Challenges in the interaction between data and society

Word Count: 2198

Table of Contents

1. Introduction

In the past ten years, there has been amount of literatures on the theme of relationship between data and society. One of the main reasons why a growing number of researchers focus on this subject is *technological determinism*. According to Paragas and Lin (2014), in this theory, technology is considered as a vital feature of society and can greatly change social communication methods such as the emergence of mobile phone and the Internet. It is, however, evident that society can also have an influence on technology. Mayer-Schonberger and Cukier (2013), for example, hold the view that all information in society can be transformed into data format, called *datafication*. This requires analysts to invent and master a set of new matching technology skills. Based on these, it can thus be argued that society and technology can interact with each other and promote a synergetic development instead of a one-way effect.

In addition, the author strongly believes that data is also a synthesis of societal factors rather than simple quantitative value. In this respect, Kitchin (2014) holds a similar view. He claims that data should be considered as a comprehensive sociotechnical assemblage and this data assemblage contains different societal elements, such as *political economy* and so forth. Moreover, it is apparent that there are an increasing number of challenges in the process of data and society mutual development. Several cases, such as the impact of non-neutral algorithm on public position, has raised questions regarding with what direction do we want society and technology to develop. In this context, the author maintains that our social world should be equality, diversified and ethical. For data scientists, it is our duty to balance the agency and try to maintain ethical practice.

The aim of this essay is to analyse the interaction between data and society in order to indicate the main challenges and provide guidance to data scientists or citizens. This essay is divided into four sections. First, it introduces the definition and philosophy of data science. Second, investigates the challenges to data science practice. The third section is responds to these challenges. Finally, a conclusion is drawn based on the issues explored below.

2. Definition and philosophy of data science

Numerous researchers provide their views on the definition of data science. Compared to others, this definition is more suitable for the content of this essay because it relates to the knowledge pyramid which is discussed more detail below:

> At a high level, data science is a set of fundamental principles that guide the extraction of knowledge from data…Data science involves principles, processes, and techniques for understanding phenomena via the analysis of data.
> (Provost & Fawcett, 2013, p. 2)

In terms of epistemology, there is, however, a controversy over how to generate and justify data science knowledge. In my view, data science outputs involve rational knowledge and logical interpretation. This is because, when facing too much information, data scientists prefer to select partial information as proxy data to make a decision, which is seen as a convenient judgement standard. The research study by O'Neil (2016), for instance, found that a young man without credit rating is much easier to get loans than a housecleaner. This is because the young man has Ph.D. or engineer classmates on Facebook, but the latter's Facebook friend might be unemployed or in jail. There is thus evidence to suggest that, based on personal perspective, data scientists tend to regard proxy data, such as Facebook friendship, as evaluation standard instead of comprehensive information. The process thus involves subjective judgement. It is, therefore, argued that data science is not socially neutral knowledge and it could cause non-neutral outputs. The following section critically discusses the major influenced factor and consequence of data science practice.

3. Challenges to data science practice

According to Simon Lindgren (2014), social structure derives from traditions or stereotypes and it can confine activity of social agent to limit their power. It is, therefore, apparent that agency, the power of different agents, is unequally distributed in society. In my view, unbalanced power is a major factor for non-neutral data science practice and non-neutral data science practice reflect bias in society.

3.1 Socio-political power

As noted by Braman (2009), there are three political forms of power: *instrumental, structural, and symbolic*. In the information age, however, a new but central power appear on the stage: *informational* power, which refers to control informational sources of other powers in order to influence citizen behaviour. There is considerable evidence to show that governments and large technology companies have this power and tend to compete for it. For example, it is generally acknowledged that most governments have surveillance programs on the ground of national security. Several governments, however, have begun to illegally collect personal communication data or even steal other state secrets such as PRISM (surveillance program). It is evident that the aim of these surveillance programs is to seize *informational* power and maintain their social structure. Similarly, another famous political example is Cambridge Analytica. The company collect personal data from social network and analyse user's political position in order to provide service for potential political and commercial clients. Even for large companies, such as Facebook or Google, they also exert *informational* power on users such as selling and displaying different political advertisement based on user's activity.

In addition, there is a tenable argument that governments realise that they are losing *informational* power due to the rise of technology company. In the Arab Spring, for example, a great number of messages are disseminated by social media, such as Facebook or Twitter, instead of traditional medium. Similarly, recent "Yellow Vest" protest in France is not organized by any labour union but participants communicate and gather through social media. These cases indicate that Internet companies seize *informational* power from the government and this is the reason why many countries have passed laws to control Internet companies. Similarly, Mozur, Scott and Isaac (2017) also claim that there is a power struggle between the Internet company and government and the latter is waking up and regaining their power. It is, therefore, argued that power has massive and conclusive impact on data science practice but even worse, affected data science practice could also cause bias and discrimination in society.

3.2 Bias

According to Kitchin (2014), knowledge pyramid bases on all kinds of data: “data precedes information, which precedes knowledge, which precedes understanding and wisdom” (p. 9). In my view, bias is potential for exist or emerge in the following process.

First, bias can emerge when using data to represent reality. As discussed above, in the *datafication* process, data scientists prefer to pick proxy data in order to make decision conveniently. There is, however, the major challenge of selecting proxy data: decision makers mainly rely on their own reasoning or intuition to judge the relationship between data and real world. In other words, the validity of data could be influenced by a personal view and someone’s attempt to use cheating proxy data to meet their target. Several global technology companies, for example, begin to use the number of star, programmer candidate win at GitHub website, as an employment standard. As a result, a part of applicants use GitHub star bot to increase star number. Similarly, many YouTuber or Instagrammer buy followers to improve their popularity so as to win advertisements. It is because they believe that they can build a fake world by modifying corresponding proxy data.

Second, there are apparent bias existed in several widely useful data mining technologies, in the process of extracting information from data. Noble (2018), for instance, finds that Google photo automatically link African Americans to “apes” or “animals” and in addition, when retrieving “monkey’s face” on Google image search, results contain the picture of First Lady Michelle Obama. In his study, Friedman (1996) defines these examples as *Preexisting Bias*, which derive from social practice and attitudes. In addition, it is evident that another category of bias: *Technical* bias also exists in algorithm decision-making on Internet content. The algorithms of advertisement, for example, prefer to recommend higher paying jobs to male than female (Datta et al., 2015). Another report supports this view is that, according to Smith (2018), a majority of Americans (74%) maintain that the content they see on social media. These are chosen by algorithms instead of editors, cannot objectively reflect public’s attitude on social issues. Accordingly, the main weakness of bias in data mining technologies is that it could provide misinformation to users.

Third, bias could emerge in the production or interpretation of visualisation, when inducing knowledge from information. In his study, Yau (2017) agrees that practitioner might tell lies or mislead by charts. Truncating value axis, for example, can show bigger changes and transforming continuous variables into categorical can hide partial significant information. In

addition, it is argued that the output of visualisation is not an accurate representation of the real world. D'lgnazio (2015) makes a valid point in that general public often mistakenly consider visualisation as fact, which have *rhetorical* power to influence them conversely.

3.3 Ethic issue

Recently, an increasing number of companies focus on the ethics of new data science technology such as Apple, who assert that they will never comprise users' privacy under any circumstance. This is because Hasselbalch and Tranberg (2016) believe that data ethics become more important for business just like eco-friendly which is not only the demand of investors but also a competitive advantage. According to Floridi and Taddeo (2016), however, apart from "the ethics of data", "the ethics of algorithms" is another major challenge. In his study, Sage (2018) found that "Yellow Vest" movement is accelerated by Facebook's news feed algorithms: the system would prefer to recommend sensationalist articles in which is full of violent and irritated content in order to increase clicks. It is apparent that this practice, using algorithms wrongly interpret data, is legal but unethical

3.4 Summary

In brief, government and large global technology companies increasingly focus on *informational* power in order to control or influence citizens' behaviour. In addition, this informational power can cause bias in society such as racism and sexism. The emergent of "Occupy Wall Street", however, suggests that citizens are dissatisfied with the increasing power of government or group and believe that they are losing autonomy. In addition, the general public have been recognised that everyone would judge events from perspective personal social status or life experience. In this instance, however, we still need to minimise the impact of implicit or unconscious bias. The following section provides guidance for citizens to cope with power and discuss the responsibility for data scientists.

4. Response to challenge

As cited above, social structure can limit agency. In my view, however, social structure and agency might not be completely opposite. There is a valid point made by Kennedy, Poell and Dijck (2015) in that structure only limit on action range rather than determine specific activity. In addition, they also claim that there is no meaning for activities of agency without social structure. It is because structure shapes agency, but agency can also react against it.

4.1 Open data and data protection law

In his study, Foucault (1980) argues that power is not only negative but also can be positive for disadvantaged agents to change the current form of power structure. There is considerable evidence to show that making public data accessible to citizens can balance strong informational power which is controlled by governments. Baack (2015), for instance, holds the view that it is possible to empower citizens by opening raw data to wider public. Other studies also report similar findings of the positive effect of open data movement for decreasing discrimination and privacy in society (Bates, 2012; Kennedy, 2015; Kitchin, 2014). In addition, it is claimed that data protection law is another countermeasure. According to General Data Protection Regulations, for example, Google and Facebook have to face up to £3.8 billion fine due to invasion of privacy (Baxter, 2018).

4.2 Responsibility for data scientist

Based on the discussion above, there are three responsibilities for data scientists to take. The first is to help citizens recognize that data science practice or knowledge is influenced by socio-political power rather than neutral socially. The second is to empower and train agents to collect, analyze and interpret data in order to correct social bias. Finally, for themselves, it is important to maintain ethical standards and critically analyze data science practice in future work or life.

5. Conclusion

To summarise, it is evident that data is a synthesis of societal factors and data and society can interact with each other. In this process, however, there are several challenges we need to recognize. First, data science knowledge is non-neutral and influenced by socio-political power in order to control our behaviour. Second, there could be bias existed in three processes: using data to represent reality, extracting information from data and interpretation of visualisation. Third, a number of data science practice is legal but unethical in view of privacy or public trust in algorithms. Facing these challenges, open data movement and law is effective, but more importantly, it is data scientists' responsibility to keep ethical technology skills, critically interpret cases, and balance the power in society.

References

Baack, S. (2015). Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism. *Big Data and Society, 2*(2), 1-11. https://doi.org/10.1177/2053951715594634

Bates, J. (2012). "This is what modern deregulation looks like": co-optation and contestation in the shaping of the UK's Open Government Data Initiative. *The Journal of Community Informatics*, *8*(2). Retrieved from http://ci-journal.net/index.php/ciej/article/view/845

Baxter, M., GDPR Report. (2018). *EU fines Google £3.8 billion, and that's without a data breach.* Retrieved December 13, 2018, from https://gdpr.report

Braman, S. (2009). *Change of State: Information, policy, and power.* Cambridge: MIT Press

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies, 2015*(1), 92-112. https://doi.org/10.1515/popets-2015-0007

D'Ignazio C., The Center for Civic Media. (2015). *What would feminist data visualisation look like.* Retrieved December 10, 2018, from https://civic.mit.edu

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philos Trans A Math Phys Eng Sci, 374*(2083), 1-5. https://doi.org/10.1098/rsta.2016.0360

Foucault, M. (1980). *Power/Knowledge: Selected Interviews and Other Writings 1972-1977*. New York: Pantheon Books.

Friedman B., & Nissenbaum H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems, 14*(3), 330–347.

Hasselbalch, G., & Tranberg, P. (2016). *Data Ethics: The New Competitive Advantage.* Frederiksberg: Publishare ApS.

Kennedy, H., & Moss, G. (2015). Known or knowing publics? Social media data mining and the question of public agency. *Big Data and Society, 2*(2), 1-11. https://doi.org/10.1177/2053951715611145

Kennedy, H., Poell, T., & Dijck, J. (2015). Data and agency. *Big Data and Society, 2*(2), 1-7. https://doi.org/10.1177/2053951715621569

Kitchin, R. (2014). *The Data Revolution: big data, open data, data infrastructures and their consequences.* London: SAGE.

Mayer-Schönberger, V., & Cukier, K. (2014). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: John Murray.

Mozur, P., Scott, M., & Isaac, M. (2017, Sept 17). Facebook Faces a New World as Officials Rein in a Wild Web. *The New York Times.* Retrieved from: https://www.nytimes.com

Noble, S. (2018). *Algorithms of Oppression: how search engines reinforce racism*. New York: New York University Press.

O'Neil, C. (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy.* New York: Crown Publishers

Paragas, F. C., & Lin, T. (2016). Organizing and reframing technological determinism. *New Media & Society, 18*(8), 1528-1546. https://doi.org/10.1177/1461444814562156

Provost, F., & Fawcett, T. (2013). *Data Science for Business.* California: O'Reilly Media.

Sage, A. (2018, December 6). How Facebook has been exploited to stir up more anger. *The Times.* Retrieved from https://www.thetimes.co.uk

Simon Lindgren. (2014, Mar 24). *E012: Structure and Agency* [Video File]. Retrieved from https://www.youtube.com/watch?v=l9cfJ_PegQE

Smith, A. (2018). *Public Attitudes Toward Computer Algorithms.* Washington: Pew Research Center.

Yau, N., Flowing Data. (2017). *How to Spot Visualization Lies.* Retrieved December 10, 2018, from https://flowingdata.com