Datafied Brains and Digital Twins

Lessons From Industry, Caution For Psychiatry

Rainey, Stephen

Citation (APA)
Rainey, S. (2022). Datafied Brains and Digital Twins: Lessons From Industry, Caution For Psychiatry. *Philosophy, Psychiatry and Psychology*, *29*(1), 29-42. https://doi.org/10.1353/ppp.2022.0005

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Datafied Brains and Digital Twins: Lessons From Industry, Caution For Psychiatry

Stephen Rainey

Datafied Brains and Digital Twins: Lessons From Industry, Caution For Psychiatry

Stephen Rainey, Ph.D.*

PPP

Abstract: This paper asks what sorts of ethical caution ought to attach to increasingly data-driven approaches to understanding the brain. This is taken to be an important question especially owing to a likely near future of neuromonitoring and neuromodulation devices with applications in psychiatry. The paper explores this by i) sketching the concept of 'digital twin,' ii) drawing a schematic picture of 'brain datafication' in general, and iii) developing a means of understanding some challenges present in datafication through the lens of digital twins. One central concern arises from the role algorithmic processing of neural recordings plays in terms of neuroscientific objectivity, with knock on effects for psychiatric ethics. Essentially, this is owing to a way in which algorithmic processing in brain data construction appears to be deductive in character, but is in fact based on a particular scheme of inductive inference. The challenges explored urge ethical caution as they concern epistemological gaps in data-centered neuroscientific progress, as well as knock-on effects for psychiatry.

Keywords: Brain data, digital twin, neurorecording, deep learning, algorithms, psychiatry, ethics

IN A CONTEXT of growing technological complexity in recording and processing brain signals, 'brain data' is increasingly available for use in neurology, psychiatry, wellness applications, and recreational activities. Brain data can control devices and ground predictions about brains and behavior. It can steer interventions in brain activity, modulating it to some desired level. But what do brain data represent? Are they a clearer means of seeing the brain's activity, or that of the mind? A data-centric approach to the brain raises distinctive philosophical and ethical challenges engendered by the datafied, neurofunctional account of cognitive and behavioral performance.

Among the variety of techniques for recording and processing the bioelectric activity of the working brain, electroencephalography (EEG) stands out as especially accessible and useful. With this technique, there is large scope for computational processing of recorded signal into varieties of data. EEG appears in clinical contexts, as well as in the consumer market. Though EEG has limitations because of poor spatial signal resolution and the effects of recording through the skull, it is the most widely used technology for recording brain activity. This is especially so in a non-clinical setting. Recordings from the brain promise direct disclosures of neurofunctional constraints upon an agent's cognition, emotions and behavior.

---

* The author reports no conflict of interest.

The disclosures are 'direct' in not requiring any intentional discussion with the agent involved. But brain recordings do not thereby 'speak for themselves.' They must be processed in order to create usable data, and be used according to some rationale.

Especially when combined with contextual data, brain data may reveal sensitive personal information. EEG recordings of electrical activity across large areas of the brain can be processed in order to diagnose diseases, such as epilepsy, but increasingly also to find other distinct brain states including neural correlates of alertness, and attention. These predictions, and a variety of others like them, are made on the basis of brain data. But exactly how predictions relate to the data on which they are based, to the brain, or to the person about whom they are made, is not clear. For instance: what do these data capture? Are we dealing with an image of the brain mediated in data, or with the brain itself as inferred from those data? Or ought we to consider the data as relating more widely to the person about whom we are to make predictions? A useful concept to explore these questions is that of a 'digital twin.'

The term 'digital twin' as it is used here is a concept borrowed from manufacturing and process-management contexts. A digital twin is a virtual version of a physical system (Kritzinger, Karner, Traar, Henjes, & Sihn, 2018). The utility of such a twin is thought of in terms of accurately modeling *in silico* a physical object or system such that information about that object or system could be derived from inserting the digital twin in simulated contexts. In manufacturing, this helps to pre-empt potential points of failure, tolerances, and identify possible efficiencies prior to actual manufacture. This saves money, and time, while promoting good foresight about the likely behavior under different circumstances of the eventual object to be made (Grieves & Vickers, 2017, p. 87).

The idea of a digital twin that serves this pre-manufacturing purpose appears very sensible. Resources need not be used in creating physical prototypes in order to discover problems. Instead, models can be made digitally that are accurate enough simulations of objects under various circumstances. Kritzinger et al. (2018) explore

further distinctions within the digital twin concept that illuminate the data dimension. They distinguish the 'digital model' from 'digital Shadow.' A digital model is a digital representation of an object without any dynamic links to data on that object. The model remains static even if the object changes. Modeling changes in the object requires the production of a new model. The digital shadow by contrast has a one-way flow of data between the state of the physical object and the digital object and so it updates dynamically with changes in the object (Kritzinger et al., 2018, p. 1017). This would most easily be imagined in terms of a digital model that responds to changes in the physical object, but does not affect the physical object when it changes. We could imagine a case of such a digital shadow as a monitoring application for a physical object: perhaps a digital shadow of a physical bridge, from which sensor data is relayed to the shadow in order to maintain real time monitoring of the stresses, movements, capacity, and so on, of that bridge.

A 'digital Twin' here is a further elaboration on this scheme, in that data moves bi-directionally between the digital twin and the physical object. Kritzinger et al. go on to suggest that,

> The digital object might also act as controlling instance of the physical object. There might also be other objects, physical or digital, which induce changes of state in the digital object. A change in state of the physical object directly leads to a change in state of the digital object and vice versa. (Kritzinger et al., 2018, p. 1017)

If we stick with the bridge example, the digital twin for that bridge might include more than monitoring. Were the physical bridge to be close to capacity and stress tolerance, for example, the digital twin might modify traffic signals to decongest the physical bridge. This control could affect the physical bridge through data biases and thresholds built into the digital twin (i.e., without human input). Moreover, a digital twin might affect further digital objects, like maintenance schedules for city council resources, in order to schedule checks following busy periods for the bridge. Digital twins connected with complicated data flows could be used to record detailed 'service records' for the real objects they twin, as well as predicting their likely

future performance (Grieves & Vickers, 2017, p. 95). This kind of picture was the stuff of 'smart cities,' touted since the 1990s (Batty et al., 2012; Caragliu, Bo, & Nijkamp, 2011).

These ideas of digital model, digital shadow, and digital twin are of obvious use in manufacturing and systems of process control, as illustrated with the bridge example and allusion to smart cities. It will be informative now to pivot discussion to the context of brain datafication and use these concepts to investigate some possible upshots from such datafication. Specifically, in what follows it will be asked which kinds of digital counterparts to brains—models, shadows, twins—best capture what is currently emerging in brain datafication. This is of particular relevance in terms of critically analyzing processes within neuroscientific discovery, and in clinical practices in present and future psychiatry that can be expected to draw heavily upon neuroscience.

DATA ON THE BRAIN

Brains are complex systems, but that nonetheless exhibit clear order. The promises of neuroscience in providing clarity on the brain are hoped to provide greater insights to human minds, rationality, behavior, and disease. In 2005, Thomas Insel and Remi Quirion wrote that, "clinical neuroscience must be integrated into the discipline of psychiatry…" and that in the future, "psychiatrists and neurologists may be best considered "clinical neuroscientists"" (2005) With this, they were advocating a movement that sought to promote the concept of psychiatry as *clinically applied neuroscience*. The thought is that using the tools of neuroscience, like functional imaging, and EEG, dysfunctional neural circuits can be identified that underlie mental disorders. Conventional definitions of mental illness have arisen heterogeneously over time, leading to vagueness and 'fuzzy boundaries.' Psychiatry, as compared with other medical sciences, lacks definitive diagnostic approaches, and treatment pathways. The evolution of various editions of *The Diagnostic and Statistical Manual of Mental Disorders* (DSM) and *International Classification of Diseases* (ICD) (see Regier, Kuhl, & Kupfer, 2013) has in large part been an attempt

to remedy this anomalous relationship between psychiatry and medicine in pursuing, 'consistent clinical descriptions of syndromes,' and 'specificity, that is ability to distinguish different types of problems' (Kirmayer & Crafa, 2014).

Using the DSM or ICD as a guide for structuring clinical interviews, psychiatrists can ensure clinical consistency. Nevertheless, psychiatric models based on medical approaches in general encounter challenges. Whereas genetics, for instance, has served to improve approaches to cancer care, the kinds of genetic-environment-context-behavioral relations experienced by any individual make the discovery of genetic bases for psychiatric conditions vastly complex. Social and cultural values, and personal and political histories and contexts, are relevant to characterizing mental illness in ways not seen in physical illness. Moreover, the DSM and ICD editions change, making for clinical practices that are dynamic over time. This prompts some, such as Insel and Quirion, to look to the brain itself to ground a robustly diagnostic and therapeutic psychiatry. In focusing on the brain as a homogenous substrate of mental illness, it is hoped to develop a precision medicine approach, divested of complicated historical heterogeneity.

With mental disorders identified as *brain disorders*, and a raft of neuroscientific approaches at hand, psychiatric disease can be targeted and treated. The vagueness and 'fuzzy boundaries' among the syndromes of the DSM and ICD are replaced with scientifically grounded definitions of brain-based pathology as mental illnesses, in a context termed the 'Research Domain Criteria' (Insel et al., 2010). Given the unique challenges of psychiatry within medical science more generally, this appears promising. But psychiatry as clinically applied neuroscience is not without its own complexity.

As with any science, neuroscience raises epistemological, methodological, and conceptual questions. Epistemologically, neuroscientific knowledge is not clear-cut as tools such like functional imaging relying upon choices among data curation techniques, and complicated statistical modeling (Poldrack, 2006; Vul, Harris, Winkielman, & Pashler, 2009). Methodologically, neuroscientists may consider their own lab-based work quite

removed from human behavior, where they are mainly concerned with close investigation of neural circuits in animal models, for example. Neuroscience in general might be thought of as being curiosity-driven, or driven by wide practical goals as laid down by research funding agencies (Baughman, Farkas, Guzman, & Huerta, 2006; Goering & Klein, 2020). Psychiatry may have something similar in its guiding principles, but it will also have a public health agenda, sensitive to sociopolitical values that neuroscience may not have pressing reasons to consider. This highlights the practical question of how to put the 'applied' in clinically applied neuroscience.

Increasingly, this involves the use of technologies that record and process brain signals, in order to make diagnostic or predictive inferences about the entire range of human behaviors and dispositions, from the neuronal to the societal level. This in turn involves the use of machine learning algorithms, and sophisticated processing of brain signals to produce usable data, with associated serious questions for clinical practice. In terms of clinical practice, technologies might increasingly come to dominate spaces of hitherto interpersonal clinical encounter, emphasizing instead neuroelectrophysiology as a homogenous indicator of mental wellness. More widely, the use of technologies in recording brain signals, processing these signals to produce brain data from which predictions might be made of a person will have implications for how humans see themselves individually and collectively. Reflection upon policy will be required in order to anticipate these anthropological and ethical implications, and to regulate where prudent.

Given the distinctions between models, shadows, and twins, it would seem that in the context of neuroscience and brain data different instances might require different treatments (Table 1). For example, a brain atlas might be expected to be a digital model of a brain. The atlas would respond to changes in knowledge about the brain, but not in real time, and without a direct data link. A neurofeedback device, on the other hand, which provided a user with an account of their own neural activity might be considered as running in terms of a digital shadow. The state of the device would change in step with changes in the brain by means of a data link (e.g., EEG electrodes to monitor electrophysiological activity). A brain–computer interface (BCI)-controlled prosthetic limb, or a neuroprosthetic device, might also be thought of as requiring a digital shadow in this sense. If we consider a neuromodulation device, such as a device that detects the onset of seizure and administers electro-stimulation to the brain in order to prevent fitting, this might operate on the basis of a digital twin. The device state would change in step with the brain, and in certain cases operate so as to change the activity of the brain based on pre-set parameters.

Recordings from the brain promise direct disclosures of neurofunctional constraints upon an agent's cognition, emotions and behavior. The disclosures are 'direct' in not requiring any intentional discussion with the agent involved. But as already mentioned above, brain recordings do not thereby 'speak for themselves' as they must be *processed* and *used* in specific ways. How processing and use proceed inevitably raises questions about how recording technologies converge with technical questions (e.g., the nature of algorithms used to classify signals into kinds), and processing intentions (e.g., as fundamental research, or as clinical decision support).

Brain Datafication

Datafication provides an ordered view of the otherwise hugely complex and inter-relating activity of the brain. We already know from neurophysiology and neuropsychology that the brain is not, in Patricia Churchland's words, a "bramble bush" of chaotic interconnections (1989, p. 99). A great deal has been discovered and codified about how the brain is organized, and about the variety of functional differentiations that can be drawn among the range of signals that can be derived from it. Indeed, this ever-growing knowledge is one motivation for the enterprise of neuroethics. With more sophisticated knowledge of how our brains constrain perception, judgment, memory, and so on, we ought to revisit philosophical accounts of areas like knowledge, ethics, and character. This seems especially so where judgment and

TABLE 1. Summary of Some Instances of Brain Datafication and Their Potential as Digital Models, Shadows, or Twins for the Brain

Instance	Digital Model	Digital Shadow	Digital Twin
Brain atlas	X	X	
Neurofeedback device		X	
Neuroprosthesis		X	
Electroceutical			X

ethics coincide, because if we can understand more about why certain decisions are made in given circumstances, accounting for the brain's activity, then we may need novel accounts of responsibility for human action (cf. Levy, 2007).

The suggestion here is that, just in the same way neuroethics recognizes the potential for neuroscience to prompt revisions to received wisdom in terms of perception and the rest, the datafication of neuroscientific knowledge prompts renewed questions for neuroethics. Datafication of the brain allows us to operationalize neuroscientific insights to the brain in new ways. This is especially the case where algorithmic processing of brain recordings and, perhaps especially, deep learning applications are in play. These might not only be seen to prompt revisiting philosophical questions, but also raise critical points about the practice of neuroscience itself, as well as related practices like psychiatry. Not least, this can be seen in terms of the ways in which data patterns are derived by deep learning that are not explicitly articulated, hence are not available for critical assessment. This is a complex area, which requires some further clarification before returning to the specifics of how analysis in terms of 'digital twins' can clarify. To that end, Figure 1 shows a simplified set of processes from research questions, through brain recording, datafication, and applications. Essential to staking out the areas of interest here, is discussion of how the loosely tagged areas A through E are inter-related.

(A) might be seen as the context of discovery for some enquiry. A research question, informed by neuroscientific data, shapes research design concerning some particular matter. A corpus of knowledge and training in neuroscience and neuropsychology will steer what kinds of questions are settled, worth asking, worth revisiting, or are novel. This theoretical background sets the scene for (B), the context of justification for discoveries, which might be seen as the conjunction of physical requirements for addressing the question under investigation. Depending on the object or phenomenon of investigation, different brain areas will be implicated. Depending on which brain areas are relevant, different types of recording techniques, types of electrodes, and recording sites, will be required in order to derive appropriate signals. Here, for simplicity's sake, (B) is given as a generic EEG experimental paradigm, for the recording of neuroelectrical activity. In this arrangement, electrical signals are recorded from the brain by some means (e.g., external EEG, intracranial grid, intracortical probe). Electrical activity generated by the neurons of the brain are recorded as a research participant goes about a task. This enables researchers to correlate overt behaviors or perceptual cues with the timings of specific electrical activity. The recordings of brain activity might be thought of as 'raw' data in that they contain more information than the specific experimental paradigm requires. It must be processed and sorted, here represented as phase (C), during which techniques are applied to extract relevant signals from the complicated raw signal. In Figure 1, this includes algorithmic processing, not excluding deep learning techniques.

Processing brain recordings using deep learning is less common than might be imagined. One reason that deep learning is not more ubiquitous
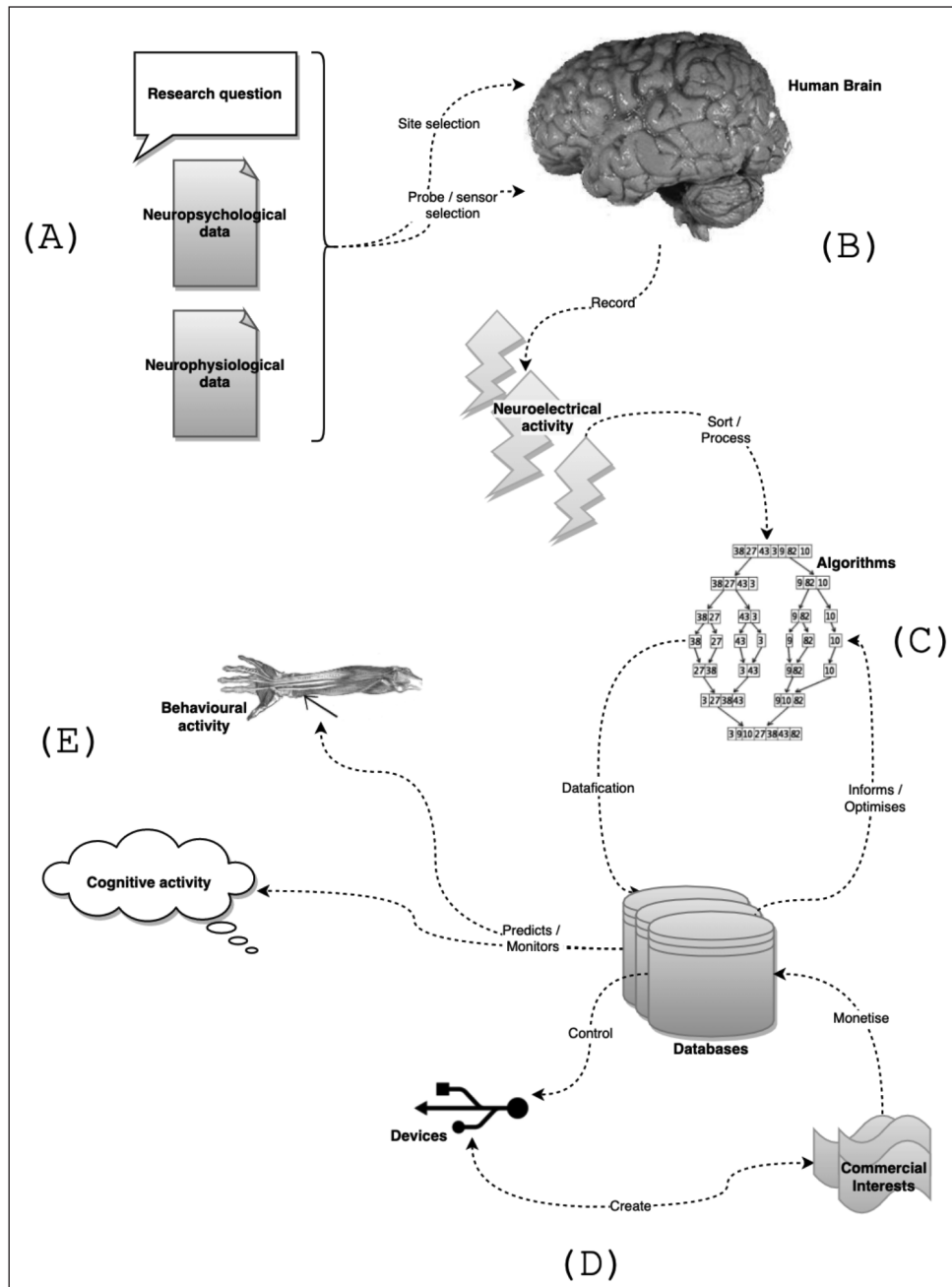
Figure 1. A set of inter-related processes attending cases of brain recording and datafication in general. (A) loosely labels the context discovery—theory and research question formation—itself constrained by wider context including research funding priorities, state-of-the-art science, political environment, etc. (B) gathers together the general experimental context, or context of justification, for the recording of brain signals given some specified research aim. (C) represents the core datafication stage, based in the processing of signals recorded in (B), which is especially salient given the growing technical complexity of neuroscientific experimentation. (D) groups together technological outputs from neuroscientific work, including devices developed on the back of discoveries, databases of neural data, and the commercial interests that can affect them. (E) signifies the phenomena of scientific interest to neuroscience in terms of the results gained, from which further theory and questions can be developed.

in neuroscience already is that it requires a huge amount of data in order to be trained (Lotte et al., 2018). Given that data is being created in large amounts continually, it might be predicted that deep learning will become more prominent, and quickly. Artificial deep neural networks have the potential to learn from raw signal features for classifying the content of recordings according to the requirements of the experimental design. This creates brain data which can, minimally, be used to predict cognitive and behavioral activity, or control devices of various sorts (Bell, Shenoy, Chalodhorn, & Rao, 2008; Birbaumer, 2006; Blain, 2019). In (D) the control of devices is salient, as this can include medical devices, like BCI-driven prostheses, as well as consumer applications, like 'brain typing' devices or 'mind' controlled toys. For this reason, (D) includes an element relating to market forces. These cluster together devices that can represent commercial opportunities, as well as the databases derived from brain signal processing. Dealing in data is a growing area of interest for many technology companies, and brain data seems especially apt for marketability (Kellmeyer, 2018).

Finally, (E) represents a cluster of potential applications for data derived from the brain especially relevant to advancing knowledge of human thought and action. Where data can serve to simplify processes of correlating neural activity and overt behavior or cognitive activity, there is potential for making advances in neuroscience and related fields like psychiatry (Churchland & Sejnowski, 2016). Especially where brain signals are processed by deep learning applications, for instance, the capacity for making strides in these kinds of areas is increased owing to the speed at which algorithms can process huge volumes of data. Moreover, deep learning applications are apt at discovering obscure patterns in ways humans are not. This can lead to a discovery approach in brain data, wherein hypotheses are not produced and tested, but are instead derived from discovered inter-relations among otherwise disparate data (cf, Toga et al., 2015).

With these short characterizations of A through E made, it remains to be seen how these areas do and how they ought to relate. (A) and (B) are unproblematic in themselves, as they constitute a standard picture of how science works. Essentially, theory informs research questions that entail the construction of testing paradigms, from which specific results or falsifications are sought. In a simplified sense, the combination of (A), (B), and (E) represents a schematic scientific whole. The results of (B) might be correlated with the outcomes in (E) and hypotheses verified or falsified on that basis. But the increasing complexity of our understanding of the brain, and the advances made in recording techniques alone make step (C) a practical necessity, and likely prompt a near future of deep learning within it. Yet this is an area of difficulty.

The data derived from processing can be seen to feed into databases, which themselves go on to inform and optimize the algorithms that process brain recordings. This opens the risk that we come to expect interesting characteristics of brain recordings to be *deduced* by algorithms, though they are themselves *inductive engines* through and through. Introducing a pseudo-deductive loop into an inductive system, obscured through the technical complexity of deep learning, could serve to foreshorten experimental sensitivity to novelty, hence to scientific advance. This would also be a risk encountered by an under-examined relation between (C) and (A), wherein pseudo-deductive operations become codified, through constituting the data derived from experiment, in the theories constraining research questions.

The data resulting from neural recordings, sorted and processed in the pursuit of addressing a research question, can become part of the corpus upon which algorithmic processing itself goes on to rely. It may be training data for future neural nets, for instance. It may be used in testing of experimental results from one lab to another, assessing reproducibility of results. The sharing of data, especially in contexts of 'open science' means not only that different research groups can test one another's results, but also that new questions can be asked using existing and growing datasets (Choudhury, Fishman, McGowan, & Juengst, 2014). But recalling the relationships among theory, experience, and curiosity that drives the process of research question formation in (A), there is a potentially fuzzy line between experimental result and settled facts in theory.

There is data curation at work owing to these interconnected steps that is not overt. It is not concealed, in some intentional sense, but it is just not available. It may come about by way of, say, the operations of a deep convolutional neural network whose layers are not open to scrutiny. If the way the relevant patterns are derived from the data with deep learning cannot be explicitly articulated, it cannot be critically assessed as one might critically assess a set of propositions constituting an argument. Instead an evaluative approach is required, that must draw upon a wider set of more general, data-ecological parameters such as the processes illustrated in Figure 1. This means that examining brain data is not a straightforward matter of interpreting data, but a matter of evaluating a complex chain of data curation, selection, processing (i.e., data construction), and then interpreting in a frame of reference itself conditioned somewhat by the same data (the A, B, C, D) complex in the diagram, roughly).

Put differently: there is no transcription from brain activity to recording that can subsequently be read off and interpreted—the very nature of the recording, in using algorithmic steps based (like feature selection, classification, signal transformation, themselves processes derived from and optimized via prior data) means the recording is already conditioned by contingent factors, including information from prior datasets.

Hypotheses behind research questions, based in settled fact and prevailing wisdom, rely on robust experimental data. These data come about through the experimental activities illustrated in (B) recorded in (C), and exhibited through results in (E). With the crypto-inductive operation of algorithmic activity contributing both to experiment and to theory, this suggests a structural problem within the hypothetico-deductive structure of (A), through (B), to (C). This is no reason to suggest the experimental endeavors are flawed, but it does offer reasons to recognize limits. Critically appraising scientific practices and limits is exactly what researchers do (Poldrack, 2011; e.g., Poldrack & Farah, 2015). But in terms of digital twins, this will underwrite a point of concern to follow below.

How (C) and (D) relate is another dimension of risk to the robustness of scientific methods like

that present (C) alone. The presence of commercial interests, in, for example, BCI devices, will produce market forces on device development. As an integral part of device function, this will include the databases derived from experiment. Brain data will thus be a focal point for commercialization, and industrial activity. The relation between this dimension of activity and (E) in which human cognition and behavior are scrutinized is due scrutiny. The obvious focal point for commercial interests and human behavior comes in terms of marketing. Better insights into human behavior and cognition allow for the development of products and services more aligned with peoples' wants and needs. This is not an unalloyed good, however, as the same mechanisms allow for better targeting of consumers and the manipulation desires in order to optimize marketing from the seller's perspective.

On this somewhat dystopian note, the relationship between (D) and (A) cannot be overlooked. With market forces and industrial activity at work in the arena of brain data, the power of money to sway research agendas cannot be ignored. Influencing scientific curiosity toward better understandings of neural phenomena with marketable potential would be a retrograde step. To be sure, such interests do align with fundamental curiosity, and political agendas underlying research funding agency decision-making about grant structures contain economic imperatives. But the possibility for a direct influence of the consumer market upon science agenda-setting is not on the same level as these sociopolitical realities.

Using Brain Data in Psychiatry

Insel and Quirion's picture of psychiatry as clinically applied neuroscience includes a central place for brain imaging as a means of identifying pathologies of the mind. The identification faulty brain circuits show the mechanisms of mental illness, on this view. But the above discussion suggests that brain datafication is an essential part of this endeavor. Data analysis by means of algorithms is not agnostic, or neutral, but comes laden with theory and prior data (Kitchin, 2014). This constitutes the loop described above, which appears to provide deductive inferences from data

to phenomena, but which is in reality a complex inductive process. A digital shadow is an *in silico* counterpart of a reality, dynamically connected via a one-way data stream. In Table 1, a neurofeedback device was considered such a shadow. But now, having considered the nature and role of brain datafication, the status of the device as a digital shadow of the reality might be due revision.

The idea of the digital shadow is passive. It is a representation of a reality on the basis of whose states the real object can be assessed, maybe changed. We can learn about the real object by looking at its digital shadow and decide about what we want to do with it based on the information we gain. But the brain data constituting the shadow are not neutral, so in fact the digital shadow is of a brain *model*. A model is a digital representation without dynamic links to input data, or a set of parameters required to create a physical instance of that which the model modelled. In this case, the neurofeedback device is a model for not the human brain, but the content of the pseudo-deductive loop as described between (B), (C), and (D) in Figure 1. This is a model, because there is no dynamic link to data—there is just the data. The dynamic links are those to the context of justification of (B), and to the optimization of processing via databases, represented in (D). Given the role of brain imaging in psychiatry as clinically applied neuroscience, this has some pressing ethical concerns.

What is uniquely ethically challenging here is the specific kind of inductive loop present where algorithms are used across the board in the sorting of signals from brains, and in their own optimization, while at the same time furnishing databases with material that goes on to frame experimental paradigms and research questions. These in turn lead to further experimentation, which involves those algorithms, further optimization, further database furnishing, and so on. This is the loop, as illustrated in Figure 1, where (C) to (D) is seen as a loop, and in (A) wherein the context of discovery is described as containing neurophysiological/neuropsychological data (i.e., the contents arising from (C) to (D)).

(C) is interposed in a complex of other processes which represent standard scientific practice (C)

appears problematically as it draws from (B) and (E), and can characterize (A), meaning it appears on both sides of the discovery/justification divide. One clear potential here is for feedback loops, as discussed in terms of data by Cathy O'Neil in *Weapons of Math Destruction* (2016). The uniquely ethically problematic implication of a datafied loop within an otherwise clinical approach to mental illness is that the loop appears to use technological means to detect dysfunction, or to *designate* brain activity as disordered. But the loop is classifying data here and now according to past data themselves constructed by that very same loop—it is not patient centered, but *model* centered.

In a paper investigating the use of EEG to distinguish brains of alcoholics from non-alcoholics in a small cohort of 20, Bae, Yoo, Lee, and Kim (2017) use graph theory to model relations among brain areas in their subjects. Their approach is explicitly data-focused, seeking causal relations among data patterns derived from brain activity. They acknowledge the nature of the models developed from brain data as not conventional models of brains, as they explore network connectivity and activity abstracted from brain signal recordings (2017, p. 770). The constructed model is different from a conventional physiological model in that it takes brain data and reconstructs patterns derivable from the processed signals captured from the specific regions of the brains, according to the physical and technical parameters of the recording methods and experimental intentions. In the case of Bae, this is external EEG, but in other cases, it might be intracranial probe, intracortical probe, and so on. The explicit acknowledgement here is that the model gained from the data is different from a normal model in terms of representational content. It represents the properties of networks observable among the data, themselves derived from brain signal recordings.

If imaging is thought to be a core means of identifying faulty neural circuits as the underlying pathology of mental illness, the imaging had better identify clearly the brain and what's gone wrong with it, for example, developmentally, or via injury. With the identification of this pseudo-deductive, crypto-inductive, data-to-data loop, it

is less clear that imaging *per se* is sufficient to do this. It is probably not the hope of many neuroscientists or psychiatrists that imaging would take on this role without a serious helping of critical and expert analysis anyway. Multi-disciplinary, expert teams will often be involved in exploring diagnoses of mental illness or neurodevelopmental disorder and imaging plays a role within a wider diagnostic decision-making process. Insel and Quirion themselves acknowledge, "that psychiatry presents to the rest of medicine a unique blend of interpersonal skills and behavioral expertise that will be increasingly needed in this era of care dominated by technology" (2005, p. 2221). In this, the loop being described does not present a terrible problem. But it is important to recognize it, especially where technology can seem to offer neutral, objective, answers. It ought to be emphasized that what seems a digital shadow of the human brain, via neural recording, is really a digital shadow of a digital brain model. The question ought not to be whether it is a good or bad model, but how the difference in *representational content* might make a difference.

The ethical import of digital twins is of greater ethical concern. A digital twin was thought of as an *in silico* representation of a real object, dynamically connected with a bi-directional data flow. This corresponded with an ability for the twin to alter the state of the real object, and vice versa. An example would be a psychiatric 'electroceutical,' which would monitor brain states and provide electrical or magnetic stimulation in order to correct unwanted states, and induce desired states. This area is certainly expected to expand (Famml, Litt, Tracey, Boyden, E& Slaoui, 2013). This might be conceivable on an analogy with electroceutical devices aimed at preventing severe epileptic fits (Kavehei, Hamilton, Truong, & Nikpour, 2019). Upon detection of a brain state characteristic on the onset of a fit, the electroceutical device delivers stimulation to the brain that arrests the fit. The neuromonitoring dimension of such a device can include artificial intelligence to predict onset of seizure, based on a model of brain function. Similar technology could detect and remedy psychiatric events. Extending this, on the idea of psychiatry as clinically applied neuroscience, neuromonitoring and neuromodulation for treating psychiatric disorders is a likely step.

The epilepsy electroceutical would use an instance of a digital twin. The idea of a psychiatric electroceutical, as an instance of a digital twin, is ethically difficult. As with the 'digital shadow' relating to the model of a brain, and not an actual brain, so too the twin. In the case of epilepsy, a physical process within the brain can be predicted in which electrical activity overwhelms normal function, causing a seizure. The causality being detected in such a case is that of action potentials in neurons, and their cascading out of control. But the activity to be detected in, say, an instance of attention deficit hyperactivity disorder, or of obsessive compulsive disorder, or of depression seems significantly different to this. While it might be argued that the basis for each is neural, so they have that much in common, the step from brain altering to mind altering is significant. If an electroceutical were to be deployed to counteract serious depression, for instance, this would involve a neuromonitoring device recording brain activity. This would be processed, and the data compared with a (statistically) normal set of brain activity. At some threshold, a depressive episode would be detected and a neuromodulatory stimulus produced to arrest it. This would be the dynamic, bi-directional data flow in action.

Because of the loop discussed above, this represents an ethical concern because a psychiatric electroceutical would be based on a digital twin whose correlate in reality was not a human brain, nor the brain of the actual patient/device user, but that was a data model of a brain. In terms of a concrete person, modulating their brain to better approximate a fit with a model might not respect their clinical need, dignity, or rights. It might improve their experience, in that it might arrest a depressive episode, but it would represent an intervention upon their mind according to a model, not a specific diagnostic procedure. This dimension of digital twins is not a problem in their industrial context. In that context,

> Digital Twin Instances could be interrogated for the current and past histories. Irrespective of where their physical counterpart resided in the world, individual instances could be interrogated for their current system state: fuel

amount, throttle settings, geographical location, structure stress, or any other characteristic that was instrumented. Multiple instances of products would provide data that would be correlated for predicting future states. For example, correlating component sensor readings with subsequent failures of that component would result in an alert of possible component failure being generated when that sensor pattern was reported. The aggregate of actual failures could provide Bayesian probabilities for predictive uses. (Grieves & Vickers, 2017, p. 95)

Here, the lens provided by data provides a basis for statistical evaluation of component failures such that reasonable hypotheses about likelihood of future failures. This allows planning for, say, maintenance and repair schedules. But what would the equivalent be in a psychiatric context? Personal, social, political, historical, economic conditions might all serve to prompt specific behaviors and attitudes in a person. All such factors would represent context for their modes of activity in the world at large, including on the neural level. Bayesian probabilities reconstructed from models of brain activity might not adequately capture the detail of a person's behaviors, in that person's context. Yet this is what the digital twin can offer, and as an instance of such a twin, an electroceutical for psychiatry would be limited. The ethical dimension is acute here because, unlike the digital shadow case, the digital twin has that dynamic, bi-directional data flow and can act as a controlling instance of the physical object. The digital twin is active, and so must be regarded cautiously in a context of decision support.

It would be an exciting development in psychiatry to herald new generations of diagnostic and therapeutic devices that could act on the brain. Considering such devices in terms of the brain data on which they would run ought to prompt reflection on the possibilities for such devices. The nature of brain datafication generates models of brain activity upon which subsequent devices operate. In developing devices that would operate as digital shadows, their basis in relating to models ought to be borne carefully in mind, and their role in decision support considered closely. In future developments of devices that would act as digital twins, this relation to models rather than

individual person's brains, ought to be paramount. To be ethically sound, such devices might require novel, and very detailed, consent procedures. Their potential use might require serious restriction. Or, technically, such device might need novel architectures in order that brain data models feature as a part of a wider, more patient-specifically trained device.

Wider Ethical Concerns

This characterization of brain datafication in general highlights areas of ethical concern where digital models, shadows, and twins are considered. Beyond these, specific issues also attend the broader scheme especially in terms of databasing and the role of commercial interests (C, D, and E in Figure 1). These can be seen in terms of the following:

1. Freedom of choice

Where the presence of commercial interests intervenes in how (C), (D), and (E) relate. Unlike the kinds of manipulation we are used to with marketing in general, those of neuromarketing would seek to operate on a sub-choice level, conditioning preferences rather than servicing them, and operating on a range of data unknown to the individual targeted (Dijck, 2014). This behavioral level concern is underpinned by deeper issues as follows.

2. Consent, and autonomy

Williamson (2019) sounds a note of caution about the already developing desire to connect brain data with diagnosis of disorders such as attention deficit hyperactivity disorder, to 'sculpt' educational performance, and to promote desirable brains states through the use of brain datasets and real time brain recording in non-clinical settings like classrooms. How practices like these could be developed ought to be carefully scrutinized as, similarly to 1, they provide an outline for influencing the basis for desire and perceived need prior to consideration.

Especially with respect to digital twins, the theoretical possibility of far-reaching neuro-

manipulation through electrostimulation ought to be considered carefully. This is especially the case where interests beyond those of health, well-being, or fundamental research are at play. While the rationale of science aiming at technology development might well be control of the natural world and the processes it consists in, this rationale could be detrimental when applied to human beings. Control of the human brain, and thereby to some extent the human mind, ought to be something approached with utmost care and caution. Academic research continues on the concept of mental liberty, or the right to mental integrity (cf. Lavazza, 2018). That notwithstanding, the role of technology research into devices like electroceuticals ought not to prompt a crisis for these ongoing ruminations. Until more is understood, and understood more deeply, development of such technologies ought to draw upon wide discussion of what's at stake, rather than narrowly upon what can be done technically.

3. MONETIZATION OF SCIENTIFIC DATA FOR PRIVATE PROFIT

While 1 and 2 refer to the potential for capital to gain instrumental influence over human desire in some respect, there is a symmetrical systemic risk. If science in general has among its aims the better understanding of nature, and this manifests in capitalist systems as a technology industry centered on manufacturing devices to control that better understood nature, this could represent a privatization of that drive. It would be a sort of asset-stripping approach to otherwise public interest science. The assets include not just intellectual property, but research capacity and infrastructure in terms of scientists and their labs.

Whereas fundamental curiosity is thought of as a major driver for scientific research, as already noted, sociopolitical and private aims also play a role especially via grant funding mechanisms. But an increasing distortion of financial incentives in favor of private companies, and especially private technology companies, represent a matter for ethical concern. The interests of 'big tech' closely align with those of fundamental brain research. But the motivation of curiosity is here replaced with instrumental control for market gain. The swaying of fundamental research according to political aims as encoded in public research funding calls (e.g., the Grand Challenges of the European Commission's Framework Programme) has at least some democratic legitimacy. Those of private funding agencies are regulated too, as per charitable giving, or rules regarding trusts, and so on. The operations of private companies are often more opaque in being answerable really only to shareholders, and in terms of the bottom line. With booming power based in vast profits, the capture of fundamental research capacity is a genuine possibility. Especially where data is the lifeblood of brain research, this possibility makes neuro-research particularly vulnerable.

Dealing with these wider issues will require concerted sociopolitical efforts especially regarding research governance. Detailed plans for science–policy interfaces would be helpful, as those modelled on 'co-responsibility,' for example and championed by von Schomberg (2020). Additionally, the social value of research may be in need of detailed discussion where novel forces including big tech come into play (cf, e.g., Ganguli-Mitra, Dove, Laurie, Taylor-Alexander, 2017). Which kinds of intellectual property organizations ought to be considered entitled to may become a salient point of reflection too, given the wider sorts of interests and the stakes discussed here and the intimacy of the human sciences at stake.

CONCLUSIONS

The central concern motivating this discussion was the role played by the algorithmic processing of neural recordings, especially in terms of neuroscientific objectivity and its knock-on effects for psychiatric ethics. The topography of brain datafication as sketched in Figure 1 served to illustrate this by suggesting the ways in which data was implicated in the construction of the data it then went on to use. This highlighted an issue for the ways in which algorithmic processing appears to be deductive in character, but is in fact based on inductive inference from datasets itself curates. This presented unique ethical issues for brain datafication in clinical application. This is

because where datafication might be seen as designating disordered brain function, it might be seen as detecting mental illness as part of a diagnostic strategy. Moreover, where a putative psychiatric electroceutical were at stake, the state toward which a brain would be stimulated would itself be based in data with this loop of construction/curation present. Such a state would not necessarily be clinically justified, but rather endorsed by a somewhat opaque technical system. For human sciences, this is ethically problematic.

There are questions in need of further investigation regarding the role of data in neuroscience, and its clinical application in psychiatry. This is most clearly the case where deep learning, or other algorithmic processing of brain recordings, becomes more central. Using ideas borrowed from manufacturing, some of the context for these questions has been highlighted. It remains to be seen how relationships among digital models, shadows, twins, human brains, cognitive activity, and physical behavior, can be made sense of. Helpful lessons can be drawn from other disciplines in which datafication has already played a major role, including genomics and sociology. There may not be a blanket answer to the questions of relations among data and what is datafied, with specific applications needed case-by-case evaluation. In general, where uncertainties exist and risks can be imagined there is a prima facie case for careful governance. This involves careful description of the uncertainties present in the field, normative analysis in order to map out the kinds of good we want to promote and challenges we wish to avoid, and a translation of the norms produced into actionable policies for the field. Given it is still relatively early days in the datafication of brains and the use of that data in practice, now is a good time to invigorate the kind of wide-ranging, inter-disciplinary, and multi-stakeholder scrutiny that can produce good governance. In doing this, development of critical evaluation can co-evolve with the datafication of relevant practices in neuroscience and its applications.

References

Bae, Y., Yoo, B. W., Lee, J. C., & Kim, H. C. (2017). Automated network analysis to measure brain effective connectivity estimated from EEG data of patients with alcoholism. *Physiological Measurement*, *38*(5), 759–773.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., . . . Portugali, Y. (2012). Smart cities of the future. *European Physical Journal Special Topics*, *214* (1), 481–518.

Baughman, R. W., Farkas, R., Guzman, M., & Huerta, M. F. (2006). The National Institutes of Health Blueprint for Neuroscience Research. *Journal of Neuroscience*, *26* (41), 10329–10331.

Bell, C. J., Shenoy, P., Chalodhorn, R., & Rao, R. P. (2008). Control of a humanoid robot by a noninvasive brain–computer interface in humans. *Journal of Neural Engineering*, *5* (2), 214.

Birbaumer, N. (2006). Breaking the silence: Brain–computer interfaces (BCI) for communication and motor control. *Psychophysiology*, *43* (6), 517–532.

Blain, L. (2019, March 8). Review: Hands-free flight with EEGSmart's mind-controlled UDrone. *New Atlas*. Available at: https://newatlas.com/udrone-mind-controlled-drone-umind-review/58791/

Caragliu, A., Bo, C. D., & Nijkamp, P. (2011). Smart cities in Europe. *Journal of Urban Technology*, *18* (2), 65–82.

Choudhury, S., Fishman, J. R., McGowan, M. L., & Juengst, E. T. (2014). Big data, open science and the brain: Lessons learned from genomics. *Frontiers in Human Neuroscience*, *8*.

Churchland, P. S. (1989). *Neurophilosophy Toward a Unified Science of the Mind Brain*. Cambridge, MA: MIT Press.

Churchland, P. S., & Sejnowski, T. J. (2016). Blending computational and experimental neuroscience. *Nature Reviews. Neuroscience*, *17* (11), 667–668.

Dijck, J. van. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society*, *12* (2), 197–208.

Famm, K., Litt, B., Tracey, K. J., Boyden, E. S., & Slaoui, M. (2013). A jump-start for electroceuticals. *Nature*, *496* (7444), 159–161.

Ganguli-Mitra, A., Dove, E. S., Laurie, G. T., Taylor-Alexander, S. (2017). Reconfiguring social value in health research through the lens of liminality. *Bioethics, 31* (2):87–96.

Goering, S., & Klein, E. (2020). Fostering neuroethics integration with neuroscience in the BRAIN initiative: Comments on the NIH Neuroethics Roadmap. *AJOB Neuroscience*, *11* (3), 184–188.

Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In F.-J. Kahlen, S. Flumerfelt, & A. Alves (eds.), *Transdisciplinary perspectives on complex systems* (pp. 85–113). New York: Springer International Publishing.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, *167* (7), 748–751.

Insel, T. R., & Quirion, R. (2005). Psychiatry as a clinical neuroscience discipline. *JAMA*, *294* (17), 2221–2224.

Kavehei, O., Hamilton, T. J., Truong, N. D., & Nikpour, A. (2019). Opportunities for Electroceuticals in Epilepsy. *Trends in Pharmacological Sciences*, *40* (10), 735–746.

Kellmeyer, P. (2018). Big brain data: On the responsible use of brain data from clinical and consumer-directed neurotechnological devices. *Neuroethics*, *14*, 83–89.

Kirmayer, L. J., & Crafa, D. (2014). What kind of science for psychiatry? *Frontiers in Human Neuroscience*, *8*.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *1* (1), 2053951714528481.

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, *51* (11), 1016–1022.

Lavazza A (2018) Freedom of thought and mental integrity: The moral requirements for any neural prosthesis. *Frontiers in Neuroscience, 12*, 82.

Levy, N. (2007). *Neuroethics*. Cambridge, UK: Cambridge University Press.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., & Yger, F. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering*, *15* (3), 031005.

O'Neill, C. (2016) *Weapons of math destruction, how big data increases inequality and threatens democracy*. London: Penguin.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10* (2), 59–63.

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron*, *72* (5), 692–697.

Poldrack, R. A., & Farah, M. J. (2015). Progress and challenges in probing the human brain. *Nature*, *526* (7573), 371–379.

Regier, D. A., Kuhl, E. A., & Kupfer, D. J. (2013). The DSM-5: Classification and criteria changes. *World Psychiatry*, *12* (2), 92–98.

Toga, A. W., Foster, I., Kesselman, C., Madduri, R., Chard, K., Deutsch, E. W., . . . Hood, L. (2015). Big biomedical data as the key resource for discovery science. *JAMIA*, *22* (6), 1126–1131.

Von Schomberg, R. (2020). In memory of Karl-Otto Apel: The challenge of a universalistic ethics of co-responsibility. *SSRN Electronic Journal*. Available at SSRN: https://ssrn.com/abstract=3515173 or http://dx.doi.org/10.2139/ssrn.3515173

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4* (3), 274–290.

Williamson, B. (2019). Brain data: Scanning, scraping and sculpting the plastic learning brain through neurotechnology. *Postdigital Science and Education*, *1* (1), 65–86.