Risks**: As an AI solution is deployed, new data coming in may change or alter the dynamics and the AI model may lose its accuracy. The AI solution may therefore provide incorrect outcomes when fresh incoming data is significantly different from the underlying data used to train the model. This could lead to poor or questionable corporate decisions or actions.
- **Mitigation**: There are three ways to deal with this: One, perform regular data profiling and data quality analysis to detect significant changes or drifts in data patterns. Two, design and deploy robust AI model metrics that are generated and tracked regularly. Three, develop escalation matrices. Depending on the deterioration of the model, the appropriate business and governance unit can be notified and decisions made about retaining/tweaking/killing the model.

- **Example**: Callsign [12] is a British firm that leverages DL (deep learning) techniques and combines biometrics, geo-location and behavioural analytics with multi-factor authentication to help clients authenticate their identities. The firm conducts intensive testing with the use of proofs-of-concept, prototypes, and peer reviews. It uses behavioural biometric models and tools to build the model's accuracy.
- **Example**: DBS Bank tracks the model metrics every month to ensure stability of the AML Filter Model. The results from the training, back-testing and validation stages are used as benchmarks for the model's metrics. These metrics are fine-tuned after deployment. The model is monitored every month and reviewed every six months by the bank's ML team. This ensures that any deviation from the pre-defined thresholds will be flagged for the team's review. These recommendations are then reviewed and approved by the governance team before deployment. DBS has also set up internal controls to keep tabs on risks.

B. Societal Issues

Some societies place more value on equity, equality, or meritocracy than others. Organisations should be mindful that user expectations on fairness and ethics vary depending on cultural and national contexts. AI systems, especially if deployed online, may reach unintended audiences and cause embarrassment or outrage in those communities or customers.

- **Example**: Some societies place a high value on affirmative action, while others believe it is unfair. A solutions developer is building an AI-powered admissions platform for sale to universities worldwide. One university may be required to admit a percentage of students based on diversity quotas, another may admit

students based entirely on grades, a third may admit students regardless of the student's financial status, while a fourth doesn't offer financial aid for needy students. The solutions developer works closely with each university to ensure each university's unique values are incorporated in the AI solution.

- Example: A study, called The Moral Machine [13], designed to explore the moral dilemmas faced by autonomous vehicles. It found that ethical principles varied between individuals and were correlated with modern institutions and deep cultural traits. You can take a quiz at the Moral Machine website [13]. The platform showcases moral dilemmas and asks you to judge which of two outcomes you think are more acceptable. You can then compare your responses with those of other people who took the quiz.

a. Localisation

Most organisations don't invest in developing AI solutions or services for just one market. The goal is usually to offer it to as many countries as possible. That's ideal, but the world is not a homogenised entity of humans. Even when solutions are localised or customised for a specific country, different states or provinces may have different norms that possibly contradict others in the same country. Can moral and legal disclaimers be provided upfront to save possible embarrassment or lawsuits?

- **Example**: A social networking company considers deploying two systems: a GDPR compliant system for EU users, and a higher revenue generating version with added features and personal data collection opt-in for all other users. Given that some users strongly support GDPR to protect individuals, while others view it as a hindrance to innovation and profitability, having two systems may not be viewed well by the online

community. The company then considers whether some users would view the move as "getting away with selling as much data as possible" and weighs the potential reputational cost against the potential revenue increase. Based on the company's research, business model, and competitive landscape, they decide it is worth deploying both systems to capture as much of the market as possible.

b. Accountability

Decisions made by humans that are based on AI-generated insights must still be made critically and ethically. That's because seemingly accurate AI models do not necessarily represent the true state of the world. Explanations, even when supported by data, cannot be used to justify actions that cause harm to humans. The explanations do not absolve organisations of responsibility. It is crucial therefore that organisations using or developing AI systems educate their employees, developers and customers accordingly.

- **Example**: Refer to an earlier example on Apple Card. Although default rates were correlated with gender, the incident caused an outrage on social media. Society did not accept the company's explanations; such attempts were met with increased hostility and fears of trusting AI systems. Society will hold organisations accountable whether decisions were made based on data or not.
- **Example**: A model with high accuracy during testing sounds promising, but it is not a guarantee of real-world accuracy. If one variable was fed into an AI model as a "feature", and the other as a "target", the model may find a correlation, despite there being none. Sampling errors, bias, hidden variables, proxy metrics and spurious correlations can all lead to

misleading results. The AI system may perform well during training using test data but fail during deployment. Figure 4 captures this point, from Spurious Correlations [14].

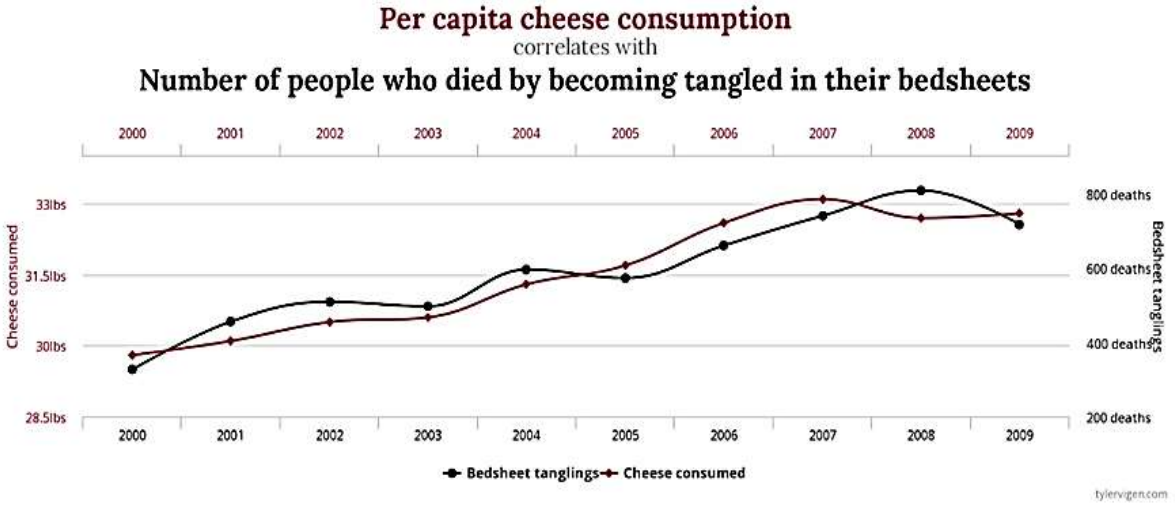

Figure 4: Spurious correlations [14]

References

[1] B. Dickson, “What is the difference between artificial and augmented intelligence?,” 4 December 2017. [Online]. Available: https://bdtechtalks.com/2017/12/04/what-is-the-difference-between-ai-and-augmented-intelligence/.

[2] A. Holmes, “ How Guardians Of The Galaxy 2 Actually Made Kurt Russell Look Young, According To James Gunn,” 10 May 2017. [Online]. Available: https://www.cinemablend.com/news/1657620/how-guardians-of-the-galaxy-2-actually-made-kurt-russell-look-young-according-to-james-gunn%23:~:text=A%20young%20actor,%20Aaron%20Schwartz,place%20Aaron's%20skin%20onto%20him..

[3] “TL;DS - 21 fairness definition and their politics by Arvind Narayanan,” [Online]. Available: https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/.

[4] “David Schubmehl,” [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=PRF003956.

[5] L. Columbus, “Roundup Of Machine Learning Forecasts And Market Estimates For 2019,” 27 March 2019. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2019/03/27/roundup-of-machine-learning-forecasts-and-market-estimates-2019/?sh=1445e9297695.

[6] H. Suresh and J. V. Guttag, “A Framework for Understanding Unintended Consequences of Machine Learning,” *arXiv*.

[7] P. Dizikes, “Study: On Twitter, false news travels faster than true stories,” 8 March 2018. [Online]. Available: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308.

[8] T. Embury-Dennis, “MAN WHO INVENTED 'LIKE' BUTTON DELETES FACEBOOK APP OVER ADDICTION FEARS,” 6 October 2017. [Online]. Available: https://www.independent.co.uk/life-style/gadgets-and-tech/facebook-inventor-deletes-app-iphone-justin-rosenstein-addiction-fears-a7986566.html.

[9] J. Langston, “Who’s a CEO? Google image results can shift gender biases,” 19 April 2015. [Online]. Available: https://www.washington.edu/news/2015/04/09/whos-a-ceo-google-image-results-can-shift-gender-biases/.

[10] “Robots 'to replace up to 20 million factory jobs' by 2030,” 26 June 2019. [Online]. Available: https://www.bbc.com/news/business-

48760799#:~:text=About%201.7%20million%20manufacturing%20jobs,million%20industrial%20robots%20by%202030..

[11] K. Hao, “Should a self-driving car kill the baby or the grandma? Depends on where you’re from.,” 24 October 2018. [Online]. Available: https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/.

[12] Callsign, “Identity Fraud Detection, Authentication & Authorization | Callsign,” [Online]. Available: https://www.callsign.com.

[13] “Moral Machine,” [Online]. Available: https://www.moralmachine.net.

[14] “Spurious Correlations,” [Online]. Available: https://www.tylervigen.com/spurious-correlations.