This version is a preprint. Please cite published version: 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), pp. 21–33.

Digital Psychiatry: Risks and Opportunities for Public Health and Well-Being

Christopher Burr, Jessica Morley, Mariarosaria Taddeo and Luciano Floridi

***Abstract*—Common mental health disorders are rising globally, creating a strain on public healthcare systems. This has led to a renewed interest in the role that digital technologies may have for improving mental health outcomes. One result of this interest is the development and use of artificial intelligence for assessing, diagnosing, and treating mental health issues, which we refer to as 'digital psychiatry'. This article focuses on the increasing use of digital psychiatry *outside* of clinical settings, in the following sectors: education, employment, financial services, social media, and the digital well-being industry. We analyse the ethical challenges of deploying digital psychiatry in these sectors, emphasising key problems and opportunities for public health, and offer recommendations for protecting and promoting public health and well-being in information societies.**

***Index Terms*— artificial intelligence, digital ethics, digital psychiatry, digital well-being, mental health.**

I. Introduction

THE incidence of common mental health disorders is rising globally. The World Health Organisation states that between 2005 and 2015 the total estimated number of people living with depression increased by 18.4% and 14.9% for anxiety disorders [1]. At the same time, the Organisation for Economic Co-operation and Development has stressed that the healthcare systems of its member states are under increasing strain from rising demand that has not been met by an increase in resource [2]. This pressure, combined with additional socioeconomic issues (e.g. unequal access to treatment), has resulted in an increased interest in the role that digital technologies may have in improving mental health outcomes [3], [4].

Digital technologies that use artificial intelligence (AI) to infer whether users are at risk of, or currently suffering from, mental health disorders are already available on the market, including public app stores and commercially-available services [5], [6]. We refer collectively to these technologies using the label 'digital psychiatry'. Digital psychiatry is now used to infer, with varying degrees of reliability and validity [7], whether an individual is suffering from depression [8]–[10], anxiety [11], autism spectrum disorder [12], post-traumatic stress disorder [13], and suicidal ideation [14].

In addition to risk assessment and diagnosis, digital psychiatry has also been used to deliver personalized treatment to individuals [15], [16]. Many of these developments have resulted from the increased availability of behavioral and biometric data, and new data streams have the potential for providing greater ecological validity for epidemiological studies [17], while also enabling new forms of predictive analytics [18]. However, such data are now also collected by institutions and organizations that are not part of formal healthcare system, for the purpose of extracting mental health insights about consumers, employees, users, students, and others. The increasing use of mental health data outside the healthcare system, alongside technological developments in digital psychiatry, raises a number of pressing ethical challenges related to individual health and well-being, as well as wider social concerns like public health (e.g. epidemiological inflation; diminishment of patient autonomy).

The goal of this article is to analyze the ethical challenges related to digital psychiatry, identifying risks and opportunities, known problems, proposed solutions, and outstanding gaps. We focus primarily on the use of digital psychiatry *outside* of clinical settings. This focus allows us to explore how the use of digital psychiatry in non-clinical settings could lead to a change in the distribution of responsibility for the maintenance of individual and public mental health. In section II, we frame the ethical challenges associated with digital psychiatry by exploring its use *within* the context of formal healthcare systems. The purpose of this framing is to emphasize the risks associated with deploying digital psychiatry outside of an environment that is governed by well-established frameworks of accountability and ethical principles and practices (i.e. a formal healthcare system). In section III, we explore sector specific uses of digital psychiatry in non-clinical settings, including education, employment, financial services, social media, and the digital well-being industry.[1] In section IV, we

This research was funded by a grant from Microsoft Research awarded to Professor Luciano Floridi.

C. Burr is with The Alan Turing Institute. He was with the Digital Ethics Lab at the Oxford Internet Institute, University of Oxford at the time this research was conducted e-mail: cburr@turing.ac.uk).

J. Morley is with the Digital Ethics Lab at the Oxford Internet Institute, University of Oxford.

M. Taddeo is with the Digital Ethics Lab at the Oxford Internet Institute, University of Oxford, and also affiliated with the Alan Turing Institute.

L. Floridi is with the Digital Ethics Lab at the Oxford Internet Institute, University of Oxford, and also affiliated with the Alan Turing Institute.

[1] The term 'digital well-being industry' is used to refer to the growing, commercial availability of software (e.g. mobile apps), hardware (e.g. smart watches), and services (e.g. online chatbots) that seek to promote the self-governance of mental health and well-being (see section III/E).

analyze further some of the key ethical opportunities and risks, and offer a series of related recommendations for protecting and promoting individual and social well-being in information societies. Finally, section V concludes the article.

II. Framing the Ethics of digital psychiatry

Machine learning algorithms are used to learn "statistical functions from multidimensional data sets to make generalizable predictions about individuals" [19, p. 91]. These models can then be used to build automated decision-making systems that can serve as decision-support tools, when there is a human-in-the-loop, or fully automated systems, when designed to bypass or replace human decision-making. Both types of systems can be used in digital psychiatry, leading to significant opportunities for improving mental health outcomes [3]. However, there are also risks to such approaches that need to be addressed, namely: diminished patient autonomy, limited model generalizability, risk of epidemiological inflation, unequal access to (or engagement with) potential support, and loss of privacy and trust (see section IV for discussion). To understand these risks better, it is helpful to note how digital psychiatry is being deployed within formal healthcare systems.

The delivery of formal healthcare is typically governed by deeply entrenched bioethical principles, norms of conduct, and regulatory frameworks that maintain professional accountability. These practices and frameworks are designed to minimize risks to patient health and well-being, among other things, and are starting to address the use of digital technologies in healthcare. For instance, the UK's Department of Health and Social Care has recently established a code of conduct for the safe and ethical deployment of AI and other data-driven technologies within the NHS [20]. Like most areas of healthcare, clinical psychiatry can be split into the following series of stages: assessment, diagnosis, and treatment. Digital psychiatry can be used in each of these stages, either to support clinicians or automate key aspects of healthcare delivery. For instance, Lucas et al. found that the use of virtual human interviewers increased the disclosure of mental health symptoms among active-duty service members, who may otherwise be unwilling to seek support due to perceived stigmatization [16]. The use of this virtual interviewer was able to improve user engagement with mental health services while also creating a novel environment to collect behavioral data to facilitate *assessment* and *diagnosis*. Furthermore, environmental and behavioral data collected by mobile devices (e.g. smartphones, wearables) can provide ecologically-valid sources of information about patients [21], [22], which could be used to extract insights about the social-determinants of their health to improve decisions about *treatment options* (e.g. social prescribing or pharmacological). The complexity and size of these datasets means that AI is often necessary to help clinicians extract meaningful insights [23].

As the above examples suggest, digital psychiatry can be used to augment and extend the abilities of healthcare professionals, automating key tasks while ensuring that human judgement remains an integral part of the process. Others have already commented on the importance of keeping humans (and society) in-the-loop when deploying AI systems in healthcare [24], but this concern takes on a particular importance in the context of psychiatry. As Burns notes,

> "[t]he diagnosis of a mental illness […] reflects a judgement, albeit often imperfect, that individuals have become different from their 'normal' selves in some fairly recognizable way." [25, p. 178]

It is crucial, therefore, that the judgment as to whether individuals have become "different from their "normal' selves" remains a responsibility of a human agent.

Furthermore, as Johnstone & Boyle note, the purpose of a diagnosis is

> "fundamentally an attempt to make sense of a person's presenting problems, to understand how they have come about and what might help". [26]

Automated decision-making systems may be well-suited to tasks like classification, but they are poorly suited to perform tasks involving, for example, therapeutic, interpersonal conversation.[2] Hence, this participatory process of *sense-making* between a patient and psychiatrist, alluded to by Johnston & Boyle, is also currently poorly reflected in digital psychiatry.

In the same vein, decisions about when, how, and whether to intervene in digital psychiatry are complicated by the introduction of new forms of data. For instance, Lehavot et al. raise a series of ethical concerns in relation to the use of suicidal postings on social media by healthcare professionals [29]. As they explain, self-disclosure of suicidal thoughts can serve a therapeutic purpose for some patients. Therefore, ill-timed or misjudged interventions, could lead to unintended harm by failing to respect a patient's perceived boundaries of privacy. These ethical challenges are far from resolved even in cases where only healthcare professionals are involved. Therefore, the use of digital psychiatry to automate similar interventions outside clinical settings are only likely to complicate matters further.

These are just a couple of instances of where the use of digital psychiatry to automate key aspects of the clinical pathway raises potential risks—many of which emerge because of key differences between human psychiatrists and digital psychiatry (see Table I for a list of important differences). These risks exist despite the aforementioned principles and practices that govern how formal healthcare is delivered, and sometimes despite retaining a human-in-the-loop. Therefore, as the use of digital psychiatry moves outside the remit of the institutional practices and frameworks of healthcare systems, we can expect additional ethical challenges to emerge.

In the next section, we introduce some of these risks in several sectors that have begun to deploy digital psychiatry: education, employment, financial services, social media, and the digital well-being industry. While some risks are sector-dependent, others are shared across sectors. For instance, the potential to automate aspects of psychiatry outside of formal healthcare systems suggests a shift in the distribution of

[2] Similar concerns have been raised in relation to embodied robotics used in assisted living and geriatrics [27], [28].

responsibility for the maintenance of public mental health. This can be seen when considering the case of employers, who previously did not have a reliable mechanism for assessing the mental health of employees, but with the development of digital psychiatry are now able to take advantage of new opportunities to support employee well-being. However, it is possible that these new opportunities are also establishing new *duties of care* and *responsibilities to intervene* (e.g. detecting an employee that is suffering from high levels of occupational stress and anxiety), which will be hard to discharge without the ethical and legal frameworks that provide necessary guidance and support.[3] These frameworks are already well-established in the case of formal healthcare systems and supported by mechanisms of legal and professional accountability that are designed to safeguard public health, but they may fail to translate outside of their original domain [30]. This is why the use of digital psychiatry outside of formal healthcare systems raises pressing ethical challenges related to health and well-being.

III. The Wider Use of Digital Psychiatry

In this section we explore several sectors where digital psychiatry is used outside of formal healthcare systems, introducing some key ethical challenges that are explored further in section IV.[4]

A. *Education*

AI has already had a significant impact on the organization and delivery of formal education [31]. In relation to digital psychiatry, two developments stand out: the use of AI to develop 'intelligent tutoring systems' that facilitate self-directed learning (e.g. delivering immediate and personalized feedback), enabling teachers to focus on tasks that machines are ill-equipped to deliver (e.g. nurturing more creative modes of learning); and the development of 'intelligent support systems' for school and university administrators who wish to identify students who may be at risk of mental health issues and their related challenges (e.g. absenteeism from bullying).

Many examples of intelligent tutoring systems exist, including online learning platforms that use AI to adapt to student performance [32], and wearables that use sensors to measure levels of attention and engagement in students and provide real-time feedback to teachers.[5] Therefore, the possibility of delivering personalized, real-time feedback and help to students who require additional support due to mental health issues is an interesting opportunity for using digital psychiatry in education. A research team at IBM characterize this opportunity in terms of a 'digital assistant'. They note how the

> "increased access students have to smart devices that have a multitude of sensors […] can provide data for better understanding of the student's context and usage patterns" [35, p. 6976].

A primary concern here is the fact that such contextual information or usage patterns has also been shown to include latent information pertaining to an individuals' mental health, such as levels of anxiety inferred from computer usage logs [11]. As such, the ability to use AI in the classroom to monitor student performance and engagement may lead to the unintended consequence of unfair treatment of those with undiagnosed mental health issues (e.g. attention deficit disorder). While this may also be framed as an opportunity for earlier assessment, this would only be true if the respective educational institution has the necessary facilities (e.g. counselling services) to ensure that the student receives appropriate support. Otherwise, additional monitoring of student performance and engagement may simply create a feedback loop that further exacerbates mental health issues, prompting increased depression, anxiety, and decreased self-esteem, which are associated with academic pressures [36].

Similar concerns also extend to intelligent support systems, which could (it has been suggested) be used by school and university administrators to identify students that may be in need of more targeted support [37], and used to predict self-harm, drug abuse and eating disorders [38]. Further examples include the use of mobile phone location data for detection of depression and anxiety in undergraduate students [39], and the analysis of social media data to detect vulnerable students [40].

When we consider that educational institutions typically have a *duty of care* towards young people, it is understandable that they would be interested in employing digital psychiatry to support this duty. However, the scope of this duty, and the (perceived) protection it affords to students can vary. A recent report commissioned by the UK All-Party Parliamentary Group on Data Analytics notes,

> "whilst data can help with this role, it can also hurt the individual when mistakes are made, potentially irreparably. The Inquiry was given an example of a student who had been mortified when their disability was revealed to their lecturer, and the comment was made that although the scale (and therefore severity) of the error might be very low for the organisation committing it, for that individual it could be life-changing" [41, p. 32].

Examples like this one highlight an ethical and practical need to consider the competing interests, values, and priorities of those who are affected by digital psychiatry. For instance, a student who suffers with mental health issues, and wishes to keep this fact private, may find the use of digital psychiatry by administrative staff to be a violation of their privacy, and judge it as a disproportionate action taken under the auspices of a duty of care.

[3] This concern applies to several areas and is discussed in more detail in section IV.

[4] These sectors are not intended to be exhaustive and our review is not based on a systematic search of the literature. However, the scope of the analysis is sufficient for our purpose to provide motivating reasons for the claim that digital psychiatry has wider public health issues and ethical implications that extend beyond formal healthcare systems.

[5] A Harvard-based technology firm have developed a product known as FocusEDU, which they describe as providing "the world's first technology that can quantify real-time student engagement in the classroom", in order to enable teachers to "track student engagement and class attention levels as they're happening" [33]. In addition, a research group at MIT's Media Lab, have also developed a similar product (AttentivU) that seeks to improve attention through real-time monitoring of a user's engagement [34]. Both devices employ physiological sensors (i.e. electroencephalography) to measure engagement.

B. Employment

Like education, there are two strands to highlight in connection with the use of digital psychiatry in employment: its use to support management practices and hiring practices (i.e. HR analytics).

Koldijk, Neerincx, & Kraaij demonstrate how unobtrusive sensors (i.e. computer logging, facial expressions, posture and physiology) can be used to collect employees' behavioral data to train automatic classifiers to infer working conditions and stress-related mental states [42]. While such techniques could support employees' well-being, we cannot ignore the potential risk of privacy violations, especially in light of a recent report about the increasing use of monitoring techniques by large firms, to track employee behaviors while at work or keystroke loggers to monitor employee communication [43]. A recent qualitative study explores how the use of digital technologies to monitor employees is related to self-reported impacts on anxiety, stress, and depression, and argues that the use of such systems impacts negatively an employee's mental health, and may lead to a loss of commitment, professional identity, and self-confidence [44]. The risks raised in relation to the use of digital psychiatry in education and the potential to create feedback loops that exacerbate mental health issues also apply here. However, employers are less likely than educators to have the necessary human resources to support employees (e.g. access to counsellors), and the risks for individual privacy remain high. For instance, continuous monitoring, could be construed as surveillance and 'policing' for compliance, and users may perceive the risk that the collected data is not just used for the purpose of promoting their well-being but also to ensure they are abiding by the 'rules' [45].

Turning to the use of digital psychiatry to support hiring practices, it has been noted that some firms are exploring how AI can improve HR-analytics by mining sources such as social media (e.g. LinkedIn profiles) [46]. This, along with the acknowledgment that "Facebook data has also been used to infer dark side [sic] personality traits, such as psychopathy […], and narcissism […]" [46, p. 14], raises substantial ethical concerns related to privacy and discriminatory profiling.

AI may also be used to infer relevant information from interview data, possibly using design strategies from gamification to structure interviews (e.g. to infer cognitive ability or problem-solving ability). Chamorro-Premuzic et al. argue that these methods could help standardise the interview process "making it more objective and cost efficient while reducing the impact of interviewer biases" [46, p. 14], while also proving beneficial to firms looking to identify the best talent. For instance, Chen et al. developed an automated system, which is able to extract relevant behavioral and psychological features from video interviews [47]. Part of this process relied on well-known linguistic assessment tools (e.g. Linguistic Inquiry and Word Count, LIWC[6]), which detects psycholinguistic properties associated with positive and negative emotions. However, this system also relies on the automated scoring of an individual's speaking skills that comprised multiple dimensions, including fluency, prosody, and pronunciation.

Although the use of intelligent systems for hiring practices is highly varied in nature, there are two specific concerns to which we can point. First, individuals with Social Anxiety Disorder (SAD) sometimes speak slower and more quietly than others [49]. It is realistic to imagine that this behavioral feature may reduce employment prospects for individuals suffering from SAD. For example, forms of algorithmic discrimination may place too high a weight on prosodic features (e.g. due to biases inherent in their training datasets or feedback loops that propagate existing biases further and increase the risk of discrimination). Second, employment can have myriad benefits for mental health and well-being, such as potentially supportive social relationships [50]. The misuse of digital psychiatry in hiring, therefore, may act as a barrier for those suffering with mental health issues, preventing them from accessing social environments that could be beneficial to their mental health and well-being.[7]

C. Financial Services

Mental health disorders can lead to serious financial difficulties and further exacerbate an individual's level of suffering. For instance, disorders due to addictive behaviors (e.g. gambling or mood disorders) are associated with compulsive buying and spending that place heavy burdens on individuals [52]. Technologies already exist to mitigate the negative impacts on one's finances, like spending blocks for customers who wish to limit purchases during specific times (e.g. when at risk from manic episodes). These technologies may provide more immediate support to individuals, offering the necessary means to manage their finances in ways that are responsive to their individual needs. However, financial services companies are increasingly looking at whether AI can be used to detect problematic behavioral patterns in transaction data before they arise. For example, the UK's Financial Conduct Authority (FCA) notes that one firm uses speech analytics software to parse calls for "triggers or clues to vulnerability, such as mention of illness, treatment, diagnosis, depression" [53, p. 81]. And, Evans notes that financial services companies are interested in the potential of combining transaction datasets with additional sources of data, such as social media usage or biometric signals (e.g. heart-rate variability) [54]. However, so far, ethical, legal, and regulatory concerns have impeded these developments [54].

In recognition of these concerns, the FCA has established a code of good practice that seeks to impose a duty of care of financial service providers towards vulnerable consumers, namely, "someone who, due to their personal circumstances, is especially susceptible to detriment [e.g. an individual with a mental health issue], particularly when a firm is not acting with appropriate levels of care" [53, p. 20].

[6] LIWC is a popular method in computational linguistics for inferring psychological information based on an individual's language use [48].

[7] It should be noted that a recent review showed how digital psychiatry techniques are still in their infancy in employment and not widely used [51]. However, the fact that these tools and techniques could enable employers to cut costs both at the level of management and hiring suggests that their use is likely to grow.

As part of their code of good practice, the FCA encourages *proactive intervention*, rather than just *reactive support*, for customers who choose to disclose information about their mental health. There are many ways in which the term 'proactive intervention' could be construed. One suggestion that the FCA offers is to "spot abnormal patterns or danger signals and act before people are actually in difficulties" [53, p. 88]. While the intention behind this suggestion may be beneficent, its lack of specificity means that it is up to firms to exercise caution when deciding how to act upon the guidance, raising two main concerns.

First, there may be no clear justification for using digital psychiatry instead of alternative strategies that uphold principles of data minimization [55] and offer support without explicit diagnosis or identification. Moreover, once identified, there is a risk that such consumers may become targets of unscrupulous actors [56]. A narrow focus on the *opportunities* that digital psychiatry offers carries a risk of creating an imbalanced incentive structure, whereby alternative strategies may be overlooked (e.g. universal design methods that aim to support and empower *all* users [57]).

Second, as Evans asks in response to the FCA's recommendation of proactive intervention: "[d]oes identifying that a customer may be at risk place an ethical obligation on a firm to take action, even if the customer has not sought support and may not know they're unwell?" [54, p. 3]

In this case too, we see a concern about how far the duty of care of financial services organizations extends, and whether they have a right to intervene.

D. *Social Media*

Social media receives a lot of attention in relation to the use of digital psychiatry techniques for suicide prevention [58], [59]. Part of the attention is due to empirical uncertainty about social media's own role in increased suicide rates among adolescents, and the differential impact on males and females [60], [61].[8]

Because many people who are suicidal are unknown to healthcare professionals, some have begun to ask what role social media can play in detecting suicidal ideation and delivering targeted prevention [63]. Here, a key aim is to use digital psychiatry to develop *risk assessment tools*. These tools can be viewed as attempting to augment and automate the traditional psychiatric pathway, perhaps by adding an earlier 'screening' stage. Risk assessment tools, as Velupillai et al. note,

> "may be powerful even if the [positive predictive value] of the predictions are low, because they can be deployed on a large scale." [64, p. 3]

The ability to operate at scale is what makes digital psychiatry so promising for deployment on social media platforms [65]. However, this is a double-edged sword, as the larger scale also compounds the potential risk of harm (i.e. greater number of inaccurate assessments). Part of this risk stems from the widespread use of natural language processing (NLP) techniques to extract relevant psychiatric features in text (e.g. risk factors such as psychiatric stressors [66]). Gkotsis et al. present findings from a NLP study, which classifies social media posts as indicative (or not indicative) of a mental health disorder, and links particular posts to specific disorders [67]. However, as they note, such a strategy may be insufficient to diagnose users with a specific mental health disorder—for instance, they could be posting about a friend or family member.

More recently, Merchant et al. showed how machine learning algorithms trained on social media posts were more accurate than those trained on demographic information, where the *ground truth* was extracted from consenting patient's electronic health records [68]. This can help overcome the previous problem of determining the ground truth, but it is unlikely to be a strategy to be deployed outside of clinical settings due to the limited access to patient records.

While there are undoubtedly many positives to social media companies using their size and influence for good in this way, there are also serious ethical concerns related to the use of digital psychiatry by social media platforms that need urgent consideration [59]. For example, there is a lack of transparency regarding how risk assessment tools are developed and operate, leading to *diminished trust* in the ability of social media platforms to use such tools in an ethical and safe manner [58]. Goggin reports that once Facebook's risk assessment tools identify a vulnerable user on the basis of their profile, two outreach actions can take place, "[r]eviewers can either send the user suicide resource information or contact emergency responders" [65]. However, it is unclear how often each of these actions takes place, or what determines whether a user is high or low risk. This lack of transparency is problematic because social media platforms can provide users with valuable spaces to enquire about issues like side effects from treatments, share coping skills, and even engage in therapeutic forms of self-disclosure, thereby feeling less isolation or stigma [29], [67]. The recurring issues surrounding the transfer of responsibility from formal healthcare systems to the private sector and the ethical challenges of determining when there is a right or duty to intervene, previously highlighted in the case of education and human resources, also apply here.

E. *The Digital Well-Being Industry*

In 2013, the World Health Organisation recommended the development and implementation of "tools or strategies for self-help and care for persons with mental disorders, including the use of electronic and mobile technologies" as a way of "improving access to care and service quality" [69].

Reports focusing on the on-going trends in the digital well-being industry note that the market is worth nearly $8 billion in the United States [70], and that there are now more than 300,000 health-related mobile apps available [71]. However, it is not clear whether the digital well-being industry delivers on promises—like improved access, real-time support, and more

[8] For instance, the World Health Organisation are working towards the global target of reducing the suicide rate in countries by 10% by 2020, and as part of their efforts are investigating the "supplementary role that the Internet and social media are playing in suicide communications" (World Health Organisation, 2014, p. 32).

empowered users—or whether the proliferation of digital technologies for personal well-being create new risks for public health, leading to additional challenges for formal healthcare systems.

Unlike the previously discussed sectors, the use of digital psychiatry within the digital well-being industry is too diffuse to categorize neatly. However, two key developments are commercially-available mobile health (mHealth) devices (e.g. mobile apps and fitness trackers) and online services (e.g. chatbots). These technologies offer users a range of services that aim to enhance a user's ability to monitor or track health-related behaviors in a more self-directed manner, by using digital psychiatry to personalize feedback and offer real-time support [5]. However, Barras also draws attention to a growing reliance on unlicensed therapists, noting concerns from healthcare professionals about the potential of a reduced quality in mental healthcare, and questioning the efficacy of the forms of motivational therapy offered by these services [15].

Similar concerns about efficacy have also been raised with respect to services that claim to offer support for mental health issues such as anxiety, schizophrenia, post-traumatic stress disorder, eating disorders, and addiction [5]. Larsen et al. evaluate the quality of health-related claims made by commercially-available mHealth apps, noting that 36% of the apps evaluated made claims about improvements in knowledge or skills to support self-management, 30% made claims about improvements in symptoms or mood, and 10% claimed the ability to diagnose or detect a mental health condition [72]. However, only two apps offered details of the direct evidence supporting these claims, and only one provided citation details to the scientific literature. Similar findings regarding the lack of empirical support exist for other commercially-available services [73], [74], raising questions about the informational value of digital psychiatry in the digital well-being industry.

The concern about efficacy becomes more pressing when we acknowledge that these products or services may also distance users from healthcare professionals, who may be able to mitigate some of the worst effects. In the UK, for example, products and technologies that are classified as medical devices are heavily regulated by the Medicines and Healthcare products Regulatory Agency. Products and services within the digital well-being industry often currently fall outside of this regulatory scope. As Anthes notes,

> "an app that claims to prevent, diagnose or treat a specific disease is likely to be considered a medical device and to attract regulatory scrutiny, whereas one that promises to 'boost mood' or provide 'coaching' might not". [5]

This is why so many services rely on unlicensed therapists (i.e. 'motivational therapist' or 'wellness coach') or hide behind disclaimers to avoid scrutiny [72].

Healthcare services and governments across the globe have responded to these developments [75], [76]. For instance, the American Psychiatric Association has designed an evaluation framework to guide informed decision-making around the use of smartphone apps in clinical care [6]. These are positive developments, but nevertheless have limited reach due to their focus on clinical contexts, rather than the wider sectors discussed in this paper.

Many of the risks analyzed throughout this section are a result of the transfer of responsibility from traditional healthcare providers to institutions, organizations, and individuals. This raises a series of questions, such as whether an organization has a duty or right to intervene on the basis of inferred information; whether an institution has the necessary resources to support individuals (e.g. students or employees); and whether an individual has the decisional capacity and health literacy to take responsibility for their mental health using commercially-available services. Table II summarizes these issues, for each the sectors identified in this section, and we explore them further in the next section.

IV. The Ethics Digital Psychiatry: Opportunities and Risks

In this section we expand on the ethical issues identified in the previous section. As these issues are not all common instances of ethical principles, values, or standards, etc., we do not attempt to organize them within an over-arching framework (e.g. mapping them to principles of bioethics or principles that govern the design and development of AI [77], [78]). We also note open questions and positive recommendations that we expect will be central to ongoing discussions of digital psychiatry.

A. Autonomy

The concept of 'autonomy' is an important concept in both bioethics [30] and moral and political philosophy more broadly [79]. While its usage varies, for present purposes it is sufficient to note that its importance and value reflects a respect for both an individual's capacity to decide and freedom of whether to decide [80]. This is because medical decisions are understood to involve value judgements (e.g. comparative quality of life assessments for two treatment options). For instance, a patient experiencing depression may be fully informed by their psychiatrist about their mental health and the options available to them in terms of recovery, but nevertheless autonomously decide to forego any treatment because their condition may be an important part of their self-identity.[9]

While these features of autonomy are important in bioethics, we also need to go beyond the perspective of patient autonomy to identify specific ethical issues that arise with the use of digital psychiatry outside of clinical settings. For instance, we can ask:

> (1) Does the passive collection of data impact a user's ability to participate in the decision-making process?
> (2) How does the use of digital psychiatry outside of a formal healthcare system change the way that autonomy is construed?

Starting with (1), we can note how passive forms of data surveillance and monitoring may create a non-transparent form of informational asymmetry, whereby the user may be unaware

[9] This acknowledgement is part of the *recovery approach* [81, p. 527], which views recovery as "a deeply personal, unique process of changing one's attitudes, values, feelings, goals, skills, and/or roles. It is a way of living a satisfying, hopeful, and contributing life even with limitations caused by illness. Recovery involves the development of new meaning and purpose in one's life as one grows beyond the catastrophic effects of mental illness."

which behavioral signals are used as input features for assessment or diagnosis. This informational asymmetry can impact autonomy in several ways. First, it limits the availability of information that an individual may use to decide how to present themselves, affecting their ability to influence how they are perceived by others [82]. Second, it may undermine users' trust in the respective system due to perceived privacy risks, which could alter the evaluation of possible, desirable actions (e.g. whether to use a support forum). Third, reliance on a pre-determined set of features that are used to assess or diagnose an individual may prevent an individual from being able to challenge decisions by appealing to other relevant features that fall outside of the system's ontology (e.g. an over-reliance on the use of linguistic features like prosody impacting hiring outcomes to the detriment of specific applicants).

Moving on to (2), non-clinical uses of digital psychiatry could lead to concerns regarding autonomy that may not arise in the context of formal healthcare—most notably questions about whether an organization has the *right to intervene* on the basis of inferred information[10]. For instance, the use of digital psychiatry to monitor the psychological well-being of employees to improve management practices may be perceived as unjustified forms of surveillance that infringe on an individual's right to mental integrity and self-governance. However, employees may choose not to challenge this use of personal data for fear of repercussions, instead suffering from the increased anxiety that the use of such systems may create. To address these questions, it is necessary to consider how different theories of autonomy—including those not developed for the explicit purpose of medical practices or relationships—may conceptualize the risks and opportunities of digital psychiatry across different sectors. There is no easy way around this, other than pursuing further research that explores how ethical principles, such as 'autonomy', require contextual specification in order to provide useful practical guidance for developers and policymakers in different sectors.[11]

B. Model Generalizability

Where predictive models have been trained on non-representative samples (e.g. limited demographic variation or sets of features in the dataset), there is a risk of embedding bias into the performance of the respective classifiers (e.g. digital psychiatry in hiring). This connects to the issue of *model generalizability*, which refers to the performance of a decision function (e.g. a machine learning algorithm trained on one dataset) to new cases or contexts (e.g. temporal, geographic, genetic, cultural, or disease related) [19]. Model generalizability is especially problematic in psychiatry, where there is widespread acceptance that different demographics experience mental health in varied ways (e.g. different social attitudes towards a particular disorder impacting severity of symptoms).

If the aim of digital psychiatry is to construct generalizable, predictive models to be deployed in an online environment, the sample data needs to be sufficiently representative of the heterogeneous population to which it will be applied.[12] Consider again the case of early screening for suicide prevention. It is important to note that differences in demographic and personal characteristics, like gender [83], lead to differences in how individuals seek help. Freeman et al. note that

> "[c]onsidering the differences in suicidal intent between males and females […], gender targeted prevention and intervention strategies would be recommended." [83, p. 1]

This is why a single model for detecting and acting upon indicators of suicidal ideation is unlikely to be suitable.

Given the impact on individual health and well-being, aside from being an empirical matter, ensuring adequate model generalizability is an inherently ethical responsibility for developers to ensure the technology they are deploying is fit-for-purpose. This is why the assessment of any new service should not be carried out in isolation from the rest of the system in which it will be embedded. For example, an app that can identify early signs of depression could be dangerous if there is not a well-functioning healthcare system to support the individual. However, as many of these digital solutions are developed by commercial companies, it is not immediately apparent who is responsible for conducting this assessment. In state-provided, single-payer systems like the UK's National Health Service (NHS), it would be the responsibility of those commissioning the service. However, if the service is provided as 'direct-to-consumer', this option may not be available. This creates a scenario in which the onus may sit with the technology providers, requiring them to accept a public health responsibility that may admittedly be viewed as a supererogatory action[13].

One viable option is for public health organizations to generate the evidence of the public health risks of commercial uses of digital psychiatry and use this evidence base as leverage to instigate partnerships with developers. For example, this could be facilitated by targeting the process of uploading apps to commercial app stores, so that developers receive a message notifying them that before they are able to upload a health-related app it must be submitted to review by the relevant health authority for the purpose of ensuring public safety. In this scenario, app stores would incur a cost in slowing the rate of uploaded apps, potentially resulting in a slight reduction in income from this stream, but would also reduce the risk of liability for supporting potentially dangerous products without taking on anything other than a signposting responsibility.

[10] Our use of the term 'right' is used in a loose sense and does not reflect the endorsement of a particular normative theory.

[11] Unfortunately, this article is not the appropriate place to expand on these complex normative questions, which will be expanded on in future work.

[12] Although individual models could be developed for specific use cases, this would not necessarily fully overcome all of the issues discussed in this sub-section.

[13] In ethics, the term 'supererogatory' is used to refer to an action that is morally praiseworthy but not obligatory. For instance, an action that goes "above and beyond" what is morally acceptable.

C. *Epidemiological Inflation*

The use of digital psychiatry outside of formal healthcare raises a possible risk regarding epidemiological inflation. The general idea is that potentially well-intentioned tools that seek to identify mental health issues could contribute to a rise in their prevalence within the population. For instance, some have stressed the possibility that the use of symptom tracker apps may worsen the very symptoms they're designed to help with (e.g. anxiety associated with sleep tracking) [84]. There is insufficient evidence, at present, to determine how widespread such effects may be for any given condition. This is because separating the truth from mere panic is complicated by myriad factors. For instance, as Makin notes,

> "[c]comparing the rate of mental-health diagnoses in today's adolescents with those in previous generations is complicated by *changes in definitions* and a greater *willingness to report* mental-health problems." [85]

The lack-of empirical certainty does not mitigate concerns entirely. Research into placebo and nocebo effects suggests that physiological changes can occur as a result of learning of one's genetic risk for disease like Alzheimer's, independent of actual genetic risk [86]. If this phenomenon extends to scenarios where a user learns about wider risk factors associated with other mental health issues, then there would be further reason to adopt a more cautious attitude to the manner in which digital psychiatry may present information to users.

There is a trade-off here between competing interests of seeking to do good (beneficence) versus seeking to avoid harm (non-maleficence). In lieu of further empirical evidence about the causal contribution of digital psychiatry to mental health issues, it is likely that policymakers and healthcare professionals will fall back on the more cautious (and pragmatic) principle of non-maleficence. This has implications for the way in which 'diagnostic' uses of digital psychiarty are calibrated. Developers are likely to follow the 'smoke detector principle' (i.e. give more 'false positives' than 'false negatives'), especially in the absence of clear guidance on liability, as it is believed that less harm can come to the individual from a false 'referral' for further treatment than from missing a potentially life-limiting symptom of mental health crisis. Without evidence as to whether these 'diagnoses' can be linked to greater levels of health anxiety, however, it is not possible to say whether such a policy is justifiable.

Healthcare policymakers should commit to funding research in this area, so that the necessary evidence can be generated to inform appropriate and proportionate governing policies. Research should include elements related to cognitive and behavioral sciences, including design constraints from human-computer interaction, to specify the way in which information is displayed impacts the relevant health outcome. This information will enable policymakers to develop detailed guidance and standards to govern the design and use of digital psychiatry. Such standards must include guidance on how liability will be dealt with if information is interpreted in a way that is harmful, or if a diagnosis is missed or over-inflated, to minimize the extent to which false positives are seen as the 'safer' option.

D. *Earlier Screening and Improved Access*

It is well-understood how early identification and treatment of mental health issues such as depression can reduce the risk of suicide [25]. If digital psychiatry can deliver earlier screening and improved access, it is certainly worth further attention, as it may improve the sort of 'just-in-time' interventions that could save lives through more efficient prioritizing of resources [87], or help reduce unnecessary distress in sectors like financial services or education. While this may justify the use of risk assessment tools, it is important to stress that their use may also lead to some loss of privacy for individual users, due to necessary data collection. Aside from risks to privacy, a question remains as to whether the operators of relevant platforms or technology have a right to intervene if their automated tools identify a user at risk. The lack of agreement upon standards or codes of conduct for practices outside formal healthcare—aside from implicit organizational norms or commitments to principles such as data minimization—will likely serve as an obstacle to reaching any agreement on this question.

Enabling risk assessment tools to provide genuine benefits to the health outcomes of individuals is difficult because when diagnoses are made outside of the formal healthcare system they act more like inefficient and ineffective referrals. While risk assessment tools might suggest to an individual that they should seek medical advice, they do not necessarily provide them with access to the appropriate support or treatment. Instead, it is entirely possible that an individual who has been assessed outside of the formal healthcare system must be assessed again in a clinical setting before they are able to access treatment. In these cases, risk assessment tools do not "improve access" for individuals in a meaningful way, while they may create a psychologically distressing situation where individuals must live with the knowledge of a potential diagnosis for an extended period of time without any support.

Even when digital psychiatry *does* improve access in a meaningful sense, it may not do so fairly. Consider Babylon's NHS England service *GP at Hand*, which provides patients with video-based GP consultation with minimal waiting times[14]. This service has been criticized for 'cherry picking' patients—improving access for those who are digitally-savvy and relatively healthy whilst worsening access for those who still want to (or can only) meet GPs in-person [89]. On the one hand, there is nothing inherently 'wrong' with providing targeted services, or with giving individuals the ability to assess themselves; on the other it is important that the service's impact on the system as a whole is considered, so that 'improved access' does not lead to new forms of health inequality. To avoid this, healthcare system governing bodies should invest in the ability to model routinely the impact of digital psychiatry services in much the same way that care pathways in hospitals can be A/B tested in a modelling environment to maximize

[14] When compared with the average 2-week waiting time for booked in-person consultations [88].

efficiency of patient flow. This would give healthcare systems the ability to take necessary actions to ensure improved access genuinely means improved access.

E. Greater Engagement and Empowerment

Digital psychiatry may lead to greater engagement and empowerment for patients. However, as Morley and Floridi note, what constitutes *empowerment* is far from clear [90]. This is because, in many instances, the concept is predicated on an often-unfounded assumption that reflection on information (for example one's medical record) will automatically result in rational action. Yet, there is a risk of harm to individuals who do not have the 'right capacities' to engage critically with health information in the moment that they are presented with it, and who may be unable to identify accurately 'wrong' or non-validated sources of information. This risk has increased as digital psychiatry techniques have begun to operate outside of formal healthcare systems which can act as information gatekeepers protecting those who lack the decisional capacity to critically assess informational value-—an acute problem in the context of mental health disorders [91]—from potentially harmful information. For instance, a study found that some commercially-available apps—outside of the gatekeeping capacity of the formal healthcare system—make false claims, such as bipolar disorder is contagious or that drinking hard liquor before bed is a good way to deal with a manic episode [92]. Thus, Burr and Morley argue that the goal of genuine empowerment requires greater attention to possible barriers to *engagement*, which prevent individuals from seeking treatment for mental health issues in the first place [93].

Given the complex nature of mental health (e.g. person-specific contextual factors such as environmental influence), however, the clinically-demonstrated efficacy of technologically-mediated interventions may be restricted to the original context in which the study was conducted [94]. Therefore, while genuine empowerment demands prior focus on the barriers to engagement, the complex nature of these barriers in relation to mental health limits the scope of this opportunity. For example, lack of social support may prove to be a significant psychological burden for some, whereas economic factors may represent more serious barriers to engagement and empowerment for others. It is for this reason that digital psychiatry interventions should not be viewed as 'silver bullets', capable of cutting through existing socioeconomic complexities and creating entirely new methods of care. This is particularly important while the evidence of their social impact at scale is lacking.

There is also an ethical concern related to the intended goal of these technologies, especially those within the digital well-being industry. It is possible that some uses of digital psychiatry (e.g. chatbots) do little more than provide *paid-for access* to digital forms of talk therapy. As such, it could be argued that these services commercialize an important social function that has typically been provided by friends and families. Instead of empowering users, these services could contribute to a potential diminishment of the inherent value of social relationships by removing the opportunity for an actual friend to engage in virtuous forms of compassionate listening, and possibly offer more insightful support due to a wider understanding of the contextual factors (e.g. lifestyle, previous experiences).

F. Real-Time Support

A well-known challenge for public health is *adherence* to treatment. This is particularly problematic in cases of on-going therapy for chronic conditions, where patients are often expected to engage in long periods of self-directed treatment without access to a medical professional. Digital psychiatry may help with this challenge if implemented into already existing, ubiquitous mobile technologies. This is why digital psychiatry techniques that can extract relevant information from ecologically-valid biometric signals are attracting increasing attention (e.g. monitoring of student engagement in classrooms).

The value-sensitive design of these technologies is key to grasp these opportunities. For example, Mohr et al. state that adherence itself is not sufficient for improved health outcomes—improvement often requires "*sustained behavior change* over many weeks or months" [94, p. 427]. However, they continue, current

> "mental health technologies are mainly didactic or informational, which might not be ideal for promoting sustained engagement and behavior change for many people". [94, p. 427]

In this respect, Peters, Calvo, & Ryan argue that contemporary research into the satisfaction of psychological needs can be used to provide valuable design constraints for engineers and developers [95]. This approach can help develop more effective behavior change strategies, which foster sustained engagement by targeting an individual's values that are intrinsically motivating. While a promising avenue to explore, Mohr et al. also give reasons to be cautious:

> "[a]lthough improved design and technology may make mental health technologies easier and more engaging to use in the future, many of today's mental health technologies require some human support from a coach or therapist to sustain engagement and obtain substantive, reliable outcomes." [94, p. 427]

This suggests that the use of digital psychiatry to automate aspects of the user experience may need to be limited, even if it results in some additional friction. Instead of ambient monitoring, services could have emergency features that connect users to immediate in-person support services, which are accessible through voice, touch etc., in order to uphold accessibility and universal design standards [57]. Push notifications, which can be disabled on smartphones to give users greater control, could also be sent to mHealth devices so that *active* users can provide feedback on if, how, or why the device is enabling them to manage their health better.[15] Over

[15] Where the apps in question are part of the wider ecosystem of formal healthcare systems, for example in the instance of the NHS Apps Library, the providers of these service can be encouraged, or contractually obliged, to provide feedback data back to the central system so that it can 'learn' what apps or services are most effective at supporting its patients.

time, this may be perceived as best practice, and providers of technologies who do not do this may be seen as less focused on achieving positive health outcomes for their users and, therefore, less trustworthy.

G. *Privacy and Trust*

As public perception of digital psychiatry techniques increases, alongside media coverage of privacy violations [96], we will likely see a greater public concern for online privacy and a diminished sense of trust in organizations that deploy digital psychiatry [97], [98].

Huckvale et al. found systematic gaps in accredited mental health app's compliance with data protection principles (i.e. unsecure transmission of sensitive or personally identifying information, lack of local encryption, lack of privacy policies) [99]. Systematic reviews like this are vital to foster compliance but may prove to be inadequate to prevent individuals from altering their online behavior on the basis of *perceived* privacy violations. This is sometimes referred to as "self-surveillance" or "chilling effects" [100], and refers to the psychological phenomena whereby a belief that one is being watched leads to altered behavior.[16] Self-surveillance can arise from a *perceived* privacy risk (legitimate or not, i.e., a violation of privacy as determined by a legal framework, or a belief that a social media platform is engaging in intrusive data collection when it is not). This highlights the importance of trust and transparency between service users and platforms [101]. These are crucial, for instance, to enable the use of media platforms and online communities for the therapeutic disclosure of personal or sensitive information.

One potential way to tackle privacy risks and lack of trust is to tighten policies related to tracking and targeted advertising [102]. Current mechanisms for creating 'safe(r) spaces' online, require greater digital literacy than the majority of people have—for example, using a VPN or the browser extensions (e.g. 'trackmenot' extension[17])—and, people struggling with significant mental health issues may find that their capacity to take these actions is reduced. Thus, the responsibility for creating these spaces should not be incumbent on individuals. Instead, it should be discharged by public institutions and social media platforms. The creation of these spaces could fall, at least in the UK, under the 'duty of care' that the British government wants to see imposed on online companies [103].

V. Conclusion

Digital psychiatry is a relatively new phenomenon, both within and outside of formal healthcare systems, which raises pressing questions concerning its impact on mental health at both an individual and social perspective (e.g. social media's impact on public health). Answers to these questions will need to be addressed before we can harness the full potential of digital psychiatry.

The analysis presented in this article identifies key ethical challenges posed by digital psychiatry, which could be avoided with more attentive design and regulation. Our analysis uncovers open questions and a series of recommendations that require further work to be addressed—most notably the question of when institutions and organizations have a right or duty to intervene. The analysis also indicates that the deployment of digital psychiatry in non-clinical settings could lead to significant public health risks. Therefore, it is vital that researchers, health professionals, technology designers, and policymakers continue to assess and evaluate how and where digital psychiatry should be deployed to maximize the benefits for individual and social well-being, while minimizing the public health risks.

References

[1] World Health Organisation, 'Depression and Other Common Mental Disorders: Global Health Estimates', World Health Organisation, Geneva, 2017.

[2] G. P. Martin, E. Sutton, J. Willars, and M. Dixon-Woods, 'Frameworks for change in healthcare organisations: A formative evaluation of the NHS Change Model', *Health Serv Manage Res*, vol. 26, no. 2–3, pp. 65–75, Aug. 2013, doi: 10.1177/0951484813511233.

[3] T. Foley and J. Woollard, 'The digital future of mental healthcare and its workforce'. Health Education England, 2019.

[4] P. Garg and S. Glick, 'AI's Potential to Diagnose and Treat Mental Illness', *Harvard Business Review*, 22-Oct-2018.

[5] E. Anthes, 'Mental health: There's an app for that', *Nature News*, vol. 532, no. 7597, p. 20, Apr. 2016, doi: 10.1038/532020a.

[6] J. Torous *et al.*, 'Towards a consensus around standards for smartphone apps and digital mental health: Towards a consensus around standards for smartphone apps and digital mental health', *World Psychiatry*, vol. 18, no. 1, pp. 97–98, Feb. 2019, doi: 10.1002/wps.20592.

[7] C. Burr and N. Cristianini, 'Can Machines Read our Minds?', *Minds and Machines*, Mar. 2019, doi: 10.1007/s11023-019-09497-4.

[8] J. C. Eichstaedt *et al.*, 'Facebook language predicts depression in medical records', *Proceedings of the National Academy of Sciences*, vol. 115, no. 44, pp. 11203–11208, Oct. 2018, doi: 10.1073/pnas.1802331115.

[9] A. G. Reece, 'Instagram photos reveal predictive markers of depression', p. 12, 2017.

[10] S. Saeb, E. G. Lattie, S. M. Schueller, K. P. Kording, and D. C. Mohr, 'The relationship between mobile phone location sensor data and depressive symptom severity', *PeerJ*, vol. 4, p. e2537, Sep. 2016, doi: 10.7717/peerj.2537.

[11] Y. Fukazawa, T. Ito, T. Okimura, Y. Yamashita, T. Maeda, and J. Ota, 'Predicting anxiety state using smartphone-based passive sensing', *Journal of Biomedical Informatics*, vol. 93, p. 103151, May 2019, doi: 10.1016/j.jbi.2019.103151.

[12] Y. Nakai, T. Takiguchi, G. Matsui, N. Yamaoka, and S. Takada, 'Detecting Abnormal Word Utterances in Children With Autism Spectrum Disorders: Machine-Learning-Based Voice Analysis Versus Speech Therapists', *Perceptual and Motor Skills*, vol. 124, no. 5, pp. 961–973, Oct. 2017, doi: 10.1177/0031512517716855.

[13] D. Leightley, V. Williamson, J. Darby, and N. T. Fear, 'Identifying probable post-traumatic stress disorder: applying supervised machine learning to data from a UK military cohort', *Journal of Mental Health*, vol. 28, no. 1, pp. 34–41, 2019, doi: 10.1080/09638237.2018.1521946.

[14] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, 'Natural Language Processing of Social Media as Screening for Suicide Risk', *Biomedical Informatics Insights*, vol. 10, p. 117822261879286, Jan. 2018, doi: 10.1177/1178222618792860.

[15] C. Barras, 'Mental health apps lean on bots and unlicensed therapists', *Nature Medicine*, Mar. 2019, doi: 10.1038/d41591-019-00009-6.

[16] For instance, chilling effects were noted in response to Edward Snowden's revelation of widespread surveillance by the NSA [100].

[17] TrackMeNot is a browser extension that helps protect web searchers from surveillance and data-profiling: http://trackmenot.io

[16] G. M. Lucas *et al.*, 'Reporting Mental Health Symptoms: Breaking Down Barriers to Care with Virtual Human Interviewers', *Frontiers in Robotics and AI*, vol. 4, Oct. 2017, doi: 10.3389/frobt.2017.00051.

[17] A. Madan, M. Cebrian, D. Lazer, and A. Pentland, 'Social sensing for epidemiological behavior change', in *Proceedings of the 12th ACM international conference on Ubiquitous computing - Ubicomp '10*, Copenhagen, Denmark, 2010, p. 291, doi: 10.1145/1864349.1864394.

[18] B. Reeves *et al.*, 'Screenomics: A Framework to Capture and Analyze Personal Life Experiences and the Ways that Technology Shapes Them', *Human–Computer Interaction*, pp. 1–52, Mar. 2019, doi: 10.1080/07370024.2019.1578652.

[19] D. B. Dwyer, P. Falkai, and N. Koutsouleris, 'Machine Learning Approaches for Clinical Psychology and Psychiatry', *Annual Review of Clinical Psychology*, vol. 14, no. 1, pp. 91–118, 2018, doi: 10.1146/annurev-clinpsy-032816-045037.

[20] Department of Health and Social Care, 'Code of conduct for data-driven health and care technology', *GOV.UK*, 19-Feb-2019. [Online]. Available: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology. [Accessed: 15-Apr-2019].

[21] D. C. Mohr, M. Zhang, and S. M. Schueller, 'Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning', *Annu. Rev. Clin. Psychol.*, vol. 13, no. 1, pp. 23–47, May 2017, doi: 10.1146/annurev-clinpsy-032816-044949.

[22] M. Khan, G. Fernandes, U. Sarawgi, P. Rampey, and P. Maes, 'PAL: A Wearable Platform for Real-time, Personalized and Context-Aware Health and Cognition Support', *arXiv:1905.01352 [cs]*, May 2019.

[23] D. S. Watson *et al.*, 'Clinical applications of machine learning algorithms: beyond the black box', *BMJ*, p. l886, Mar. 2019, doi: 10.1136/bmj.l886.

[24] J. Morley and L. Floridi, 'How to Design a Governable Digital Health Ecosystem', *SSRN Journal*, 2019, doi: 10.2139/ssrn.3424376.

[25] T. Burns, *Our Necessary Shadow: The Nature and Meaning of Psychiatry*. London: Penguin, 2014.

[26] L. Johnstone and M. Boyle, 'The Power Threat Meaning Framework: An Alternative Nondiagnostic Conceptual System', *Journal of Humanistic Psychology*, p. 002216781879328, Aug. 2018, doi: 10.1177/0022167818793289.

[27] M. Coeckelbergh, 'Health Care, Capabilities, and AI Assistive Technologies', *Ethical Theory and Moral Practice*, vol. 13, no. 2, pp. 181–190, Apr. 2010, doi: 10.1007/s10677-009-9186-2.

[28] A. Fiske, P. Henningsen, and A. Buyx, 'Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy', *J Med Internet Res*, vol. 21, no. 5, p. e13216, May 2019, doi: 10.2196/13216.

[29] K. Lehavot, D. Ben-Zeev, and R. E. Neville, 'Ethical Considerations and Social Media: A Case of Suicidal Postings on Facebook', *Journal of Dual Diagnosis*, vol. 8, no. 4, pp. 341–346, Nov. 2012, doi: 10.1080/15504263.2012.718928.

[30] T. L. Beauchamp and J. F. Childress, *Principles of biomedical ethics*, 7th ed. New York, N.Y.: Oxford University Press, 2013.

[31] N. Sclater, A. Peasgood, and J. Mullan, 'Learning Analytics in Higher Education: A review of UK and international practice', Jisc, Apr. 2016.

[32] J. Kay, P. Reimann, E. Diebold, and B. Kummerfeld, 'MOOCs: So Many Learners, So Much Potential ...', *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 70–77, May 2013, doi: 10.1109/MIS.2013.66.

[33] BrainCo, 'FOCUSEDU', *brainco.tech*, 2019. [Online]. Available: https://www.brainco.tech/product/focusedu. [Accessed: 12-Jun-2019].

[34] N. Kosmyna, C. Morris, U. Sarawgi, T. Nguyen, and P. Maes, 'AttentivU: A Wearable Pair of EEG and EOG Glasses for Real-Time Physiological Processing', in *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, Chicago, IL, USA, 2019, pp. 1–4, doi: 10.1109/BSN.2019.8771080.

[35] R. Kokku, S. Sundararajan, P. Dey, R. Sindhgatta, S. Nitta, and B. Sengupta, 'Augmenting Classrooms with AI for Personalized Education', in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6976–6980, doi: 10.1109/ICASSP.2018.8461812.

[36] R. Beiter *et al.*, 'The prevalence and correlates of depression, anxiety, and stress in a sample of college students', *Journal of Affective Disorders*, vol. 173, pp. 90–96, Mar. 2015, doi: 10.1016/j.jad.2014.10.054.

[37] R. J. Watson and J. L. Christensen, 'Big data and student engagement among vulnerable youth: A review', *Current Opinion in Behavioral Sciences*, vol. 18, pp. 23–27, Dec. 2017, doi: 10.1016/j.cobeha.2017.07.004.

[38] R. Manthorpe, 'Artificial intelligence being used in schools to detect self-harm and bullying', *Sky News*, 21-Sep-2019. [Online]. Available: https://news.sky.com/story/artificial-intelligence-being-used-in-schools-to-detect-self-harm-and-bullying-11815865. [Accessed: 24-Sep-2019].

[39] P. I. Chow *et al.*, 'Using Mobile Sensing to Test Clinical Models of Depression, Social Anxiety, State Affect, and Social Isolation Among College Students', *J Med Internet Res*, vol. 19, no. 3, p. e62, Mar. 2017, doi: 10.2196/jmir.6820.

[40] L. R. Shade and R. Singh, '"Honestly, We're Not Spying on Kids": School Surveillance of Young People's Social Media', *Social Media + Society*, vol. 2, no. 4, p. 205630511668000, Nov. 2016, doi: 10.1177/2056305116680005.

[41] J. Tindale and O. Muirhead, 'Building Ethical Data Policies for the Public Good: Trust, Transparency and Tech', Policy Connect, UK, 2019.

[42] S. Koldijk, M. A. Neerincx, and W. Kraaij, 'Detecting Work Stress in Offices by Combining Unobtrusive Sensors', *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 227–239, Apr. 2018, doi: 10.1109/TAFFC.2016.2610975.

[43] E. Saner, 'Employers are monitoring computers, toilet breaks – even emotions. Is your boss watching you?', *The Guardian*, 14-May-2018.

[44] B. Skinner, G. Leavey, and D. Rothi, 'Managerialism and teacher professional identity: impact on well-being among teachers in the UK', *Educational Review*, pp. 1–16, Jan. 2019, doi: 10.1080/00131911.2018.1556205.

[45] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[46] T. Chamorro-Premuzic, R. Akhtar, D. Winsborough, and R. A. Sherman, 'The datafication of talent: how technology is advancing the science of human potential at work', *Current Opinion in Behavioral Sciences*, vol. 18, pp. 13–16, Dec. 2017, doi: 10.1016/j.cobeha.2017.04.007.

[47] L. Chen *et al.*, 'Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm', in *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, Tokyo, Japan, 2016, pp. 161–168, doi: 10.1145/2993148.2993203.

[48] J. W. Pennebaker, R. L. Boyd, K. Jordan, and Blackburn, K, 'The development and psychometric properties of LIWC2015.', University of Texas at Austin, Austin, TX, 2015.

[49] V. Silber-Varod, H. Kreiner, R. Lovett, Y. Levi-Belz, and N. Amir, 'Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals', presented at the Speech Prosody 2016, 2016, pp. 1211–1215, doi: 10.21437/SpeechProsody.2016-249.

[50] M. Modini *et al.*, 'The mental health benefits of employment: Results of a systematic meta-review', *Australas Psychiatry*, vol. 24, no. 4, pp. 331–336, Aug. 2016, doi: 10.1177/1039856215618523.

[51] J. H. Marler and J. W. Boudreau, 'An evidence-based review of HR Analytics', *The International Journal of Human Resource Management*, vol. 28, no. 1, pp. 3–26, Jan. 2017, doi: 10.1080/09585192.2016.1244699.

[52] T. Richardson, M. Jansen, and C. Fitch, 'Financial difficulties in bipolar disorder part 1: longitudinal relationships with mental health', *Journal of Mental Health*, vol. 27, no. 6, pp. 595–601, Nov. 2018, doi: 10.1080/09638237.2018.1521920.

[53] Financial Conduct Authority, 'Consumer Vulnerability', Financial Conduct Authority, Feb. 2015.

[54] K. Evans, 'Financial transactions data, AI and mental health: The challenge', *Money and Mental Health Policy Institute*, Jan-2019. [Online]. Available: https://www.moneyandmentalhealth.org/wp-content/uploads/2019/01/FCA-financial-transactions-data-discussion-note-footnotes-at-end.pdf. [Accessed: 16-May-2019].

[55] ICO, ‘Principle (c): Data minimisation’, *Information Comissioner’s Office*, 24-Apr-2019. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation/. [Accessed: 24-Sep-2019].

[56] K. Evans, ‘Money and Mental Health submission to the Department for Digital, Culture,Media and Sport’s consultation on the Centre for Data Ethics and Innovation’, *Money and Mental Health Policy Institute*, 05-Sep-2018. [Online]. Available: https://www.moneyandmentalhealth.org/wp-content/uploads/2018/09/20180905-Centre-for-Data-Ethics-and-Innovation-Consultation-Money-Mental-Health-response-.pdf. [Accessed: 24-Sep-2019].

[57] M. Toboso, ‘Rethinking disability in Amartya Sen’s approach: ICT and equality of opportunity’, *Ethics Inf Technol*, vol. 13, no. 2, pp. 107–118, Jun. 2011, doi: 10.1007/s10676-010-9254-2.

[58] T. Basu, ‘Facebook wants to fight teen suicide. Experts aren’t sure they’re doing it right’, *MIT Technology Review*, 11-Sep-2019. [Online]. Available: https://www.technologyreview.com/s/614301/facebook-wants-to-fight-teen-suicide-experts-arent-sure-theyre-doing-it-right/. [Accessed: 24-Sep-2019].

[59] N. N. Gomes de Andrade, D. Pawson, D. Muriello, L. Donahue, and J. Guadagno, ‘Ethics and Artificial Intelligence: Suicide Prevention on Facebook’, *Philos. Technol.*, vol. 31, no. 4, pp. 669–684, Dec. 2018, doi: 10.1007/s13347-018-0336-0.

[60] House of Commons and Science and Technology Committee, ‘Impact of social media and screen-use on young people’s health’, House of Commons Science and Technology Committee, Fourteenth Report of Session 2017–19, Jan. 2019.

[61] D. A. Ruch, A. H. Sheftall, P. Schlagbaum, J. Rausch, J. V. Campo, and J. A. Bridge, ‘Trends in Suicide Among Youth Aged 10 to 19 Years in the United States, 1975 to 2016’, *JAMA Netw Open*, vol. 2, no. 5, p. e193886, May 2019, doi: 10.1001/jamanetworkopen.2019.3886.

[62] World Health Organisation, ‘Preventing Suicide: A Global Imperative’, World Health Organisation, Luxembourg, 2014.

[63] R. Bruffaerts *et al.*, ‘Treatment of suicidal people around the world’, *Br J Psychiatry*, vol. 199, no. 1, pp. 64–70, Jul. 2011, doi: 10.1192/bjp.bp.110.084129.

[64] S. Velupillai *et al.*, ‘Risk Assessment Tools and Data-Driven Approaches for Predicting and Preventing Suicidal Behavior’, *Front. Psychiatry*, vol. 10, p. 36, Feb. 2019, doi: 10.3389/fpsyt.2019.00036.

[65] B. Goggin, ‘Inside Facebook’s suicide algorithm: Here’s how the company uses artificial intelligence to predict your mental state from your posts’, *Business Insider*, 06-Jan-2019. [Online]. Available: https://www.businessinsider.com/facebook-is-using-ai-to-try-to-predict-if-youre-suicidal-2018-12. [Accessed: 11-Jun-2019].

[66] J. Du *et al.*, ‘Extracting psychiatric stressors for suicide from social media using deep learning’, *BMC Med Inform Decis Mak*, vol. 18, no. S2, p. 43, Jul. 2018, doi: 10.1186/s12911-018-0632-8.

[67] G. Gkotsis *et al.*, ‘Characterisation of mental health conditions in social media using Informed Deep Learning’, *Sci Rep*, vol. 7, no. 1, p. 45141, Apr. 2017, doi: 10.1038/srep45141.

[68] R. M. Merchant *et al.*, ‘Evaluating the predictability of medical conditions from social media posts’, *PLoS ONE*, vol. 14, no. 6, p. e0215476, Jun. 2019, doi: 10.1371/journal.pone.0215476.

[69] World Health Organisation, ‘SIXTY-SIXTH WORLD HEALTH ASSEMBLY: Comprehensive mental health action plan 2013–2020 (WHA66.8)’. 27-May-2013.

[70] D. Agarwal, J. Bersin, and G. Lahiri, ‘Well-being: A strategy and a responsibility 2018 Global Human Capital Trends’, *Deloitte Insights*, 28-Mar-2018. [Online]. Available: https://www2.deloitte.com/insights/us/en/focus/human-capital-trends/2018/employee-well-being-programs.html. [Accessed: 11-Jun-2019].

[71] The Lancet Digital Health, ‘An app a day is only a framework away’, *The Lancet Digital Health*, vol. 1, no. 2, p. e45, Jun. 2019, doi: 10.1016/S2589-7500(19)30031-7.

[72] M. E. Larsen *et al.*, ‘Using science to sell apps: Evaluation of mental health app store quality claims’, *npj Digit. Med.*, vol. 2, no. 1, p. 18, Dec. 2019, doi: 10.1038/s41746-019-0093-1.

[73] M. E. Larsen, J. Nicholas, and H. Christensen, ‘A Systematic Assessment of Smartphone Tools for Suicide Prevention’, *PLoS ONE*, vol. 11, no. 4, p. e0152285, Apr. 2016, doi: 10.1371/journal.pone.0152285.

[74] D. R. Bateman *et al.*, ‘Categorizing Health Outcomes and Efficacy of mHealth Apps for Persons With Cognitive Impairment: A Systematic Review’, *J Med Internet Res*, vol. 19, no. 8, p. e301, Aug. 2017, doi: 10.2196/jmir.7814.

[75] A. Ferretti, E. Ronchi, and E. Vayena, ‘From principles to practice: benchmarking government guidance on health apps’, *The Lancet Digital Health*, vol. 1, no. 2, pp. e55–e57, Jun. 2019, doi: 10.1016/S2589-7500(19)30027-5.

[76] P. Henson, G. David, K. Albright, and J. Torous, ‘Deriving a practical framework for the evaluation of health apps’, *The Lancet Digital Health*, vol. 1, no. 2, pp. e52–e54, Jun. 2019, doi: 10.1016/S2589-7500(19)30013-5.

[77] L. Floridi *et al.*, ‘AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations’, *Minds and Machines*, vol. 28, no. 4, pp. 689–707, Dec. 2018, doi: 10.1007/s11023-018-9482-5.

[78] A. Jobin, M. Ienca, and E. Vayena, ‘The global landscape of AI ethics guidelines’, *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[79] G. Dworkin, *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press, 1988.

[80] A. Sen, *The Idea of Justice*. London: Penguin, 2010.

[81] W. A. Anthony, ‘Recovery from mental illness: The guiding vision of the mental health service system in the 1990s.’, *Psychosocial Rehabilitation Journal*, vol. 16, no. 4, pp. 11–23, 1993, doi: 10.1037/h0095655.

[82] A. Sen, ‘Well-Being, Agency and Freedom: The Dewey Lectures 1984’, *The Journal of Philosophy*, vol. 82, no. 4, pp. 169–221, 1985, doi: 10.2307/2026184.

[83] A. Freeman *et al.*, ‘A cross-national study on gender differences in suicide intent’, *BMC Psychiatry*, vol. 17, Jun. 2017, doi: 10.1186/s12888-017-1398-8.

[84] M. C. Marill, ‘Why Tracking Your Symptoms Can Make You Feel Worse’, *Wired*, 30-May-2019.

[85] S. Makin, ‘Searching for digital technology’s effects on well-being’, *Nature*, 28-Nov-2018. [Online]. Available: https://www.nature.com/articles/d41586-018-07503-w. [Accessed: 23-May-2019].

[86] B. P. Turnwald, J. P. Goyer, D. Z. Boles, A. Silder, S. L. Delp, and A. J. Crum, ‘Learning one’s genetic risk changes physiology independent of actual genetic risk’, *Nat Hum Behav*, vol. 3, no. 1, pp. 48–56, Jan. 2019, doi: 10.1038/s41562-018-0483-4.

[87] C. Odgers, ‘Smartphones are bad for some teens, not all’, *Nature*, vol. 554, no. 7693, p. 432, Feb. 2018, doi: 10.1038/d41586-018-02109-8.

[88] J. Kaffash, ‘Average GP waiting times remain at two weeks despite rescue measures’, *Pulse Today*, 02-Jun-2017. [Online]. Available: http://www.pulsetoday.co.uk/your-practice/practice-topics/access/average-gp-waiting-times-remain-at-two-weeks-despite-rescue-measures/20034534.article. [Accessed: 01-Jul-2019].

[89] A. Downey, ‘Funding model for GP at Hand “not appropriate”, Ipsos Mori review finds’, *Digital Health*, 29-May-2019. [Online]. Available: https://www.digitalhealth.net/2019/05/gp-at-hand-funding-not-appropriate-ipsos-mori/. [Accessed: 01-Jul-2019].

[90] J. Morley and L. Floridi, ‘Against Empowerment: How to Reframe the Role of mHealth Tools in the Healthcare Ecosystem (Draft)’. 2019.

[91] T. Hindmarch, M. Hotopf, and G. S. Owen, ‘Depression and decision-making capacity for treatment or research: a systematic review’, *BMC Medical Ethics*, vol. 14, no. 1, Dec. 2013, doi: 10.1186/1472-6939-14-54.

[92] J. Nicholas, M. E. Larsen, J. Proudfoot, and H. Christensen, ‘Mobile Apps for Bipolar Disorder: A Systematic Review of Features and Content Quality’, *J Med Internet Res*, vol. 17, no. 8, p. e198, Aug. 2015, doi: 10.2196/jmir.4581.

[93] C. Burr and J. Morley, ‘Empowerment or Engagement? Digital Health Technologies for Mental Healthcare’, Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3393534, May 2019.

[94] D. C. Mohr, K. R. Weingardt, M. Reddy, and S. M. Schueller, ‘Three Problems With Current Digital Mental Health Research . . . and Three Things We Can Do About Them’, *PS*, vol. 68, no. 5, pp. 427–429, May 2017, doi: 10.1176/appi.ps.201600541.

[95] D. Peters, R. A. Calvo, and R. M. Ryan, ‘Designing for Motivation, Engagement and Wellbeing in Digital Experience’, *Front Psychol*, vol. 9, May 2018, doi: 10.3389/fpsyg.2018.00797.

[96] K. McVeigh, ‘Samaritans Twitter app identifying user’s moods criticised as invasive’, *The Guardian*, 04-Nov-2014.

[97] M. Taddeo, ‘Trust in Technology: A Distinctive and a Problematic Relation’, *Know Techn Pol*, vol. 23, no. 3–4, pp. 283–286, Dec. 2010, doi: 10.1007/s12130-010-9113-9.

[98] M. Taddeo, ‘Trusting Digital Technologies Correctly’, *Minds & Machines*, vol. 27, no. 4, pp. 565–568, Dec. 2017, doi: 10.1007/s11023-017-9450-5.

[99] K. Huckvale, J. T. Prieto, M. Tilney, P.-J. Benghozi, and J. Car, ‘Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment’, *BMC Med*, vol. 13, no. 1, p. 214, Dec. 2015, doi: 10.1186/s12916-015-0444-y.

[100] J. W. Penney, ‘Chilling effects: Online surveillance and Wikipedia use’, *Berkeley Tech. LJ*, vol. 31, p. 117, 2016.

[101] M. Taddeo and L. Floridi, ‘The Debate on the Moral Responsibilities of Online Service Providers’, *Sci Eng Ethics*, vol. 22, no. 6, pp. 1575–1603, Dec. 2016, doi: 10.1007/s11948-015-9734-1.

[102] D. Beer, D. J. Redden, D. B. Williamson, and D. S. Yuill, ‘What is online targeting, what impact does it have, and how can we maximise benefits and minimise harms?’, Centre for Data Ethics and Innovation, 2019.

[103] HM Government, *Online Harms White Paper*. UK: Open Government, 2019.