Needles in Haystacks: Law, Capability, Ethics and Proportionality in Big-Data Intelligence-Gathering

Julian Richards, Centre for Security and Intelligence Studies, University of Buckingham

Edward Snowden's revelations in 2013 about the nature and scale of data-gathering in the two intelligence agencies, the NSA in the US and its UK sister agency, GCHQ, added a new dimension to a debate already underway about the transformation of intelligence-gathering in a Big Data age. There is no doubt that Snowden's revelations provided the bow-wave of a fundamentally critical and anti-state stance on the intelligence questions in hand. Such a critical stance is reflected in much of the academic literature about Big Data. In this chapter, such concerns are critically appraised.

The core theme is much of the critical discourse is that of "panoptic panic" described by Lyon (2014, p.6), in which the citizens of liberal democracies increasingly feel threatened by the descent of their societies into an Orwellian dystopia of "mass surveillance" (itself, a contested notion about which more below). There is a sense of creeping anxiety to match the supposedly creeping powers of state intelligence gatherers, whereby increasingly sophisticated techniques can become more pre-emptive and predictive in their stance, with all the attendant risks of misdiagnosis of miscreants leading to wrongful imprisonment, or worse. We are moving, say the critics, into the sort of "pre-crime" nightmare that has been the realm of movies hitherto (Minority Report, Robocop). And this is to say nothing of the risks of a state abusing its increasingly powerful capabilities for political or corrupt purposes.

Also at the core of the critical debate is the perennial question about the balancing of security against freedom in a modern liberal democracy. As transnational threats such as organised crime and terrorism take root, are we at risk of becoming a paranoid "risk society" in Ulrich Beck's conceptualisation (Beck, 2002)? As the state capitalises on sophisticated data-gathering and data-modelling opportunities, are we doing the work of the enemies of democracy for them by turning full circle into a Stasi-like surveillance society? In Beck's analysis, prediction and risk are interrelated concepts: fear of the latter leads to an increasing desire for capability in the former, to attempt to block risks before they happen (Kerr and Earle, 2003, p.68). Such fundamental questions about the desired nature of modern democratic society could hardly be more important, and Big Data may be posing new and particularly challenging questions in the debate.

Epistemological issues

Much of the academic discussion in this area grapples the epistemological shift that Big Data supposedly demands. Big Data is not just about handling data in traditional ways but at much greater volumes. Instead, normative thinking suggests that Big Data opens up a whole new set of techniques and capabilities which move data analysis away from a retrospective building of links between entities and more towards more pre-emptive and predictive possibilities. The whole paradigm of analysis and its outcomes start to be redefined.

In the criminal intelligence sphere, such possibilities were noted relatively early in the 1990s in the shape of "intelligence-led policing", which evolved into "predictive policing". Here, the idea was that data could be used much more intelligently to focus and direct increasingly pressured surveillance and interdiction resources by modelling criminal patterns spatially and temporally. These, in turn, could be used to predict future crime activities and patterns. In a world of public sector cutbacks, the dividends could be significant.

The concept made perfect sense, although subsequent experience has shown this process to be more difficult than it initially seemed. It may be the case that at the root of the problem is a cultural issue within Western policing, whereby senior managers schooled in traditional, pre-information age techniques for tackling crime have not yet understood the potential offered by manipulation of data for intelligence benefits (with some honourable exceptions). For some of these managers, data analysis produces little more than pretty pictures and maps to adorn mandated strategic crime reports.

By the turn of the 21st century, however, the debate has intensified through the seemingly exponential rise in available data and the possibilities for analysing it, as metropolitan societies move much more resolutely into the information age. This explosion in data is being matched by substantial developments in computing power, and in the ability of states and organisations to not only store more data, but to run highly sophisticated analytical techniques against them. In policing and in intelligence more generally, there is evidence that the possibilities of predictive policing are being reappraised.

Of course, we now know through Snowden's revelations that the big Signals Intelligence (Sigint) agencies on either side of the Atlantic were working away throughout this period to capitalise on the new opportunities and expand capabilities to the very limits of computing capability. At the heart of the revelations is the question of the scale of data gathering. In the US, the newly exposed Prism programme revealed an industrial-scale collection of internet metadata from major commercial

internet service providers. Meanwhile, across the Atlantic, the Tempora programme revealed that GCHQ was tapping into trunk trans-global fibre optic cables transiting its shores, in order to amass enormous databases of international communications activity.

There are questions here of the intelligence rationale being adopted by the US and UK states (and, indeed, undoubtedly by numerous other states). There are two components to this question. First, and perhaps most significantly, is the way in which national security conceptualisation and policy have changed through the turn of the century, and particularly in the post-Cold War and post-9/11 era. The terrorist attacks in the US in September 2001 ushered in the War on Terror, and, with it, an avowedly more pre-emptive and offensive-realist security stance in the US and amongst its closest Western allies. President Bush's first State of the Union address after the attacks captured the mood:

> *I will not wait on events, while dangers gather. I will not stand by, as peril draws closer and closer. …. We can't stop short. If we stop now–leaving terror camps intact and terror states unchecked–our sense of security would be false and temporary. History has called America and our allies to action, and it is both our responsibility and our privilege to fight freedom's fight. (CNN, 2002)*

Here we can see a rationale for a more forward-reaching national security stance, in which the transnational dangers of a globalised world are met upstream. From an intelligence-gathering perspective, this is likely to mean two things. First, the supposedly transnational nature of threats such as contemporary terrorism mean that communications and connections between individuals and organisations will be made across global networks. It will not be enough, most probably, to restrict interception of communications to limited, domestic datasets, such "targets" and threats are likely to be operating in the civil sphere. The connection between terror camps in failed states and terrorist attackers in the Western world is likely to be one embedded in particular civilian communities, whose communications will pass on public networks rather than in isolated and dedicated military or diplomatic channels, as might have characterised the targets of primary interest in the past.

If our states are to chase-down such targets, therefore, they will need to dip their hands into public, civil communications and data footprints. In former President Bush's terms, such a task was not just a nice-to-have, but a matter of duty and responsibility for the world's greatest power and its closest allies to "fight freedom's fight". This was a task of the loftiest importance. Thus, the needs and values of democratic societies are balanced against the methods with which to deliver security. And the equation had changed after 9/11. As the UK Prime Minister at the time, Tony Blair, described it: the attacks in New York and Washington had changed at a stroke the "calculus of risk" (Richards, 2012).

Needles and haystacks

The second important component to the post-9/11 change in policy, however, is the methodology adopted to tackle the supposedly changed and enhanced threats posed in the new era. In this question, states could arguably be accused of using old thinking to tackle a new problem.

Policing and intelligence organisations will often use the old analogy of searching for needles in haystacks. In contemporary society, it is assumed that a very small percentage of individuals are plotting to do us serious harm. But those individuals (and their planning) are lost within a vast morass of civil society. The trick is to find who they are and to extract them deftly from the welter of innocent citizens around them, without compromising democratic norms and expectations in disproportionate ways.

As communications behaviour expands and becomes more complex, however, the haystacks (in terms of data generated by modern citizens) start to become much bigger, and much more diverse. Think of the way in which our communications behaviour has changed, and particularly the number of mechanisms any of us use on a daily basis to communicate with others. Not that long ago, the fixed landline telephone was the only method of doing so. Now, many of us use a range of messaging applications for different purposes and different circles of people, with the applications sometimes numbering into double figures. For the intelligence agencies, each of these channels of communication are a potential new haystack in which the sought-after needles may be located.

There is also a retrospective component which seems to be at the centre of modern security threats in metropolitan societies. Intelligence agencies sometimes characterise this as “target discovery” or “target development”. Perhaps not unreasonably, when a serious incident occurs such as the bombing of marathon runners in Boston or the attack on the Charlie Hebdo magazine offices in Paris, the immediate question asked is whether the perpetrators were “on the radar” of the security services. With depressing regularity, it usually transpires that the perpetrators were indeed known to the authorities in some shape or form, but that investigations had not recognised the immediacy of threat or prioritised counter-action highly enough.

The tactical and indeed political challenge for such agencies in the wake of these incidents is twofold. First, there is the immediate tactical question of who else might be closely connected with the perpetrators of the attack and who might have been involved in the planning of these, and possibly future attacks. It appears that the only way we know of how to achieve this (a point to which I will return) is to quickly build the social networks of the perpetrators from available data, and not just

those pertaining to the aftermath of the attacks, but ideally those that have already happened in the crucial pre-attack planning period. This will allow the authorities to home-in and possibly arrest other suspects.

With Big Data and sophisticated analytical techniques, there is also the theoretical possibility of working horizontally rather than vertically, and reaching sideways into the discovery if other networks of individuals who are planning attacks but have not yet carried them out. It is assumed that groups of individuals planning criminal or terrorist activities interact with one another in especially covert and obscure ways, so as to avoid the attention of the authorities. If such specific and unusual patterns of communication could be captured as an algorithm, this could be washed against the wider dataset of public communications to try to spot groups of individuals behaving in similar ways. This opens up the possibility of more pre-emptive and predictive intelligence analysis.

However, the critics pose the question: is the needle-in-a-haystack approach the wrong thinking for a new problem, especially if it entails a massive intrusion into civil liberties through the collection and databasing of industrial-scale quantities of public communications data? The first question here is partly whether mass "collection" of data can be appropriately characterised as mass "surveillance". In much of the media and commentary about what Snowden revealed, the former is often described as the latter (see for example Lyon, 2014, p.3). Florid language suggests that new Big Data capabilities being proposed or exercised by the state allow for a panoptic nightmare in which unseen state officials sit in a room and examine each and every communication and interaction with the internet made by every man, woman and child. In the UK, the Liberal Democrat MP, Julian Huppert, said of the proposed Data Communications Bill (which was defeated in the House of Lords on its first pass) that the bill would give "huge powers to the Home Secretary to require information to be kept about every phone call you make, text you send, facebook image you like, and anything else." (Huppert, 2012).

Such a connecting of everyday activities by you and I with the surveillance activities of the covert state is designed to emphasise the supposedly Orwellian and repressive nature of the modern state in a Big Data age. The French philosopher, Gilles Deleuze, suggested that there was no longer a difference between the toughest and most tolerant of regimes, since, within both, "liberating and enslaving forces confront one another" (1992, p.4). The Foucauldian "society of discipline" becomes the "society of control".

There are a number of risks identified by the critics. At the core of the argument is the basic question of "proportionality" (to use the language of the European Convention on Human Rights) when a state appears to gather and store the vast majority of the public's communications, even if (as discussed

below) a human eye never looks at more than a fraction of it. The fact is, it could be argued, that the data has been gathered and public privacy compromised.

Furthermore, As Lyon (2014, p.9) suggests:

> *The needle-and-haystack argument carries with it a high probability of false positives, which do matter immediately and intensely because the likelihood is high of harm to specific persons.*

The second risk of the needle and haystack approach is therefore that either corruption or incompetence, neither of which are unknown in official milieux, could lead to the wrong person being criticised, arrested, or worse. Even in less dramatic circumstances, individuals could find their future employment prospects severely compromised if they manage to find themselves on the "wrong list", even if the listing happened many years ago. Similarly, insurance could become costly or even unattainable for some.

We can think here of the case of Lotfi Raissi, the Algerian flight instructor who was the first person to be arrested in connection with the 9/11 attacks. Information subsequently revealed that basic social network analysis, in which Raissi had been in communications contact with most of the 9/11 hijackers, had sealed his fate as a suspect. After many months in jail and the threat of dire consequences if the US managed to extradite him from London where he had been arrested, Raissi managed to sue the government for wrongful arrest and be awarded a six-figure sum in damages.

Part of the anxiety here arises from comparing commercial applications of Big Data analysis to the state security sphere. We know that major corporations are increasingly modelling consumer behaviour using large-scale data analysis to good commercial effect. But the science is relatively untested at the time of writing and there are instances of things going wrong. The case of Google Flu Trends is an interesting one. Here, data derived from search terms entered into Google was used to predict the spread of influenza in the US. Despite successful earlier analysis, the 2013 modelling drastically overestimated peak flu levels (Butler, 2013). For sure, this will probably lead to further refinement and development of the algorithm to ensure better performance, rather than the scrapping of the whole approach. But the question could reasonably be asked as to whether the application of such techniques in a national security context could not sometimes lead to more dire consequences for certain individuals and serious breaches of human rights. Do we know enough about these approaches yet to trust them in sensitive security contexts?

Zwitter argues that the questions posed by Big Data and its technological underpinnings may be outrunning the ethicists (2014). If Big Data does indeed mean that the paradigm of intelligence-gathering is fundamentally changed by the current developments, then ethical questions of what is

right and wrong in a modern liberal democracy in this area may have been thrown uncomfortably open. What, for example, does a rightful expectation of privacy mean in a world of comprehensive social networking?

The same dilemma could be said to apply to questions of law. In the UK, the parliamentary oversight committee, the Intelligence and Security Committee (ISC) has conducted a wide-ranging Privacy and Security Inquiry in the wake of the Snowden leaks. While this process has not yet completed at the time of writing, it is likely that one of its more substantive outcomes will be to suggest that the main law in the UK governing intelligence-gathering activities, the Regulatory and Investigatory Powers Act (RIPA), will need to be reviewed and overhauled, not least as it was drafted during the 1990s when modern internet-based communications were only just beginning to emerge at scale.

RIPA allows for derogation from the relevant parts of the Human Rights Act (HRA) that govern the right to privacy, under authorised activities where the interests of national security or threats of serious crime are relevant. Oversight of the government's activities in these areas is provided by a set of independent commissioners (such as the Interception of Communications Commissioner), the ISC, and an Investigative Powers Tribunal (IPT) to which complaints of unlawfulness can be made by any member of the public.

As discussed, the government has also tried to put in place a Data Communications Bill, which would mandate communications service-providers to store and make available on request communications data (that is, metadata rather than communications content) in support of intelligence operations. Following the defeat of this Bill in the House of Lords, and in response to a European Court of Justice ruling that the 2009 Data Retention Directive was invalid, the UK government passed emergency legislation in 2014 called the Data Retention and Investigatory Powers Act (DRIPA) to address, it said, problems of maintaining capability in the face of heightened threats to security from terrorism.

A number of critics have challenged the effectiveness of these oversight and legislative mechanisms in ensuring a proper balance between security and liberty. On the passing of DRIPA, fifteen leading technology law experts wrote and open letter to the House of Commons, lamenting a "serious expansion of the British surveillance state", which , they said, was in potential breach of European law (The Guardian, 15 July 2014).

The defeated 2012 Data Communications Bill is regularly referred to as the "Snooper's Charter" by critics. One of the leading critics, and the lynch-pin in opposition to the bill, is the former leader of the Liberal Democrats and Deputy Prime Minister, Nick Clegg. He has characterised opposition to the bill as being a problem of proportion. While he claims to support the need for strong security in the face

of a changing threat, Clegg protests that "it is not a very British thing to confer or imply guilt on the whole of the nation by retaining records of every website everyone has visited over the course of a year" (BBC, 18 January 2015). Here again we see the conceptual notion that mass collection of data (and, in this case, metadata rather than the actual content of messages) compromises the human rights of the citizens of a modern liberal democracy, despite the protestations from the security services that they are merely amassing the haystacks so that they have a better chance of finding the needles. What is right or wrong here is an ethical question, but also a legal one in terms of defining and indeed updating the powers of the intelligence agencies.

It is also the case that the oversight mechanism in the UK is not necessarily trusted by all. The Guardian newspaper, which has been one of the primary outlets for Snowden's leaks, was quick to point out that the Investigative Powers Tribunal's ruling against GCHQ in February 2015 for failing to make known to ministers the extent of its use of capabilities revealed by Snowden up to December 2014, was not only "unlawful", but was the first ruling against an intelligence agency in the IPT's fifteen-year existence. Leaving aside asinine debates about the difference between the terms "unlawful" and "illegal" (despite repeated implications to the contrary in The Guardian, GCHQ has not yet been found to have broken any laws) it is clearly the case in much of the critical coverage of these issues that the ISC committee and the IPT lack credibility for being too much a part of the establishment.

Conclusions - countering the critics

In academic terms, the debate about the ethics of Big Data for intelligence-gathering are somewhat complex in their characterisation. It is probably fair to say that critical debates of the "panoptic panic" kind are probably in the ascendant, bolstered by widespread media commentary from civil libertarians. Underpinning the critical discourse is a fundamental distrust of state agencies to do the right thing ethically or indeed legally; a view given extra impetus by the scale and breadth of some of the capabilities revealed by Snowden. The critics would probably argue that the state has the whip-hand as it holds the powers, and can also do so in secret. Were it not for Snowden, argue the critics, none of this would have come to light.

It could be argued, however, that there are some important counter-arguments to place before the critics. Firstly, there is much evidence from those with an intimate working knowledge of how the intelligence agencies operate – including the ISC committee, for example – that operations are conducted with a scrupulous attention to proper authorisation and accountability. Certainly, the IPT's ruling against GCHQ was embarrassing and will cause questions to be asked amongst its management,

but the fact remains that there has still been no evidence whatsoever that GCHQ has broken any law in any of its recent operations. One of its former directors, Sir David Omand, has written that the culture of compliance with the law has become firmly embedded in the agency's daily working culture (Omand, 2012). While the ISC and IPT may be criticised for never (until very recently) delivering any verdict critical of the intelligence machinery, it is doing a disservice to the members of those bodies to suggest fundamental corruption in their activities. At the same time, bureaucracies do sometimes make mistakes.

The slippery and sometimes deliberate mutation in public discourse of large-scale collection of data into "mass surveillance" is misleading and unhelpful. Certainly, there are entirely reasonable questions to be asked about proportionality, and we should never lose sight in a liberal democracy of the risks of an erosion of our values in the face of creeping surveillance powers. But to suggest that the intelligence agencies are scrutinising each and every interaction we make on a daily basis and looking for "thought-crimes" is fanciful if not to say ridiculous. There is much evidence to suggest that a tiny proportion – less than 1 percent – of all collected data is ever reviewed by a human analyst. Such a comprehensive intrusion into privacy for such small returns may have very pertinent ethical questions attached to it, but to suggest that these processes are akin to a reborn Stalinist Great Terror or Cultural Revolution are historically and morally inaccurate and inappropriate.

Part of the problem is a technical one, in terms of the best way to find the needles in the haystacks. Critics will point out that the traditional methods of databasing vast amounts of data are surely outmoded in an age of greatly enhanced analytical tools and capabilities. It must be possible to pinpoint targets of interest – when they need to be targeted – in much more focused ways that minimise collateral intrusion to the extents that Snowden has described?

However, experts and insiders will often suggest that there is, as yet, no known better way of finding targets within the morass of public communications. Furthermore, key targets can often only be found through their connection with other key individuals as and when they surface. And retrospective analysis following target discovery is a critical part of the process. Without some form of data retention, it is difficult to see how discovery of the planning activities of newly-emerging targets of interest can ever be achieved. Without using current techniques, it is difficult to see how we could ever answer a question as to the circle of people that the Charlie Hebdo attackers came into contact with, when they were planning their attacks. We would only ever be able to look forwards and react to the next attack rather than try to pre-empt any nascent attack plans.

None of this is to say that new solutions to such problems will never be found, which may allow us to adopt less intrusive and more effective mechanisms of target discovery. We cannot know what we do

not yet know, and the pace at which computing technology is developing clearly means that any predictions are unwise. In a sense, this is the promise and opportunity of Big Data: that it will not just give us more, but will fundamentally change the way we think about and address such difficult questions.

Bibliography

BBC. *Nick Clegg defends opposition to 'snoopers' charter'*. (18 January 2015). From http://www.bbc.co.uk/news/uk-politics-30870442 accessed 17 July 2015

Beck, U. "The Terrorist Threat: World Risk Society Revisited". *Theory, Culture and Society*, 19/4 (2002)

Butler, D. "When Google got flu wrong". *Nature,* 494/7436 (2013)

CNN. *Bush State of the Union address*. From http://edition.cnn.com/2002/ALLPOLITICS/01/29/bush.speech.txt/ accessed 17 July 2015

Deleuze, G. "Postscript on the Societies of Control". *October*, 59 (1992)

Guardian, The. *Academics: UK 'Drip' data law changes are 'serious expansion of surveillance'*. (15 July 2014). From http://www.theguardian.com/technology/2014/jul/15/academics-uk-data-law-surveillance-bill-rushed-parliament accessed 17 July 2015

Huppert, J. *Communications Data Bill cannot proceed*. Liberal Democrat Voice (11 December 2012). From http://www.libdemvoice.org/julian-huppert-mp-writes-the-data-communications-bill-as-it-is-simply-cannot-proceed-32103.html accessed 17 July 2015

Kerr, I., and Earle, J. "Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy". *Stanford Law Review Online*, 66/65 (2013)

Lyon, D. "Surveillance, Snowden, and Big Data: Capacities, consequences, critique". *Big Data and Society* (July-December 2014)

Omand, D. *Securing the State*. London: C. Hurst and Co. (2012)

Richards, J. *A Guide to National Security: Threats, Responses and Strategies*. Oxford: Oxford University Press (2012)

Zwitter, A. "Big Data Ethics". *Big Data and Society* (July-December 2014)