Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Re-identification of Anonymised Data

This input paper was prepared by Dr Ian Oppermann

Suggested Citation

Oppermann, I (2018). Re-identification of Anonymised Data. Input paper for the Horizon Scanning Project “The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing” on behalf of the Australian Council of Learned Academies, www.acola.org.

1. INTRODUCTION

Future Smart Services for homes, factories, cities, and governments rely on sharing of large volumes of often personal and sensitive data between individuals and organisations, or between individuals and governments. The benefits from more easily sharing data is the ability to create locally optimised or highly personalised services based on preference and choice, as well as developing efficiencies and savings from understanding of the wider network of users and providers. Despite these potential benefits across a range of people centred services and infrastructure, data sharing remains a challenge.

The main challenges are centred on concerns about unintended consequences of sharing data from appropriate use and interpretation, errors or unauthorised disclosure or use of data, and concerns about adherence to privacy legislation.

Secondly, aggregation of individual data is a common approach used to reduce the risk of personal disclosure within a data set. A key challenge for sharing data sharing is that ***there is currently no way to unambiguously determining if there is personal information within aggregated data***. Consequently, different techniques and Output levels for aggregated data are used across organisations depending on a perceived risk associated with the data being shared. The implications of this are profound when thinking of the utility and use cases which are affected by the level of aggregation.

Thirdly, concerns raised by Privacy advocates as the capability for analysing data increases. When the number of data sources used to create and deliver a service or address a policy challenge increases to hundreds or thousands of data points, the complexity of the problem may rapidly exceed the ability for human judgement alone to determine if the integrated data (or the insights generated from them) contain personally identifying information.

Personal data covers a very wide field and is described differently in different jurisdictions. The definition is always very broad and in principle, covers any information that relates to an identifiable, individual living (or within 30 years of death in some states). Many regulatory frameworks rely on a “reasonable” test.

The ambiguity about the presence of personal information in sets of data highlights the limitations of most existing privacy regulatory frameworks. The inability of human judgment to determine “reasonable” likelihood of reidentification when faced with large numbers of complex data sets limits the ability to appropriately apply the regulatory test.

2. DATA SHARING FRAMEWORKS

In September 2017, the Australian Computer Society (ACS) released a technical whitepaper which explored the challenges of data sharing[1]. The paper highlighted that, a fundamental challenge for the creation of smart services, is addressing the issue of whether a set of data sets contains personally identifiable information. Determining the answer to this question is a major challenge as the act of combining data sets creates information. The paper further proposed a modified version of the “Five Safes” framework for data sharing which attempts to quantify different thresholds for “Safe”.

[1] See ACS website, available online https://www.acs.org.au/content/dam/acs/acs-publications/ACS_Data-Sharing-Frameworks_FINAL_FA_SINGLE_LR.pdf (Accessed 9th August 2018)

2.1 Modified "Five Safes" Framework

The 2017 whitepaper introduced several conceptual frameworks for practical data sharing including an adapted version of the "Five Safes" framework[2]. Many organisations around the world including the Australian Bureau of Statistics use the Five Safes framework to help make decisions about effective use of data which is confidential or sensitive. The dimensions of the framework are:

Safe People – refers to the knowledge, skills, and incentives of the users to store and use the data appropriately. In this context, 'appropriately' means 'in accordance with the required standards of behaviour', rather than level of statistical skill. In practice, a basic technical ability is often necessary to understand training or restrictions and avoid inadvertent breaches of confidentiality; an inability to analyse data may lead to frustration and increases incentives to 'share' access with unauthorised people.

Safe Projects – refers to the legal, moral, and ethical considerations surrounding use of the data. This is often specified in regulations or legislation, typically allowing but limiting data use to some form of 'valid statistical purpose', and with appropriate 'public benefit'. 'Grey' areas might exist when 'exploitation of data' may be acceptable if an overall 'public good' is realised.

Safe Setting – refers to the practical controls on the way the data is accessed. At one extreme researchers may be restricted to using the data in a supervised physical location. At the other extreme, there are no restrictions on data downloaded from the internet. Safe settings encompass both the physical environment (such as network access) but also procedural arrangements such as the supervision and auditing regimes.

Safe Data – refers primarily to the potential for identification in the data. It could also refer to the sensitivity of the data itself. It may also refer to the quality of the data and the conditions under which it was collected.

Safe Outputs – refers to the residual risk in publications from sensitive data.

The Five Safes framework is relatively easy to conceptualise when considering the extreme cases of 'extremely' Safe although it does not unambiguously define what this is. An extremely Safe environment may involve researchers who have had background checks, projects which have ethics approval and rigorous vetting of Outputs. Best practice may be established for such frameworks, but none of these measures is possible to describe in unambiguous terms as they all involve judgement.

The adapted model explores different, quantifiable levels of "Safe" for each of People, Projects, Setting, Data and Outputs as well as how these different "Safe" levels could interact in different situations. Figure 1 shows the dimensions of the adapted "Five Safes" framework taken from the 2017 ACS Technical whitepaper.

[2] *Five Safes: designing data access for research*, T. Desai, F. Ritchie, R. Welpton, October 2016, http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/b691218a6fd3e55fca257af700076681/$FILE/The%20Five%20Safes%20Framework.%20ABS.pdf (Accessed 9th August 2018)

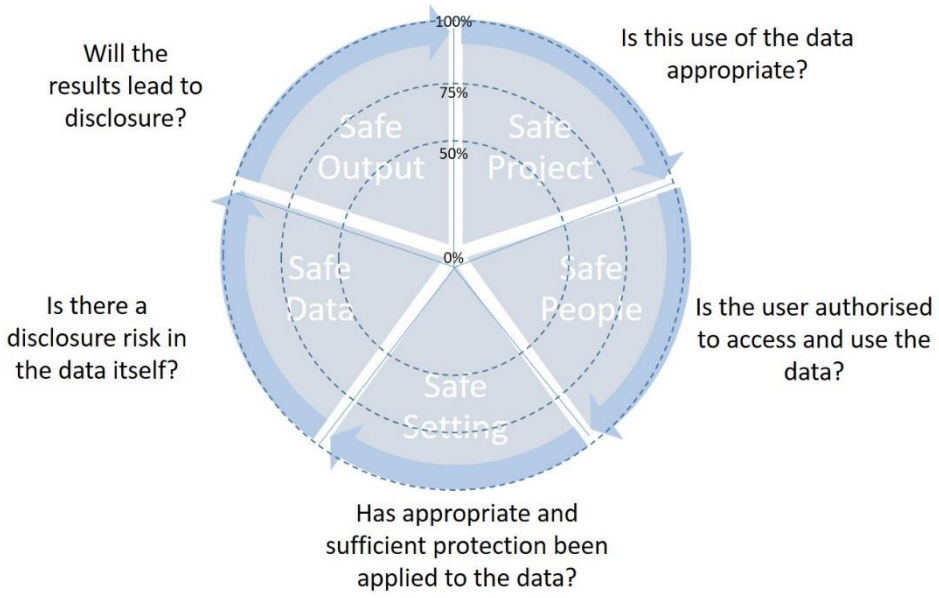

Figure 1. Modified Five Safes Framework

3. DEALING WITH AI - WHAT HAPPENS WHEN THE "PEOPLE" ARE "ALGORITHMS"?

3.1 Adapting the "Five Safes" Framework

In the world of AI, the "Safe People" may be replaced with algorithms which process data supplied for analytical purposes (such as clustering or classification) or for purposes of delivering a smart service (such as smart lighting, or smart message routing). The environment an algorithm operates in may be very different to a human researcher, and the restrictions and scrutiny placed on an algorithm may be far more intrusive than those which can be applied to a human researcher. Consequently, some of the implicit assumptions in the Five Safes framework need to be re-examined.

The Five Safes is a system model and so is intended to be considered in the context of all the elements. The answer to whether a researcher (or algorithm) is permitted to access a data set assumes that all other necessary conditions are in place. Supposing secure facilities do not exist; then this does not seem like an appropriate way to use the data. However, this does not mean the questions of whether a researcher should have access to the data changes; only that the proposed solution as a whole is not acceptable – in this case because of a failure of the "Safe Setting" dimension.

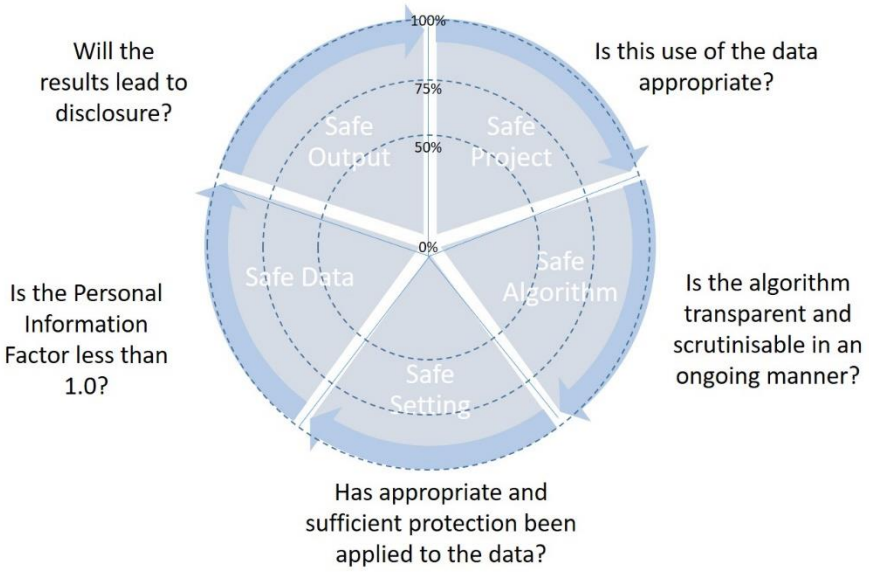

Figure 2. Five Safes Framework for Algorithms

Safe Algorithms – For an artificially intelligent algorithm, the behaviours and associated access conditions can be enforced under many circumstances but will need supervision if adapting over time. Any biases which develop also need to be monitored.

Safe Projects – still refers to the legal, moral, and ethical considerations surrounding use of the data. 'Grey' areas might exist when exploitation of data may be acceptable if an overall public good is realised or with consent from the person who is provided the project outcome (knowledge) or who benefits from the AI driven service. The Safeness of the project that an algorithm undertakes should be known before application of the algorithm to the data.

Safe Setting – When the "researcher" is an algorithm, the operating environment can be locked, disconnecting the algorithm from other sources of input. This does not however allow for any biases in the algorithm itself from being evaluating or the implications of these being understood.

Safe Data –When the "observer" is an algorithm, the context which the algorithm brings to the data can be strictly limited through limiting access to other data sets, strictly limiting the Personal Information Factor to be less than 1.0.

Safe Outputs – There is a distinct difference to be further examined as to a single discrete Output from an algorithm and something that feeds an operational loop (such as a steering algorithm, or cruise control algorithm).

In practice, the Project undertaken by AI may be very small compared to the scope undertaken by a human researcher. Consider for example the use of Monte Carlo analysis[3] which consists of repeated evaluations of an environment under different sampled values of random variables. Each "Project" is small however the results of thousands of small projects may be merged to create a deeper understanding of a process or system.

The framing questions to be considered include:

[3] See for example https://en.wikipedia.org/wiki/Monte_Carlo_method (Accessed 9th August 2018).

- Is it possible to apply the “Safes” framework when the researcher is an algorithm?
- Is it possible to determine 75%, 50% or 25% safe levels for aspects of the model for an algorithm?
- Could, for example, a 100% safe level for an algorithm be described and combined with a 25% safe setting?

4. CONCLUSIONS

The underpinning concepts of the Five Safes framework are significantly stretched when “person” or “researcher” is extended to an artificially intelligent algorithm. The basic considerations of the risk framework however remain including the “Safe People” and “Safe Projects” dimensions. “Safe Algorithms” may have been peer reviewed to detect bias, and constantly monitored as they develop. “Safe Projects” may be extended to consider the real-world implications of a steering or braking decision of a self-driving vehicle.

One fundamental difference when considering the operation of an algorithm is that it may train on a set of data and then continually adapt or “learn” post training during the operational phase. The Safes framework implies distinct, discrete discovery-oriented analytics projects rather than continuous operational loop: a discrete project carried out by a person, who releases results, which inform those who operationalise a service or system. If the Five Safes was a continuous process where Outputs fed directly into a next loop of projects, the evaluation of Safe People, Safe Projects and Safe Data would need to be automated.

The potential for continuous “learning” by algorithms introduces distinct challenges. It has been cited numerous times that AI is prone to amplify sexist and racist biases from the real world [4] [5] and potentially evolve to positions well beyond those intended by developers. A Safe Algorithm needs to be constantly monitored for their Safe Level – which may change over time or be recalibrated.

One of the implications that can be drawn from the discussion of the framework is that several of the dimensions are highly dependent on judgement. “Safe Projects” are particularly depended on a judgement-based evaluation of risk. Whilst frameworks may be developed to help decision making in these areas, there is no unambiguous way to determine quantified levels of ‘safe’ for this dimension.

“Safe Setting” is largely depended on restrictions applied at a technology and governance level.

The “Safe Outputs” dimension brings us back to the heart of the challenge of data driven analytics. The human context of recipients of the results of the data analysis project or the AI driven service. For a project outcome, the challenge relates to the ability of any human (or algorithmic) recipient of these Outputs to find additional data in the wider world to combine with the Outputs of the data analysis project. For the recipient of the AI driven service, the challenge relates to the responsibility of the real-world Outputs of the service.

[4] See for example New Scientist, “Discriminating algorithms: 5 times AI showed prejudice”, April 2018 https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/ (Accessed 9th August 2018)
[5] See for example TechRepublic, “Why Microsoft's 'Tay' AI bot went wrong”, March 2016 https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/ (Accessed 9th August 2018)