Fusing Algorithms and Analysts: Open-Source Intelligence in the Age of 'Big Data'

Christopher Eldridge, Christopher Hobbs and Matthew Moran[1]

This is an Accepted Manuscript of an article published by Taylor & Francis in *Intelligence and National Security* on 13 December 2017, available online: http://www.tandfonline.com/doi/full/10.1080/02684527.2017.1406677

Abstract

In the age of 'Big Data', the potential value of open-source information for intelligence-related purposes is widely recognized. Of late, progress in this space has increasingly become associated with software that can expand our ability to gather, filter, interrelate, and manipulate data through automated processes. While the trend towards automation is both innovative and necessary. However, techno-centric efforts to replace human analysts with finely crafted algorithms across the board, from collection to synthesis and analysis of information, risk limiting the potential of OSINT rather than increasing its scope and impact. Effective OSINT systems must be carefully designed to facilitate complementarity, exploit the strengths, and mitigate the weaknesses of both human analysts and software solutions, obtaining the best contribution from both. Drawing on insights from the field of cognitive engineering, this article considers at a conceptual level how this might be achieved.

Introduction

In August 2013, a large-scale chemical weapons attack took place in Ghouta agricultural belt in the suburbs of Damascus, Syria. Described by UN Secretary General Ban Ki-moon as the 'most significant confirmed use of chemical weapons against civilians since Saddam Hussein

[1] Department of War Studies, King's College London

used them' in Halabja in 1988, a United Nations Mission including experts from the Organisation for the Prohibition of Chemical Weapons (OPCW) and the World Health Organisation (WHO) quickly confirmed the use of the nerve agent sarin in the attack.[2] The incident was the latest in a series of chemical weapons attacks allegedly perpetrated by the embattled Assad regime, but the Ghouta incident eclipsed past incidents in terms of scale. Preliminary assessments estimated that some 1,400 people were killed in the attack, including over 400 children.[3]

In response to the Ghouta incident, the governments of the United States, France and the United Kingdom, among others, published intelligence assessments that documented some of the evidence against the Assad regime. These reports were significant in that they served to anchor respective national discussions on the prospect of military intervention in Syria. The reports were also notable for their reliance on open sources, that is to say, information drawn from relevant, publicly-accessible sources whatever their medium of dissemination. Indeed, this was one of the first occasions that the role of open sources was so extensively credited in intelligence assessments of such high importance, and it provided an insight into contemporary perceptions regarding the role and value of open-source information, both within the intelligence sector and beyond. The US assessment, for example, acknowledged its reliance on a "significant body of open-source reporting", including "videos; witness accounts; thousands of social media reports from at least twelve different locations in the Damascus area; journalist accounts; and reports from highly credible nongovernmental organizations".[4]

---

[2] "Syria crisis: UN report confirms sarin 'war crime'", *BBC News*, 16 September 2013, http://www.bbc.co.uk/news/world-middle-east-24113553. See also "United Nations Mission to Investigate Allegations of the Use of Chemical Weapons in the Syrian Arab Republic. Final Report", United Nations General Assembly, Sixty-eighth session, Agenda item 33, 13 December 2013, http://undocs.org/A/68/663.

[3] "Government Assessment of the Syrian Government's Use of Chemical Weapons on August 21, 2013", Office of the Press Secretary, The White House, 30 August 2013, https://obamawhitehouse.archives.gov/the-press-office/2013/08/30/government-assessment-syrian-government-s-use-chemical-weapons-august-21.

[4] The White House Office of the Press Secretary, "Government Assessment of the Syrian Government's Use of Chemical Weapons on August 21, 2013," August 30, 2013, https://obamawhitehouse.archives.gov/the-press-office/2013/08/30/government-assessment-syrian-government-s-use-chemical-weapons-august-21.

The potential value of open-source information for intelligence-related purposes is now widely recognised.[5] Over the last two decades, technical developments have facilitated both a dramatic increase in the amount of information available online and a major shift in how it is published. Internet users no longer simply consume information, they produce it, using social media to engage with a potentially global audience. On one hand this presents an opportunity for organisations seeking to gain or maintain an investigative or analytical edge. In the age of 'Big Data', the sheer volume of information available online offers unprecedented analytical opportunities in areas ranging from terrorism to nuclear proliferation. On the other hand, this vast and fluid information environment poses an overwhelming challenge: the volume, format, accessibility and quality of information are all variables that are constantly changing.

An open-source analyst can perhaps be likened to a gold prospector, carefully sifting through the informational mud of the World Wide Web in search of those nuggets of information that help to shed light on a particular problem or puzzle. Certainly, a focused individual combining an analyst's mindset with subject-matter expertise and a knowledge of the various tools and techniques that can make online searches more efficient can draw much from the ether using the 'prospector' approach. But progress is inevitably slow and the analyst is battling against a flood of information that is rising at an incredible rate.

As a consequence, progress in this space has, of late, increasingly become associated with software that can use automated processes or visualisation capabilities to expand the analyst's ability to gather, filter, interrelate, and manipulate data. Yet while tools for exploiting large datasets have been successfully applied in areas such as economics and the natural

[5] Former Open Source Center Director Douglas Naquin argued in 2010 that "An organization that invests in open source today is akin to an individual who invested in Google in its first year. OSINT has always been an integral component in intelligence, but in five years, I believe the value proposition can only increase. An organization with an appreciation for OSINT's value and potential will be the most effective in the future". See Director of Central Intelligence, *INTellingence: Open Source Intelligence* Washington, DC: Director Open Source Center, 2013, https://www.cia.gov/news-information/featured-story-archive/2010-featured-story-archive/open-source-intelligence.html.

sciences, using computerised tools to collect and analyse OSINT poses particular challenges. Intelligence-relevant information is both incredibly diverse in nature and highly context-specific. Information that is valuable in one context may be insignificant in another. Moreover, in addition to the trend analysis characteristic of other uses of Big Data, intelligence or other investigative work often seeks to identify *every* available, relevant piece of data; the devil is almost always in the detail here. Further, the credibility of every document must be assessed, taking the author's motivations and expertise into account. The nuanced judgements necessary to make these types of decisions are likely to be beyond the capabilities of even the most advanced software solutions for some time to come.

Against this background, this article has two objectives. First, the analysis explores the issues faced by those seeking to exploit open-source information for intelligence purposes at a time when information overload has become the norm. To do this, we consider how this area of activity might be affected by the opportunities and challenges offered by the "Big Data" phenomenon. While the role of open-source information in intelligence has received considerable academic attention in recent years, there has been little in the way of thorough, critical discussions about how software tools can and should be integrated into open-source collection and analysis. Second, we draw on insights from the field of cognitive engineering to inform a discussion of how, at a conceptual level, the design of systems that facilitate complementarity, exploit the strengths, and mitigate the weaknesses of both human analysts and software solutions, obtaining the best contribution from both in an integrated system, might be achieved. The field of cognitive engineering offers a wealth of experience in the design and understanding of human-computer partnerships that has not, as yet, been considered or applied in the intelligence context.

The Nature of Open-Source Intelligence

The use of open-source information to support intelligence activities and other investigative research is not new; intelligence work has long relied on a combination of secrets and publicly-available information.[6] While the exploitation of open-source information for intelligence purposes is well-established, however, there has always been a certain amount of confusion regarding the nature of OSINT. Is open-source intelligence simply another way of describing various sources used as part of a conventional research process? What is distinct about the approach and contribution here? Much of the confusion relates to the fact that, as Michael Warner suggested, the term intelligence is often understood as "that which states do in secret to support their efforts to mitigate, influence, or merely understand other nations (or various enemies) that could harm them".[7] The term OSINT thus appears to be an oxymoron; the open-source premise sits uncomfortably alongside the secrecy that the term 'intelligence' evokes.[8]

Before considering the impact of 'Big Data'-related developments, therefore, it is first necessary to clarify what we mean by open-source intelligence and how it contributes to the broader intelligence function.[9] The first point to note here is that open source intelligence is an outlier to the conventional strands of intelligence, and its definition is the subject of some debate.[10] A useful starting point is the definition offered by Wirtz and Rosenwasser, who describe (OSINT) as the "insight gleaned from publicly available information that anyone can

[6] As early as 1947, former CIA Director Allen Dulles claimed in testimony to the Senate Committee on Armed Services that open sources could "supply us with over 80 percent […] of the information required for the guidance of our national policy". Cited in Stevyn Gibson, "Open source intelligence", *The RUSI Journal* (2004), Vol.149, No.1, p.20. See also Laura Calkins work on the history of US-UK open source intelligence cooperation: Laura Calkins, "Patrolling the Ether: US–UK Open Source Intelligence Cooperation and the BBC's Emergence as an Intelligence Agency, 1939–1948", *Intelligence and National Security* (2011), Vol.26, No.1, pp.1-22.

[7] Michael Warner, "Sources and Methods for the Study of Intelligence," in *Handbook of Intelligence Studies*, ed. Loch K. Johnson (Routledge, 2007), p.17.

[8] This issue forms part of the discussion in Hamilton Bean, *No More Secrets: Open Source Information and the Reshaping if US Intelligence* (Santa Barbara: Praeger, 2011).

[9] James J. Wirtz and Jon J. Rosenwasser, "From Combined Arms to Combined Intelligence: Philosophy, Doctrine and Operations", *Intelligence and National Security* (2010), Vol.25, No.6, p.736.

[10] Ibid.

access by overt, non-clandestine or non-secret means to satisfy an intelligence requirement".[11] They go on to note that "using this expansive definition, information that is proprietary but acquired by legal means, e.g. certain law enforcement or industry data, falls under the rubric of open source intelligence". Beyond this, we would highlight the political or security significance of OSINT and argue that the legality of the collection process should be linked to a robust ethical approach. Significantly, this information is not self-generated by the actor engaged in collection and is subject to particular information access choices and budgetary parameters that vary across individuals and organisations.[12]

As the above example of Syria demonstrates, OSINT has become a vital information stream in intelligence work. Alan Dupont noted in 2003 that estimates of the amount of US intelligence derived from unclassified, publicly-available sources ranged between 40 and 95 percent, although 80 percent is the figure most frequently referenced; more recent articles make similar claims.[13] As we will discuss below, there is an enormous—and constantly growing—volume of information available in the public domain. But OSINT offers important advantages beyond its utility as a source of information for intelligence analysts. In Dupont's view, one of the more important advantages of open-source intelligence is its cost-effectiveness: expertise from beyond as well as within the intelligence complex can relatively easily be tapped and exploited, thus "freeing up scarce resources for tasks where OSINT is less able to contribute;" similarly, open-source information is simpler to disseminate to facilitate collaboration and

[11] Wirtz and Rosenwasser, "From Combined Arms to Combined Intelligence: Philosophy, Doctrine and Operations", p.736.

[12] Specific examples of open-source information in our definition include information openly available on internet sites; information on "members only" sites to which an actor has legitimate access; information stored in databases to which the actor subscribes; and information, such as publicly-available satellite imagery, for which the actor has paid for access. The definition excludes information created as a result of the actor's own activities (inspections carried out by a treaty verification organisation, for example) or obtained via hacking, deception, or espionage. Of course, actors have many options regarding, for instance, databases to which they subscribe, and different actors make different purchasing decisions. This has consequences for the specific corpus of open-source information available to each actor.

[13] Alan Dupont, "Intelligence for the Twenty-First Century", *Intelligence and National Security* (2003), Vol.18, No.4, p. 26.

reporting.[14] A practical example is provided by Monica Den Boer, who explained in a 2015 article that EU states and agencies have recently increased the amount of open-source information that they share via a restricted-access website because it can be collected without breaking the law.[15]

Clearly, information gleaned from open sources has long had the potential to make a significant contribution to the broader process by which intelligence is generated. When sufficient resources are made available — not only financial but also human, structural, organisational, and information resources — the collection and analysis of open-source information can make important contributions to investigative activities of many types. Reporting by the *Guardian* newspaper provides a recent example from the world of journalism. *Guardian* reporters sifted through the enormous "WikiLeaks" data set to isolate information related to the use of improvised explosive devices (IED) in Iraq. According to the *Guardian*, between 2004 and 2009, there were about 7,500 IED-related attacks, and another 8,000 unexploded IEDs were identified and cleared. Using this data, the *Guardian* reporters were able to create a visualisation of how the use of IEDs changed over time, and in which areas of the country.[16] What, then, do the recent, dramatic changes in the ether mean for the work of the open-source analyst? What is the significance of the rise of 'Big Data'?

'Big Data': Opportunities and Challenges

[14] Alan Dupont, "Intelligence for the Twenty-First Century", *Intelligence and National Security* (2003), Vol.18, No.4, p. 26.
[15] Monica Den Boer, "Counter-Terrorism, Security and Intelligence in the EU: Governance Challenges for Collection, Exchange and Analysis," *Intelligence and National Security* (2015), Vol.30, No. 2-3, p. 412.
[16] Simon Rogers, "Wikileaks Data Journalism: How We Handled the Data," *The Guardian*, January 31, 2011, sec. News, https://www.theguardian.com/news/datablog/2011/jan/31/wikileaks-data-journalism.; Nathan Yau, "Infographics: Winds of Change", *The Economist*, July 06, 2013, sec. News, https://www.economist.com/news/books-and-arts/21580446-revolution-taking-place-how-visualise-information-winds-change

In recent years, the term 'Big Data' has become a buzzword of sorts in discussions relating to developments online. Despite its ubiquity in academic and popular discourse, however, the term is poorly understood and often ill-defined. It has become a popular way to refer to the dramatic increase in recent years of the amount of data available online. But the concept represents more than vast quantities of information. McAfee and Brynjolfsson argue that the concept comprises three core factors: volume, variety and velocity.[17] The issue of volume is perhaps the one most frequently highlighted in popular commentary due to the staggering figures involved. According to a prediction published in 2014 by an IT analysis company, the "digital universe" doubles in size every two years and will reach 44 zettabytes by 2020.[18] Indeed, for Kevjn Lim, it is the process of digitalisation that is the main driver behind the Big Data phenomenon. In 2000, 25 percent of the world's stored information was digital; by 2013, it was over 98 percent.[19] Equally important for understanding the nature of 'Big Data', however, is the variety of the data sources: "Big data takes the form of messages, updates, and images posted to social networks; readings from sensors; GPS signals from cell phones, and more".[20] Furthermore, "many of the most important sources of Big Data are relatively new".[21] It is only since the mid-2000s that social networks such as Facebook and Twitter began to gain real momentum. Finally, the speed of data creation is regarded by many as even more important than the volume. On this issue of velocity, McAfee and Brynjolfsson note, "Real-time or nearly real-time information makes it possible for a company to be much more agile than its competitors".[22]

[17] Andrew McAfee and Erik Brynjolfsson, "Big Data: The Management Revolution," *Harvard Business Review*, October 1, 2012, https://hbr.org/2012/10/big-data-the-management-revolution.
[18] A zettabyte is 1 X 10^{21} bytes, i.e. 1 billion terabytes. The prediction is from Dell EMC, "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things," April 2014, https://www.emc.com/infographics/digital-universe-2014.htm.
[19] Kevjn Lim, "Big Data and Strategic Intelligence," *Intelligence and National Security* 31, no. 4 (June 6, 2016): 619–35.
[20] McAfee and Brynjolfsson, "Big Data.", 63
[21] Ibid.
[22] Ibid.

This argument is stems from the business context but the logic applies across a range of fields; access to real-time information flows opens up unprecedented analytical opportunities.

Thinking about Big Data from an intelligence perspective, Degaut identifies several important characteristics of open-source intelligence work in what he terms the "information revolution" environment.[23] These include interactivity, or the ability to connect seemingly isolated actions or events; proliferation of new and unknown actors and information sources, increasing the importance of robust procedures for assessing credibility; and feedback, meaning that policymakers receive information in short time frames from many quarters, with which "traditional" intelligence agencies are expected to compete.[24] Degaut also relates this to the issue of "speed," noting that while Big Data does indeed facilitate near-real-time global dissemination of news and other information, it also encourages policymakers' expectations that they will receive relevant, analysed intelligence in near-real-time.[25]

The phenomenon of Big Data presents considerable opportunities and challenges to information-focused activities across the government and private sectors. Consider, for example, the analytical possibilities associated with social media, arguably the best-known class of 'Big Data' and certainly "the largest body of information about people and society that we have ever had".[26] Advertising agencies, for example, can purchase access to usage data from social media and internet platforms and then use tailored analytical software to identify trends that may provide a commercial advantage. Such analyses do not seek to identify specific people or activities, but rather aggregate vast numbers of discrete data events to characterise, say, purchasing patterns over time. Sociologists, anthropologists, and other scientists also use this

[23] Marcos Degaut, "Spies and Policymakers: Intelligence in the Information Age", *Intelligence and National Security* (2016), Vol. 31, No. 4, pp. 515-17.
[24] Marcos Degaut, "Spies and Policymakers: Intelligence in the Information Age", *Intelligence and National Security* (2016), Vol. 31, No. 4, pp. 515-17.
[25] Marcos Degaut, "Spies and Policymakers: Intelligence in the Information Age", *Intelligence and National Security* (2016), Vol. 31, No. 4, pp. 515-17.
[26] David Omand, Jamie Bartless, and Carl Miller, "#Intelligence" (DEMOS, 2012). p.15, 25.

kind of data to study patterns of human behaviour, economic factors, etc. On this point, Omand, Bartlett and Miller argue that research based on data obtained from social media sources could contribute to the understanding of phenomena such as "thresholds, indicators and permissive conditions of violence; pathways into radicalisation; …analysis of how ideas form and change; and investigation of the socio-technical intersections between online and offline personae".[27] For this sort of aggregate research, the possibilities presented by the internet are nearly endless and expanding daily. Furthermore, software companies have developed powerful tools that can perform the collection, processing, and analysis that enables such research at speed. Some of the potential here was predicted by academics writing in the mid-2000s on the implications of technological advancement for the collection, analysis and production of intelligence. Dupont, for example, noted the significance of developments in "high-speed data processing" as he lamented "how little academic attention has been devoted to the changes that are taking place in the technology, management and integration of the intelligence systems that [would] underpin any 'revolution in military affairs'".[28] Yet nobody could have predicted the scale of the increase in information flows, nor the extent to which the boundaries between reality and the online world would become blurred by our willingness to project ourselves into the ether.

At the same time, these opportunities are counterbalanced by significant challenges. From a technical perspective, the sheer scale of these enormous data sets and the fact that they are increasingly 'unstructured' – drawn from diverse and often unconventional sources and therefore not sharing a common organisational structure – makes them difficult to process using traditional database and software techniques. Customised data analytics approaches are required to aggregate and manage data effectively. This leads to another problem: "the technology landscape in the data world is evolving extremely fast" and thus to effectively

[27] Ibid.
[28] Alan Dupont, "Intelligence for the Twenty-First Century", *Intelligence and National Security* (2003), Vol.18, No.4, p.15.

harness the analytical potential of Big Data requires a flexible, responsive and innovative technical approach.[29] Beyond these technical issues, actors engaged in this space must also consider the legal and ethical implications of data capture and analysis.[30]

For intelligence agencies and other investigative organisations, the analytical opportunities mentioned above are equally valid. Lim argues, for example, that much of the utility of 'Big Data' for intelligence analysis lies in the possibilities of trend analysis for helping to establish that an event or pattern is occurring or has occurred. Tailored algorithms can also "reveal anomalies contained in large datasets when compared to historical data and are applicable to counter-terrorism and other police, security and defence intelligence scenarios. Algorithms 'crawl' through data sources identifying inconsistencies, errors and fraud".[31] Beyond this, modern communication technology instantly brings news reports from around the world to the investigator's desk. Privately-owned satellites take thousands of images of the earth that can be purchased and analysed. Databases offer information that can be used to assess trade flows and search for proliferation-relevant trafficking.[32] Human activity is being documented, and those records are being made public, at a rate far beyond any other time in history. All of these developments open or expand avenues of enquiry for the analyst or investigator.

There are, however, additional challenges to be considered. Principal among these is the difficulty of targeted information gathering described in the previous section. If we continue with our example of trend analysis, the value of this approach in identifying patterns can also be limited in the context of efforts to find the proverbial 'needle in the haystack', particularly

[29] Eric Spiegel, "Six Challenges of Big Data," *WSJ*, March 26, 2014, https://blogs.wsj.com/experts/2014/03/26/six-challenges-of-big-data/.
[30] For a variety of perspectives on these issues, see Anno Bunnik et al., *Big Data Challenges - Society, Security, Innovation and Ethics* (Basingstoke: Palgrave MacMillan, 2016).
[31] Neil Couch and Bill Robbins, "Big Data for Defence and Security," Occasional Paper (Royal United Services Institute, September 2013).
[32] See, for example, Jennifer Webster et al., "PNNL Strategic Goods Testbed: A Data Library for Illicit Nuclear Trafficking, PNNL-SA-102611," *Proceedings of the Information Analysis Technologies, Techniques and Methods for Safeguards, Nonproliferation and Arms Control Verification Workshop*, May 12, 2014, 168.

when the ultimate goal is to identify *every* highly-relevant bit of data that is available in the open sources. The big picture view provided by trend analysis can prevent analysts from seeing discrete pieces of information that make up the mosaic of intelligence. Closely linked to this is the challenge posed by volume. On one hand, there may be more information for the analyst to find, but on the other, the nuggets of value are surrounded by ever increasing quantities of irrelevant data.

This point also links to some of the new questions prompted by 'Big Data', particularly about the distinction between, and comparative value of, objective and subjective knowledge. More data is not the same thing as better data and does not automatically lead to more credible results. Users of Twitter, for example, do not comprise a statistically representative sample of the global population. The increase in the amount of available data does not obviate the need for "subjective" interpretation.[33] Indeed Hollnagel and Woods contend that, "the belief that more data or information automatically leads to better decisions is probably one of the most unfortunate mistakes of the information society".[34]

Rise of the Machines: The Turn to Automated Approaches

Even the brief overview above makes clear that 'Big Data' presents myriad opportunities and challenges to open-source analysts. In our view, there is no question as to *whether* analysts need the assistance of well-designed applications to effectively exploit this flood of data. Technological developments gave rise to 'Big Data' and they must also help harness its potential. Rather, the question is *how* technology can be best applied to the problems at hand**.** How should new applications be integrated with existing tools and information already collected? What are the costs and benefits of adoption?

[33] Danah Boyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication and Society* 15, no. 5 (2012): 667–69.
[34] Eric Hollnagel and David D. Woods, *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering* (Taylor and Francis, 2005).

In this new information environment, governments and the private sector initially focused on enhancing collection activities. Examples here range from assistance with improving the efficiency of internet searches to the agglomeration of data from multiple sources, both internal and external. Increasingly, however, the focus has shifted towards analysis. At stake here is the ability of algorithmic software to capitalize on advances in machine learning and, operating within certain parameters, to make sense of the information gathered. An earlier point referenced the manner in which private sector companies analyse data from social media to guide marketing campaigns. Another example from this sector is the research and development in natural language generation software that has enabled coherent and relevant narratives to be drawn from the enormous data sets gathered in the ether. This technology has, in turn, been used by media companies such as the Associated Press and Forbes to increase by an order of magnitude the number of stories they can produce in certain subject areas, resulting in greater revenues.[35] In this context so-called 'robot journalists' can provide near real time reporting and generate personalized targeted stories for readers from different demographics.[36] Ultimately, this application of automated approaches contributes to the sense of omnipresence and field leadership that media organisations crave in the 24-hour news context.

Clearly, the implications for intelligence agencies are significant. Automated approaches that are capable of both collecting and making sense of information can serve as a powerful force multiplier for analytical efforts. In theory, the extended analytical reach provided by these tools allows for increased contextual awareness, while the potential for data correlation may offer the possibility of a certain measure of prediction. The IARPA (Intelligence Advanced Research Projects Activity) OSI (Open Source Indicators) programme,

[35] Automated Insights, "Automation Helps AP Publish 10 Times More Earnings Stories," January 29, 2015; McAfee and Brynjolfsson, "Big Data."
[36] The Slow Journalism Company, "Delayed Gratification - Rise of the Robot Journalist," August 15, 2014, http://www.slow-journalism.com/from-the-archive/rise-of-the-robot-journalist.

launched by US intelligence organisations in 2012, is a good example in this regard. The OSI programme seeks to develop ways and means to anticipate significant societal events through the automated analysis of a diverse set of open-source information, including patents, scientific and technical information, social media, search engine queries, public webcams, satellite imagery and the media.[37] OSI funds a variety of projects that utilize artificial intelligence tools and concepts such as probabilistic logic to make predictions about events such as "civil unrest, political elections, economic crises and disease outbreaks".[38] One of these is the EMBERS (Early Model Based Event Recognition using Surrogates) project, an industry-university partnership seeking to "beat the news" with a system designed to "continually monitor data sources 24x7, mine them to yield emerging trends, and process these trends into forecasts". The EMBERS project has a particular focus on attempting to forecast civil unrest across 10 Latin American countries, but such an approach could potentially be applied elsewhere if it is successful.[39]

It is clear that well-designed applications are needed to effectively exploit the flood of open-source information. But the implications of algorithmic advances extend far beyond the applied processes of collection and analysis. This marked technical turn also raises important questions that touch at the heart of what OSINT is understood to be. Is the conventional understanding of OSINT still relevant in the age of 'Big Data'? Does the future of OSINT lie wholly in automation? Is the role of the analyst diminished with each technical advance?

The answers to these questions are not immediately obvious, not least because some of the core issues at stake are often obscured by a false dichotomy. With the rise of 'Big Data' has come the notion of or desire for end-to-end technological solutions, encompassing collection,

[37] Jason Matheny, "Open Source Indicators: Intelligence ARPA" (Office of the Director of National Intelligence, 2013), https://www.aaas.org/sites/default/files/Jason_Matheny_AAAS_FBI.pdf.
[38] Ibid.
[39] Naren Ramakrishnan et al., "'Beating the News' with EMBERS: Forecasting Civil Unrest Using Open Source Indicators," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14 (New York, NY, USA: ACM, 2014), 1799–1808.

processing and analysis. This, in turn, presents the effective exploitation of open-source analysis as an either-or choice between analysts and algorithms. When considering OSINT, there is a strong temptation to think that it will only be able to retain and augment its value if finely crafted algorithms replace humans across most of the OSINT enterprise, from the collection and exclusion of information to the integration and synthesis of different data streams.

Yet there is a fundamental flaw in this technologist logic: software solutions cannot yet replicate the complex process of human judgement or the nuanced insights that come from deep subject matter expertise.[40] On this point, Hare and Coghill note that the rapidly-developing field of artificial intelligence is making great strides in cognitive activities.[41] The recent success of Google's artificial intelligence division in programming a machine to master the Chinese game of Go without help from human players is a good example in this regard. In developing an agent that engages in *tabula rasa* learning, where the programme becomes its own teacher and there is no need for human intervention, it is claimed that the programme is "no longer constrained by the limits of human knowledge". Yet advances here are offset by the challenges of moving beyond specific and highly structured domains. Hare and Coghill argue convincingly that until what they term "artificial general intelligence" is developed —and they predict that this is still several decades in the future — humans will still be required to generate hypotheses or generate potential answers to complex questions. In the shorter term, analysts will increasingly use computers to test hypotheses, utilising "combinations of tools to build models and interrogate

[40] The debate on the potential for formal systems to achieve human-like intelligence is a longstanding one. A good example here is the debate around mathematician Kurt Gödel's 'Incompleteness Theorem'. Authors such as Lucas and Penrose have used Gödel's work to support their argument that artificial intelligence can never fully match the depth and complexity of the human mind, an view that has been strongly challenged by others. See for example Chapter 17 - Prospects of Artificial Intelligence in Mariusz Flasinski, Introduction to Artificial Intelligence (Springer, 2016).

[41] Nick Hare and Peter Coghill, "The Future of the Intelligence Analysis Task", *Intelligence and National Security* Vol. 31, no. 6 (2016): 865.

intelligence data".[42] Simply put, "Big data's power does not erase the need for vision or human insight".[43]

Analysts versus Algorithms

In a recent article Kevjn Lim succinctly summarised the strengths and weaknesses of automated approaches, noting that "Big Data analytics shift the focus of inquiry from *causation* to *correlations"*.[44] Big Data tools can provide new insights into strategic trends and anomalies, but may reveal little about *why* they are happening. Software packages are continually being refined to manipulate ever greater volumes of available data, but Lim argues convincingly that "as powerful as these platforms may be, and even assuming that data points are all geospatially and time-tagged, they still require that the analyst know *specifically what to look for*".[45] In Lim's view, Big Data analytics are clearly a force multiplier, but they must complement (or even remain subservient to), and not replace, "subject-matter expertise and causality-driven theoretical models".[46] In fact, the exploitation of Big Data *increases*, rather than decreases, the need for human judgement and expertise. Lim argues, "If the volatility of the human subject creates cognitive challenges for the area studies expert, the incorporation of Big Data demands, more than ever, a greater margin of manoeuvre for human intuition and a 'standard of judgment'".[47]

It is important to note that the logic underpinning this argument has been in circulation for some time. Dupont, for example, argued over a decade ago that "If intelligence collection in the twenty-first century is likely to be dominated by smart machines, intelligence assessments

[42] Nick Hare and Peter Coghill, "The Future of the Intelligence Analysis Task", *Intelligence and National Security* Vol. 31, no. 6 (2016): 865.
[43] McAfee and Brynjolfsson, "Big Data.", 7
[44] Lim, "Big Data and Strategic Intelligence.", 622
[45] Ibid.
[46] Ibid. 629-32
[47] Ibid. 622

will still reflect the perspicacity of human minds. No amount of raw data can substitute for an insightful human analyst able to discern the critical policy or operational significance of an event, action or trend which may be hidden within a mass of confusing and contradictory information".[48] Yet the rise of big data, combined with advances in our ability to capture and interrogate these vast swathes of data, has given rise to a techno-centric culture where expectations are almost wholly rooted in technological advancement. The significance of the human element to OSINT has been eroded, and in this new environment little attention has been devoted to addressing the implications of this imbalance.

There exist multiple examples where over-reliance on algorithms has posed problems or resulted in failure. One of the best known is that of Google Flu Trends (GFT), an initiative launched in 2008 with the goal of predicting flu outbreaks weeks ahead of the US Centre for Disease Control and Prevention (CDC), thereby providing potentially life-saving insights.[49] Drawing on Google search data, GFT sought to relate user queries to flu cases, in order to "nowcast" changes in propensity.[50] However, over its seven years of operation GFT missed multiple nonseasonal pandemics while "persistently overestimating flu prevalence".[51] These failings were attributed to "big data hubris" and the commercially-oriented dynamics of Google's search algorithm, with GFT exchanging its image as "poster child of big data" for one as the embodiment of its limitations, resulting in the programme's shutdown in 2015.[52]

The challenges experienced by GFT did not result in any negative health effects, as treatment planning continued to be based on CDC analysis. However, in other cases the poorly-planned adoption of automated approaches has had more serious consequences. In the insurance

[48] Dupont, "Intelligence for the Twenty-First Century.", 22.
[49] Miguel Helft, "Google Uses Web Searches to Track Flu's Spread," *New York Times*, November 11, 2008, http://www.nytimes.com/2008/11/12/technology/internet/12flu.html.
[50] David Lazer and Ryan Kennedy, "What We Can Learn From the Epic Failure of Google Flu Trends," *WIRED*, October 1, 2005, https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/.
[51] David Lazer et al., "The Parable of Google Flu: Traps in Big Data Analysis," *Science* 343, no. 6176 (March 14, 2014): 1203–5.
[52] Lazer and Kennedy, "What We Can Learn From the Epic Failure of Google Flu Trends."

sector, for example, a number of companies have sought to adopt automation to speed up processes such as claim handling. In the case of one company, AIG, this was also seen as a means of cutting personnel costs and recouping the considerable losses suffered by the company in 2015.[53] However, this embrace of automation failed spectacularly, in large part because the company was primarily focused on cost-saving measures and reductions in personnel numbers, and did not give due consideration to how automation would integrate with and complement existing work processes. This resulted in the application of automation to inappropriate areas and the adoption of software that had not been fully road-tested.[54] Moreover, somewhat ironically, the problem was exacerbated by the very personnel cuts that automation was supposed to facilitate as staff reductions meant there were less analysts available to support integration of the new automated tools.[55] This case is unlikely to be unique to AIG, given the perceived promise of automated approaches and the considerable pressures within the private sector to cut costs while increasing profit margins.

Similar challenges are being experienced in intelligence circles where the development, procurement, and implementation of automated analysis technology has emerged as a top priority. However, its potential impact is being limited by what Couch and Robins describe as an "imbalance in the investment in collectors and in the tools to support its analysis, rendering analysts incapable of taking into account all available sources when performing their assessment".[56] This at a time when, the authors argue, the expansion of available data necessitates an increase, not a relative decrease, in the number of analysts. Simply put, there "is a need for tools to reduce the volume of material that analysts must assess, allowing them to

[53] http://www.businessinsurance.com/article/00010101/NEWS06/311089985/Quarterly-loss,-Icahn-keep-things-roiling-at-AIG
[54] Interview with former AIG analyst who worked on this issue, 14th July 2017.
[55] Ibid.
[56] Couch and Robbins, "Big Data for Defence and Security.", 9-10.

focus on those likely to be the most fruitful".[57] Ultimately, Couch and Robins argue that "the skill of the future analyst is likely […] to concentrate more on configuring sophisticated search tools to which subject-matter experts can then apply their experience, intuition and human judgement".[58]

Frank concurs, noting that "While it is unlikely, and perhaps undesirable, for machines to replace human analysts in the production of strategic intelligence, the balance of labour between analysts and machines are likely to change in the near future… Analytic tradecraft should experiment with new approaches to dividing labour between man and machines in order to help analysts make better use of the data that is available to them, check for unarticulated and unexamined assumptions, and generally stress test the breadth and depth of mental models in order to ensure the completeness of their analytic efforts and products."[59]

Further insights regarding the limitations of automation can be gleaned from other disciplines. The work of philosophers such as Arthur Kuflik, for example, highlights additional factors that must be kept in mind when considering the transfer of responsibility for completing tasks, in whatever measure, from humans to computers.[60] Kuflik considers two arguments that might be used against such transfers in specific situations: first, that "our own finiteness and fallibility" as humans impose limits on what human-designed technology can achieve; and second, that there are certain personal tasks (such as solving a puzzle) that it would be inappropriate for computers to perform.[61] Kuflik further argues that the humans who design, programme, or control computers bear the ultimate moral responsibility for the consequences of decisions taken by computers, an important point in the context of recent debate on

[57] Couch and Robbins, "Big Data for Defence and Security.", 9-10.
[58] Ibid.
[59] Aaron Frank, "Computational Social Science and Intelligence Analysis," *Intelligence and National Security* Vol. 32, no. 5 (2017): 593.
[60] Arthur Kuflik, "Computers in Control: Rational Transfer of Authority or Irresponsible Abdication of Autonomy?," *Ethics and Information Technology* 1, no. 3 (September 1, 1999): 173–84.
[61] Ibid.

autonomous systems.[62] Decisions made by computers are "implementational and not fundamental".[63] These issues of responsibility, fallibility and control further strengthen the argument for the enduring significance of the open-source analyst. How, then, should we approach the exploitation of 'Big Data' for intelligence or investigative purposes? How can new technical developments in computer science be integrated with existing techniques and methods in the arena of intelligence and investigation? How can the potential of both analysts and algorithms be maximised? We contend that the focus on technology and the hype surrounding big data has constrained thinking and debate around these issues. There has been little effort to look at how the effective integration of analysts and machines might be achieved, even at a conceptual level. In particular, there is a need to look beyond disciplinary boundaries and consider the lessons that research in other fields may hold for those seeking to maximise the potential benefits of OSINT.

Cognitive Engineering and Strategic Partnerships

In our view, the first step in this process is to seek to counter what appears to be an emerging false dichotomy of analyst versus algorithm.[64] Human analysts working with open sources in the intelligence field require the assistance of computer systems, databases, and visualisation tools to extract useful information from the deluge of data they face every day. However, the analyst's skill, expertise, and judgment remain irreplaceable. Indeed, while the value of these attributes may have been highlighted in the past, they are now both under threat and more important than ever in the face of the enormous and growing body of data, including relevant data that must be assessed in order to put open-source information to its best use for intelligence

[62] For a comprehensive discussion of the issue in a related context, see Alex Leveringhaus, *Ethics and Autonomous Weapons* (Palgrave MacMillan, 2016).
[63] Kuflik, "Computers in Control.", 174
[64] This view was supported in discussions with a number of UK government officials with responsibility for developing open-source intelligence capabilities. Moreover, the emergence of this dichotomy was largely attributed to the influence of private sector companies with a commercial interest in technical solutions.

and related purposes. It is thus more productive to approach the challenges posed by 'Big Data' and the new information landscape from the perspective of strategic partnerships between humans and computers. Framed this way, the concept appears odd, yet the idea is well-established in the field of cognitive engineering, and developments here hold relevance for those seeking to break new ground in the collection and analysis of open-source information.

Cognitive engineering emerged as a field of study in the 1980s as computers became increasingly powerful and the range of uses to which they might be put, in control systems for example, began to dramatically expand.[65] The seminal paper published by Woods and Roth in 1988 argued that computer technology "offers new kinds and degrees of machine power that greatly expand the potential to assist and augment human cognitive activities in complex problem-solving worlds, such as monitoring, problem formulation, plan generation and adaptation, and fault management".[66] The principal question then was similar to the one the OSINT community faces today, namely how to "deploy the power available through new capabilities for tool building to assist human performance".[67]

A key concept that gained momentum in the field is that of the "Joint Cognitive System", which emphasizes the importance of understanding a system composed of humans and computers as one system.[68] The entire system has a task, or a set of tasks, to accomplish, and the design of the system should consider how those tasks can be most effectively completed, putting all components to their best use. This is different, for example, than understanding the "system" as the set of technologies that humans simply use; i.e. "disintegrating" the joint system into its constituent elements. In the view of the engineers, the focus in system design should be on *function* rather than structure and on *co-agency* rather than

[65] D. D. Woods and E. M. Roth, "Cognitive Engineering: Human Problem Solving with Tools," *Human Factors* 30, no. 4 (August 1, 1988): 415–30.
[66] Ibid.
[67] Ibid.
[68] Hollnagel and Woods, *Joint Cognitive Systems*., ch.1.

disintegration.[69] If designed carefully, the performance of such a system can be "greater than the sum of the performance of each of its component parts".[70]

This is vividly illustrated by an example from the world of chess, where machines have been dominant for more than 20 years following World Champion Garry Kasparov's loss to IBM's Deep Blue supercomputer in 1997, in a match famously labelled "the brain's last stand".[71] Following his defeat, Kasparov and others sought to explore how human strategy, insight and intuition could be combined with computers' advantages in calculating power and "remembering" details. This led to the development of 'Advanced Chess', a partnership between man and machine designed to "increase the level of play to depths never before observed", resulting in games with both "perfect tactical play and highly meaningful strategic plans".[72] In 2005, the first Freestyle Chess Tournament was held, where teams of humans and computers competed against one another. The result of this inaugural competition, however, was a shock to the chess community. Amidst a host of 'centaurs' – the label given to the human/computer partnerships – featuring chess grandmasters, the winning team "consisted of two young New England men, Stephen Cramton and Zackary Stephen (who were comparative amateurs, with chess rankings down around 1,400 to 1,700) and their computers".[73] The reason these less experienced and less talented chess players won was simple: "Cramton and Stephen were expert at collaborating with computers. They knew when to rely on human smarts and when to rely on the machine's advice".[74]

---

[69] Ibid.
[70] Ian Macleod, "Cognitive Quality in Advanced Crew Training System Concepts: The Training of the Aircrew-Machine Team," in *Contemporary Ergonomics 1984-2008: Selected Papers and an Overview of the Ergonomics Society Annual Conference*, ed. Philip D. Bust (CRC Press, 1996), 49.
[71] Steven Levy, "What Deep Blue Tells Us About AI in 2017," *WIRED*, May 23, 2017, https://www.wired.com/2017/05/what-deep-blue-tells-us-about-ai-in-2017/.
[72] Jose A. Fadul, *More Lessons in Chess and in Life* (Lulu Press, 2011)., p 144
[73] Clive Thompson, *Smarter Than You Think: How Technology Is Changing Our Minds for the Better* (Harper Collins, 2013)., p.14.
[74] Ibid.

This example from the chess world holds great relevance in terms of the need for effective integration if maximum benefits are to be extracted from man-machine systems. In designing joint cognitive systems, boundaries "must be made explicit, both between the system and its environment and between the elements of the system…the boundary clearly depends both on the purpose of the analysis and on the purpose of the [joint cognitive system]".[75] However, when defining the boundaries, i.e. allocating functions, between the human and technological elements of the system, it is important to note that "function allocation cannot be achieved simply by substituting human functions by technology, nor vice versa, because of fundamental differences between how humans and machines function and because functions depend on each other in ways that are more complex than a mechanical decomposition can account for".[76]

A recent article in the *Journal of Cognitive Engineering and Decision Making* provides a list of requirements for successful function allocation. In our view, this list offers a useful model for any effort designing and implementing a joint system of analysts and computers to facilitate open-source intelligence work. First, each individual system component, or agent, must be allocated functions that it is capable of performing. This can be accomplished, for example, by comparing each required function of the system to the capabilities of each agent and assigning the function to the agent most capable of carrying it out. Second, each agent must be capable of performing its collective set of functions under realistic operating conditions. In other words, individual agents should not be assigned more tasks than they can reasonably complete, or tasks that contradict or interfere with one another. Third, the function allocation must be realisable with reasonable teamwork. Teamwork is key, not only because it enables agents to accomplish tasks they cannot complete on their own, but also because it serves to coordinate tasks across agents. So teamwork should be planned for and facilitated. Fourth, the

[75] Hollnagel and Woods, *Joint Cognitive Systems*., p.67.
[76] Ibid. p.122-123.

allocation must support the dynamics of the work. This involves, for example, anticipating actions taken by one agent that have an effect on actions taken by another. Finally, function allocation should be the result of deliberate decisions, for example as part of the broader engineering design process.[77] This sort of nuanced approach offers a valuable framework for effective human-computer partnerships that comprehensively exploits the processing power of computers and uses this to complement human attributes such as judgement, rather than replace them with poor alternatives.

The Value for OSINT

Clearly, existing technologies and tools under development can be extremely useful in helping intelligence analysts make the best use of the opportunity presented by the immense and rapidly expanding body of available open-source information. But it is crucial that a careful, critical approach be taken to the question of which tools should be used and how they should be integrated. Odom warns that, "striving to do as much analysis as possible by technical means, especially with software algorithms, systems designers have too often promised more than they can deliver…more realism about what can and cannot be 'fused', combined with the greater computer literacy of younger analysts, has brought progress. Still emphasis on 'processing' software, often meaning a considerable degree of machine-generated analysis, cannot replace a lot of brain work and labor on the part of analysts."[78]

So how might automated tools most effectively be integrated into the OSINT process? At stake here are two discrete issues: the particular attributes and strengths of various tools (both new and established) that can support the work of analysts; and the implementation of these tools as part of a broader systematic approach to effective OSINT.

[77] Karen M. Feigh and Amy R. Pritchett, "Requirements for Effective Function Allocation: A Critical Review," *Journal of Cognitive Engineering and Decision Making* 8, no. 1 (March 2014): 23–32.
[78] William E. Odom, "Intelligence Analysis", *Intelligence and National Security* Vol. 23 No. 3 (2008): 325.

With regard to tools, notable instruments include the continuous collection of information based on targeted searches and from key sources using, among other things, increasingly refined data scrapers, the translation of foreign language information, the elimination of duplicates, and the tagging and indexing of sources to enable efficient internal searching.[79] Tools can also be used to identify and record relationships amongst key entities and across different information sources, serving to direct the attention of the analyst to useful correlations and patterns. Indeed, such visualisation capabilities can be of great value, because they 'remember' individual facts (such as a specific relationship between two entities) and 'remind' the analyst of them. On this point, however, it is important to note that unrealistic expectations for software can be problematic. Visualisation is not the same thing as analysis. Human analysts with subject-matter expertise in the organisation's area of responsibility, as well as an understanding of the organisation's processes and context, are required to assess the relevance of a visualisation and determine what its content's meaning and significance are for the work.[80]

As discussed earlier a key challenge for open source analysts is information overload. This is compounded by research demonstrating that during the intelligence process, even experienced analysts may quickly narrow their focus to a particular "set of documents on which they based their analysis", leading to crucial information potentially being missed.[81] Automated tools have the potential to play an important role here, for example through "exploratory searching in order to allow analysts to have a better sense of how their samples relate to what is potentially available".[82] Related to this is the automated categorisation of different documents

[79] On this point, see A.J. Rockmore, "Catalyst Entity Extraction and Disambiguation Study Final Report" (IARPA, 2008)., p.2. (marker "Unclassified/For Official Use Only")
[80] Arpad Palfy, "Bridging the Gap between Collection and Analysis: Intelligence Information Processing and Data Governance," *International Journal of Intelligence and CounterIntelligence* 28, no. 2 (April 3, 2015): 365–76; Phil Nolan, "A Curator Approach to Intelligence Analysis: International Journal of Intelligence and CounterIntelligence" 25, no. 4 (August 29, 2012): 786–94.
[81] Emily Patterson, David D. Woods, and David Tinapple, "Aiding the Intelligence Analyst: From Problem Definition to Design Concept Exploration," Interim Report (United States Airforce Laboratory, March 2001).
[82] Ibid.

from low to high value based on user-defined criteria, which may of course vary from task to task. Here attributes might include the document length, writing style, percentage of original vs. replicated content, publication data, and the nature of the source (official vs. unofficial).

While consideration of tools and their function is important, the more pressing issue that this article has sought to address is how these are implemented as part of a broader OSINT system that is robust, efficient and equipped to deal with the needs of particular organisations. Clearly, this will vary according to objectives and across different organisations. It is imperative, however, that the current and future needs and practices of specific actors drive the decision-making process here. This requires, for example, that the designers have a detailed understanding of current day-to-day procedures in place for the exploitation of open-source materials, as well as the higher-level context of the analytical and intelligence-related goals for the work. This imperative also necessitates the direct, ongoing involvement of open-source analysts and collectors in the system-design process. And, as argued above, the process of task allocation should be thorough, detailed, and realistic.

When this type of considered approach is not used, various pitfalls can present themselves. Changes in work practices are an excellent example. The integration of new software into daily work routines may necessitate considerable changes in existing procedures. Such changes may well constitute a positive development, but only if they are supported with appropriate resource levels, including training, and—importantly—only if the net result of the effort is higher-quality analysis or a greater amount of high-quality analysis. Failure to take this into account can result in, for example, analytical staff spending their time carrying out mechanistic tasks, such as tagging documents, rather than analysing their content. Worse, the

organisation can struggle through the disruption to work practices only to learn a new approach that yields no appreciable increase in the quality of analysis.[83]

In practical terms, the design of a system that fuses the skills of analysts with the processing power of computers must begin with a comprehensive review of analytical needs and goals of the actor concerned. This must be accompanied by a detailed mapping of current work flows and processes that will better position the actor to identify areas and tasks that could benefit from technological intervention, a rebalancing in favour of the human analyst, or both. Only *after* this step is complete should specific new software options be considered. In designing a new or upgraded system, the key concepts from Cognitive Engineering can be helpful, in particular the conceptual focus on function rather than structure and on co-agency rather than disintegration. The process of proper function allocation as identified in the cognitive engineering literature is useful as well.

When assessing what specific computerized systems to adopt decision-makers will have to weight the strengths and weaknesses of a wide variety of tools. There is a now a crowded market of third party applications relevant to the work of the intelligence analyst.[84] These claim to offer powerful, albeit potentially expensive, solutions to open source intelligence challenges. However, as was highlighted earlier in the case of AIG, the integration of generic tools into an existing system can be difficult. Furthermore vendors of proprietary software often adopt a "lock-in" approach that makes it difficult to switch to a rival package at a later date.[85] Alternatively, organisations can utilise open-code software, which is often inexpensive or even

[83] Arpad Palfy, "Bridging the Gap between Collection and Analysis"; Annette Mills and Trevor Smith, "Knowledge Management and Organizational Performance: A Decomposed View," *Journal of Knowledge Management* 15 (2011): 156–71.
[84] Examples include i2 Analyst's Notebook (https://www.ibm.com/us-en/marketplace/analysts-notebook), Palantir (www.palantir.com), and Recorded Future (www.recordedfuture.com)(all sites accessed 12 November 2017). For a summary of some of the different tools available see: Baiju, NT. 'Top Business Intelligence (BI) tools in the market', Big Data Made Simple, http://bigdata-madesimple.com/top-business-intelligence-bi-tools-in-the-market/ (25th October 2017);
[85] Kevin Xiaoguo Zhu and Zach Zhizhong Zhou, "Lock-In Strategy in Software Competition: Open-Source Software vs. Proprietary Software", *Information Systems Research*, 32, Issue 2 (June 2012): 536 – 545.

cost-free, modifying the code to meet their own needs. Or they can develop their own in-house solutions. This has the advantage of being truly bespoke, although it may lack the functionality and analytical power of commercial alternatives. Realistically, organisations will likely opt for a mixture of these approaches, ideally as part of an evolving system that is capable of integrating new capabilities as they become available and as the information landscape changes.

Conclusion

Recent developments in the online information environment mean that software-based approaches now form a core element of the OSINT endeavour. Without the support of automated processes, analysts would simply be unable to cope with the deluge of information swirling around the ether. Yet while OSINT enthusiasts must engage with algorithmic approaches, it is important that they not be overwhelmed by them. The trained analyst has a level of knowledge, expertise and judgement that cannot be coded. Effective exploitation of open-source information requires that we fuse the capabilities of analysts and algorithms, even as we continue to respect the line that separates them.

The significance of this argument is clear, yet relatively little academic attention has been devoted to the question of how effective fusion might best be achieved. In this paper we have sought to advance thinking on this front by drawing on insights from the field of cognitive engineering. Our concern here has been to show how, at a conceptual level, lessons from this field can support the effective design of OSINT systems that capitalise on the combined strengths of both machines and human analysts, while mitigating their respective weaknesses. It is our hope that this paper will set the path for additional work in this area, for there is considerable scope for further research, particularly empirical work around OSINT process models within different sectors and in organisational environments subject to varying constraints. Only through obtaining a detailed understanding of how analysts are currently

utilizing open source information for investigative or intelligence purposes can automated tools be tailored to their needs.

Bibliography:

Bean, Hamilton. *No More Secrets: Open Source Information and the Reshaping of U.S. Intelligence*. Santa Barbara, Calif: Praeger, 2011.

Boer, Monica Den. "Counter-Terrorism, Security and Intelligence in the EU: Governance Challenges for Collection, Exchange and Analysis," *Intelligence and National Security* 30, No. 2-3, (2015): 402-419.

Boyd, Danah, and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication and Society* 15, no. 5 (2012): 667–69.

Bunnik, Anno, Anthony Cawley, Michael Mulqueen, and Andrej Zwitter. *Big Data Challenges - Society, Security, Innovation and Ethics*. Basingstoke: Palgrave MacMillan, 2016.

Couch, Neil, and Bill Robbins. "Big Data for Defence and Security." Occasional Paper. Royal United Services Institute, September 2013.

Dell EMC. "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things," April 2014. https://www.emc.com/infographics/digital-universe-2014.htm.

Dupont, Alan. "Intelligence for the Twenty-First Century", *Intelligence and National Security* 18, No.4 (2003): 15-39.

Fadul, Jose A. *More Lessons in Chess and in Life*. Lulu Press, 2011.

Feigh, Karen M., and Amy R. Pritchett. "Requirements for Effective Function Allocation: A Critical Review." *Journal of Cognitive Engineering and Decision Making* 8, no. 1 (March 2014): 23–32.

Flasinski, Mariusz. *Introduction to Artificial Intelligence*. Springer, 2016

Frank, Aaron. "Computational Social Science and Intelligence Analysis," *Intelligence and National Security* 32, no. 5 (2017): 579-599.

Gibson, Stevyn. "Exploring the Role and Value of Open Source Intelligence." In *Open Source Intelligence in the Twenty-First Century: New Approaches and Opportunities*, edited by Christopher Hobbs, Matthew Moran, and Daniel Salisbury. Liverpool: Palgrave MacMillan, 2014.

Hare, Nick and Coghill, Peter. "The Future of the Intelligence Analysis Task", *Intelligence and National Security* 31, no. 6 (2016): 858-870.

Helft, Miguel. "Google Uses Web Searches to Track Flu's Spread." *New York Times*, November 11, 2008. http://www.nytimes.com/2008/11/12/technology/internet/12flu.html.

Hobbs, Christopher, and Matthew Moran. "Armchair Safeguards: The Role of Open Source Intelligence in Nuclear Proliferation Analysis." In *Open Source Intelligence in the Twenty-First Century : New Approaches and Opportunities*, edited by Christopher Hobbs, Matthew Moran, and Daniel Salisbury. Liverpool: Palgrave MacMillan, 2014.

Hollnagel, Eric, and David D. Woods. *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*. Taylor and Francis, 2005.

Insights, Automated. "Automation Helps AP Publish 10 Times More Earnings Stories," January 29, 2015. https://automatedinsights.com/blog/automation-helps-ap-publish-10-times-more-earnings/.

Kuflik, Arthur. “Computers in Control: Rational Transfer of Authority or Irresponsible Abdication of Autonomy?” *Ethics and Information Technology* 1, no. 3 (September 1, 1999): 173–84.

Lazer, David, and Ryan Kennedy. “What We Can Learn From the Epic Failure of Google Flu Trends.” *WIRED*, October 1, 2005. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. “The Parable of Google Flu: Traps in Big Data Analysis.” *Science* 343, no. 6176 (March 14, 2014): 1203–5.

Leveringhaus, Alex. *Ethics and Autonomous Weapons*. Palgrave MacMillan, 2016.

Levy, Steven. “What Deep Blue Tells Us About AI in 2017.” *WIRED*, May 23, 2017. https://www.wired.com/2017/05/what-deep-blue-tells-us-about-ai-in-2017/.

Lim, Kevjn. “Big Data and Strategic Intelligence.” *Intelligence and National Security* 31, no. 4 (June 6, 2016): 619–35.

Macleod, Ian. “Cognitive Quality in Advanced Crew Training System Concepts: The Training of the Aircrew-Machine Team.” In *Contemporary Ergonomics 1984-2008: Selected Papers and an Overview of the Ergonomics Society Annual Conference*, edited by Philip D. Bust, 49. CRC Press, 1996.

Matheny, Jason. “Open Source Indicators: Intelligence ARPA.” Office of the Director of National Intelligence, 2013. https://www.aaas.org/sites/default/files/Jason_Matheny_AAAS_FBI.pdf.

McAfee, Andrew, and Erik Brynjolfsson. “Big Data: The Management Revolution.” *Harvard Business Review*, October 1, 2012. https://hbr.org/2012/10/big-data-the-management-revolution.

Mercado, Stephen C. “Sailing the Sea of OSINT in the Information Age — Central Intelligence Agency.” *Studies in Intelligence* 48, no. 3 (2004).

Mills, Annette, and Trevor Smith. “Knowledge Management and Organizational Performance: A Decomposed View.” *Journal of Knowledge Management* 15 (2011): 156–71.

Nolan, Phil. “A Curator Approach to Intelligence Analysis: International Journal of Intelligence and CounterIntelligence” 25, no. 4 (August 29, 2012): 786–94.

Odom, William E. “Intelligence Analysis”, *Intelligence and National Security* 23 No. 3 (2008): 316-332.

Office of the Press Secretary, The White House. “Government Assessment of the Syrian Government’s Use of Chemical Weapons on August 21, 2013,” August 30, 2013. https://obamawhitehouse.archives.gov/the-press-office/2013/08/30/government-assessment-syrian-government-s-use-chemical-weapons-august-21.

Omand, David, Jamie Bartless, and Carl Miller. “#Intelligence.” DEMOS, 2012.

Palfy, Arpad. “Bridging the Gap between Collection and Analysis: Intelligence Information Processing and Data Governance.” *International Journal of Intelligence and CounterIntelligence* 28, no. 2 (April 3, 2015): 365–76.

Patterson, Emily, David D. Woods, and David Tinapple. “Aiding the Intelligence Analyst: From Problem Definition to Design Concept Exploration.” Interim Report. United States Airforce Laboratory, March 2001.

Ramakrishnan, Naren, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, et al. “‘Beating the News’ with EMBERS: Forecasting Civil Unrest Using Open Source Indicators.” In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1799–1808. KDD ’14. New York, NY, USA: ACM, 2014.

Rockmore, A.J. “Catalyst Entity Extraction and Disambiguation Study Final Report.” IARPA, 2008.

Rogers, Simon. “Wikileaks Data Journalism: How We Handled the Data.” *The Guardian*, January 31, 2011, sec. News. https://www.theguardian.com/news/datablog/2011/jan/31/wikileaks-data-journalism.

Spiegel, Eric. “Six Challenges of Big Data.” *WSJ*, March 26, 2014. https://blogs.wsj.com/experts/2014/03/26/six-challenges-of-big-data/.

Tait, Amelia. “Human Journalists Hate Robot Journalists, Says New Report.” *New Statesman*, March 2, 2017. http://www.newstatesman.com/science-tech/technology/2017/03/human-journalists-hate-robot-journalists-says-new-report.

The Slow Journalism Company. “Delayed Gratification - Rise of the Robot Journalist,” August 15, 2014. http://www.slow-journalism.com/from-the-archive/rise-of-the-robot-journalist.

Thompson, Clive. *Smarter Than You Think: How Technology Is Changing Our Minds for the Better*. Harper Collins, 2013.

Warner, Michael. “Sources and Methods for the Study of Intelligence.” In *Handbook of Intelligence Studies*, edited by Loch K. Johnson. Routledge, 2007.

Webster, Jennifer, Luke Erikson, Christopher Toomey, and Valerie Lewis. “PNNL Strategic Goods Testbed: A Data Library for Illicit Nuclear Trafficking, PNNL-SA-102611.” *Proceedings of the Information Analysis Technologies, Techniques and Methods for Safeguards, Nonproliferation and Arms Control Verification Workshop*, May 12, 2014, 168.

Wirtz, James. *Understanding Intelligence Failure: Warning, Response and Deterrence*. Routledge, 2017.

Woods, D. D., and E. M. Roth. “Cognitive Engineering: Human Problem Solving with Tools.” *Human Factors* 30, no. 4 (August 1, 1988): 415–30.

Wirtz, James J. and Rosenwasser, Jon J. “From Combined Arms to Combined Intelligence: Philosophy, Doctrine and Operations.” *Intelligence and National Security*, Vol.25, No.6 (2010): 725-743.

Zhu, Kevin Xiaoguo and Zhou, Zach Zhizhong. “Lock-In Strategy in Software Competition: Open-Source Software vs. Proprietary Software”, *Information Systems Research*, 32, Issue 2 (June 2012): 536 – 545.