The Ethics of Given-off versus Captured Data in Digital Social Research

Josh Cowls
Oxford Internet Institute
1 St Giles
Oxford, United Kingdom OX1 3JS
+44 (0)1865 612777
josh.cowls@oii.ox.ac.uk

Ralph Schroeder
Oxford Internet Institute
1 St Giles
Oxford, United Kingdom OX1 3JS
+44 (0)1865 287224
ralph.schroeder@oii.ox.ac.uk

ABSTRACT

This paper proposes new terminology to enhance understanding of how big data can be used for research, in both commercial and academic contexts. We distinguish between data as given-off and data as captured, and draw on insights from interviews conducted with researchers using such data to elaborate on this distinction. We conclude with a series of recommendations for research design and conduct, based on this re-conceptualization of 'data' and 'capta'.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues – *Ethics, Human safety, privacy*

General Terms

Design, Human Factors, Theory.

Keywords

Big data, personal data, privacy, social media, social research

1. INTRODUCTION

The deluge of data that is generated online has galvanized a wide range of research with human activity as the focus of enquiry – often unbeknownst to the 'participants' themselves. Much of the data is of high value for social scientific enquiry – for example, content created on the most popular social networking sites. For similar reasons but different purposes, these data are of substantial financial value to the platforms which host and hold the content. The recent backlash against the high-profile study of emotional contagion on Facebook [1, 2] points to the uneasy collaboration between corporations and academics in understanding the nature of this data, and the often divergent purposes to which this understanding may be put.

The ethical consequences of using this data – whether for academic or commercial aims – are further complicated by the association between the human subjects of research and the data that they produce. The term 'data' emerges from the Latin 'dare', 'to give'; but in many cases, the human subjects of this research are unaware that, and might be unwilling to allow, the data they generate to be used in this way. Legal and ethical frameworks for the use of this social information may be strengthened by instead deeming it 'capta', or that which is taken or captured[1].

This paper begins by expanding upon this terminology of 'data' and 'capta'. The following section draws on interviews conducted with researchers at the forefront of this area of research (in both academic and commercial settings), to examine whether and how this reconceptualization might operate in practice. The final section discusses the implications for design of future research in light of the preceding theoretical and practical insights. The overall aim is to enable researchers to consider the relevant aspects of various types of data in this area which, we argue, cannot be confined purely to ethical or legal considerations.

2. RECONCEPTUALISING (SOME) DATA AS 'CAPTA'

In the modern world, we are surrounded by technologies and tools designed to capture, in digital form, the signals – be these physical, social or emotional – that we give off to our wider environment. This, of course, is data – and because it is created by us it is also, to varying degrees, about us. From the way data is thus conceptualized, the use of the word data makes sense etymologically: data is the plural of datum, which in turn is the past participle of the Latin verb dare, to give, so a datum is literally 'something given.'

Yet some of the ethical implications of conceptualizing data in this way are unsettling, even if we take a very narrow view of what constitutes data that we as individuals create (excluding data which is generated about us but not by us, for example health records.) This stems from the fact that a far greater proportion of our actions are now recorded in a digital – and thus systematically analyzable – form. For many people, the primary modes of communication have shifted online; consumer electronic devices track our physical movements with unprecedented granularity; and many more of our economic transactions are tied to us as individuals. In this paper, our examples are confined to social media and information seeking online, though we will come back to the broader variety of digital data in the conclusion.

As a result of these technological and social changes, it is harder to conceive of much of the information about ourselves that we generate as being truly data, as 'something given' or 'given off'. In many cases, the information which is created by us, about us, could be better thought of as 'capta' or 'something taken', from the Latin verb 'capere'. There are various factors which justify this reconceptualization of personal data on ethical grounds.

Firstly, due to the vastly greater diversity and sheer degree of data that can be collected about human behavior, individuals are now much more likely to be unknowing data creators. For example, much of the communication which takes place on social networking sites is asynchronous, meaning that communication data must be stored by the networking platforms for later retrieval.

[1] The use of 'capta' as an opposition to 'data' was first offered by Peter Checkland [3], though is used in a different sense here.

This data is rendered valuable because many leading social media platforms use targeted advertising based on this accrued personal data as the basis for their revenue stream, which would be impossible were the data not captured. In this sense, at the point of performing a recordable action – be it writing a tweet or making a credit card purchase – individuals are far less likely to know that their actions will be recorded and analyzed as data. In the interests of creating a 'pure', unadulterated dataset, is it usually not helpful to data collectors to make this explicit (even if the legal terms and conditions provide clarity here, this clarity is unlikely to be something that individuals have an adequate understanding of).

Secondly, and perhaps more applicable to a time when large-scale data collection projects have gained substantial public and press attention, even if individuals know that their actions will be captured as data, they seldom know how it will be used. This relates to the diversity of data which can now be collected as well as the diversity of uses to which it can be put. Neither, necessarily, do data collectors themselves exactly know when or how data they collect will ultimately be used in analysis; the immediate priority is simply collecting and storing the data for later use.

Thirdly, the linking of different datasets relating to the same person or people can create significant added value for those who hold the data. In many cases it is possible to use metadata connected to different data sources to (re) identify individuals and tie together different sources of data about the same person. This further obscures an individual's awareness of when, how and by whom the data she generates will be analyzed – and what the real-world consequences of that analysis may be.

For these reasons, in many cases it may be necessary to reconceptualize data as capta, something taken. Where users don't know that an activity will be systematically recorded for analysis, or don't know in which ways it will be analyzed, or for what purpose, or in conjunction with what other personal information, it seems more appropriate to reclassify this information as capta rather than as something that is 'given' (in both senses of the word) or 'given off'. The lack of control over, or even knowledge of, the uses to which data is put suggests a shift in the relationship between data subject and data holder in the current landscape.

In the following section, we reify this reconceptualization by offering case studies of research utilizing data and capta in both academic and private sector contexts.

3. DATA AND CAPTA IN PRACTICE: PERSPECTIVES FROM CASE STUDIES

3.1 Data and capta in comparative perspective

As part of the project Accessing and Using Big Data to Advance Social Science Knowledge, funded by the Alfred P. Sloan Foundation, we interviewed more than 125 researchers who have used big data in academic, policy-making, private and not-for-profit settings. Here, we present a small sub-sample of interview responses from this project which relate specifically to their views of the status of different types of data and how they should be conceptualized as being captured from or given off by data subjects.

We can begin with an unconventional (compared with other researchers) view that was ventured by Hal Varian, chief economist at Google and emeritus professor at the University of California. Varian discussed how Google's new Consumer Surveys platform offers an alternative approach to standard advertising models based on personal data:

Varian: I'll tell you about one system that is going to be very important in social science and people are only just starting to realize it, and that's the issue of surveying. So as you know conventional survey techniques are under a lot of stress. People are over-surveyed, they don't get any benefits of it, response rates are going down and down, you're lucky to get single-digit response rates these days. [A website which has] implemented Google Consumer Surveys says 'sorry, subscribers only, but we will let you in if you answer this ten-second survey.' … So you've got a motivated user, one who wants to access some content, and they're willing to access the survey, because it means they're getting access to this content. The survey taker likes this because he's getting answers to his survey question, and the publisher likes it because they're paid for people to view this content. Well, here you get a 70 per cent response rate: it's really quite remarkable.

At first glance, this appears to be an 'ideal type' case of data, in relation to the definition provided above; through surveys, individuals volunteer explicit, specific responses. This is a sharp contrast to other more common forms of personal data, such as social media posts, which are volunteered spontaneously, in a socially sensitive context. However, Varian went on to explain how survey responses are combined with data from other sources to provide greater context to the responses given:

Interviewer: And what else do we know about those users? Does Google aggregate other information around them that can then be transmitted around that data?

Varian: So what happens is you get their inferred gender and age, and that's basically built from the website, and then you get their location and so from the location you can pin down their expected income by zip code or by geographic location. So you know where they're coming from, that tells you something about income, you know male/female, and you know ages.

Varian's response demonstrates some of the complexities involved with the distinction between data and capta: in the case of Google Consumer Surveys, survey responses – which could comfortably be characterized as data – are combined with information better understood as capta: a user's location, automatically obtained from their IP address, as well as demographic information based on inference. What we also see is the value of the data: they are not about identifiable individuals, but for Google the value of combining a survey with generic characteristics of users creates value by creating a profile of the media behaviors of populations.

The blurry definitions of data and capta in practice were also in evidence in our interview with Axel Bruns, a Professor at Queensland University of Technology. Bruns has conducted a large number of political science studies using content published on Twitter. He reflected on some of the ambiguities involved with how it should be conceptualized:

Bruns: But, you know, in theory at least at some point it's quite possible every public tweet ever made will be available. So at that point privacy, in a sense, kind of goes out the window, you're still hidden within the archive but if it's searchable someone could search for everything you've ever tweeted. Now, the maximalist [approach is] I guess to say, well, in the end everything will be public anyway. And of course, if your account is public rather than private on Twitter, then at some point it has been public [and] anyone, without having to be a Twitter user, can go to the website – you can look at my account and you can see what I've tweeted, at least the recent tweets. So, the ephemerality and the invisibility of Twitter is probably overrated in that sense: certainly any public tweet is literally public and may also be, of course, in Google or Bing or some other search engine index. [But] there's the expectation that users might realistically have of how visible their tweets will be as they're tweeting. So if I know I've ten followers and I'm tweeting I probably don't expect for my tweet to be on the front page of the Times or whatever, I don't expect my tweet to go really beyond those ten users. If I'm, you know, David Cameron or Julia Gillard or some major politician, I probably expect that all of my tweets will be archived somewhere. And then there's a lot in the middle, there's a big, big grey area in the middle, and a lot of this also has to do not just with how many followers I've got but am I tweeting to hashtags, am I tweeting to very visible hashtags, if I'm a heavy contributor to, let's say, Queensland flood hashtag that was covering our floods here.

Bruns draws attention to the specific context in which thoughts are expressed on social media, and highlights how that context might be neglected if these contributions are used in future analysis – which could give users far greater exposure to their content than originally anticipated. In this sense, by 'following the data' [4], we see how it can be transformed, between creation, collection and analysis. And unlike with Varian's use of data from Google, Bruns' points about Twitter highlights the lack of awareness among both researches and data subjects about data that may become harmful in an unintended way.

A case where this is even clearer is that of Google search queries. In comparison to survey responses, it is easier to classify Google search queries as capta. Seth Stephens-Davidowitz, a quantitative analyst at Google and a Contributing Op-Ed Writer to the *New York Times*, uses the Google Trends platform to investigate sensitive topics such as racism and sexual orientation. He characterized this resource as a boon for social science research:

Stephens-Davidowitz: Anything of a sensitive nature, whether it's racism, child abuse, sexual behaviors of various sorts, drugs, all these topics – the traditional data sources are very, very limited. And this is an area where I surmised, and I think I'm right on this, that Google would be really good at: online, [when] you need information, you have an incentive to tell the truth, [so] if you get off on racist jokes and you want to see them then you've got to type it into Google, so you really have a clear incentive to tell the truth and to tell what you really want. And it's anonymous and everything; all the data is anonymous data, so people are very honest and open.

Here, Stephens-Davidowitz emphasizes the utility of this resource specifically because it is capta, or information analyzed without a user's knowledge; if a user knew that their search queries would be analyzed in this way, where someone might be less willing to conduct such a search, particularly when the query would be so inflammatory in any kind of social environment. (Though it is important to note, as Stephens-Davidowitz does, that the identity of the individuals themselves is kept anonymous.) In this case, captured data without the users' knowledge that it is being captured is valuable precisely because these data cannot be easily obtained in other ways.

3.2 Lessons from academic and business contexts

The cases above emerge from both academic and business research contexts, and it is worth bearing this distinction in mind as we turn to some other cases with lessons across this divide. Fil Menczer, Professor of Informatics and Computer Science at Indiana University was recently authorized to release a large dataset containing around 13 billion click requests collected at his institution, Indiana University. This authorization followed months of negotiation with his institution:

Menczer: we had long interactions with [Indiana University's Human Subject Committee], and discussed the issues and what was the data format and so on, and so at the end of this process, they determined that this was not human subject data because there were no identifiers, and so therefore they gave the green light. So then this went back to the university, to the Policy and Technology Office, and we then discussed with them a protocol to make this data available for research.

Interviewer: Do you feel you are being subject to different impositions in terms of the privacy than a company would be that wanted to release its data to another company?

Menczer: Certainly, there is no doubt that universities have much stronger regulations, and they have fewer resources from the legal perspective to deal with this stuff. [But] the terms of service of Indiana University are actually quite good, compared to normal terms of service that you would find from a corporation. If you actually read it, it is actually written in a way that can be understood. I was very surprised! … I think that the university is really trying very hard to be a good citizen, and to make things clear, and so it makes it quite clear. It also is very restrictive, it does not say we can do whatever we want with the data. But companies generally will have that kind of language.

This case shows some of the obstacles set up in an academic setting to protect against the over-exposure of users. It is notable that in the University's view, this resource did not relate to 'human subjects', as the click requests were not specifically tied to individuals, which made the process of releasing the dataset somewhat smoother. But Menczer's experience highlights how in an academic setting, institutional oversight can help protect against harm to individuals where potentially sensitive information is involved.

As Menczer suggests, the situation is often different in the case of corporations. Yet Wojciech Gryc, the Chief Executive Officer of Canopy Labs, a company which makes predictive analytics accessible to small- and medium-sized businesses, discussed how the profit motive of corporations can also be harnessed to protect against harm to individuals:

Gryc: if I had to summarize why people give us data, because a lot of them have to trust us and we do our utmost to make sure that everything is encrypted and secure. But there is always a risk and I think at a certain point the main reason that people are willing to give us data

> versus maybe other groups or companies is that we can directly tie you giving us that data to potential increases in revenue. So the business value is very clear and obvious. Whereas taking the risk, I'm going to use the academic example, if you're going to work with a PhD student then you give them the data, there is a very small chance that it will ever come into something where you increase profits by 10% and there is the potential chance that the data is going to be lost. And the last thing you ever want to do is send an email that says 'We lost your data.'

As his company's success is dependent on keeping data secure, Gryc argues that this provides a stronger incentive to do so than in an academic setting. Gryc also had wider suggestions for how corporations might be more transparent with users about how personal information is held and used:

> Gryc: I would say make sure that the collection you are doing is something that you're not really hiding but you're also not being blatant about it. … I don't know if this is just a North American thing but now for a lot of websites, when you visit them it says we have a cookie policy and just click x, right? But that's a good way to just let people know that, 'Hey look, data is being collected, there's nothing intense going on, just know'.

The question of protecting the reputation of research and of analyzing user data cuts across academic and commercial uses of data, and indicates the obvious point that it is in the best interests of researchers to be as transparent as possible about collecting data. However, as we have seen, this transparency may go against the value that these data provide precisely because they are not knowingly volunteered.

4. DISCUSSION AND CONCLUSIONS

A maximalist position of complete transparency about collecting digital data seems on the face of it to be a straightforward option that protects data subjects. Such a position is also being put forward in some existing and proposed laws, though data privacy has been labelled as being more of an aspiration rather than a reality [5]. This debate is likely to continue in the coming years and at the same time unlikely to be settled in a definitive way. One implication for researchers is to consider the wider context in which data are collected and used, and to proceed against this wider background rather than merely working within what is permissible.

The implications of this injunction for design include that, apart from considering the immediate ethical and legal aspects of big data research on social media, researchers should be informed by a wider toolkit of factors which includes:

1. An understanding of data subjects' attitudes towards what is given off and what is captured.

2. An understanding of the various ways that researchers think about the value of data and attitudes to its exploitation, and how this fits into changing picture of the implications of data-driven research in an environment that has recently changed significantly.

3. To take into account the broadest possible perspective of the role of digital data in society, weighing costs and benefits (for an analysis for the era before social media, see [6]). This is a tall order, and we have presented only a fraction of interview material and of this broader context here for reasons of space. In the full presentation, we will offer a more systematic typology of current big data practices, researcher attitudes, and what is known about public perceptions of data uses more generally.

There are many experiences from academic and business settings which offer mechanisms by which researchers may better appreciate and mitigate against valid concerns around issues like privacy, transparency and consent. There is also a need, however, to not just collect and derive best practice from these, but also to establish a coherent framework about digital data. Here we have focused on social media, but the line between these and other sources of data are likely to become increasingly blurred (consider mobile phone records and location and purchasing data on the same device).

Currently, in view of public debates that have arisen particularly due to social media research, it is clear that there are emerging mismatches in understandings of data and the sensitivities surrounding them. While ethical and legal guidelines will take time to catch up in providing safeguards against misuses of data, it will be necessary to encourage wider knowledge that can address these mismatches. This knowledge will be based on aggregating not just public attitudes towards data uses, but also collecting the views of researchers about the value of data in different settings: after all, aligning their views, which are currently disparate, with wider and changing societal norms, will be a necessary precondition for resolving any tensions arising from research. The combination of public attitudes, researcher views, and a wider analysis of data-driven research in society, will go some way towards enhancing the mechanisms for bringing about a relationship of enhanced trust between those who engage in digital data research and the society which can benefit from it.

5. ACKNOWLEDGMENTS

We gratefully acknowledge the Alfred P. Sloan Foundation for the funding provided in support of this project, and the invaluable work of our colleagues Eric Meyer and Linnet Taylor.

6. REFERENCES

[1] Kramer A, Guillory J and Hancock J (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences. 111 (24): 8788-90.

[2] Schroeder, R (forthcoming) Big Data and the Brave New World of Social Media Research. Big Data and Society.

[3] Checkland, P and Howell, S (1998) *Information, Systems and Information Systems: Making Sense of the Field.* Chichester, West Sussex: John Wiley & Sons.

[4] Borgman, C (forthcoming) *Big Data, Little Data, No Data: Scholarship in the Networked World.* Cambridge: MIT Press.

[5] Greenleaf G (forthcoming) Sheherezade and the 101 data privacy laws: Origins, significance and global trajectories. Journal of Law, Information & Science.

[6] Rule J (2007) *Privacy in Peril: How We are Sacrificing a Fundamental Right in Exchange for Security and Convenience.* New York: Oxford University Press.