Philosophy & Technology (2021) 34:1905–1928
https://doi.org/10.1007/s13347-021-00492-1

RESEARCH ARTICLE

Digital Phenotyping: an Epistemic and Methodological Analysis

Simon Coghlan[1] · **Simon D'Alfonso**[1]

Received: 27 April 2021 / Accepted: 4 November 2021 / Published online: 11 November 2021

Abstract

Some claim that digital phenotyping will revolutionize understanding of human psychology and experience and significantly promote human wellbeing. This paper investigates the nature of digital phenotyping in relation to its alleged promise. Unlike most of the literature to date on philosophy and digital phenotyping, which has focused on its ethical aspects, this paper focuses on its epistemic and methodological aspects. The paper advances a tetra-taxonomy involving four scenario types in which knowledge may be acquired from human "digitypes" by digital phenotyping. These scenarios comprise two causal relations and a correlative and constitutive relation that can exist between information generated by digital systems/devices on the one hand and psychological or behavioral phenomena on the other. The paper describes several modes of inference involved in deriving knowledge within these scenarios. After this epistemic mapping, the paper analyzes the possible knowledge potential and limitations of digital phenotyping. It finds that digital phenotyping holds promise of delivering insight into conditions and states as well producing potentially new psychological categories. It also argues that care must be taken that digital phenotyping does not make unwarranted conclusions and is aware of potentially distorting effects in digital sensing and measurement. If digital phenotyping is to truly revolutionize knowledge of human life, it must deliver on a range of fronts, including making accurate forecasts and diagnoses of states and behaviors, providing causal explanations of these phenomena, and revealing important constituents of human conditions, psychology, and experience.

Keywords Digital phenotyping · Philosophy · Psychology · Science · Epistemic · Wellbeing · Ethics

✉ Simon Coghlan
simon.coghlan@unimelb.edu.au

Simon D'Alfonso
dalfonso@unimelb.edu.au

[1] School of Computing & Information Systems, Faculty of Engineering and Information Technology, The University of Melbourne, Victoria, Australia

1 Introduction

Digital phenotyping, some claim and hope, will revolutionize understanding of human wellbeing and experience (Goodday & Friend, 2019; Huckvale et al., 2019; Insel, 2017; Jain et al., 2015; Torous et al., 2016; Wisniewski et al., 2019). In a single day, a huge number of personal data points, far more data than in a neurological image or lab test, can be captured from our interactions with the Internet and ubiquitous portable digital devices like smartphones (Bhugra et al., 2017). The analyzed data, it is said, will reveal our external behaviors and internal states (Torous et al., 2016). Although data from our moment-to-moment activities may appear trivial and mundane (Martinez-Martin et al., 2018), some argue that it could yield valuable knowledge of hidden current and future ill-health.

What are we to make of the claim and hope that digital phenotyping could revolutionize our understanding of human health and experience? In this paper, we adopt a broad philosophical vantage point to investigate the epistemic nature of digital phenotyping in relation to its alleged promise. This epistemological mapping and analysis extends a small but growing body of recent philosophical work on the subject (Birk & Samuel, 2020; Burr & Cristianini, 2019; Burr et al., 2020; Loi, 2019; Mulvenna et al., 2021; Sharon, 2017; Stanghellini & Leoni, 2020; Tekin, 2020). Unlike most of the literature to date on philosophy and digital phenotyping, which has focused on its ethical aspects, this paper focuses on its epistemic and methodological aspects.

Enthusiasm for digital phenotyping often stems from ostensibly game-changing advances in mental health (Fisher & Appelbaum, 2017; Wang et al., 2016). Some believe that digital tools that "seamlessly interact, learn, and grow with users" (Mohr et al., 2017, p. 41) could prove more clinically penetrating than neuroscience (Insel, 2017) and generate new categories of mental illness (Martinez-Martin et al., 2018). According to a Lancet commission report, digital phenotyping may enable the replacement of "phenomenologically derived descriptions" or subjective self-reports with "objective behavioral data" and more "reliable definitions" and thereby improve human health (Bhugra et al., 2017, p. 41).

It may be argued, however, that digital phenotyping's promise goes beyond insights into health to encompass human experience more broadly. Expectations have been raised that digital phenotyping will produce very accurate images of us (Burr & Cristianini, 2019) and might even originate a "new science of behavior" (Huckvale et al., 2019, p. 9) with far-reaching insights into human psychology (Montag et al., 2020). It is also hoped that digital phenotyping will contribute to neurology and mental forensics (Pirelli et al., 2016) and generate insight into beliefs, emotions, values, intelligence (Kosinski et al., 2013), aptitudes, attitudes, and political or sexual orientations (Burr & Cristianini, 2019). Cognate neologisms such as personal informatics, personal sensing (Mohr et al., 2017), psycho-informatics (Yarkoni, 2012), and reality mining (Eagle & Pentland, 2006) hint at this promise of novel and deep knowledge and understanding.

These claims and hopes suggest two valuable possibilities for digital phenotyping. One possibility is a *wellbeing value*, resulting from more informed and effective interventions to improve human lives, including enhancing "P4 medicine"—medicine that is predictive, preventative, participatory, and personalized. The other possibility is a *pure knowledge value*, due to an enriched and deepened understanding of individuals, groups, or humanity itself.[1]

Unsurprisingly, optimism for digital phenotyping has elicited some cautionary responses. These responses include ethical concerns about digital phenotyping, including de-personalizing medicine (Prainsack, 2017), pathologizing behavior (Birk & Samuel, 2020), risking improper consent (Montag et al., 2020), facilitating data insecurity and surveillance (Gooding, 2019; Martinez-Martin et al., 2018), and abusing trust (Kosinski et al., 2013). Other concerns are connected to digital phenotyping's epistemic or knowledge potential and value. This value has been questioned. For example, some argue that digital phenotyping is overhyped (Tekin, 2020) and that much more research and empirical validation is required to gauge its epistemic worth (Carr, 2020). Indeed, some writers caution about the risk of digital phenotyping in reducing complex mental and social phenomena to numbers and associations (Stanghellini & Leoni, 2020) or to digital markers of superficial behavior which could engender an impoverished picture of human experience (Birk & Samuel, 2020).

In this paper, we do not aim to settle *empirical* questions about digital phenotyping's value. Instead, the paper's chief contribution lies in illuminating the methodology and epistemic nature of digital phenotyping. Below, we characterize digital phenotyping (Sect. 2) and present a knowledge taxonomy of digital phenotyping with four distinguishable components (Sect. 3). This tetra-taxonomy helps us to explore possible epistemic implications, advantages, and limitations of digital phenotyping (Sect. 4).

Salient issues identified and contributions made in these sections are (1) the distinction between two types of causal relations, correlation, and constitution in digital phenotyping and their relation to deductive, inductive, and abductive inference; (2) the distinction between diagnostic and prognostic digital phenotyping determinations; (3) the idea of digital phenotyping distortion effects such as observer effects; (4) the contrast between the practical predictive value of digital phenotyping and its ability to produce knowledge that illuminates human life and experience; (5) the potential importance of unsupervised machine learning for generating novel insight; and (6) the epistemic challenge of inferring "subjective" conditions or states and the role of first-person disavowals of those conditions or states.

In these ways, we shed critical light on the idea that digital phenotyping might provide a new or enriched understanding of human wellbeing, experience, and nature. Ultimately, the analysis can inform empirical studies and also be used by those interested in important social and ethical questions raised by this emerging technological approach.

[1] We note that "pure" knowledge could in time also lead to wellbeing benefits.

2 Characterizing Digital Phenotyping

We can start by noting that *pheno* means manifesting; an organism's phenotype is its biological characteristics, which partly arise from its genotype. A "disease phenotype" is a disease's manifestation, whereas the "endotype" is the disease mechanisms. Such an endotype ordinarily cannot be discerned from the phenotype but is hidden and thus requires illuminating through investigation and analysis. Analogously, digital data derived from various digital devices (it is thought) are linked to human conditions and states. Determining exactly what certain digital data signals—what, one might say, the underlying "endotype(s)" might be—requires investigation and analysis; this is where digital phenotyping comes in. A few definitions will now help to set the scene.

Digital sensing is the activity of digital devices in collecting and storing data about individuals. Digital sensing allows "passive, continuous, quantitative, and ecological measurement-based care" (Martinez-Martin et al., 2018) and can often yield much more data, of a finer-grain, than methods such as periodic questionnaires, interviews, tests, and observations. Electronic activities and digital sensors (Saeb et al., 2015) include accelerometers, GPS, Bluetooth, phone calls, gyroscope, barometer, light sensors, microphone, voice and text capture (e.g., on social media), skin conductance, gestural sensing, email use, web browsing, and interaction with screens (e.g., swiping, typing, locking, unlocking) (Birk & Samuel, 2020). Digital devices other than smartphones that can obtain data include wearables, portable EEG and ECGs, biochips, environmental sensors in IoT, PCs, and tablets. (Bhugra et al., 2017).

Authors distinguish, sometimes differently and conflictingly, between *active* and *passive* data. Active data, we shall say, are obtained by direct input from users in response to *prompts* for that data. Questions periodically posed by devices to individuals (e.g., "how are you feeling?")—known as ecological momentary assessments (EMAs)—are one example. Another example is that of users responding to a prompt to produce speech which is then acoustically analyzed (Torous et al., 2016). Passive data includes unprompted data received from sensors. Some of these sensors are deliberately designed to passively collect data related to health (e.g., ECG monitors, blood sugar sensors); others are embedded in ubiquitous devices and used opportunistically (e.g., accelerometers, Bluetooth, GPS tracking).

We may, however, distinguish *interactive* data from active and passive data. Interactions could include swiping, tapping, talking, and web searching and may be content-free (e.g., metadata about interactions) or content-rich (e.g., semantic content in social media interactions) (Martinez-Martin et al., 2018). Interactions may be more effective than passive sensors at capturing information about factors like hand motor function and, especially when semantic content is present and analyzed using natural language processing, about state of mind. We introduce this distinction between active and interactive data partly because it may have further important implications. For example, explicit prompts for information (as opposed to merely interacting with a device) may affect a person's state of mind in special ways, including in ways that distort the inferences we make about them.

The resulting aggregation of unstructured data from digital sensing is variously called the digital exhaust, fingerprint, trace, or footprint. "Tracking" metaphors suggest a means of insight, while "exhaust" metaphors suggest useless raw data. However, while much (or most) of the digital exhaust may be uninformative, it may also contain important information, perhaps involving complex arrangements of diverse data forms. To mirror the particular biological language of this domain (i.e., "phenotyping", its link to "genotype", and so on), we could refer to the aggregate of digital data collected about a person as their *digitype*.[2]

Digital phenotyping is variously defined (Martinez-Martin et al., 2018). One definition is the process of analyzing and making useful sense of digital data, e.g., extracting meaningful patterns from it (Mohr et al., 2017). Analysis could occur with or without artificial intelligence (AI), such as using machine learning to classify and draw inferences about behaviors (e.g., sleep patterns) and psychology (e.g., mood disorders) (Ware et al., 2020). Such analysis might be done for, say, health, education, employment, insurance, and military purposes.

Which precise characterization we give of digital phenotyping depends on what is of primary interest or value to us. When Jain et al. introduced the term "digital phenotype", they were focused on human[3] health interventions (Jain et al., 2015). Others further narrow the focus to psychiatry and clinical psychology (Burr et al., 2020; Insel, 2017). Indeed, digital phenotyping is most associated with measuring or identifying human behaviors and psychological states rather than other characteristics. However, if our main interest is in human experience as such, the definition we give of digital phenotyping may be correspondingly widened.

Philosopher Michele Loi aligns the "digital phenotype" closely with Richard Dawkins's extended phenotype. Just as a wider understanding of an animal's phenotype exceeds biological characteristics to encompass environmental interactions (think Beaver dams), so too might the "digital phenotype" encompass other aspects of the relation between humans and digital sensing technologies, such as human culture and society (Loi, 2019). Dawkins's extended phenotype comprises species-typical behaviors involved in (evolutionary) feedback loops with genetic characteristics of members of that species. This phenotype springs from causal interactions between DNA, behaviors, and environments. For Loi, the human digital phenotype analogously "consists of digital information *produced by humans* and *affecting humans*" (Loi, 2019, p. 157, italics original). Specifically, it concerns causal relationships between digital devices/digital information and cultural and social creation. Digital phenotyping might accordingly be seen as the process of determining a "digital phenotype" in this sense.

The notion that elements of digital sensing and its data could affect our behavior and states is, as we shall see, important. Nonetheless, since we want a characterization of digital phenotyping that allows for the possibility of producing greatest

[2] To continue the metaphor (and as we shall see): As "DNA phenotyping" involves making inferences from the genome, digital phenotyping involves making inferences from the digitype.

[3] We might note here that digital sensing and phenotyping could potentially also provide greater understanding of *non-human animals* and their health and welfare (Buller et al., 2020).

epistemic value, we should think of it as a process that potentially applies to the entire digitype to infer a *range* of potentially important information from it (including, but surpassing, information that might in turn affect individuals or cultures). It is, furthermore, difficult or impossible to predict in advance of exploring digital phenotyping just what elements of the digital "exhaust" (if any) will prove most illuminating of human experience.

Torous et al. define digital phenotyping as the "moment-by-moment quantification of the individual-level human phenotype in-situ using data from smartphones and other personal digital devices" (Torous et al., 2016). Although continuous sensing is a salient feature, digital phenotyping may also involve a digital biopsy (Fisher & Appelbaum, 2017, p. 6). Some authors confine the "digital phenotype" to data obtained from digital sensing, or, more narrowly still, from smartphones (Onnela Lab, 2017). A more expansive conception could perhaps also include digitized information from other sources, such as digitized records of X-rays, doctor visits, and birth certificates. Mohr et al. point to integration of personal sensing data with clinical and genomic databases "to deepen our understanding" of health and wellbeing (Mohr et al., 2017, p. 42). Thinking more widely still, integrating information from epigenetics, microbiomics, neurology, physiology, and psychological observation may well further increase digital phenotyping's epistemic power. This inclusion, however, would expand the definition very significantly beyond how it is currently and typically understood.[4]

Consolidating the above thoughts: the digitype is the aggregate of an individual's digital data from various devices (and *perhaps* digital data from other sources such as medical records) upon which the process of digital phenotyping is made. Roughly, digital phenotyping is the process of attempting to draw epistemically valuable inferences about the states or conditions of individuals or groups from the digitype. In the next section, we explain how digital phenotyping can involve different kinds of logical inferences and several types of relation existing between data and information on the one hand and psychological (or physical) properties on the other. This analysis enables us to give a more complete characterization of digital phenotyping in the conclusion.

3 Tetra-taxonomy of Digital Phenotyping

Data signals provided by devices and sensors can be used to make predictions and determinations about individuals and groups. Such determinations can occur in different ways. Below, we outline a taxonomy of digital phenotyping possibilities that pertain to determining psychological properties, broadly construed. This approach provides a basis for understanding the potential or value of digital phenotyping.

[4] Perhaps we could distinguish *pure* and *impure* forms of digital phenotyping. A pure form would involve analyzing, maybe through AI or machine learning, data obtained only from digital sensing that is used to make evaluations. An impure (or hybrid (Loi, 2019)) form would incorporate (many) other digital data inputs.

The taxonomy identifies the following epistemic scenarios and characteristics: (1) psychological property causes data/information features; (2) information feature(s) causes psychological property; (3) data/information features are correlated with psychological property; and (4) information features (co-)constitute psychological property.

Some preliminary explanations of this taxonomy are needed. Captured *digital data* are associated with and used to derive *information features*. For example, GPS data may generate information about the frequency and duration of someone's visits to various locations. Digital data and information features may be used to infer psychological properties, and vice versa. *Psychological properties*[5] include human traits, moods, behaviors, states, attitudes, orientations, feelings, conditions, and illnesses.

Our taxonomy results from asking: In what ways can psychological properties be connected with information features and their associated digital data? Inferences to psychological properties occur through *analysis*, including by machine learning (ML) pattern analysis of data. *Inferences* as logical types may, as we will see, involve or approximate deduction, induction, and abduction. Deduction comprises inferences that follow logically necessarily from premises; induction involves inferences derived from particular (often numerous) phenomena; and abduction involves inferences based on best explanation of phenomena. In fact, inferences in digital phenotyping are typically *predictive* or *probabilistic* (even when deductive inference is involved—see below) rather than certain. We can now detail the four-fold taxonomy.

3.1 Psychological Property Causally Affects Data/Information Feature(s)

This possible scenario (and the next one) in the taxonomy involves causal relations. In this first possibility, the causal direction of influence runs from psychological property to information feature (Fig. 1). For example, psychological property [insomnia] might cause information feature [sudden midnight phone usage]. Similarly, property [depression] could cause feature [phone battery not being charged]. The digital phenotyping analyst infers the properties from the data or information features. There will, no doubt, generally be more than one possible cause for an information feature. Additional causes may include other psychological features or non-psychological events. The aim in making this type of inference is to collect as much relevant information as possible to support a stronger inference from the data/information feature(s) to the psychological property that caused them.

A key logical inference type in this scenario is *abduction* (Psillos, 2011). Here, we have some information features for which we seek the best explanation from > 1

[5] We would point out that *physical* conditions (e.g., cardiac disease) and conditions may also be inferable from digital data/information. However, our main focus in this paper is on psychological insight (even when that is obtained from e.g. physiological data)—which reflects certain of the bolder claims about human experience made about digital phenotyping and identified in the Introduction above. Note, though, that our definition of psychological properties is broad, e.g., it includes behaviors.

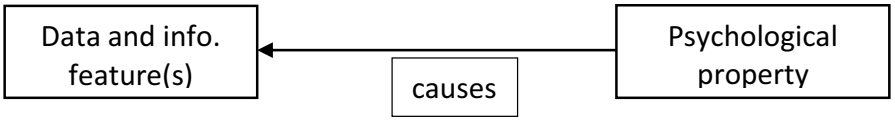

Fig. 1 Psychological property that causally affects data/information feature(s)

possible cause. Seeking the best causal explanation from among alternatives involves basing the (abductive) inference on additional knowledge, whether scientific or commonsense understanding. Aside from abduction, the analyst may sometimes "deductively" infer the absence of a property from the absence of a feature(s). This can occur when the psychological property necessarily or almost always causes certain features or combinations of features such that their absence logically excludes their "diagnosis". Nonetheless, both abductive and "deductive" inferences here are generally probabilistic, since the strength of the inference depends on the presence and extent of the causal connection and the weight and accuracy of the data and the information features associated with the data—and these may not be known with certainty. Furthermore, not all psychological properties are necessarily or always connected with certain information features.

Some illustrations will help explain the probabilistic nature of these determinations. It might be the case that, say, depression often causes a person to charge their phone or socially interact less frequently. But this connection is not a certainty, since there are many other causes of this behavior (e.g., laziness, distraction). Furthermore, the digital device may not be in usage or a required sensor could be turned off. Also, some depressed people do not change their recharge patterns at all. Hence, the absence of phone charging does not guarantee depression's presence, and the presence of phone charging does not guarantee depression's absence.

Nonetheless, sometimes the causal connection running from property to feature may be strong or reliable enough to make inferences with practical certainty. For instance, if a certain psychomotor disorder (almost) always causes a certain pattern of smartphone screen interaction, then the absence of that pattern from the user's screen interaction data would effectively imply (deductively) that the user lacks that disorder.[6] Similarly, the presence of highly distinctive information feature(s) may sometimes allow a practically certain (abductive) inference that a person has a certain corresponding psychological property that causes those distinctive features.

3.2 Information Feature(s) Causally Affect Psychological Property

In this second possibility, the causal direction is reversed (Fig. 2). Here, the information feature or what it represents causes the property. For example, the way someone uses a certain app might have a causal effect on their psychological state. The

[6] This ruling out of "diagnoses" reflects the hypothetico-deductive method of science and clinical medicine.

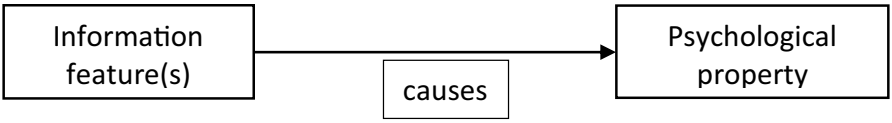

Fig. 2 Information feature(s) that causally affect psychological property

information feature extracted from the data in this example is the manner and extent of that phone usage, and it is this that causes (or causally affects) the psychological property. Another example is answering too many negative EMA questionnaire items (active data), which might cause negative psychological states. In these cases, the phone usage precedes the condition, rather than being an effect of the condition. Thus, data/information features can in theory not only be "manifestations of biologic disease" (Jain et al., 2015) but sometimes causes of them.

The relevant type of inference involved here tends to be *induction*. For example, an inductive inference might be based on observing that there have been a significant number of cases where a certain type of individual (or group) who uses a certain app beyond a threshold subsequently exhibits some condition. Here, we might inductively infer that another individual (or group) who uses the app similarly will develop that condition. The above examples involve active and interactive data, but purely passive data too may enable induction. For example, when information extracted from passive data reveals that many phone users who engage in certain behaviors or combinations of behaviors acquire a psychological property, an inductive inference that the behaviors are causes of the property might be made.

3.3 Correlations Exist Between Information Features and Psychological Properties

In this third possibility, there is the presence of correlation but no direct causation between information feature and psychological property (Fig. 3). Nonetheless, features can still be used to infer properties under non-causal correlation, insofar as the features carry relevant information linked to causes or causal influences (Dretske, 1981). To illustrate: many people who live in a certain area might have a certain condition. Suppose this condition is actually caused by some other factor, say, lead poisoning from the water. Living in the area does not (directly) cause the condition, and having the condition does not cause living in the area. Nonetheless, geolocation data can be used to infer, abductively or inductively, that these individuals have the condition. Sometimes inferences (based on data-property correlations) to causes of properties can theoretically be made. For example, a drastic fall causing amnesia in

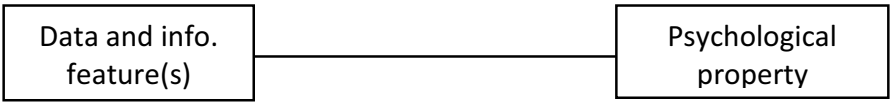

Fig. 3 Correlations between data/information and psychological property

an individual might be inferred (abductively in this case) from accelerometer data. The fall is the common cause of both accelerometer changes and amnesia. Although the amnesia does not cause the accelerometer data or vice versa, the accelerometer-based information correlates with the amnesia and the (abductive) inference to the cause is made.

3.4 Information Features Are Constitutive of Psychological Properties

In this fourth possibility, information features are or represent constituents of the behavior or state itself (Fig. 4). In fact, certain conditions might actually be defined by information features extracted from digital data. As an example, suppose some notion of problematic smartphone usage were conceptualized. This conceptualization might then be measured, and diagnostic determinations made, by quantifying smartphone usage and determining the presence of the condition once certain usage thresholds are exceeded. This might occur for, say, addictive phone-based behaviors (Fisher & Appelbaum, 2017).

Another example involves cognitive or psychomotor tests. Here, certain forms of (active or interactive data) testing could be devised on smartphone devices, such that to have a certain condition is partly defined by the resultant data or the features they represent. In this sense, the captured features constitute or co-constitute the condition. One inference type involved here could be deduction: if a psychological property is exhaustively defined by certain digitally derived features—or else is defined by those digitally-derived features *plus* other features that are known—then, the data/features can be used to deductively infer the psychological property. In contrast, when certain data or features partly but not entirely constitute a psychological property (Fig. 4) and when other constitutive features are not known—such that we cannot *deduce* the property from the digital data or features—the mode of inference may be abduction.

4 Analysis of Digital Phenotyping's Potential

The above taxonomy describes four distinctive scenarios and types of knowledge acquisition using digital phenotyping. These scenarios can bear on an approach's epistemic power for enhancing wellbeing and pure knowledge. But could digital phenotyping take us much beyond what we already have? Since we already have

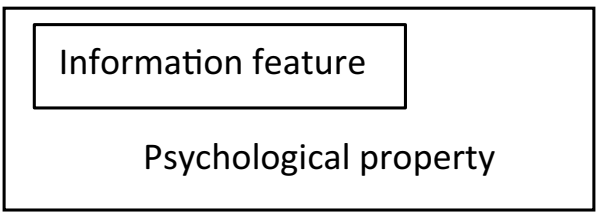

Fig. 4 Information features that are constitutive of psychological property

many illuminating scientific theories, interventions, and tools, we may wonder how digital phenotyping might relate to existing scientific modes of inquiry into human experience and practical responses to it, in ways that are genuinely helpful, illuminating, or even revolutionary. We turn now to this question.

4.1 Non-intrusiveness, Convenience, Efficiency, and Practicality

Minimally, digital phenotyping could support and facilitate mental health monitoring by providing a ubiquitous and economical means of monitoring compared to traditional and resource-limited in-person assessments and psychometric testing (a staple of modern psychology). Given robust and reliable predictive models, where some input features based on digital footprint data predict with sufficient accuracy some psychometric score that an individual would receive, digital phenotyping data and models could serve as a proxy for traditional testing and psychometric scoring, a setup that would be both very convenient and of substantial practical benefit. What exactly the input signal represents or indicates philosophically speaking in such digital phenotyping models evokes questions concerning the philosophical foundations of psychometric testing, itself a rich topic (Maul et al., 2016).

Although it may be *theoretically* possible for massive and continuous data collection and processing to be carried out by humans, it is often not feasible. No human analyst could reliably match natural language processing's (NLP) ability to continuously trawl through and process social media posts. Moreover, processing millions of records from sensors (such as screen touch movements) and discerning patterns from this data is humanly impossible. In addition, digital phenotyping is generally less obtrusive than traditional monitoring since it can be performed in a natural, in situ setting. Monitoring can be seamlessly integrated into an individual's life, in a sense becoming an extension of them.

Inferences to psychological properties can be "diagnostic" or "prognostic". Digital phenotyping may improve the efficiency of both. "Diagnostic" refers to findings that, perhaps in combination with other information, may indicate (Saeb et al., 2015) the occurrence a of property with or without information about, for instance, its severity and intensity—for instance, minute changes in swiping patterns may indicate occurrent subtle, mild, or early stages of mental illness. "Prognostic" findings indicate potential future events. For example, certain data may be used to calculate risks of future behavior (e.g., criminal recidivism), while other data (e.g., concerning mild cognitive impairment) may allow forecasting of responses to treatment for illnesses (e.g., major depression) (Huckvale et al., 2019).[7] Prognostication and diagnosis might be improved in virtue of the input of readily obtainable passive and interactive data, and also active data which involves users' efficiently entering information or self-reports (Mohr et al., 2017, p. 38). This may in theory allow much earlier intervention and better prevention of mental illness (Bhugra et al., 2017).

[7] Note that despite the forward-looking temporal requirement, prognostication can be undertaken retrospectively to make predictions about (then) future states.

We can distinguish two levels of data collection and analysis: individual and general. Digital phenotyping could be highly personalized (Mohr et al., 2017, p. 35). For example, it could be adapted to learn about a person's specific habits, idiosyncrasies, and activities, and to calibrate inferences to those variables, potentially enabling digital precision healthcare. Baselines for specific individuals could be established which would then be used to map deviations from the "norm" for specific individuals (D'Alfonso et al., 2018). Within-individual changes could be tracked, and, where relevant, tailored interventions for that individual could be made (Huckvale et al., 2019). Such tailored interventions, delivered via smartphone apps and informed by sensing data, have come to be known as ecological momentary interventions (Balaskas et al., 2021) or just-in-time adaptive interventions (Nahum-Shani et al., 2018), and they offer another advantage of digital phenotyping based on personal digital devices: the tool which is used to detect something about an individual can also be used to deliver, at opportune times, therapy modules/exercises to address detected health issues.

Digital phenotyping might also be non-personalized. Digital sensing data and other data from perhaps thousands or millions of individuals could be gathered for more basic scientific work. Such Big Data could, for example, be collected in a central repository or "large-scale phenotyping databank" (Huckvale et al., 2019, p. 8) and made accessible to researchers. Apart from examining humans in general, digital phenotyping might be employed on a specific collective. For example, it might help gauge a particular population's mental health by probing social media and other online activity. In this way, digitally gleaned and massed individual information could potentially enable (e.g., via AI processing) new understandings about the general and the collective. In turn, these more general understandings might provide insight into specific individuals.[8] We discuss AI and machine learning (ML) below.

4.2 Accuracy and Unwanted Effects

Although the term "predictive" broadly applies to digital phenotyping evaluations (e.g. Martinez-Martin et al., 2018), predictions can, as we noted, be prognostic or diagnostic. Forecasting (say) future disease or likely progression (as opposed to indicating a current disease) may, as we also said, be more beneficial overall for human wellbeing, assuming that it is accurate and that effective preventative

[8] Population psycho-analytics, which stay as inferences about the population, should be clearly distinguished from using a population to train an ML model that is then applied to individuals. The dominant machine learning paradigm would develop models based on training sets (i.e., sample populations), and then run future individuals through the model to make a prediction about those individuals. The objective is for the model to be sufficiently robust and accurate that it can be applied to subsequent individuals in a confident enough manner, although it will never be 100% accurate. A case study at the individual level could be profitable, but this would be somewhat contrary to the *clinical* goal of digital phenotyping prediction, which aims to make predictions about an individual once data about them is received. Case studies by contrast receive data about an individual and then subsequently analyze and make sense of that data, after the fact. Perhaps the two could be combined, i.e., make inferences about an individual using ML model as data comes in, and subsequently complement those inferences with a case study.

steps are available. Mohr et al. say that "GPS features appear to predict depression many weeks in advance" (Mohr et al., 2017, p. 30). However, because it relates to events which do not yet exist, prognostication may sometimes be less accurate than predicting existing events, since other events may intervene in the course of things to prevent those events from occurring. That said, the reliability of predictions will vary from case to case.

There is little doubt that significantly more empirical work is required to demonstrate digital phenotyping's accuracy and reliability. To demonstrate it, several conditions should be met. First, it is important to ensure that, among the digital exhaust, sufficient representative and good quality data enter into analytic phenotyping systems. Second, AI and ML judgments must, where relevant, be linked to relevant "ground truths" (Merchant et al., 2019). Third, and relatedly, digital phenotypists must be cognizant of various social and scientific understandings of wellness and psychology (Dwyer et al., 2018), because certain conceptions of both pathological and non-pathological states may differ between cultures and/or be scientifically controversial. Fourth, digital phenotyping must accommodate the fact that there may be variation in the expression of various states between different individuals (Delude, 2015). Fifth, AI judgments must be subjected to valid, reproducible, and extensive controlled scientific studies. These can take many years in some cases.

We should also recognize possible forms of observer effect in cases of digital phenotyping, which could cause misleading predictions. A term most commonly associated with physics, the observer effect concerns the disturbance of a system by the act of observing or measuring that system. The "digital phenotyping observer effect" can be explained as follows. A thermometer measures human temperature, yet generally has no causal effect on the temperature itself. Similarly, footprints are the consequences and indicators of walking but have no effect on the act of walking itself. In contrast to these monodirectional indicators and epiphenomena, smartphones and other digital devices, which despite sometimes efficiently and non-intrusively capturing information, may nonetheless sometimes have subtle or even overt effects that interfere with accurate inference-making.

Potential observer effects might occur at different levels. One direct or explicit way in which smartphones could conceivably alter what they measure is in the use of ecological momentary assessments (EMA). For example, notifications requesting active data input from an individual might itself cause psychological changes, such as irritation about being tracked by clinicians or researchers, which influence that data. However, there appear to have been few studies that investigate how EMAs of mental health symptoms may influence the symptoms that they measure—a phenomenon that is also known as assessment reactivity (van Ballegooijen et al., 2016).

There is also the possibility of a Hawthorne effect, recognized in psychology, in which the belief of being watched alters behavior. People who feel observed may sometimes "spin" information about themselves or act more self-consciously (Penney, 2016). Torous et al. report that their digital phenotyping platform "gives only very minimal feedback to the subject in order to avoid behavior change that could result from this feedback" (Torous et al., 2016). Although information distortions

caused by (say) EMAs versus self-consciousness (as just described) are different, they nonetheless roughly fall under the umbrella of "observer effects".[9]

Aside from any digital phenotyping observer effect distortions that interfere with diagnosis or prognostication, digital sensing of an individual could itself also partly cause conditions or states to emerge in that same individual (Mulvenna et al., 2021). This could occur in both good and bad ways. Some note that digital phenotyping might improve a person's wellbeing by increasing emotional self-awareness and control over their psychological health (Chandrashekar, 2018; Simblett et al., 2019). In contrast, Burr et al. refer to "epidemiological inflation", whereby a symptom tracker app, for example, may exacerbate the very symptoms it is designed to monitor and/or remediate (Burr et al., 2020). Consider how a sleep tracking app may worsen sleep-related anxiety in an insomniac.

Another example of an undesirable causal effect is internet or smartphone addiction resulting from over-engaging with devices (Harris et al., 2020). Moreover, *erroneous* information from digital phenotyping that is delivered to the subject might precipitate the very condition it "falsely" predicts/indicates. Consider the self-fulfilling prophecy of being told, and then believing, that you have or will develop depression or anxiety, which you might not otherwise have developed.[10] Conversely, digital prediction of a future or developing state or condition could enable an individual to take remedial or preventative steps and hence falsify the forecast, to the individual's advantage. There is thus a range of distortion effects and various deleterious and beneficial implications of digital phenotyping to be considered.

4.3 Causes vs. Correlations

Perhaps digital phenotyping may sometimes accurately (e.g., with high sensitivity and specificity) discern occurrent or future states, behaviors, conditions, and responses to interventions. This discernment need not only occur via inferences based on *causes* (Saeb et al., 2015) of such psychological properties. Rather, it may also occur by discerning correlations (type 3 in the tetra-taxonomy). In other words, captured information may simply represent non-causal proxies for actual causes of states and behaviors. Some authors claim that digital phenotyping only determines correlations, not causes (e.g. Stanghellini & Leoni, 2020). However, if digital phenotyping could *also* identify causes of important properties (see type 2 in the tetra-taxonomy), then in an important sense, it would have more epistemic power. This is for least two reasons.

First, identifying causes of properties (prima facie) promises more effective interventions to improve wellbeing. Although identifying reliable correlations can be

[9] Besides a digital phenotyping observer effect, we could also note the "platform effect" (Loi, 2019), whereby features of an interface (e.g. Google's auto-completion function for search queries), rather than the sense of being observed, distorts the information obtained digitally (Malik & Pfeffer, 2000).

[10] Clearly, the delivery of such information to an individual should be done cautiously or phrased in such a way that it does not instill certainty of impending mental ill-health.

diagnostically and prognostically powerful, intervening to influence proxies of negative conditions may have no or minimal positive impact.[11] By contrast, intervening at the level of direct causes may be more effective. Second, identifying causes of traits, behaviors, or conditions arguably often yields deeper insight into human experience. Causal understandings generally provide greater explanatory power than correlation-based understandings. Compare, for example, being able to accurately predict anxiety via subtle proxy indicators of feelings and beliefs, with the insight into causal feedback loops between belief and feeling/behavior that is provided by cognitive-behavioral science. The former provides (very) useful knowledge, but the latter deepens our knowledge and understanding of the *nature* of anxiety.

There are both individual and general types of explanation here. Causal understanding of an individual's condition (e.g., knowing the causes of severe anxiety for X) is important for promoting individual wellbeing. However, generalizable causal knowledge gives not just insight into an individual's situation, but also into important human problems and states and even, when the insight is profound, into human nature. Such general knowledge arises from data about individuals, but it relies upon amassing such data and making sense of it. In this regard, the technological possibilities in digital phenotyping may take on additional value. Huckvale et al. write that:

> the causal and temporal relationships between cognitive dysfunction and affective symptoms is itself an open research question that is amenable to exploration using phenotyping. The potential feasibility of discreet, continuous digital phenotyping in young adults offers a route to address the specific call for longitudinal studies that can assess if and how cognitive symptoms precede the peak onset of depression in the mid-late twenties (Huckvale et al., 2019, p. 4).

The reference here to longitudinal studies reminds us that properly determining the causes of physical or psychological phenomena must follow a legitimate scientific method that involves, for example, control groups, randomization, and replicability. Thus, there are differences in how we should treat digital phenotyping as a means of determining causes when compared to its use in determining correlations that make accurate predictions. The scientific method for finding causes has different requirements to, say, a statistical or machine learning method of finding associations.

Yarkoni and Westfall claim that a psychological understanding of humans consists in both explaining (understanding causal relations) and predicting (forecasting) human behavior (Yarkoni & Westfall, 2017). Psychological science, they argue, has wrongly neglected prediction in favor of explanation. They point out that best explanation is not necessarily the best predictor of behavior. We could add to their point the idea that improving prediction may also advance scientific causal explanations.

We can agree with Yarkoni and Westfall that both prediction (in the sense of forecasting) and causal explanation are foundational to psychology (Yarkoni & Westfall,

[11] Note the possibility that some correlating features will turn out to be *distal* causes. For example, a demographic fact like living in a suburb that happens to contain a lead-polluted lake can be a distal (or indirect) cause of lead-poisoning which is proximately (or directly) caused by the lake and the lead in it.

2017). However, it is important to understand that whether explanation or prediction (forecasting) provides the deepest psychological insight is partly a *value* question. The answer can vary with *what* is being causally explained (e.g., why I'm scared of spiders vs. the basis of morality) or predicted (e.g., whether I can overcome my fear of spiders vs. whether humans will eradicate selfishness). Some kinds of knowledge are more valuable or more momentous than others. Pure science, we can remember, concerns itself with understanding the basic or ultimate causes of phenomena at different levels of organization (e.g., atomic vs. chemical vs. biological vs. psychological). Illuminating reality in this way, and not simply making predications without understanding of underlying causal relations, is foundational to science's value.

To be sure, applied sciences, like medicine or conservation biology, are often interested in accurate prediction by any means for the purposes of intervention. And, as we indicated, prediction is a component of the verification of causal theories in the pure sciences. Furthermore, and conversely, causal explanation in the applied sciences is a means to achieve better prediction and successful interventions. Prediction is indeed a worthy scientific and practical goal. However, even the applied sciences aim at explanation or causal understanding for its own sake: such understanding is a key part of their epistemic value. Accordingly, digital phenotyping would fall short of generating a "new science of behavior" if it proved to have significant predictive but little explanatory power—or at least, it would then lack a key characteristic belonging to both pure and applied sciences. We return to the issue of epistemic value soon.

4.4 Constitutive Elements

We have acknowledged that digital phenotyping *might* produce significant improvements in both forecasting and diagnosing psychological properties. For example, digital sensing might detect very subtle behavioral changes or patterns predictive of dementia well before the onset of any of its intrinsically troubling features. These may be changes or patterns which are intrinsically trivial or that appear mundane when viewed independently of what they foreshadow. Such predictive power, we noted, does not necessarily give insight into the nature of, say, dementia. However, we also said that digital phenotyping might help identify some causes of human states and behaviors. Individual and especially general (e.g., Big Data-generated) knowledge of *causes* of the above types has explanatory power that could potentially provide insight into human nature. So too could knowledge and understanding of the *constitutive* elements of conditions/psychology (type 4 in the taxonomy). Both causal and constitutive elements are particularly important in understanding human behavior and psychology.

Causal and constitutive factors can denote the same or different items. Thus, SARS-CoV-2 is both constitutive and causative of COVID-19, whereas (apparently) an historical population of horseshoe bats is causative but not constitutive of that illness. In the psycho-ethical domain, subconscious contempt and fear may partly constitute a person's attitude towards certain individuals or groups, as well as causing that person to act differently towards them. Although some causes of states can also

be constituents of those states, not all causes are constitutive. Temporally remote causes are an example. Furthermore, some constituents of psychological properties need not be causes of those properties. For example, paranoia is constitutive of forms of schizophrenia but does not cause them.

To highlight the possible value of non-causal constitutive determinations, consider a seminal (non-digital) example from psychological science. Despite its widely accepted scientific shortcomings, many psychologists and philosophers recognize that Freud's theory of the unconscious delivers profound insight into the human mind. This it does both by identifying potential causes of psychological states and behaviors and by identifying constituents of them. In fact, those constitutive elements are arguably more important for human knowledge than the causal elements in Freudian theory. That is partly because certain key causal aspects of that theory (e.g., Oedipal and psychoanalytic elements) are scientifically problematic. Nonetheless, constitutive Freudian elements can reconfigure the way we understand ourselves—not just as individuals, but as a psychologically distinctive class of animals (Stevenson et al., 2013).

So, the theory's value is arguably found less in the (scientifically contested) proposition that below consciousness features of the mind can by some mechanism cause, for example, central kinds of mental disorder, and more in the claim that submerged psychological dimensions illuminate and explain vital aspects of human experience. Certainly, these elusive mind features may be considered important partly because they help explain certain causal effects, such as damaging behaviors. Yet knowledge of pervasive unconscious phenomena can be revelatory of human nature whether or not those phenomena imply causal mechanisms amenable to scientific investigation.

Could digital phenotyping provide insight into the constituents of psychological properties? Certainly, there are some conditions or states that are necessarily tied closely to elements of the digitype. Unusually prolonged time spent on smartphones, measured by those same devices, may be constitutive of smartphone addiction. However, we may once again ask whether digital sensing essentially records data connected to a range of properties the constituents of which are knowable independent of digital sensing. The possibility that close (if practically difficult) observation of people would yield the same information might temper the hope that digital phenotyping augurs an epistemic advance or revolution. Conceivably, however, the approach may identify very subtle patterns, correlations, causes, or constitutive factors that are otherwise very hard or impossible to detect. In this sense, digital phenotyping might provide a process that differs from current approaches not just in *degree*, but also in *kind*.

For example, some previously unknown, subtle, complex features might be discerned from, say, various patterns of interactions with smartphones, perhaps combined with other passive, active, or interactive data. Moreover, it is at least possible that machine learning, especially of the *unsupervised* kind, could generate new diagnostic or psychological categories. Supervised machine learning involves a labelled training set, where for each element of the set, a collection of input features (information points extracted from a digital footprint) is paired with some labelled outcome (Russell & Norvig, 2021). These labelled outcomes can be categorical (e.g., individual has depression or not) or continuous (e.g., individual received a score

of X on a certain psychometric measure for depression). Based on this initial training data, a supervised machine learning model is then constructed that can predict further instances of this outcome (dependent variable) based on the input features (independent variables) alone. Thus, supervised ML-based digital phenotyping is premised on pre-established psychological categories or measures.

Unsupervised machine learning, by contrast, receives input features but no labelled outputs to associate them with. Rather, unsupervised techniques determine similarities between the input elements (which consist of features) and then separate these input elements into separate groups based on this similarity determination. For example, suppose that the input consisted of a set of pairs (x, y), with x and y being numbers between 1 and 100 quantifying two properties. An unsupervised algorithm such as *K-means clustering* (Mulvenna et al., 2021) might receive this set and determine that there are three clusters in the underlying structure of the data: pairs with low x and y values, pairs with low x and high y values, and pairs with high x and low y values. This is all done with numerically rigorous method.

Now, suppose that a set of (x, y) pairs were inputted into such an algorithm, with each pair providing certain quantitative information for an individual with depression. If three such salient clusters appeared, then in a sense this generates three different depression subgroups. These cluster outputs would need to be interpreted by mental health experts to determine their clinical utility or relevance, but suppose, for example, that it was found that among this group of depressed individuals, some benefited from medication A, some from medication B, and some from medication C. What is it about each individual that makes them benefit from a particular medication? Now suppose that those who benefited from medication A fell into the first cluster, those who benefited from medication B fell into the second cluster, and those who benefited from medication C fell into the third cluster. Given these results, we now have an informed method, based on clusters generated by unsupervised machine learning, to prescribe medication to new depressed individuals based on the clustering classification they fall under. More generally, such unsupervised methods could potentially lead to the establishment of subtypes of depression and, beyond this, of other psychological categories.

Another important point to make about the constitutive elements of states, behaviors, or conditions concerns first-person versus third-person perspectives. Imagine that digital phenotyping strongly indicates (predicts) that person P now has psychological property χ. This determination, let us imagine, is based on analysis of the combination of digital data from the individual and Big Data from many individuals. We are supposing here that this data constitutes a third-person perspective: it does not involve self-reports or assertions by the subject. What then should we say if P denies or repudiates χ as determined by that instance of digital phenotyping?

One possibility is that the digital phenotyping verdict is correct, and person P is mistaken (Fisher & Appelbaum, 2017). For example, P, though sincere in her repudiation, may be in denial. Further, it is well recognized that the accuracy of first-person perspectives can be tainted by a range of factors including forgetting, lying, misremembering, misinterpreting, providing incomplete information, withholding, feeling guilt or shame, and impression management (Fisher & Appelbaum, 2017).

Some individuals may deliberately confound digital investigation to avoid a diagnosis or assessment.

Another possibility under this scenario, however, is that diagnosis χ as determined by digital phenotyping under the above conditions, even when based on extensive information, is false. Consider two ways in which this could occur. The first way is that χ is false independent of P's perspective on the matter. Suppose digital phenotyping, perhaps based on drinking, eating, toileting, ambulating, and sleeping patterns, determines that P has diabetes. P's denial that they have diabetes has no weight in the assessment that the digital phenotyping falsely diagnosed diabetes. Instead, the digital phenotyping diagnosis is proved wrong by other "objective" data, such as blood tests. This can also be true in mental illness: a psychiatric patient's self-report may be at odds with reality. As Thomas Insel emphasizes:

> Many patients realize, just as we learned from thermometry, that they cannot trust their subjective experience. Just as people with diabetes learn that every moment of lethargy is not hypoglycemia and people with hypertension learn that every headache does not mean elevated blood pressure, people with mental disorders are asking for something more objective to help them to manage their emotional states, distinguishing joy from the emergence of mania and disappointment from a relapse of depression (Insel, 2019, p. vi).

However, the idea that we might bypass or downplay subjective experience and self-reporting, though important, must be significantly qualified. For the second of the two ways that digital phenotyping may get it wrong is that the determination χ is false wholly or partly in virtue of P's own perspective—namely, in this case, their repudiation of χ. We might suggest that this second way applies to certain psychological states or conditions, such as despair, sadness, depression, happiness, and love, at least to some extent and in some cases. Like other third-person assessments, digital phenotyping (when it makes such assessments) may provide important information about a state or condition. But it does not follow that self-reports or avowals lack a constitutive role, sometimes a vital one, in such assessments. Consider the following argument from Birk and Samuel (see also Stanghellini & Leoni, 2020) that using digital phenotyping to infer mental qualities dangerously courts reductivism of essentially social phenomena and could generate faulty conclusions:

> …loneliness is not an *objectively measured quality* but rather one's *self-interpreted social situation*…Loneliness is rarely static, and may indeed fluctuate from situation to situation. Thus, we would argue, there are some states that cannot be inferred purely from passively sensed data (Birk & Samuel, 2020, p. 8, italics original).

This passage is illuminating, but it requires an amendment. Loneliness is not necessarily entirely reducible to "one's self-interpreted social situation." Loneliness may be constituted both by self-interpretation or avowals of loneliness *and* by other conditions and features. In addition, inductive or abductive

inferences can be accurate even though their grounds are only proxies rather than constituents (such as those avowals) of the inferred property. Even if some featural grounds of the inference are weak or misleading grounds when considered individually—e.g. because they involve measuring smartphone location rather than the users' location (Birk & Samuel, 2020, p. 8) in some assessment of, say, whether an individual is lonely—the collected mass of diverse data (e.g., websites visited, patterns of speech, location data) and their appropriate analysis (e.g., by powerful machine learning techniques) may ameliorate the tendency of individual features to result in misleading inferences.

Nonetheless, it is vital to understand that at least some states or conditions have the peculiar characteristic of being *partly constituted* by first-person avowals or disavowals. In this way, certain first-person (dis)avowals carry a peculiar epistemic weight for observers even though they are defeasible (as in the case of denial). Yet if a person is not in denial or in some other special state, and if other important conditions or features obtain, then what that person says about certain mental states of theirs can be *authoritative* as to the presence, absence, and nature of those states. A person's values, attitudes, appetites, and sexual and political orientations are of this type, but so too may be psychological states like loneliness, depression, hopelessness, and despair.

Hence, it is essential that digital phenotyping sometimes takes into account first-person beliefs and assertions in order to accurately assess certain psychological properties. Of course, when a subject falsely asserts or self-reports due to denial and insincerity, a conflicting conclusion produced by digital phenotyping may enlighten observers and sometimes even the (highly attentive) subject herself (Metz, 2018). But in the absence of those special qualifying conditions, certain conflicting self-reports and assertions must be taken seriously, since they may legitimately unsettle even apparently solid digital determinations. Hence, the hopeful notion that "phenomenologically derived descriptions" and subjective self-reports may be replaced by "objective behavioral data" in digital phenotyping (Bhugra et al., 2017, p. 41) is an oversimplification.

Nevertheless, having marked those cautions related to constitutive features of psychological properties, we can now point out that digital phenotyping might incorporate those first-person features into its determinations. And, as we discussed, digital phenotyping based on sophisticated analytic techniques (especially machine learning on Big Data) may potentially deliver insights into the nature (causes and constituents) of known or new psychological properties. Whether it comes good on this potential remains to be seen.

5 Conclusion

In this paper, we responded to growing claims and hopes that digital phenotyping will significantly advance or even revolutionize human wellbeing and knowledge of human life, especially regarding human behavior and psychology. In our response, we characterized digital phenotyping as the process of drawing inferences from the digitype, or the aggregate of an individual's active, passive, and interactive digital data obtained from various devices. Our analysis enables a more complete definition of digital phenotyping to now be given: Digital phenotyping is the process of

attempting to draw from the digitype(s) epistemically and practically valuable inferences about states or conditions, often psychological, of individuals or groups; it can involve deductive, inductive, and/or abductive inferences that occur under conditions variously involving correlative, constitutive, and (two kinds of) causal relations that exist between digital data and information on the one hand and psychological (or physical) properties on the other.

The paper's aim was *not* to settle empirical questions about digital phenotyping's value—for that involves empirical rather than philosophical methodologies—but rather to reveal the modes of epistemic insight into wellbeing and psychological and physical phenomena that digital phenotyping may or may not provide. Our tetra-taxonomy and the subsequent analysis it gave rise to found that digital phenotyping has at least the potential to deliver not only practical, efficient, convenient, and non-intrusive information acquisition, but possibly also valuable insights related to both diagnosis and prognostication and to the discernment of causes, correlations, and constituents of states and conditions.

At the same time, our analysis pointed to risks and qualifications that must be made regarding digital phenotyping, including the possibility of digital phenotyping distortion effects like the observer effect, the importance of empirical validation, the need to differentiate causes, constituents, and correlations, and the way in which conflicting first-person reports can affect (without undermining in every instance) the accuracy of some psychological determinations and predictions.

In the end, only empirical studies can show whether the high promise of significantly improved human wellbeing and of major advances in knowledge are realized by digital phenotyping. Even so, empirical studies and further analyses need to grasp the philosophical nature, potential, and problems of this method. Although it is beyond the scope of this paper to explore, the above understanding of digital phenotyping should be useful in addressing certain social and ethical questions. Ethical analyses, for example, should appreciate the epistemic potential of digital phenotyping to increase behavioral insight and even to generate new psychological categories. By the same token, moral investigation should recognize, among other things, the possibility of distorting digital phenotyping observer effects, overconfidence in digital phenotyping determinations, and confusion about the nature of associated inferences. We hope that this epistemic and methodological analysis of digital phenotyping helps stimulate and guide further empirical, philosophical, and socio-ethical study on this novel emerging use of technology.

Acknowledgements We thank an anonymous reviewer for their very helpful feedback and advice. The authors declare no competing interests.

References

Balaskas, A., Schueller, S. M., Cox, A. L., & Doherty, G. (2021). Ecological momentary interventions for mental health: A scoping review. *PLoS ONE, 16*(3), e0248152. https://doi.org/10.1371/journal.pone.0248152

Bhugra, D., Tasman, A., Pathare, S., Priebe, S., Smith, S., Torous, J., Arbuckle, M. R., Langford, A., Alarcón, R. D., Chiu, H. F. K., First, M. B., Kay, J., Sunkel, C., Thapar, A., Udomratn, P., Baingana, F. K., Kestel, D., Ng, R. M. K., Patel, A., … Ventriglio, A. (2017). The WPA-Lancet Psychiatry

Commission on the Future of Psychiatry. *The Lancet Psychiatry*, *4*(10), 775–818. https://doi.org/10.1016/S2215-0366(17)30333-4

Birk, R. H., & Samuel, G. (2020). Can digital data diagnose mental health problems? A sociological exploration of 'digital phenotyping.' *Sociology of Health & Illness*, *42*(8). https://doi.org/10.1111/1467-9566.13175

Buller, H., Blokhuis, H., Lokhorst, K., Silberberg, M., & Veissier, I. (2020). Animal welfare management in a digital world. *Animals, 10*(10), 1779. https://doi.org/10.3390/ani10101779

Burr, C., & Cristianini, N. (2019). Can machines read our minds? *Minds and Machines, 29*(3), 461–494. https://doi.org/10.1007/s11023-019-09497-4

Burr, C., Morley, J., Taddeo, M., & Floridi, L. (2020). Digital psychiatry: Risks and opportunities for public health and wellbeing. *IEEE Transactions on Technology and Society, 1*(1), 21–33. https://doi.org/10.1109/TTS.2020.2977059

Carr, S. (2020). 'AI gone mental': Engagement and ethics in data-driven technology for mental health. *Journal of Mental Health, 29*(2), 125–130. https://doi.org/10.1080/09638237.2020.1714011

Chandrashekar, P. (2018). Do mental health mobile apps work: Evidence and recommendations for designing high-efficacy mental health mobile apps. *MHealth*, *4*(3), Article 3. https://mhealth.amegroups.com/article/view/18848

D'Alfonso, S., Carpenter, N., & Alvarez-Jimenez, M. (2018). Making the MOST out of smartphone opportunities for mental health. *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, 577–581.

Delude, C. M. (2015). Deep phenotyping: The details of disease. *Nature, 527*(7576), S14–S15. https://doi.org/10.1038/527S14a

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. MIT Press.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Eagle, N., & Pentland, A. S. (2006). Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing, 10*(4), 255–268.

Fisher, C. E., & Appelbaum, P. S. (2017). Beyond googling: The ethics of using patients' electronic footprints in psychiatric practice. *Harvard Review of Psychiatry*, 1. https://doi.org/10.1097/HRP.0000000000000145

Goodday, S. M., & Friend, S. (2019). Unlocking stress and forecasting its consequences with digital technology. *Npj Digital Medicine, 2*(1), 1–5.

Gooding, P. (2019). Mapping the rise of digital mental health technologies: Emerging issues for law and society. *International Journal of Law and Psychiatry, 67*, 101498. https://doi.org/10.1016/j.ijlp.2019.101498

Harris, B., Regan, T., Schueler, J., & Fields, S. A. (2020). Problematic mobile phone and smartphone use scales: A systematic review. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00672

Huckvale, K., Venkatesh, S., & Christensen, H. (2019). Toward clinical digital phenotyping: A timely opportunity to consider purpose, quality, and safety. *Npj Digital Medicine, 2*(1), 1–11. https://doi.org/10.1038/s41746-019-0166-1

Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behavior. *JAMA, 318*(13), 1215–1216. https://doi.org/10.1001/jama.2017.11295

Insel, T. R. (2019). Foreword for Digital Phenotyping and Mobile Sensing. In H. Baumeister & C. Montag (Eds.), *Digital phenotyping and mobile sensing: New developments in psychoinformatics.* (UniM INTERNET resource). Springer; UNIVERSITY OF MELBOURNE's Catalogue. https://doi.org/10.1007/978-3-030-31620-4

Jain, S. H., Powers, B. W., Hawkins, J. B., & Brownstein, J. S. (2015). The digital phenotype. *Nature Biotechnology, 33*(5), 462–463. https://doi.org/10.1038/nbt.3223

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Loi, M. (2019). The digital phenotype: A philosophical and ethical exploration. *Philosophy & Technology, 32*(1), 155–171. https://doi.org/10.1007/s13347-018-0319-1

Malik, M. M., & Pfeffer, J. (2000). Identifying platform effects in social media data. *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, 241–249.

Martinez-Martin, N., Insel, T. R., Dagum, P., Greely, H. T., & Cho, M. K. (2018). Data mining for health: Staking out the ethical territory of digital phenotyping. *Npj Digital Medicine, 1*(1), 1–5. https://doi.org/10.1038/s41746-018-0075-8

Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement, 79*, 311–320. https://doi.org/10.1016/j.measurement.2015.11.001

Merchant, R. M., Asch, D. A., Crutchley, P., Ungar, L. H., Guntuku, S. C., Eichstaedt, J. C., Hill, S., Padrez, K., Smith, R. J., & Schwartz, H. A. (2019). Evaluating the predictability of medical conditions from social media posts. *PLoS ONE, 14*(6), e0215476. https://doi.org/10.1371/journal.pone.0215476

Metz, R. (2018). *The smartphone app that can tell you're depressed before you know it yourself*. MIT Technology Review. https://www.technologyreview.com/2018/10/15/66443/the-smartphone-app-that-can-tell-youre-depressed-before-you-know-it-yourself/

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology, 13*(1), 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

Montag, C., Sindermann, C., & Baumeister, H. (2020). Digital phenotyping in psychological and medical sciences: A reflection about necessary prerequisites to reduce harm and increase benefits. *Current Opinion in Psychology, 36*, 19–24. https://doi.org/10.1016/j.copsyc.2020.03.013

Mulvenna, M. D., Bond, R., Delaney, J., Dawoodbhoy, F. M., Boger, J., Potts, C., & Turkington, R. (2021). Ethical issues in democratizing digital phenotypes and machine learning in the next generation of digital health technologies. *Philosophy & Technology*. https://doi.org/10.1007/s13347-021-00445-8

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-Time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine, 52*(6), 446–462. https://doi.org/10.1007/s12160-016-9830-8

Onnela Lab. (2017, July 21). *Digital phenotyping and beiwe research platform*. Onnela Lab. https://www.hsph.harvard.edu/onnela-lab/beiwe-research-platform/

Penney, J. W. (2016). Chilling effects: Online surveillance and Wikipedia use. *Berkeley Technology Law Journal, 31*, 117.

Pirelli, G., Otto, R. K., & Estoup, A. (2016). Using internet and social media data as collateral sources of information in forensic evaluations. *Professional Psychology: Research and Practice, 47*(1), 12.

Prainsack, B. (2017). *Personalized medicine: Empowered patients in the 21st century?* (Vol. 7). NYU Press.

Psillos, S. (2011). An explorer upon untrodden ground: Peirce on abduction. In D. M. Gabbay, S. Hartmann, & J. Woods (Eds.), *Handbook of the History of Logic* (Vol. 10, pp. 117–151). North-Holland. https://doi.org/10.1016/B978-0-444-52936-7.50004-5

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach.* (Fourth edition.). Pearson.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research, 17*(7), e175. https://doi.org/10.2196/jmir.4273

Sharon, T. (2017). Self-tracking for health and the quantified self: Re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare. *Philosophy & Technology, 30*(1), 93–121. https://doi.org/10.1007/s13347-016-0215-5

Simblett, S., Matcham, F., Siddi, S., Bulgari, V., Pietro, C. B. di S., López, J. H., Ferrão, J., Polhemus, A., Haro, J. M., Girolamo, G. de, Gamble, P., Eriksson, H., Hotopf, M., Wykes, T., & Consortium, R.-C. (2019). Barriers to and facilitators of engagement with mHealth technology for remote measurement and management of depression: Qualitative analysis. *JMIR MHealth and UHealth, 7*(1), e11325. https://doi.org/10.2196/11325

Stanghellini, G., & Leoni, F. (2020). Digital phenotyping: Ethical issues, opportunities, and threats. *Frontiers in Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.00473

Stevenson, L. F., Haberman, D. L., & Wright, P. M. (2013). *Twelve theories of human nature*. Oxford University Press.

Tekin, Ş. (2020). Is big data the new stethoscope? Perils of digital phenotyping to address mental illness. *Philosophy & Technology*, 1–15.

Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2016). New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health, 3*(2), e16. https://doi.org/10.2196/mental.5165

Torous, J., Wisniewski, H., Bird, B., Carpenter, E., David, G., Elejalde, E., Fulford, D., Guimond, S., Hays, R., Henson, P., Hoffman, L., Lim, C., Menon, M., Noel, V., Pearson, J., Peterson, R., Susheela, A., Troy, H., Vaidyam, A., … Keshavan, M. (2019). Creating a digital health smartphone app

and digital phenotyping platform for mental health and diverse healthcare needs: An interdisciplinary and collaborative approach. *Journal of Technology in Behavioral Science*, *4*(2), 73–85. https://doi.org/10.1007/s41347-019-00095-w

van Ballegooijen, W., Riper, H., Cuijpers, P., van Oppen, P., & Smit, J. H. (2016). Validation of online psychometric instruments for common mental health disorders: A systematic review. *BMC Psychiatry, 16*(1), 45. https://doi.org/10.1186/s12888-016-0735-7

Wang, R., Aung, M. S. H., Abdullah, S., Brian, R., Campbell, A. T., Choudhury, T., Hauser, M., Kane, J., Merrill, M., Scherer, E. A., Tseng, V. W. S., & Ben-Zeev, D. (2016). CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 886–897. https://doi.org/10.1145/2971648.2971740

Ware, S., Yue, C., Morillo, R., Lu, J., Shang, C., Bi, J., Kamath, J., Russell, A., Bamis, A., & Wang, B. (2020). Predicting depressive symptoms using smartphone data. *Smart Health, 15*, 100093. https://doi.org/10.1016/j.smhl.2019.100093

Wisniewski, H., Henson, P., & Torous, J. (2019). Using a smartphone App to Identify clinically relevant behavior trends via symptom report, cognition scores, and exercise levels: A case series. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00652

Yarkoni, T. (2012). Psychoinformatics: New horizons at the interface of the psychological and computing sciences. *Current Directions in Psychological Science, 21*(6), 391–397. https://doi.org/10.1177/0963721412457362

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393