DHQ: Digital Humanities Quarterly
2023
Volume 17 Number 2

The Explainability Turn

David M. Berry <d_dot_m_dot_berry_at_sussex_dot_ac_dot_uk>, University of Sussex https://orcid.org/0000-0002-7737-5586

Abstract

How can we know what our computational infrastructures are doing to us? More to the point, how can we have any confidence that their effects on our minds are positive rather than negative? Certainly, it is the case that digital infrastructures combined with spatial and temporal organisation create forms of digitally-enabled structures that serve to change the cognitive capacity of humans. How then to assess these new digital infrastructures and machine learning systems? One of the most difficult tasks facing the critical theorist today is understanding the delegation and prescription of agency in digital infrastructures. These are capital intensive systems and hence tend to be developed by corporations or governments in order to combine multiple systems into a single unity. The systems they build are often difficult if not impossible to understand and require the public to trust but not to be able to verify the system decisions. In contrast, recent moves to assuage worries over the opaque and threatening potential of computation have been partially addressed through a new legal right to challenge algorithms and their decisions. This requirement, termed “explainability,” I suggest might contribute to tool criticism within digital humanities for investigating and potentially challenging these assemblages and creating a potential for democratic contestation.

Introduction

How can we know what our computational infrastructures are doing to us? More to the point, how can we trust that algorithms and related technologies do not have a detrimental effect? As technologies make up more of our digital environment, they not only provide tools for thought, but they also shape and direct the very way we think. The move from relying on books to understand a topic to using the internet to research a topic is profoundly different, not only in terms of the acceleration in access to information, but also in the reliance on “surfing” and “searching” for information. These are different cognitive modes, or styles of thinking (see [Hayles 2007] [Hayles 2010]). Digital infrastructures combined with spatial and temporal organisation create forms of digitally enabled structures that serve to change the cognitive capacity of humans (see for example [Hutchins 1995]). In 1981, Steve Jobs, then CEO of Apple, famously called computers “Bicycles for the Mind”, implying that they augmented the cognitive capacities of the user, making them faster and more capable [Jobs 1981, pp. 8–9]. But others are not so positive, with writers such as Nicholas Carr worrying that they might also undermine and fragment the possibility for thought. As Carr wrote 1

> over the past few years I've had an uncomfortable sense that someone, or something, has been tinkering with my brain, remapping the neural circuitry, reprogramming the memory. My mind isn't going — so far as I can tell — but it's changing. I'm not thinking the way I used to think. [Carr 2008]

Similarly, Bernard Stiegler [Stiegler 2015] [Stiegler 2018] has argued that the programming industries have a vested interest in changing the way individuals think to make possible a new digital consumption economy. In this paper, I examine this new situation and social responses to computational infrastructures that can now be seen to weaken historical practices of cognition. I use the term cognition to represent not just the cognitive processes of the human mind, but to include a more substantive notion which includes not only thinking, but also feeling and projecting. In particular, I understand cognition as a synthetic faculty in the application of reason which opens the possibility for a decision. The aim is to begin to account for the way in which this synthetic faculty is being automated by algorithmic processing such that the human cognitive ability to connect factors into an explanation becomes increasingly deficient. When connected into contemporary digital infrastructures, rather than acting as bicycles for the mind, these technologies replace certain cognitive functions of the mind. Being owned and controlled by corporate organisations they tend to weaken explanatory and critical thinking and instead nudge and influence human behaviour in directions that are profitable.[1] When incorporated into digital platforms these technologies can be combined to create distraction spaces for what we might call *frenetic passivity* to repress critical cognitive activity or *thought*. Revelations from industry insiders and researchers of behavioural nudging and manipulation techniques have been widely documented and have served to prompt public calls for more regulation over these systems (see [Zuboff 2019] [McNamee 2019]). We have also seen changes to the regulatory environment as the public has become increasingly uneasy about these automated systems (for example, see [EU n.d.] [European Parliament 2022]). 2

Digital technologies substitute artificial analytic capacities that bypass and replace the synthetic function of reason. Algorithms often overtake human cognitive faculties by shortcutting individual decisions by making a digital “suggestion” or intervention. The most obvious example of this is Google Autocomplete on the search bar which tries to predict what a user will type before they have completed a sentence – and make it easy for the user to just click that rather than thinking through what they are writing. This technology has also been rolled-out to Gmail, where Google will write a user's emails by predicting what it thinks the user might be planning to write. Similarly, recent breakthroughs in artificial intelligence, such as GPT-3 also create long-form, remarkably competent, written texts based on a similar automated capacity. These techniques are increasingly being incorporated into many aspects of computer interfaces through design practices that predict, persuade, or nudge particular behavioural outcomes. For example, Apple devices often “know” where you are due next by consulting your calendar and auto-calculating your route to the next event and warning the user of the minimum time for them to arrive – sometimes even cautioning the user to leave immediately. These technologies use the mobilisation of processes of selecting and directing activity through the automation of data from information collected from millions of users [Malabou 2019, 52]. As Noble has noted, 3

> What each of these searches represents are Google's algorithmic conceptualizations of a variety of people and ideas. …Google's dominant narratives reflect the kinds of hegemonic frameworks and notions that are often resisted by women and people of color. Interrogating what advertising companies serve up as credible information must happen, rather than have a public instantly gratified with stereotypes in three-hundredths of a second or less [Noble 2018, 50].

We are witnessing the social being transformed by digital technologies that transform individuals' thinking towards “operational” or instrumental thought. But it is also important to remain alert to the social dimension beyond the level of the individual so that we are attentive to the relationship between social being and consciousness. This includes the reconfiguring of social life through the technical infrastructures of computational mediation which themselves privilege individualistic ways of framing and understanding the world. A consequence of which, as Geert Lovink has noted, is that “there is no ‘social’ anymore outside of social media.” Merleau-Ponty earlier warned us, 4

> Thinking “operationally” becomes a sort of absolute artificialism, such as we see in the ideology of cybernetics, where human creations are derived from a natural information process, but which is itself conceived on the model of human machines. If this kind of thinking takes over humanity and history, and if, pretending to be ignorant of what we know about humanity and history through contact and through location… then we enter into a cultural regimen in which there is neither truth nor falsehood concerning humanity and history, then we enter into a sleep or nightmare from which nothing would be able to awaken us [Merleau-Ponty 2007, 352].

Whilst I do not have the space to rehearse all the arguments that inform this paper (but see [Berry 2011] [Berry 2014] [Daston 2022, p. 147–150]), it can be seen that resituating the cognitive processes of thought within the concrete reality of the increasingly “smart” infrastructures that surround us changes not only how we think but also our relationship to the decisions that are taken on our behalf. My aim is to use the way in which infrastructural logics of computation decentre and overtake modes of thought, for example by undermining concentration, focus and attention, to examine the way in which this leads to a situation that undermines trust in systems. This is to critically assess attempts to assuage worries over the opaque and threatening potential of computation through a new right to challenge algorithms and their decisions called *explainability*. I will later suggest new critical practices are possible within digital humanities for investigating and potentially contesting these technologies by taking on board and extending this notion. [2] 5

I believe that the *General Data Protection Regulation 2016/679* (GDPR) [GDPR 2016] can help us to understand this new problematic. When instantiated in national legislation it has created a new right in relation to automated algorithmic systems that requires the *controller* of an algorithm to supply an explanation of how a decision was made to the user (or *data subject*) – what we might call the *social right to explanation*. The GDPR is a regulation in EU law on data protection and privacy for citizens within the European Union and the European Economic Area.[3] The GDPR creates a new kind of subject, the “data subject” to whom a right to explanation (amongst other data protection and privacy rights) is given. The notion of a *data subject* has a range of very specific and unique rights as a *natural person*, which distinguishes them from an artificial intelligence, machine-learning system, algorithm or indeed a corporation. This definition creates what we might call a post-posthuman subjectivity by creating and reinforcing a boundary between humans, corporations and machines.[4] Additionally, it has created a legal definition of processing through a computer algorithm [GDPR 2016, art.4]. In consequence, this has given rise to a notion of explainability which creates the right “to obtain an explanation of [a] decision reached after such assessment and to challenge the decision” [GDPR 2016, recital 71].[5] It has been argued that this regulation mandates a requirement for a representation of the processes of computation used in an automated decision, the calculative model, for example, and for it to be presented to the data subject on request ([Goodman and Flaxman 2017] [Selbst and Powles 2017] , cf. [Wachter et al. 2017]).[6] It is crucial however to understand that this is not just an issue of legal rights, this has also created a normative demand for a social right to explanation. 6

This debate has had implications for artificial intelligence systems with the assumption that they might have to have the capacity to provide a self-description. This has become known as the problem of explainability for artificial intelligence research, and has led to the emergence of the subfield of Explainable Artificial Intelligence (XAI). Although the GDPR is limited to the European Union, in actuality it is likely to have global effects as it becomes necessary for global companies to standardise their software products and services but also to respond to growing public disquiet over these systems (see also [Darpa n.d., n.d.] [Sample 2017] [Kuang 2017]). [7] This has also become part of a wider public discourse. Explanation was one of the *rights* outlined in an *algorithmic bill of rights* published in 2019, for instance, which argued that 7

> we have the right to be given explanations about how algorithms affect us in a specific situation, and these explanations should be clear enough that the average person will be able to understand them... "The terms of service for an AI application — or any service that uses algorithmic decision-making processes — should be written in language plain enough that a third grader can comprehend it... It should be available in every language as soon as the application goes live." [Samuel 2019].

Consequently, Explainable AI has become known as transparent AI because it attempts to design AI systems whose actions can be easily understood by humans. These new AI systems are designed to produce more "explainable models, while still maintaining a high level of learning performance" and prediction accuracy thus helping humans to "understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" [Gunning 2017]. This means that XAI systems should have to have the ability to explain their rationale, characterise their strengths and weaknesses, and convey an understanding of how they will behave in the future in order to strengthen their public accountability. These requirements pose a very difficult challenge to the developers of these systems and remain aspirational in AI system design. 8

One of the key drivers for the attention given to explainability has been a wider public unease with the perceived bias of algorithms in everyday life, especially in the rise in automated decision processes and the calls for accountability in these systems (see [Sample 2017] [Kuang 2017]). Many of these debates foreground the question of the future of humanity and the kinds of societies that these technologies create the conditions for. These implications are increasingly discussed in the media and in politics, particularly in relation to a future dominated by technologies which are thought to have huge social consequences. Computation combined with artificial intelligence and machine learning has raised challenging questions about creativity, post-work futures, mass unemployment, AI controlled drone systems, and surveillance capitalism amongst other impacts. These are important issues, but here I drill down to focus on the cognitive and explanatory issues. 9

The discussion I wish to present in this paper is largely speculative. My aim is to explore how the cognitive capacities of humans might be strengthened by developing that capacity for explanatory modes of thought through the use of explainability as a critical concept. It seems to me that we have two issues that are interesting to consider. Firstly, the GDPR requires digital technologies, such as automated decision systems (ADS), to be explainable in some sense and therefore pose a problem of representation.[8] Secondly, interpretation problems stem from a difficulty in translating a highly complex processual system that does not immediately lend itself to easy explanation for a number of difficult reasons. Explanation has nonetheless become expected as part of the political and legislative response to concerns over algorithmic inequality, bias and the opaqueness of computational systems. 10

In the first section of this paper, I seek to outline the contours under which this critique becomes urgent by an initial examination of cognitive infrastructures. In the second section, I turn to think about the concept of explainability and its potential for developing a possible tactic in response to the wider toxicity generated by algorithmic governance. The aim is to offer an immanent critique of the notion of explainability. By immanent critique, I refer to an approach drawn from the Frankfurt school, whereby the internal terms and concepts within a system are examined in relation to the reality of the claims they make about and the actuality of the world. Thus, computational systems are justified both discursively and in terms of their internal logics and yet there are contradictory tendencies in these supposedly univocal systems. Discourse and algorithms become a technique to exercise power, for example through *nudging* strategic behaviour for shaping the labour, both physical and mental, of users in specific digital environments, however these behavioural techniques do not always produce the desired effect. Although behavioural logics of control operate in our everyday lives which are subject to algorithmic management from increasingly prevalent hyper-individualised capillaries of power, there remain spaces of contestation. The justificatory move to explainability as a panacea for these systems is therefore an important diagnostic site for interrogating algorithms' power and ubiquity. [9] 11

1. Thinking Infrastructures

One of the most difficult tasks facing the critical theorist today is understanding the delegation and prescription of agency in digital infrastructures. Due to their size and complexity these infrastructures are capital intensive systems and hence tend to be developed by corporations or governments in order to combine multiple systems into a single unity. In this form they point towards a unification 12 of multiple grammars within a system of communication, such that they converge on a single ontology or technical stack. This tendency eventually allows for an underlying infrastructure to be commoditised as an external product in its own right, such as shown with the Amazon Web Services (AWS) system. AWS was originally created for Amazon's internal purposes as a corporate retentional system. Since 2006 it has become a key infrastructure with an annual income of its own of $17.1 billion (8% of Amazon’s annual revenues in 2017) and is used by customers and even competitors for various forms of so-called cloud computing. These infrastructures can be understood as systemic, themselves made up of a number of component layers, but nonetheless constituting a distinct digital totality and increasingly structured through the data architecture made possible through the implementation of edge, core and cloud compute. Edge devices, such as smartphones, feed data into core (on-premises large computing data centres) for algorithmic processing, or to cloud (off-premises shared data servers) to run AI models or complex operations. This network topology is often called the edge-to-core-to-cloud pipeline for efficiently processing data, moving data to algorithms located where the processing power is best located.

13 The patterning of these layers of computation into vast laminated systems creates what I call *infrasomatization* (see [Berry 2016]). This notion draws on the work of Bernard Stiegler who has pointed to Alfred J. Lotka's and Nicholas Georgescu-Roegen's notion of *exosomatization* as a crucial means of understanding computational capitalism (see [Bobulescu 2015] [Stiegler 2016, p. 95–96]).[10] Exosomatization and endosomatization were developed by Lotka and Georgescu-Roegen in their work on ecological economics and by Karl Popper in relation to what he called objective knowledge (see [Lotka 1925] [Georgescu-Roegen 1970] [Georgescu-Roegen 1972] [Georgescu-Roegen 1978] [Popper 1972]). Exosomatization can be understood as the use of tools (from Greek *exō* meaning “outside”), whereas endosomatization is the evolutionary adaptation of bodies into claws, nails, shells, etc. (from Greek *endon* meaning “within”), *soma*, of course is from the Greek *sōma* meaning “body.”

14 Whilst these have been important contributions, by introducing a third term, *infrasomatic*, I want to argue that we should move beyond a binary of either endosomatic or exosomatic. I think this notion captures the reticular nature of specific forms of digital technologies, which create new non-human agencies and, potentially, unpredictable entropic effects – so infrasomatization combines the notion of using software, information and automation to create infrastructures. To concentrate on the notion of infrasomatization, is to try to understand the particularity of how algorithms are deployed as a new form of cognitive infrastructure. That is, algorithms are not just exosomatizations, not just the production of tools or instruments. Infrasomatizations are created by the combination of other infrastructural systems. Indeed, infrasomatizations rely on a complex fusion of endosomatic capacities and exosomatic technics leading to what Berns and Rouvroy call algorithmic governance [Berns and Rouvray 2013] and Stiegler has called the automatic society [Stiegler 2016]. Infrasomatizations can be thought of as social-structuring technologies – inscribing new forms of the social (or, in a neoliberal register, sometimes the “anti-social”) onto the bodies and minds of humans and their institutions. They are made to be always already poised for use, to be configured and reconfigured, and built into particular constellations that form the underlying structures for the creation of social subjects. Infrasomatizations have an obduracy that can be mobilised to support specific instances of thought, rationality and action. So, for example, in the case of social media, the technical infrastructure introduces a new element overtaking and reconfiguring social relations through a new grammar of communication prescribed by these technologies. This results in changes in social relations and consequently social being. Infrasomatizations can be understood to operate in a similar manner to an infra-law, which Foucault described as,

> extend[ing] the general forms defined by law to the infinitesimal level of individual lives; or they appear as methods of training that enable individuals to become integrated into these general demands. They seem to constitute the same type of law on a different scale, thereby making it more meticulous and more indulgent [Foucault 1995, 222].

15 Infrasomatizations similarly have the capacity to operate across different scales with remarkable fidelity, from micro-targeting of nudges, to aggregated groups or “universes” of individuals which can be manipulated simultaneously. The term *infrasomatization* also gestures toward a kind of gigantism, the sheer massiveness and interconnectedness of fundamental computational technologies and resources. The infrastructural dimension of these infrasomatizations means that they can be scaled to the level of planetary technics, as their physical location, particularly when presented as computational abstractions such as notions of compute, can be strategically placed (and moved) dynamically and geographically. Compute, in this sense, is an abstract unit of computation which tends to be priced at a particular level by cloud server companies so one can purchase a certain capacity of computation. The cloud infrastructures' size contrasts with the phenomenological experience of the minuteness or ephemerality of the kinds of personal devices that are increasingly merely interfaces or gateways to underlying “smart” infrasomatic systems. For example, we might consider how technologies of location are made possible by the geospheric locative satellites, in particular GPS, but also extrapolation from WiFi, camera and audio data. Location is as crucial to the development of infrasomatizations as is the machine-learning of abstract patterns in data. This is because location provides important context, and this context enables smarter abductions to be made with data, assuming as it does that a specific piece of information, practice or action makes more sense within a particular place.[11] This is manifested in a dual structure which has a physical and logical geography often encoded simultaneously

into infrasomatizations. The first kind of location that are understood within the computational systems of infrasomatizations tend to be place-poor, lacking an understanding of the specificity of place and tend towards a calculative, instrumental Cartesian representation of space. This is in marked contrast to the phenomenological experience of place infused with mood, texture, relationships and materiality [Evans 2015].[12] The second is a technical geography overlaid onto this grid, as noted above, with the division into engineering data and processing distribution over a system division of edge, core and cloud. Although these are invisible to the user, this new secondary tripartite division of the computational is arguably more important and increasingly saturates everyday life, due to the infrasomatic distribution of processing and analysis that this structure requires and makes possible.[13]

A process of cybernetic feedback, where the system is able to self-monitor across this geography means that infrasomatic systems strengthen and grow. For example, the computational capacities of Amazon's infrastructural systems increase their reach and power from the use of its client's computational practices and metadata. An infrasomatization thereby learns from its usage, which creates an amplification loop which eventually cements its functionality as a computational necessity – it becomes “smart.” It knows when to move computational capacity from cloud to core, when to move compute resources into specific geographic locations and when to cache data requests across the system's geographic spread. Hence by extrapolating and scaling these learning systems, a private corporate retentional system becomes first a regional and then planetary one. This is commonly referred to as *Infrastructure As A Service* (IAAS). One of the key elements towards understanding these large-scale infrasomatizations is that they tend towards a logic of value extraction. That is, that their size and scale create a tendency that is manifest in the algorithms that make up these systems towards data capture and its intensification towards the maximisation of rent-seeking behaviour. This is largely a logic dictated precisely from the fact that many of these systems tend towards monopoly or oligopolistic behaviour – what Peter Thiel infamously referred to as a move from “zero to one” [Masters and Thiel 2015]. This is because as a disparate collection of digital subsystems is subsumed within a larger totality, the utility of this system eventually becomes overwhelming and cost-effective such that moving or exiting an infrastructure is increasingly prohibitive. This creates the possibility for monopoly rent on the infrastructure and hence drives the tendency toward gigantic informational systems, and monopoly-oriented corporations. Thus, the principal means of value extraction enacted in these infrasomatizations tends to be through the control of multiple monopolies at different layers of the technical stack. It goes without saying that this is an extremely profitable means of extracting value creating new forms of powerful companies, such as the FAANG corporations (Facebook, Apple, Amazon, Netflix, Google). 16

Within popular culture, a wider social concern with algorithms can be seen in the social media which have been used to highlight the inexplicable ways in which people's lives have been affected by an algorithmic decision. Sometimes these discussions reflect a confusion by users over the distinction between *noise*, where the decision is affected by incomplete or inaccurate data causing inconsistency or no decision being made, and *bias*, where the accuracy of a decision has been swayed by a predetermined or computed result affected by human biases (see [Jaume-Palasi 2018]). These have been used to justify a need for explanation to help the public understand algorithms. Bias in computer systems usually derive from either (1) data-driven bias, where the biases are embedded in the data itself, (2) bias through interaction with humans, for example Microsoft’s Tay chatbot which developed a fascist conversation style (3) emergent bias, for example through likes and shares, (4) similarity bias, where filter bubbles can emerge, and (5) conflicting goals bias, where stereotypes have been used in the development of the software in particular ways [Hammond 2016]. Indeed, there are now many documented cases where algorithmic decision processes have discriminated against people on the basis of their names, their home address, gender or skin colour [Buranyi 2017] [Eubanks 2017] [Noble 2018].[14] This is reflected in an “anxiety felt by those who fear the potential for bias to infiltrate machine decision-making systems once humans are removed from the equation” [Casey et al. 2018, 4]. It is in this context that public disquiet has risen in relation the perceived unfairness of these, often unaccountable, automated algorithmic systems. 17

So Facebook, for example, has created an infrasomatization for capture and exploitation of the social graph, particularly digital identity, through its social network and the creation of facial recognition systems such as Detectron [Facebook 2019]. Google similarly has created infrasomatizations for the various functions of search, compute, storage and databases, networking, big data, and cloud AI, identity and security, Internet of Things (IoT), API platforms, and location services, such as Google Maps. These are often built extremely quickly and issues of bias are rarely considered as part of this engineering effort. This phase of digital transformation is easily missed as it takes place behind the interface in proprietary corporate environments, and as such is a non-visual dimension of a computational mode of development. Through these infrasomatic logics, cultural practices are captured and rearticulated through grammars of action which can be used to describe, and then build infrasomatizations that may become cultural monopolies in their own right. For example, the contemporary emergence of a vast social system structured around social media which inculcates a craving for “likes”, “followers”, “subscribers” and “views” directly connected to a political economy of advertising, marketing and consumption is only the most obvious contemporary manifestation of this process. 18

The implications of this new system of exploitation is the creation of constellations of infrasomatizations that can be mobilised into de facto monopolies in specific imbrications. This, I would argue is a better way to understand these computational structures rather than the notion of “platforms” that tends to use a self-description favoured by companies in Silicon Valley itself, and therefore hides more 19 than it reveals. These infrasomatic systems are able to extract rent or tolls to pass data and calculations around a system – whether measured in terms of compute, data traffic or time. The forms of data they carry, even if only manifest as abstract metadata, are in themselves extremely valuable. Even if a particular customer of the infrasomatization may expressly prohibit the harvesting of their own data they cannot control secondary data produced as a result of their interaction on the system. This infrasomatic data offers another means of value extraction, both in terms of predicting the future growth of these infrastructures, but also the potentials for new circulations of data and logic for profit. By use of these multiple “data exhausts”, the owners of these infrasomatizations are able to capture trends, identify social tendencies and patterns, and to reincorporate this knowledge into their infrasomatic ecology, and depending on the corporation, feed this information back into circulation to amplify these tendencies in a profitable direction. Within Silicon Valley this is understood as the capacity of a digital company to create a “moat” which prevents competitors from disrupting their business model, rather like a castle with a moat surrounding it to prevent attack and capture. The creation of an infrasomatic layer is, therefore, not just a digital logic, it is also a business logic. These logics reinforce each other, creating a structure that, given enough physical computing infrastructure can scale at an exponential pace, and thereby capture value and create a kind of dependency in its customers and users very quickly.

The need to convert this raw data from its digital logic into a business logic has consequently resulted in major breakthroughs in artificial intelligence, particularly machine learning, through the creation of classification and filtering systems modelled on brain structures, and the underlying neurons. Consequently, we see a growing use of computational systems to abstract, simplify and visualize the amount of “Big Data” that is being collected. A side-effect of this has been to reinforce a tendency towards causal and statistical models to map, understand, and interpret complex social and cultural phenomena. For example, in 2008 Chris Anderson famously announced the “End of Theory” as he claimed the data deluge had made the scientific method obsolete. Indeed, he argued that “we can stop looking for models, instead we can analyze data without hypotheses”. He further argued that we can “throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot” and that “with enough data, the numbers speak for themselves” [Anderson 2008]. But of course, this shift to statistical explanation is not neutral, rather it is linked to the emergence of a political economy specific to the computational. In this data-based accumulation regime social life is transformed into calculable and predictable social trends which may be manipulated and channelled. The most striking example of this regime is the use of Facebook data by the company Cambridge Analytica which they argued could create psychographic models which could then be “nudged” to influence behaviour. The alleged result of these techniques includes the Brexit referendum result and the election of Donald Trump as president in 2016 [Guardian 2018]. Although their efficacy remains contested these nascent techniques are continually refined and improved and moved from a communicational terrain to a cognitive one. 20

Computation today means to be in the middle of things, it is no longer an end, but rather a means, a passage-way between two points: from dumb to smart. In becoming smart devices, computational systems transform everyday life into what can be thought of as a vast oil field of data, awaiting extraction by a new set of digital cultural industries. It is of no surprise that FAANG (Facebook, Apple, Amazon, Netflix and Google), the leaders of the technology industry, are racing to create the technologies for their vision of a digital life. Mathematician and architect of supermarket giant Tesco's Clubcard, Clive Humby, described data as the new oil in 2006 [Palmer 2006]. It is increasingly clear that we are now in the middle of an oil rush at the centre of which lies our lives. As Wired explains, “like oil, for those who see data's fundamental value and learn to extract and use it there will be huge rewards”[Toonders 2014]. Humby further argues that “data is just like crude. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value”. But it is not just the one-off collection of data, it is the iterative gathering of data, repeated again and again that creates the conditions for these possible insights. The oil fields of life will not soon be spent, instead they will yield greater and greater quantities of data, from which more profit can be earned.[15] 21

This extractive metaphor serves not only Silicon Valley but also inspires governmental policy. For example, Meglena Kuneva, European Consumer Commissioner, has without blinking, described personal data as “the new oil of the internet and the new currency of the digital world” [Kuneva 2009]. The UK Office for National Statistics has argued that “if data is the new oil, open data is the oil that fuels society and we need all hands at the pump” [Davidson 2016]. What makes data into open data, is that it is free of intellectual property restrictions that prevent it from being used by others by publishing constraints, such as copyright, or that it is owned exclusively by its creators. Open data, like open access publications and open source before them, grants a corporation the right to dice up and remix data. When you use your smartphone, or a smart object, the first thing that has to be clicked is the agreement to let companies extract and use this data. As the New York Times argues, 22

> Personal data is the oil that greases the Internet. Each one of us sits on our own vast reserves. The data that we share every day — names, addresses, pictures, even our precise locations as measured by the geo-location sensor embedded in Internet-enabled smartphones — helps companies target advertising based not only on demographics but also on the personal opinions and desires we post online [Sengupta 2012]

.

23 These claims reflect what we might call a *cult of data-ism* and a renunciation of the extended and important role of critical reason and theoretical thinking in modern society. But this data-ism extends beyond the mere collection of data and its analysis. Data that is collected in “data lakes” can be used to formulate behavioural and predictive logics which can provide useful interpretative and calculative advantages to corporations. They then provide the opportunity for algorithmic interventions – what I call *algoventions* – into patterns of behaviour or thought. These practices have been increasingly extended across society, but possibly the most intensive and ambitious use of these infrasomatic technologies takes place in so-called *smart cities*. Here the city is built from the ground up to facilitate the data capture and feedback loops to make possible a management and organisational control layer over the city giving top-view to city officials, but also selectively sharing data with corporations and individuals to use in their everyday activities. For example, public transport usage and problems can be collected, aggregated and circulated back to the users of the public transportation system to provide them with early-warnings of issues, propose alternate routes, or to alert them to major outages in a system. Smart cities, and their underlying infrasomatizations, are strongly coupled to geolocation data, indeed, the grid of the city is a key abstract principle upon which the data about a city is projected. By unifying multiple data streams derived from smart infrastructures, smart city computers can create realtime digital twins which attempt to create and thereby impose a data-centric spatial logic onto city life. These systems are tasked with classifying, understanding, and predicting future states of the digital twin of the city, that is the city as a gigantic finite-state machine built on the collection of massive amounts of civic, corporate and personal data.

24 Through a combination of these techniques, infrasomatizations are created which produce smart technologies that act as gateways that open out to spatial forms of organisation that delegate a locative-calculative model onto the user, structuring the world in terms of an index of spaces that are given relational properties within the row and column structure of the underlying tables and databases. We therefore need to contest this new social pattern and develop alternative visions – part of which can be through creating new tools and new modes of working with technology, but we also need tool criticism and a research programme dedicated to understanding the algorithmic condition.[16]

2. The Explainability Turn

25 It is clear that in the context of infrasomatizations, the first important question we need to consider is what counts as an explanation. Indeed, explanations are generally considered to be able to tell us how things work and thereby giving us the power to change our environment in order to meet our own ends. In this sense of explanation then, science is often supposed to be the best means of generating explanations [Pitt 1988, 7]. So, with a stress on the importance of explanation, the GDPR makes it a criterion of adequacy for satisfactory use of algorithmic decision systems in the European Union, and thereby legitimating their use in a multitude of settings. Thus, explainability and the underlying explanation are linked to the question of justification. So, what then is an explanation?

26 Hempel and Oppenheim [Hempel and Oppenheim 1988] argue that an explanation seeks to “exhibit and to clarify in a more rigorous manner”. Some of the examples they give include whole temperature reading from a mercury thermometer, which can be explained using physical properties of the glass and of mercury which has been rapidly immersed in hot water. Similarly, they present the example of an observer of a row boat where part of the oar is submerged under water and appears to be bent upwards [Hempel and Oppenheim 1988, 10]. An explanation therefore attempts to explain with reference to general laws. Mill argues that “an individual fact is said to be explained by pointing out its cause, that is, by stating the law or laws of causation, of which its production is an instance” and that “a law or uniformity in nature is said to be explained, then another law or laws are pointed out, of which that law itself is, but a case, and from which it could be deduced” [Mill 1858]. Similarly, Ducasse argued in 1925 that “explanation essentially consists in the offering of a hypothesis of fact, standing to the fact to be explained as case of antecedent to case of consequent of some already known law of connection” [Ducasse 2015, 37]. Hempel and Oppenheim therefore argue that an explanation can be divided into its two constituent parts, the *explanadum* and the *explanans*,

> By the explanandum, we understand the sentence describing the phenomenon to be explained (not the phenomenon itself); by the explanans, the class of those sentences which are adduced to account for the phenomenon [Hempel and Oppenheim 1988, 10].

27 In this sense of an explanation, the explanandum is a logical consequence of the explanans. The explanans itself “must have empirical context, that is, it must be capable, at least in principle, of test by experiment or observation,” which creates conditions for testability. However, this causal mode of explanation can become inadequate in fields concerned with purposive behaviour, as with infrasomatic digital systems.

28 In this case it is common for reference to purposive behaviour, such as in so-called machine behaviour, to be given in relation to

“motivations” and therefore for teleological rather than causal explanation. Thus, the goals sought by the system are required in order to provide an explanation. Teleological approaches to explanation may also make us feel that we really understand a phenomenon because it is accounted for in terms of purposes, with which we are familiar from our own experience of purposive behaviour. One can, therefore, see a great temptation to use teleological explanation in relation to AI systems, particularly by creating a sense of an empathetic understanding of the “personalities of the agents.” So, a proposed explanans might sound suggestively familiar, but “upon closer inspection proves to be a mere metaphor, or to lack testability, or to include no general law, and therefore to lack explanatory power” [Hempel and Oppenheim 1988, 17]. In relation to explanation, therefore, explainability needs to provide an answer to the question “why?” Scriven argues that “the right description is the one which fills in a particular gap in the understanding of the person or people to whom the explanation is directed”. This can be seen as the value of explainability as “closing the gap in understanding (or rectifying misunderstanding)” [Scriven 1988, 53].

It is clear that the concept of explainability, and the related practices of designing and building explainable systems, have an underlying theory of general explainability, but also a theory of the human mind. These two theories are rarely explicitly articulated in the literature, and I want to bring them together to interrogate how explainability cannot be a mere technical response to the contemporary problem of automated decision systems, but actually requires philosophical investigation to be properly placed within its historical and conceptual milieu. 29

The next important move is to connect the concept of explanation to automated decision systems and the explanations that they can provide. As shown above, writers such as Friedman have argued that explanation is almost always explanation of laws as a general regularity or pattern of behaviour more typical of the physical sciences [Ruben 2016, 4]. But far too many discussions of explanation assume that what can be said about *scientific explanation* exhausts what of interest there is that can be said about *explanation*. However, in relation to algorithmic systems what we tend to be talking about is what Ruben has called “singular explanation” Additionally, explanation is ambiguous as it may refer to the product or to a process, so as Bromberger points out, an “explanation may be something about which it makes sense to ask: How long did it take? Was it interrupted at any point? Who gave it? When? Where? What were the exact words used? For whose benefit was it given?” (Bromberger, quoted in [Ruben 2016, 6]). The other form of explanation “may be something about which none of the [previous] questions make sense, but about which it makes sense to ask: Does anyone know it? Who thought of it first? Is it very complicated?” [Ruben 2016, 6]. So, in speaking of an explanation one might be referring to an act of explaining, or to the product of such an act. 30

It certainly seems to be the case that the right to explanation that is being developed in relation to the GDPR is chiefly interested in the idea of an explanatory product. Thus, an “explanatory product” can be characterised solely in terms of the kind of information it conveys, no reference to the act of explaining being required. The question therefore becomes, what information has to be conveyed in order to have explained something? So in terms of the requirements given, the function of explanation is that explanation should enable us to understand why something has happened within an automated decision system. Crucially, this connection between an explanatory product and the legal regime that enforces it has forced system designers and programmers to look for explanatory models that are sufficient to provide legal cover, but also at a level at which they are presentable to the user or data subject. It is also uncertain if the “right is only to a general explanation of the model of the system as a whole ('model-based' explanation), or an explanation of how a decision was made based on that particular data subject's particular facts ('subject-based' explanation)” [Edwards and Veale 2018, 4]. This is not an easy requirement for any technical system, particularly in light of the growth of complicated systems of systems, and the difficulty of translating technical concepts into everyday language. It might therefore be helpful to think in terms of full and partial explanation, whereby a partial explanation is a full explanation with some part left out. That is, that, while presenting a complicated system of automated decision systems, it is likely pragmatically that explanations will assume an explanatory gap, assuming that the data subject is in possession of facts that do not need to be repeated. It will be interesting to see if the implementation of these systems results in an explanatory pragmatism, and how the legal system responds. 31

This of course leads to the danger of creating persuasive explanations rather than transparent explanations or a pragmatic explanation drawing on the notion of a “good enough” explanation. It also raises questions related to the over-simplication of explanations or misleading explanations and how one might challenge them or even question their underlying explanatory model.[17] This difficulty might explain the recent turn towards explainability through the notion of machine behaviour, drawing on insights drawn from research on humans and animals applied to machines.[18] These researchers argue, 32

> in the context of machines, we can ask how machines acquire (develop) a specific individual or collective behaviour. Behavioural development could be directly attributable to human engineering or design choices…. [or] a machine may acquire behaviours through its own experience. For instance, a reinforcement learning agent trained to maximize long-term profit can learn peculiar short-term trading strategies based on its own past actions and concomitant feedback from the market… In the study of animal behaviour, adaptive value describes how a behaviour contributes to the lifetime reproductive fitness of an animal. In the case of machines, we may

> talk of how the behaviour fulfils a contemporaneous function for particular human stakeholders [Rahwan et al. 2019, 480].

So underlying the concept of explainability is the assumption that algorithms are themselves explainable and following from that, that algorithms are something that can be explained to a human. This further assumes that the interpretative activity that humans are capable of can be mobilised to understand algorithms, or at least their active computational dimension. But the concept also assumes that there exists what we might call a general algorithmic explainability, in other words that all computational processes can be rendered as an explanation, and therefore explained with recourse to a translation into the discursive or symbolic order in which humans can interpret what an algorithm is doing. This therefore gestures to a theory of the human mind whereby subjective experience is capable of undertaking interpretative work and thereby of creating meaning out of an explanation of a given algorithm. But in cases where the infrasomatic systems are progressively undermining this kind of cognitive skill, this reveals a contradiction in the notion of explainability – humans might struggle to understand explanations and might therefore require cognitive support from visualisation systems created to support that capacity. 33

There is an assumption that provided we know all the factors that influenced an automated decision, whether directly or indirectly, we must be able to comprehend the movement of states within which an automated system must move on the occasion of a certain event, set of data, or calculation. However, this assumption is rather ambitious in that it assumes a lot of background, contextual or tacit knowledge and a particular level of cognitive capacity. Renz [Renz 2018, p. 4] describes this mode of apprehending and understanding realistic rationalism, arguing that “a realistic rationalism must be able to make plausible that everything that is or that happens can in principle be grasped or comprehended — that every being is, to use a traditional term, intelligible”. This might imply that an algorithm is different from its explanation, and that the algorithm exists prior to its explanation. This also has methodological implications such that an explanation must be able to secure the intelligibility of the automated process within the concepts already understood by the human interpreter. 34

These requirements raise difficult issues for designers of algorithmic decision systems as they might be impossible to implement, even on systems that seem relatively simple on the surface. As discussed above, a major justification is the growing public concerns over biases, whether intentional or not, being built into an algorithmic or machine-learning system. So, the new *right to explanation* has been mobilised as an attempt to mitigate these worries but also put in place legislative means to seek redress for them through the GDPR. But this does not necessarily mean that the actual algorithm need be provided, nor details of the processing steps outlined. Thus, this is increasingly a representational challenge – how to represent an algorithmic decision to a data subject. In effect, the processing might be presented as a simplified model, or explanation, that shows the general contours of the algorithm used in a particular case to an assumed reader, an increasingly cognitively sophisticated user who can understand the explanation.[19] 35

I call this the *Explainability Turn*. It is a genuinely interesting question as to the extent to which explainability will be able to mitigate the public anxieties manifested when confronted with opaque automated decision systems. The scale of the challenge represented by the requirement to provide an explanation seems to me to be under-appreciated, and clearing the grounds for even thinking about this problem cannot be overstated. It nonetheless seems clear that the notion of explainability has been derived from an epistemological insight informed by debates over how scientific activity itself can be explained. However, computers and their algorithms are not so easily fitted into the derivations of general laws that explanation seems to require, and this assumption therefore remains an interesting aporia in the notion of explainability. 36

Hence, I argue that thinking about explanation in relation to algorithms needs to be informed by the humanities which can enrich these debates, for example by deepening the meaning of explainability with what I call *understandability*. That is, rather than providing descriptions purely from the domains of a formal, technical and causal model of *explanation* (dominant in the sciences), these technologies would benefit from critical approaches that take account of *understanding*, more common in the humanities and social sciences (see [Berry 2011] for an earlier discussion of this, see also [Connolly 2020]). The notion of explanation needs to be interrogated by the humanities, and particularly the concept of explainability it gives rise to. This is increasingly relevant to the growing public visibility of humanities and the potential for the use of machine learning in related fields, such as digital humanities. Therefore, this is an area that digital studies and digital humanities could make an important contribution both in thinking about their own work and the impact of algorithms, but also working in conjunction with other fields. 37

Unfortunately, due to limitations of time I do not have time to discuss further here. But if it is the case that infrasomatizations create cognitive infrastructures that proletarianise our cognitive faculties creating anti-thought, overtaken thought and non-thought, then explainability creates a potential way of bringing back into visibility these issues. For the user these infrasomatizations are experienced through smart-phones and tablets which close the loop from within the brain to the outside environment, such that the aperture of thought is mediated and compressed. Hence, the capacity for the human brain to perceive that algorithms are organizing their thoughts, or even to perceive that algorithms are at work, is impaired, if not destroyed – human reason is thereby diminished and made susceptible to persuasion and propaganda as demonstrated by the Cambridge Analytica scandal that continues to 38 reverberate. These systems aim to directly influence the practice of cognition as it has been historically constituted. New retentional and protential systems are therefore directly implicated in a process of transforming the way in which we create the conditions for cognition , directly subverting, and in extreme cases replacing elements of cognitive processes in human thought and experience.

3. Conclusion

Part of the responses we need to develop are through thinking about infrastructures differently. We might, for example, seek to develop new logics for what Bernard Stiegler has called a *contributory economy* as an alternative form of political economy for digital society [Stiegler 2018]. I agree that strategies such as these are crucial to create a safe-harbour for critical reason and hence to enable the contestation or transformation of infrastructures into new possibilities and thereby create the conditions for a new epoch. In order to do this I have argued previously that we need to undertake a programme of criticism with respect to the computational and particularly its manifestation in digital capitalism [Berry 2014]. But we need to go further and seek to understand and challenge the way in which “smart” infrastructures recast certain regulatory or legal limitations into ineffective measures from which they are able to extract excessive amounts of profit and exhaust the wider economy creating new forms of structural poverty and inequality. The combination of new “smart” technologies and the social right to explanation that explainability makes possible opens up a potential for a new critical space of what we might call *tool criticism* and the development of a wider literacy for a general public sharing increased anxiety about the effects of these automated systems. This seems to me exactly the kind of expertise that humanists and social scientists are highly skilled at and who could therefore help inform the debate over explainability.[20] 39

It is here, I argue, that theory and its development is crucial to understand the contemporary computational situation through the confrontation of the object with its own concept. We need to develop an approach that refuses to ignore and smooth over contradictions and contradictory claims. Computational societies continue to embody interaction based on deception and distortion (in other words, as ideology), and which can often be translated unreflexively into algorithmic forms. The cult of data-ism is a turn away from the project of seeking to understand society and culture through the application of critical reason in human affairs towards a data-deterministic world. It is problematic to erect an abstract and metaphysical standard by which human action and society can be judged – yet the cult of data-ism makes such a claim and works hard to produce and reproduce this new data-centric milieu. Algorithms and data must be subject to citizens' power to contest and challenge this new form of authority and it is here that the concept of explainability offers a novel potential. Indeed, as a critical concept it might contribute to concrete examples of computationalism by drawing on critical theory and transforming explanation and explainability into critical practices. This further enables us to challenge the cult of data-ism and an administrative approach to thinking about algorithms and instead to suggest different ways of being in a digital age. 40

The digital world is not a static object; it is a highly dynamic and relational system which is in constant movement and undergoing continual change. For example, it is quite remarkable to note that the internet has never been taken off-line in order to be upgraded or changed, rather it is built through accretions and replacements that are slotted into or onto the existing system structure whilst it is still “running”. This is an important aspect to understanding the always-on nature of these new infrasomatic systems. It also makes understanding the material specificity of algorithmic systems extremely important, and helps to show why an analysis that focused only on the “data” or “content” of an infrasomatic or infrastructural system would be insufficient. We need to challenge the voracious appetite for data which extends to all aspects of life and is often accompanied by a cult of data-ism expressed through the cyber-libertarian notion that “information wants to be free”.[21] 41

We might note that in advanced capitalist societies, economic anarchy is interwoven with rationalization and technology to create fewer chances for mental and reflective labour. Under such conditions, the values of instrumental reason are accorded a privileged status since they are embodied in the concept of rationality itself. The confounding of calculation with rational thinking implies that whatever cannot be reduced to number is illusion or metaphysics. As a result, the conditions are created for a greater susceptibility of society to demagogic discourses and charismatic forms of power and a weakening of the potential for individuation. This forms part of the wider significance of infrasomatizations and how we need, more than ever, social critique and critical thinking under contemporary conditions. Indeed, behind the ideological claims of data science and related approaches, particularly in Silicon Valley, this fetishism of calculation and computation is dominant. In spite of its efforts to reflect the object of analysis in terms of the manifest forms of development, such as here with algorithms, critical theory depends in its analysis on particular historical conditions. 42

It is crucial to maintain a dynamic distinction between social processes and resultant social forms of commodity fetishism that make up the underlying political economy of the new digital milieu. Institutional and ideological formations are not simple reflections of an economic base; instead, work has to be done to understand both culture and economy in relation to the growing use of computation. In the context of computation it requires that we need to consider the specific historical ideas and practices within which we experience algorithms and in which they are made and remade. We must, therefore, examine the particular historical conditions that give the present its shape in relation to the specific material and ideological formations that algorithms introduce into the social and 43 economic conditions of society. Explainability, and the explanations it might give rise to, seems to me to offer a particularly rich potential for contributing to this project. This means that we need to critique an ahistorical notion of the “algorithm” and critically interrogate metaphors and analogies used in explainability that are necessary to explain but are not sufficient for understanding the instantiation of algorithmic forms.

44 One potential response then, is that on the ruins of critical reason a new sense of the gradients of cognition must be understood – what exactly are the faculties of the mind that are directly undermined or replaced by infrasomatizations? The ruins must be uncovered to create new values, new standards, new defences, to create situated identities and critical spaces for defending against the onslaught of the algorithmic giants of the 21st century. Weapons for the weak will be needed to push back this colonisation of public and private reason. The only way for there to be critical reason in a digital age, will be if it is rebuilt on these ruins. The digital humanities can contribute a new research programme to interrogate the political and technical digital monopolies that invade our lives. I suggest that the first stages will be through the mobilization of a critical concept of explainability, the second through the creation of new tools, and lastly through the theorisation of a critique of computational reason.

Notes

[1] Many technology companies rely on techniques developed in casinos to nudge behaviour to maximise profitability, such as creating addictive experiences and by disarming the will of the user. Using techniques such as “Trigger, Action, Reward and Investment” these systems help create addition to a particular product (see [Schüll 2014] [Eyal 2014]).

[2] We might contrast the idea of *explainability*, which is intended to create explanations, with the notion of *observability* developed by Rieder and Hoffman (2020) and what Lipton (2017) and others have called *interpretability*. Rieder and Hoffman argue that “observability emphasises the conditions for the practice of observing in a given domain ... We therefore position observability as an explicit means of, not an alternative to regulation” [Rieder and Hofman 2020, p. 3–10]. I seek to explicitly link explainability to critique, whereas *observability* is developed as an administrative concept to aid in regulatory and policy outcomes. Interpretability is closer to my idea of explainability as aiding human understanding of algorithmic models and software [Lipton 2017].

[3] Following the GDPR, in the UK, the enabling legislation for the European GDPR is the *Data Protection Act 2018*.

[4] It appears that the idea is that only a *natural person* may ask for an explanation, preventing algorithms or corporations from requesting an explanation from other algorithms or corporations.

[5] Whilst non-binding, the Recitals “dissolve ambiguity in the operative text of a framework”, and they provide a critical reference for future interpretations [Casey et al. 2018, 17].

[6] [Wachter et al. 2017] argue that the “GDPR does not, in its current form, implement a right to explanation, but rather what we term a limited 'right to be informed'” although this has been contested in the literature as their argument rests on a rather narrow reading of the effects of Recital 71 (see [Edwards and Veale 2018]. But nonetheless “the GDPR’s right of access only grants an explanation of automated decision-making addressing system functionality, not the rationale and circumstances of specific decisions” [Wachter et al. 2017, 19]).

[7] It is important to note that although this paper has focussed on the GDPR, explainability was also part of a Darpa research programme in 2016 (DARPA-BAA-16-53). More information can be found here: https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

[8] This means that algorithmic systems are required to provide their processing descriptions under this “right to explanation” and potentially giving rise to a critical field such as Explainable Digital Studies – XDS.

[9] These issues are explored in depth in the work of [Irani 2015] who focuses on how social conflict is mediated through particular assemblages of algorithmic systems.

[10] Exosomatization and endosomatization have been deployed by Stiegler to think about human augmentation and digital technologies, particularly in relation to the anthropocene and the counter-entropic move towards a *neganthropocene* (see for example, [Stiegler 2015] [Stiegler 2018]).

[11] Equally important is the overlaying of computational and therefore calculable layers over the physical environment. These layers are crucial for next

generation infrasomatizations, using maps and other locative technologies.

[12] For example, in terms of the technical transformation of place we might consider the softwarization of the home – a site of so-called micro-location. Its conversion into an algorithmic space is a process which is now well under way and which involves transforming dumb things into smart objects through the use of artificial intelligence. But AI cannot function without data, large amounts of data, to help them understand the world. Smart devices need to watch and record us, harvesting vast quantities of data, so that our every activity can be captured by sensors and cameras embedded within them. One of the more contentious recent examples is the proposal by Amazon to build a surveillance-as-a-service system. In this patented system, the company aims to use its network of delivery drones to keep watch over customers' houses using location data to form a flying Neighbourhood Watch drone system. It is suggested that customers could request that Amazon's drones visit their property hourly, daily, or weekly, and the drones would look for signs of break-ins, such as smashed windows, doors left open, and intruders lurking on people's property ([Porter 2019]). The patent further suggests that drones could be equipped with night vision cameras and microphones to expand their sensing capabilities [USPTO 2019a]. This is in addition to an earlier patent application envisions using a combination of Amazon Ring doorbell cameras and facial recognition technology to build a system that could be used to match images of people who show up at your door to a “suspicious persons” database ([USPTO 2019b], [Meek 2019]). It goes without saying that these activities produce useful raw data in vast quantities and for which more intensive surveillance systems are being built.

[13] This highlights the importance of the relationship between the instrumental imposition of location, understood technically as geo-fencing, against that of what Bernard Steigler is increasingly referring to as locality, a counter-computational politics of place (see http://internation.world).

[14] This has even resulted in families and groups being deliberately separated by algorithms for profit, or AI “scans” for a babysitter with “respect and attitude” [Harwell 2018].

[15] See also Jim Balsillie who argued that

> “data is not the new oil – it’s the new plutonium” and that “data at the micro-personal level gives technology unprecedented power to influence ... Amazingly powerful, dangerous when it spreads, difficult to clean up and with serious consequences when improperly used” [Balsillie 2019].

[16] Academia is itself in the middle of a digital revolution, the outlines of which are still only dimly perceived. For example, open access licenses create the data foundations for gigantic systems of surveillance to be built to monitor, manage and control academic labour. University management are enthusiastically building new collection systems using these open access licenses as their foundations (often with the tacit approval of academic faculty, librarians and researchers). This situation is happening right under the noses of academics who are swayed by moralistic arguments about participation and the sharing of knowledge, but which will actually result in the bypassing of historical and hard-won principles of academic freedom built on the notion that academic labour means that the copyrights belong in the first instance to the scholar, not to the university. This was originally developed as a practice to protect the rights of academics who could choose where to publish their work without limitation. These rights are now carelessly discarded with little critical thought as to the unintended consequences of a restriction of publication into open access venues. One of the most immediate effects is that for the first time in history, universities can, without restriction, build monitoring systems for publication at a very fine granularity because they do not have to worry about infringing academic rights of publication. These systems create accounting logics, themselves linked to performance monitoring, and eventually a policing function over academic labour. Open access has thereby become a political doxa and a technical system of organisation and management.

[17] For example, the user might be able to challenge an explanation or appeal to a higher authority if it were considered inadequate.

[18] One is tempted to assume that some developers believe that behavioural models of automated systems will be easier to describe, perhaps as black-boxed input-output models, or that a user will naturally find these descriptions more comprehensible. There is some similarity between “machine behaviour” approaches and the thinking behind the idea of so-called “counter-factual explanations” proposed by [Wachter et al. 2018], which assumes that by changing the input conditions a counter-factual output can be presented to the user. They argue that

> counterfactuals bypass the substantial challenge of explaining the internal workings of complex machine learning systems. Even if technically feasible, such explanations may be of little practical value to data subjects. In contrast, counterfactuals provide information to the data subject that is both easily digestible and practically useful for understanding the reasons for a decision, challenging them, and altering future behaviour for a better result [Wachter et al. 2018, 860]

Note how this conveniently avoids the problematic of explanation of the underlying algorithm and instead resituates the responsibility for changing “behavioural” outcomes onto the individual. This looks less like a “right to explanation” than a means to avoid the social responsibilities on “data processors” implicit in explainability by creating a “minimal form of explanation”. Which even they have to concede “counterfactuals may be insufficient in themselves” [Wachter et al. 2018, 883].

[19] This also raises questions about the potential for what we might call explainability regress, whereby explanations are sought for the explanation and so on ad infinitum. Until these cases are tested in practice, it is difficult to know what the limitations will be in relation to explanations provided by a system.

[20] Additionally, digital humanists tend to be familiar with technical systems and the questions raised by understanding and interpretation more generally, for example in the discussions about hermeneutics, understanding and practices of close and distant reading.

[21] It is worth reflecting on the naivety of some proponents of open access who extol the virtues of free information without connecting it to its genesis in cyberlibertarian modes of thought (see [Golumbia 2016].

Works Cited

Anderson 2008 Anderson, C. (2008) “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” *Wired*. Accessed 18/12/2015 at: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Balsillie 2019 Balsillie, J. (2019) “Jim Balsillie : 'Data is not the new oil – it's the new plutonium'”, *Financial Post*, https://business.financialpost.com/technology/jim-balsillie-data-is-not-the-new-oil-its-the-new-plutonium

Berns and Rouvray 2013 Berns, T. and Rouvroy, A. (2013) “Gouvernementalité algorithmique et perspectives d'émancipation : le disparate comme condition d'individuation par la relation?”, accessed 14/12/2016, https://works.bepress.com/antoinette_rouvroy/47/download/

Berry 2011 Berry, D. M. (2011) *The Philosophy of Software*. London: Palgrave.

Berry 2014 Berry, D. M. (2014) *Critical Theory and the Digital*. New York: Bloomsbury.

Berry 2016 Berry, D. M. (2016) “Infrasomatization.” *Stunlaw* at: http://stunlaw.blogspot.co.uk/2016/12/infrasomatization.html.

Bobulescu 2015 Bobulescu, R. (2015) “From Lotka's biophysics to Georgescu-Roegen's bioeconomics.” *Ecological Economics* 120, 194–202.

Buranyi 2017 Buranyi, S. (2017) “Rise of the racist robots – how AI is learning all our worst impulses”, *The Guardian*, https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses

Carr 2008 Carr, N. (2008) “Is Google Making Us Stupid? What the Internet is doing to our brains”, *The Atlantic*, https://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868/

Casey et al. 2018 Casey, Bryan and Farhangi, Ashkon and Vogl, Roland, (2018) “Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise” (February 19, 2018). *Berkeley Technology Law Journal*, Available at SSRN: https://ssrn.com/abstract=3143325

Connolly 2020 Connolly, R. (2020) “Why Computing Belongs Within the Social Sciences”, *Communications of the ACM*, August 2020, Vol. 63 No. 8, Pages 54-59

Darpa n.d. Darpa (n.d.) *Explainable Artificial Intelligence (XAI)*, https://www.darpa.mil/program/explainable-artificial-intelligence

Daston 2022 Daston, L. (2022) *Rules: A Short History of What We Live By*, Princeton University Press

Davidson 2016 Davidson, R. (2016) “Open Data is the new oil that fuels society”, Office for National Statistics, https://blog.ons.digital/2016/01/25/open-data-new-oil-fuels-society/

Ducasse 2015 Ducasse, C. J. (2015) “Explanation, Mechanism and Teleology”, in *Truth, Knowledge and Causation*, Routledge.

EU n.d. EU (n.d.) “Are there restrictions on the use of automated decision-making?”, https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/dealing-citizens/are-there-restrictions-use-automated-decision-making_en

Edwards and Veale 2018 Edwards, L. and Veale, E. (2018) “Enslaving the algorithm: from a 'right to an explanation' to a 'right to better decisions'?”, January 2018. *Publication in IEEE Security and Privacy*, https://pureportal.strath.ac.uk/files-asset/72824599/Edwards_Veale_SPM_2018_Enslaving_the_algorithm_from_a_right_to_an_explanation_to_a_right_to_better_decisions.pdf

Eubanks 2017 Eubanks, V. (2017) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, St Martin's Press.

European Parliament 2022 *European Parliament (2022) EU Digital Markets Act and Digital Services Act explained*, European Parliament, https://www.europarl.europa.eu/news/en/headlines/society/20211209STO19124/eu-digital-markets-act-and-digital-services-act-explained

Evans 2015 Evans, L. (2015) *Locative Social Media Place in the Digital Age*, Palgrave Macmillan.

Eyal 2014 Eyal, N. (2014) *Hooked: How to Build Habit-Forming Products*, Portfolio Penguin

Facebook 2019 Facebook (2019) *Detectron*, https://research.fb.com/downloads/detectron/

Foucault 1995 Foucault, M. (1995) *Discipline and Punish: The Birth of the Prison*, New YOrk: Vintage Books 1995.

GDPR 2016 GDPR (2016) *General Data Protection Regulation, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016*, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1528874672298anduri=CELEX%3A32016R0679

Georgescu-Roegen 1970 Georgescu-Roegen, N. (1970/2011) “The Entropy Law and the Economic Problem”, in Bonaiuti, M. (Ed.), *From Bioeconomics to Degrowth: Georgescu-Roegen's 'New Economics' in Eight Essays*, London: Routledge Studies in Ecological Economics, pp. 49–57.

Georgescu-Roegen 1972 Georgescu-Roegen, N., (1972/2011). “Energy and Economic Myths”, in Bonaiuti, M. (Ed.), From *Bioeconomics to Degrowth: Georgescu-Roegen's 'New Economics' in Eight Essays*, London: Routledge Studies in Ecological Economics, pp. 58–92.

Georgescu-Roegen 1978 Georgescu-Roegen, N., (1978/2011) “Inequality, Limits and Growth From a Bioeconomic Viewpoint”, in Bonaiuti, M. (Ed.), *From Bioeconomics to Degrowth: Georgescu-Roegen's 'New Economics' in Eight Essays*, London: Routledge Studies in Ecological Economics, pp. 103–113 (2011).

Golumbia 2016 Golumbia, D. (2016). “Marxism and Open Access in the Humanities: Turning Academic Labor against Itself”, *Workplace*, 28, 74-114.

Goodman and Flaxman 2017 Goodman, B. and Flaxman, S. (2017) “European Union Regulations on Algorithmic Decision- Making and a ‘Right to Explanation’”, *AI Magazine*, 38(3):50–57, 2017.

Guardian 2018 Guardian (2018) “The Cambridge Analytica Files: A year-long investigation into Facebook, data, and influencing elections in the digital age”, *The Guardian*, https://www.theguardian.com/news/series/cambridge-analytica-files

Gunning 2017 Gunning, D. (2017) *Explainable Artificial Intelligence (XAI): Programme Update*, https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Hammond 2016 Hammond, K. (2016) “5 Unexpected Sources of Bias in Artificial Intelligence”, *Tech Crunch*, https://techcrunch.com/2016/12/10/5-unexpected-sources-of-bias-in-artificial-intelligence/

Harwell 2018 Harwell, D. (2018) “Wanted: The 'perfect babysitter.' Must pass AI scan for respect and attitude”, *The Washington Post*, https://nuzzel.com/sharedstory/11232018/washingtonpost/wanted_the_perfect_babysitter_must_pass_ai_scan_for_respect_and

Hayles 2007 Hayles, N. K. (2007) “Hyper and Deep Attention: The Generational Divide in Cognitive Modes”, *Profession*, 13, 187-199.

Hayles 2010 Hayles, N. K. (2010) “How We Read: Close, Hyper, Machine”, *Ade Bulletin*, Number 150, 62-79.

Hempel and Oppenheim 1988 Hempel and Oppenheim (1988) “Studies in the Logic of Explanation”, in Pitt, J.C. (ed.) *Theories of Explanation*, Oxford University Press.

Hutchins 1995 Hutchins, E. (1995) *Cognition in the wild*. MIT Press.

Irani 2015 Irani, L. (2015) “The cultural work of microwork”, *New Media and Society* 17(5): 720-739.

Jaume-Palasi 2018 Jaume-Palasi, L. (2018) “Blessed by the algorithm: Computer says NO!”, https://media.ccc.de/v/froscon2018-2307-keynote

Jobs 1981 Jobs, S. (1981) “When We Invented the Personal Computer…,” *COMPUTERS and PEOPLE Magazine*, July-August 1981.

Kuang 2017 Kuang, C. (2017) “Can A.I. Be Taught to Explain Itself?”, *The New York Times*, https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

Kuneva 2009 Kuneva, M. (2009) “Keynote Speech”, *Roundtable on Online Data Collection, Targeting and Profiling*, http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm

Lipton 2017 Lipton, Z. I. (2017) “The Mythos of Model Interpretability”, https://arxiv.org/pdf/1606.03490.pdf

Lotka 1925 Lotka, A.J. (1925) *Elements of Physical Biology*. William and Wilkins Company, Baltimore.

Malabou 2019 Malabou, C. (2019) *Morphing Intelligence: From IQ Measurement to Artificial Brains*, Columbia University Press.

Masters and Thiel 2015 Masters,B. and Thiel,P. (2015) *Zero to One: Notes on Start Ups, or How to Build the Future*, Virgin Books.

McNamee 2019 McNamee, R. (2019) *Zucked!*, Penguin Press.

Meek 2019 Meek, A. (2019) “Amazon-owned Ring has reportedly been spying on customer camera feeds”, https://bgr.com/2019/01/10/ring-camera-customer-feeds-accessed-creepy-privacy-violation/

Merleau-Ponty 2007 Merleau-Ponty, M. (2007) “Eye and Mind”, in Toadvine, T. and Lawlor, L. (Eds.) *The Merleau-Ponty Reader*, Northwestern University Press.

Mill 1858 Mill, J. S. (1858) “Of the Explanation of the Laws of Nature”, in *A System of Logic*, New York, Book III, Chapter XII, Section 1.

Noble 2018 Noble, S. U. (2018) *Algorithms of Oppression*, New York University Press.

Palmer 2006 Palmer, M. (2006) “Data is the New Oil”, http://ana.blogs.com/maestros/2006/11/data_is_the_new.html

Pitt 1988 Pitt, J.C. (1988) *Theories of Explanation*, Oxford University Press.

Popper 1972 Popper, K. (1972) *Objective Knowledge: An Evolutionary Approach*, Oxford: University of Oxford Press.

Porter 2019 Porter, J. (2019) “Amazon patents 'surveillance as a service' tech for its delivery drones”, https://www.theverge.com/2019/6/21/18700451/amason-delivery-drone-surveillance-home-security-system-patent-application

Rahwan et al. 2019 Rahwan, I. and Cebrian, M. and Obradovich, N. and Bongard, J. and Bonnefon, JF. and Breazeal, C. and Crandall, J. and A. Christakis, N. and Couzin, I. and Jackson, M. and R. Jennings, N. and Kamar, E. and M. Kloumann, I. and Larochelle, H. and Lazer, D. and McElreath, R. and Mislove, A. and C. Parkes, D. and Pentland, A. and Wellman, M.. (2019). “Machine Behaviour”. *Nature*. 568. 477-486. 10.1038/s41586-019-1138-y.

Renz 2018 Renz, U. (2018) *The Explainability of Experience: Realism and Subjectivity in Spinoza's Theory of the Human Mind*, Oxford University Press.

Rieder and Hofman 2020 Rieder, B. and Hofmann, J. (2020) “Towards platform observability”, *Internet Policy Review*, 9(4). https://doi.org/10.14763/2020.4.1535

Ruben 2016 Ruben, D. H. (2016) *Explaining Explanation*, Routledge.

Sample 2017 Sample, I. (2017) “Computer says no: why making AIs fair, accountable and transparent is crucial”, *The Guardian*, https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial

Samuel 2019 Samuel, S. (2019) “10 things we should all demand from Big Tech right now”, *Vox*, https://www.vox.com/the-highlight/2019/5/22/18273284/ai-algorithmic-bill-of-rights-accountability-transparency-consent-bias

Schüll 2014 Schüll, N. D. (2014) *Addiction by Design: Machine Gambling in Las Vegas*, Princeton University Press.

Scriven 1988 Scriven, M. (1988) “Explanations, Predictions, and Laws”, in Pitt, J.C. (Ed.) *Theories of Explanation*, Oxford University Press

Selbst and Powles 2017 Selbst, A. D. and Powles, J. (2017) “Meaningful Information and the Right to Explanation”. *International Data Privacy Law*, 7(4):233–242.

Sengupta 2012 Sengupta, S. (2012) “Should Personal Data Be Personal?”, *The New York Times*, https://www.nytimes.com/2012/02/05/sunday-review/europe-moves-to-protect-online-privacy.html

Stiegler 2015 Stiegler, B. (2015) “Power, Powerlessness, Thinking, and Future”, *Los Angeles Review of Books*, https://lareviewofbooks.org/article/power-powerlessness-thinking-and-future/#!

Stiegler 2016 Steigler, B. (2016) “The New Conflict of the Faculties and Functions: Quasi-Causality and Serendipity in the Anthropocene.” *Qui Parle: Critical Humanities and Social Sciences*, Volume 26, Number 1, June 2017, pp. 79-99

Stiegler 2018 Stiegler, B. (2018), *The Neganthropocene*, Open Humanities Press, http://openhumanitiespress.org/books/download/Stiegler_2018_The-Neganthropocene.pdf

Toonders 2014 Toonders, Y. (2014) “Data is the New Oil of the Digital Economy”, *Wired*, https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/

USPTO 2019a USPTO (2019a) “Image creation using geo-fence data”, USPTO, http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2andSect2=HITOFFandp=1andu=%2Fnetahtml%2FPTO%2Fsearch-bool.htmlandr=1andf=Gandl=50andco1=ANDandd=PTXTands1=10313638.PN.andOS=PN/10313638andRS=PN/10313638

USPTO 2019b USPTO (2019b) “Generating Composite Facial Images Using Audio/Visual Recording and Communication Devices”, USPTO, https://www.aclunc.org/docs/Amazon_Patent.pdf

Wachter et al. 2017 Wachter, S., Mittelstadt, B., and Floridi L. (2017) *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*. *International Data Privacy Law*, 7(2):76–99, 2017.

Wachter et al. 2018 Wachter, S. and Mittelstadt, B. and Russell, C. (2018). “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR”. *Harvard journal of law and technology*, 31. 841-887.

Zuboff 2019 Zuboff, S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Profile Books.