LSE Research Online

Jannis Kallinikos and Ioanna Constantiou
Big data revisited: a rejoinder

Article (Accepted version)
(Refereed)

http://eprints.lse.ac.uk

This paper is identical to one to appear on JIT 30/1 (March 2015) addressing some of the issues the commentators (Lynne Markus, Youngjin Yoo, and Stephanie Woerner & Barbara Wixom) to our original JIT article *New games, new rules: big data and the changing context of strategy* (forthcoming, JIT 30/1) have raised.

Big data revisited: A rejoinder

Jannis Kallinikos
Department of Management,
London School of Economics

Ioanna Constantiou
Department of IT Management,
Copenhagen Business School

Abstract

We elaborate on key issues of our paper *New games, new rules: big data and the changing context of strategy* as a means of addressing some of the concerns raised by the paper's commentators. We initially deal with the issue of *social data* and the role it plays in the current data revolution. The massive involvement of lay publics as instrumented by social media breaks with the strong expert cultures that have underlain the production and use of data in modern organizations. It also sets apart the interactive and communicative processes by which social data is produced from *sensor* data and the technological recording of facts. We further discuss the significance of the very *mechanisms* by which big data is produced as distinct from the very attributes of big data, often discussed in the literature. In the final section of the paper, we qualify the alleged importance of *algorithms* and claim that the structures of data capture and the architectures in which data generation is embedded are fundamental to the phenomenon of big data.

Management strategy, organizations and society

Big data raises issues extending far beyond management strategy that forms the key subject of our original paper, published above. In different ways, Lynne Markus and Youngjin Yoo put this view forward in their respective commentaries. We cannot but fully agree on this point. In fact, our own article concludes with a whole section (that is, *Postscript on big data, human behaviour and management*) in which we sought to outline some of the wider organizational and institutional

implications of big data and the socio-technical apparatus of which big data is part and parcel.

These ideas notwithstanding, it felt nonetheless worthwhile to enter the problematization of big data from a rather circumscribed point of view. Strategy offered such a suitable entry point. Via this route, we thought, we would be able to circumnavigate some of the most unproductive simplifications in the literature on the subject and help develop arguments specific enough to advance our knowledge in the field of IS, and Organization and Management Studies. The critical examination of the entire prescriptive framework of strategic planning seemed, in addition, to provide a suitable lens for approaching the study of the organizational implications of the relevant developments. This last, in turn, offered a good point of departure for considering some of the wider institutional issues associated with the production and commercialization of online data that we attempted in our *Postscript* section of our original piece.

There are certainly other burning issues big data raises that transcend the institutional framework of IS and management. The significance of big data for privacy and personal integrity, surveillance and democracy are vital and recurring questions across various social science fields. Other issues touch upon long-standing epistemological conundrums, being tied to modes of scientific inquiry and the role data as encoded evidence plays in this process (see e.g. Gitelman, 2013; Ekbia et al., 2014). The list can no doubt be expanded. To the degree that big data is tied to automation and machine-to-machine interaction (Arthur, 2011), it may be claimed to have serious implications for employment and the long-term distribution of skill profiles and cultural knowledge across the social fabric (Lanier, 2014). In her thoughtful commentary, Lynne Markus rightly points out some of these issues. There are, however, scholarly limits with respect to what can be accomplished by a journal article that would better be accommodated.

These concerns about the scope of our article aside, there are a few other issues that the commentaries to our article single out and discuss that touch upon the

core of our argument. One of the most important points of dissent relates to the significance we attribute to social data. Our claim, in particular, that data produced by massive online trafficking on social media platforms captures much of the essence of the phenomenon of big data may appear as an unnecessarily narrow take on the matter. There is more to big data, it would seem, than just the social data created by massive user engagement in the digital ecosystems of Facebook, Twitter and other's major social media platforms. In different ways, Lynne Markus and Youngjin Yoo challenge us on this point. A similar view also seems to run though much of Stephanie Woerner and Barbara Wixom's commentary. Closely related to this important matter is another contentious issue that pivots around the role of algorithms and automated decision-making in the big picture in which big data is embedded. In what follows, we consider these cardinal issues in succession.

Expert cultures and social data

In our original article, we stressed the significance of social data but never maintained that social media is the only important channel or arrangement by means of which large data volumes are produced and disseminated. Other widespread modes of massive data generation (mentioned in our original paper) are represented by default recordings such as those captured by logs, increase use of RFID systems and sensors, CCTV, and automated records such as those for instance produced by many financial transactions and the multiple checks and balances they implicate. This picture is reinforced, as Yoo rightly points out in his commentary, by the recent diffusion of wearable, mobile and smart devices able to capture and record the whereabouts of large and heterogeneous user populations. Placed against this background, social media platforms emerge as just one channel, admittedly important, through which big data is produced and disseminated.

Why do we then insist on stressing the role of social media platforms and the data they generate in our original paper? Even though social media platforms may not be the sole producer of massive data, they still seem to us to represent the most typical example of socio-technical arrangements in which a new *para-*

digm of data generation and use is foreshadowed. Key elements of that paradigm are clearly shown in the making of daily patterns of interaction and communication to a pivotal means of data generation and data commercialization. Online participation as organized by social media platforms makes everyday life and the daily habits of users the central span of the activities that sustain data generation.

Such a shift, we feel, constitutes a significant breach of the prevailing practices by which data has hitherto being produced and relayed in organizational and institutional life. Two prerequisites on which such a change rests have far reaching implications that are easy to overlook. *First*, data generation is lifted out of the prevailing expert-dominated cultures by which the information needs of practice fields have been defined and data collected and stored. This is the outcome of the fundamental fact of making online participation and the activities of large, shifting, heterogeneous and dispersed populations of users (mostly lay people) the drivers and carriers of data generation. *Second*, the trivial concerns, habits and whereabouts characteristic of a diffuse and informal human everyday are made the target of elaborate real time recording. Through online participation, they enter the circuit of commercial use or exploitation in remarkably different ways from those of traditional marketing and the sampling, and largely *ex-post*, procedures on which the traditional mapping of vernacular attitudes has mostly been based (see e.g. Napoli 2010).

As alluded to above, the path-breaking nature of these developments and the far-reaching implications they carry can only be contemplated against the background of the strong expert cultures[1] and knowledge specific pursuits of data generation and use that have underlain the operations of corporations and the state (Desrosières, 1999; Porter, 1995; Giddens, 1990). It can further be appreciated by the contrast between, on the one hand, the diffuse everyday of online participation and, on the other hand, the legal and formal framework that has governed economic transactions and employment relations (Kallinikos, 1995,

---

[1] That is, statisticians, economists, accountants, marketers, lawyers, medical experts and the like.

2004, 2010). Placed against such a backdrop, social media platforms could be viewed as important means of social engineering. The changes they introduce go some way towards bringing together the separate orders of a diffuse and spontaneous daily life with the formal and legal nature of economic transactions and business pursuits (Gellner, 1983, 1996).

This shift towards datafying and commercializing the everyday is most clearly shown in the way social media platforms organize online participation with the view of generating *en masse* lay opinions and minute behavioural choices (such as 'liking', 'tagging', 'sharing', 'following') on just about everything (Alaimo and Kallinikos, 2015; Van Dijck, 2013). This move away from the expert-dominated culture in data generation, use and re-use instrumented as voluntary online participation of large and heterogeneous lay crowds keeps on redrawing the institutional map in which markets and organizations have been embedded for more than a century. Shorn of this looming and path-breaking change, big data tends to degenerate to just a bunch of techniques and tools for handling large data volumes in distributed information environments, a sort of data warehousing in the age of the Web.

This last is not a small challenge but it is one that continues rather than breaks with longstanding practices and traditions of data generation and use. Much of the commentary of Woerner and Wixom's and the critique they direct towards our original piece assume that big data raises issues that are possible to deal with by recourse to established methods and techniques. Their claim that "companies are not replacing their business strategy toolboxes but rather using existing toolboxes more efficiently" assumes that big data largely concerns matters of data volume, availability and, possibly, heterogeneity. This is a widespread assumption and one of the principal motivations as to why we have chosen strategy as the entry point for discussing the organizational and management implications of big data in our original paper. Surely, the science of statistics will not be reinvented by the trends we point out. There is no doubt either that part of the developments associated with big data reinforces established techniques of calculation and data crunching as much as it casts them into new contexts of rel-

evance and use. Current business practices confirm that organizations expand their use of big data whenever this is feasible while developing new models and techniques for dealing with data that are not readily subsumable under the calculative regime of existing techniques and tools (Brynjolfsson and McAffee, 2014; Varian, 2010).

Yet, social data is a different kind of beast for reasons we have been at pains to analyze in our original article and briefly touch upon in this rejoinder of ours. The data that online participation produces has a scope (that is, ordinary life) that contrasts sharply with the circumscribed and often carefully defined data pursuits of specialized practice domains. It is, in addition, being generated by the users themselves through the encoding that social platforms mediate of human habits, trivial attitudes and expressive acts, on a continuous and real time basis (Alaimo and Kallinikos, 2015). In this respect, *sensor data* and *social data* differ remarkably. These are some of the key points that emerge out of the issues we bring forward in our original article and in the present rejoinder.

Furthermore, social media platforms raise puzzling questions as to what sort of organizational beasts corporations such as Facebook, Twitter or YouTube are. Let us just remind here that the commercial viability of social media platforms is a direct function of the user populations by which they are constituted as social entities and the datafication of the activities of these populations. Market share has of course always been a strong indicator of market position and organizational strength. But the user base of social media platforms is not coextensive with the old and venerable concept of market share. Users are as much producers as targets and consumers of data and the services data crunching enables. It is true that some of the comments of Woerner and Wixom's do touch upon some of these issues, as when they are discussing the theme of data monetization and, more importantly, the theme of digital transformation. Rather than developing business prescriptions, our own article sought to shed light into this dimly perceived territory upon which current practices of calculation and data use, and the organizational forms they implicate, are reframed by the ways social media platforms recast the game of data generation, use and commercialization.

Mechanisms of data generation

It is for all these reasons that we in our original paper stressed, time and again, the significance of the *mechanisms* of data generation and dissemination that social media platforms help to set in place. Mechanisms, it should be noted, refer to and are indicative of the *processes* and causal links through which social *outcomes* are produced. In this respect, mechanisms provide the generative matrices of whatever attributes social outcomes acquire (Abbott, 2001; Elder-Vass 2011). It is against this backdrop that we have repeatedly claimed that the distinctive profile of the developments associated with big data owes much to the way social media platforms organize and instrument (that is, the mechanisms) the generation and use of online data.

In fact, the attributes of big data repeatedly mentioned in the literature (3, 4 or more Vs) and critically reconstructed in our original article (that is, data that are agnostic, heterogeneous and unstructured, trans-semiotic with real time relevance) are the result of the very mechanisms by which data is produced in the online environments of social media platforms. Thanks to these mechanisms we have been at paints to outline, a significant part of data generation and use in the contemporary Web shifts away from the prevailing data practices that have under long time been marked by the ubiquity of experts and the formality of socio-economic life. It is these mechanisms by means of which social media platforms produce and relay data in organizational and institutional life as distinct from the computational challenge of managing large data volumes in distributed environments that seem to us to signal an epochal change. In other words, the emphasis in our original paper has been on the generative processes through which data are produced and used rather than on the sheer attributes of big data. It goes without saying that whatever attributes (e.g. agnostic, unstructured, heterogeneous etc) big data acquires must be described in precise or adequate ways that make their relation to the processes that generate them clear and evident. Perhaps, we should have been more straightforward on this point in our original paper.

Granted their importance, these mechanisms are currently given an interesting tweak by recent developments coinciding with the social diffusion of wearable intelligent devices. It is reasonable to expect that the diffusion of digital wearables (as distinct from mobile devices) would mark the far-reaching datafication of the most-minute aspects of daily living and bring a new data revolution in which every pulse, as it were, of life is recorded and made the object of calculation and commercialization. In this respect, the massive involvement of digital wearables in social life both reinforces the comprehensive mapping of the everyday we here associate with social media platforms and, at the same time, breaks with it, in the sense of bypassing the encoding of social interaction and communication as the principal mode of datafication.

On a broad frontier, we agree on this with Youngjin Yoo and the insightful ways through which he approaches these critical matters. However, given the nascent nature of this technology, it is not yet clear how the patterns of its use and the legal frameworks that will regulate the production and utilization of data by wearable devices will come to crystallize. It is also important to be clear on the differences introduced by wearable devices: whereas social media platforms stylize and standardize the preferences, expressive acts and habits of online living, wearable devices extend beyond sociality into the realm of physical and biological life of individuals (see also Yoo, 2010). In a sense, these developments reinforce the difference between sensor data and social data that we pointed out earlier in this rejoinder. The issues are, no doubt, complex enough to warrant a separate and lengthy treatment.

Beyond algorithms

The scale and complexity of big data and the entire apparatus (techniques and institutions) in which big data is embedded inevitably increase the scope of automation and the processes that become blackboxed in automated sequences. As an inevitable outcome of these developments, many of the operations underlying the production and mining of big data remain at a remove from the concerns and activities of social groups and beyond immediate social inspection and control. It is therefore reasonable to be concerned by these trends and whatever

biases or predilections they may embody. We are in full sympathy of these issues raised by Lynne Markus' commentary and her worries with respect to the nature of these developments and their social impact (see e.g. Gillespie, 2014; Kallinikos et al., 2013).

Two broad sets of operations are increasingly subject to automation in the case of big data. The first concerns all those functions by which big data is generated and stored such as sensor and RFID data, CCTV records, location data of various kinds, logs and other default modes of data capture. Such automation even extends to the means through which the numerable choices of social media users (most notably, the clicks) are assigned and stored into distinct data fields. The second form of automation entails the operations by which data is aggregated, processed and calculated. A conspicuous means of this latter form of automation involves the development and use of algorithms, whereby a set of fixed procedures (each procedure made of a finite number of steps) is applied to produce a *calculable* outcome out of a range of numerically expressed relations between data items.

Algorithms are pervasive in automated systems. However, not all automation is algorithmic calculation in the sense of entailing a fixed procedure for producing a numerical output out of available data inputs. In addition to the methods of data capture mentioned above, other tasks that fall outside what is usually subsumed under the notion of algorithmic calculation involve the definition and specification of formal relations between data items and fields. Datawork of this sort entails the development and specification of logical and architectural relations (Rosenfeld and Morville, 2002; Weinberger, 2007) that may involve automated sequences and elaborate computations yet are not properly described as algorithms. Risking the opening of Pandora's box, we will in this rejoinder draw a distinction between, on the one hand, algorithms, and on the other hand, the broader category of software automation. There is much more to programming and software engineering than the development and design of algorithms. In the context of the issues that concern us here, this complexity can, to a certain degree, be pinned down through the wieldy distinction between *data structures*

and *algorithms* (Manovich, 2001; Wenger, 1997). Data structures represent systematic and stratified arrangements through which data is captured and ordered while algorithms provide a fundamental means through which such data is computed.

The conception of algorithms and the function they fulfil in the various contexts of computing are easier to grasp than the invisible and more diffuse work of the structures and arrangements through which data is produced and made available for algorithmic calculation. The cardinal importance of data structures emerge against the background of the painstaking work their development requires. It is on the basis of such elaborate datawork that data are delivered in formats amenable to algorithmic calculation and, more generally, machine processing (Busch, 2011; Hanseth and Lyytinen, 2010). As Gillespie (2014) recently argued "algorithms are inert, meaningless machines, until paired with databases upon which to function. A sociological inquiry into an algorithm must always grapple with the databases to which it is wedded." Albeit a bit too strong, our claim in our original paper that "algorithms without data are mathematical fictions" echoed precisely this fundamental condition.

The complex relationship and mutual implication of data structures and algorithms are cast in new light by social media platforms. Data structures in social media are critically contingent on the stylization of social interaction and the far-reaching standardization of the activities of data entry and data capture such stylization affords (Alaimo, 2014). Online platform participation is organized along a number of stylized activities (e.g. 'tagging', 'sharing', 'following' 'liking') that segment, compartmentalize and standardize online behaviour. This, in turn, allows user participation to be recorded as discrete and granular choices (expressed as clicks) that can be counted, aggregated and processed (Alaimo and Kallinikos, 2015; Van Dijck, 2013). An essential part of the services social media platforms provide (such as recommendations and personalization services) are heavily made by the counting, aggregation and processing of clicks and the discrete data every click generates (Alaimo, 2014).

Social media platforms therefore feature elaborate architectural arrangements through which communal interaction and daily living are transformed into data ready to enter the circuits of calculation and so called personalization. Thus seen, data architectures represent pervasive mechanisms of standardization and, by extension, automation at work. In procuring the data and delivering them in suitable formats, data structures and architectures both circumscribe and transcend the functioning of algorithms. It is these fundamental and fine-grained arrangements that provide the essential and primary grid through which the bulk of granular data are registered into standardized data fields that are subsequently used for all forms of data manipulation, including algorithmic calculation. In some fundamental sense, social media platforms can be viewed as huge *interaction machines* rather than algorithms (Wenger, 1997).

It is against this broader background of operations that we attributed algorithms a derivative role in the larger machine, as it were, at work that in the case of social media platforms stylizes social interaction, standardizes it as discrete data tokens and aggregates and processes them. Of course, algorithms represent an important component of this complex network of operations and the datawork it implicates. As our earlier remarks on the point demonstrate, we are in a broad agreement with Youngjin Yoo as regards the complex and self-propelling loop in which algorithms and data structures are embedded (see also Gillespie, 2014; Manovich, 2001). The loop is one whereby the bounty of data calls for algorithmic techniques of data reduction whose outcome contributes to the production of new data while providing, at the same time, the very rationale for further data capture.

These ideas notwithstanding, the claims tied to the putative potency of algorithms may indeed lead to an unduly simplification of the complexity on the basis of which data is procured in standardized formats and made calculable in the very first place. They may thus obscure the significance data structures play in delivering data in formats that make it compatible with the functional requirements of algorithmic 'sensing' and calculation. It was for that reason we proposed algorithms to be just the last step or, in any case, just a step in an elaborate

network of operations, structures and mechanisms through which data is harvested, piled up and made ready for calculation.

This claim of ours could perhaps acquire further resonance by associating it with the historical experience that the building of expert systems provides and the significance which "knowledge representation" is supposed to have in expert systems as compared with the standardized workings (standard logical models of inference) of the "inference engines". Without adequate knowledge representation, inference engines cannot reach far. There are of course notable differences between the open and ongoing context of the Web in which algorithms work and the more controlled environment of expert systems. But the analogy is not without merit. In this sense, one could claim that algorithms without data are mathematical fictions, inert machines, as it were, or at best *standing possibilities* awaiting realization through data ingestion. Such a claim does not render algorithms irrelevant or unimportant. It does situate their workings in a complex ecosystem of relations, data links and processes within which they acquire their proper meaning, relevance and usefulness.

References

Abbottt, A. D. (2001). *Time Matters: On theory and method*. Chicago: The University of Chicago Press.

Alaimo, C. (2014). Computational Consumption: Social media and the construction of digital consumers, http://etheses.lse.ac.uk/975/1/Alaimo_Computational_Consumption.pdf Department of Management, London School of Economics.

Alaimo, C. & Kallinikos, J. (2015). "Encoding the Everyday: Social data and its media apparatus", in Sugimoto, C., Ekbia, H. and Mattioli M. (eds.) *Big Data is not a Monolith: Policies, Practices, and Problems,* Cambridge, MA: The MIT Press (forthcoming).

Arthur, B. W. (2011). The Second Economy, *McKinsey Quarterly*, October, http://www.mckinsey.com/insights/strategy/the_second_economy (accessed 11 May 2014).

Brynjolfsson, E. and McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, New York: WW Norton & Company.

Busch, L. (2011). *Standards: Recipes for reality*. Cambridge, Ma: The MIT Press.

Desrosières, A. (1998). *The Politics of Large Numbers: A history of statistical reasoning*, Cambridge, MA: Harvard University Press.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V.R., Tsou, A., Weingart, S. and Sugimoto, C.R. (2014). Big Data, Bigger Dilemmas: A Critical review, *Journal of the American Society for Information Science and Technology* (forthcoming).

Elder-Vass, D. (2011). *The Causal Power of Social Structures: Emergence, structure and agency*. Cambridge: Cambridge University Press.

Gellner, E. (1983). *Nations and Nationalism*. Oxford: Blackwell.

Gellner, E. (1996). *Conditions of Liberty: Civil society and its rivals*. London: Penguin.

Giddens, A. (1990). *The Consequences of Modernity*. Cambridge: Polity Press.

Gillespie, T. (2014). The Relevance of Algorithms, in Gillespie, T., Boczkowski, P. J. and Foot. K. (eds), *Media Technologies: Essays in communication, materiality, and society*. Cambridge, MA: The MIT Press.

Gitelman, L. (2013). *Raw Data Is an Oxymoron*. Cambridge, MA: The MIT Press.

Hanseth, O. and Lyytinen, K. (2010). Design Theory for Dynamic Complexity: The case of building internet, *Journal of Information Technology*, 25(1): 1-19.

Kallinikos, J. (1995), Cognitive Foundations of Economic Institutions: Markets, organizations and networks revisited, *Scandinavian Journal of Management*, 11(2): 119-137.

Kallinikos, J. (2004), The Social Foundations of the Bureaucratic Order, *Organization*, 11(1): 13-36.

Kallinikos, J. (2011). "Bureaucracy under Siege: On information, collaboration and networks". In S. Clegg, M. Harris & H. Höpfl (Eds.), *Managing Modernity: Beyond bureaucracy?* (pp. 130–153). Oxford: Oxford University Press.

Kallinikos, J., Hasselbladh, H. and Marton, A. (2013). Governing Social Practice: Technologies and institutions. *Theory and Society*, 42/4: 395-421.

Lanier, J. (2014). *Who Owns the Future?*. New York; NY: Simon and Schuster.

Manovich, L. (2001). *The Language of New Media*. Cambridge, MA: The MIT Press.

Napoli, P. M. (20100. *Audience Evolution: New technologies and the transformation of media audiences*, New York: Columbia University Press,

Porter, T. M. (1995). *Trust in Numbers: The pursuit of objectivity in science and public life*, New Jersey: Princeton University Press.

Rosenfeld, l. and Morville, P. (2002). *Information Architecture for the World Wide Web*, O'Reilly Media Inc.

Van Dijck, J. (2013). *The Culture of Connectivity: A critical history of social media*, Oxford: Oxford University Press.

Varian, H. R. (2010). Computer Mediated Transactions. *The American Economic Review*, 1-10.

Weinberger, D. (2007). *Everything Is Miscellaneous: The Power of the New Digital Disorder*, New York: Times Books.

Wenger, P. (1997). Why Interaction is more Powerful than Algorithms. *Communications of the ACM*, 45(5): 81-

Yoo, Y. (2010), Computing in Everyday Life: A call for research on experiential computing, *MIS Quarterly* 34(2): 213-231.