Empowerment or Engagement? Digital Health Technologies for Mental Healthcare

Christopher Burr[1] and Jessica Morley[1]

[1]Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, OX1 3JS, United Kingdom
Correspondence email: christopher.burr@oii.ox.ac.uk

Abstract

We argue that while digital health technologies (e.g. artificial intelligence, smartphones, and virtual reality) present significant opportunities for improving the delivery of healthcare, key concepts that are used to evaluate and understand their impact can obscure significant ethical issues related to patient engagement and experience. Specifically, we focus on the concept of empowerment and ask whether it is adequate for addressing some significant ethical concerns that relate to digital health technologies for mental healthcare. We frame these concerns using five key ethical principles for AI ethics (i.e. autonomy, beneficence, non-maleficence, justice, and explicability), which have their roots in the bioethical literature, in order to critically evaluate the role that digital health technologies will have in the future of digital healthcare.

Keywords: digital health technology; empowerment; patient engagement; mental health; artificial intelligence; bioethics.

Introduction

The way that healthcare services are set to operate is likely to change drastically over the next decade as a result of key digital health technologies (DHTs) (e.g. telemedicine, wearables and smartphones, artificial intelligence, and genomics). Some of these technologies are being deployed within formal healthcare settings and are already impacting the way that patients access healthcare services (e.g. telemedicine and digital therapies), how they are monitored or diagnosed (e.g. sensors/wearables, smartphones, social media), and how healthcare services are governed and administered (e.g. electronic health records, machine learning) (The Topol Review Board, 2019). Other technologies are being used by individuals in more informal ways, embedded within their daily activities as part of a more personal concern for self-tracking of health and well-being (Lupton, 2016).

In this paper, we explore the ethical impact of some of these key technologies and the concepts used to critically evaluate them, focusing primarily on their role in mental healthcare in the United Kingdom—though many of the issues we discuss are applicable to wider healthcare services. According to the Adult Psychiatric Morbidity Survey (NatCen Social Research, 2016), one in six adults surveyed in England in 2014 met the criteria for a common mental disorder (CMD).[1] The World Health Organisation (WHO) also notes that depression is the single largest contributor to global disability and a major contributor to suicide deaths, which number close to 800 000 per year. As part of NHS England's long-term plan, significant investment for mental health services has been promised, with data and technology set to play a central role in transforming their delivery (NHS England, 2019). This investment is vital, as mental healthcare is in urgent need of new approaches, and digital technologies, such as artificial intelligence (AI), will likely have a critical role in easing the burden that mental health conditions have on individuals and society. Furthermore, this specific focus is important from a parity of care perspective, in order to ensure that the opportunities associated with DHTs are equally distributed. However, aiming for parity does not mean that we should assume the implications, both positive and negative, of the increasing use of DHTs are equal. Mental healthcare poses unique ethical challenges due to the need to consider wider psychological and social factors, many of which interact with biological factors in complex ways that are not fully understood[2]. We approach these challenges from the perspective of a broader concern about the nature of *patient empowerment*—a concept that has received a large amount of attention in recent years (Chiauzzi et al., 2016; Spencer, 2015; Bravo, Barr, Scholl, Elwyn, & McAllister, 2015)—and in relation to the key technologies identified as having a central role to play in the delivery of mental healthcare services.

In section 1, we discuss the idea that technology can empower service users to take charge of their own digitally-mediated care, supported by myriad streams of user-generated data and co-curated with various DHTs, including AI. This idea has caught the attention of many developers, stakeholders, and policy makers, but the empowerment narrative rests on some questionable conceptual and ethical foundations. We will argue that genuine empowerment depends on the prior removal of certain barriers to *engagement*, which patients suffering from a variety of mental health conditions face. To support this argument, in section 2, we adopt a bioethical perspective in order to critically evaluate the role that DHTs play in removing these barriers, as well as the

[1] The report defines a CMD as comprising different types of depression and anxiety, which cause marked emotional distress and interfere with daily function, but do not usually affect insight or cognition. CMDs are typically contrasted with major psychiatric disorders, such as schizophrenia (NatCen Social Research, 2016).

[2] For instance, the acknowledgement that "neurobiology does not fully account for the emergence of mental distress" formed the basis of one of the criticisms brought against the DSM-V in an open letter signed by 15000 individuals of 50 professional organisations (Kamens, Elkins, & Robbins, 2017, p. 682).

possible unintended consequences that arise from their implementation. In section 3, we stress that if harder governance measures are adopted to protect people from the unintended consequences that present the highest level of risk, these measures must be developed in a way that is tolerant of value pluralism. In section 4, we conclude with a brief summary of the main points discussed in the article.

1 Mental Health and Empowerment

A recent review commissioned by the previous UK Secretary of State for Health and Social Care, Jeremy Hunt, explores how technological developments are likely to impact the future of healthcare in the NHS (The Topol Review Board, 2019). Included alongside this review is an individual report that focuses specifically on mental healthcare and the key DHTs[3] that are identified as likely to have a significant impact over the next 20 years (Foley & Woollard, 2019). The role and scope of these technologies differs widely but the report notes that they "have the potential to reduce the administrative burden, allow treatment in more convenient settings, and *empower* patients and their carers to take on some of the tasks currently performed in the clinic" (Foley & Woollard, 2019, p. 25, emphasis added).

The use of the term 'empower' here is important, and reflects a growing emphasis and usage of the concept, most notably within the literature discussing digital health and well-being (Burr, Taddeo, & Floridi, 2019; Morley & Floridi, 2019). In the case of mental healthcare, a significant challenge for promoting empowerment is the fact that certain psychiatric disorders impact the individual's decisional capacity, affecting their choice of whether to engage with some service (e.g. an online CBT programme), or, more broadly, restricting their ability to make their own healthcare decisions.[4] Different disorders impact decisional capacity in myriad ways. For instance, in a review of the medical ethical and empirical literature on depression and decisional capacity, Hindmarch et al. (2013) found that being in a depressive episode impacts an individual's ability to *appreciate* the significance of information that may be relevant to healthcare decisions. In other words, information that may be treated similarly from a quantitative perspective (i.e. it is of equal quality and quantity) is not always the same from a qualitative perspective (in terms of meaning) (Floridi, 2010). The latter perspective depends on the individual who is consuming the information, as well as the prior beliefs they bring to bear on the information, how it is perceived

[3] The 13 DHTs the report identifies are telemedicine, sensors/wearables, smartphones, digital therapies, social media, genotyping microarrays, neuroimaging, electronic health records and patient health records, healthcare data collections, natural language processing, artificial intelligence, virtual reality (VR), and augmented reality (AR)

[4] Typically, decisional capacity is divided into four sub-categories: the capacity to express a choice, the ability to understand relevant information, the ability to appreciate the significance of the information, and the ability to reason with the information (Charland, 2015; Grisso & Appelbaum, 1998).

and what affordances arise between the user and their environment (Nagy & Neff, 2015). These types of assessment and considerations are important for identifying the specific barriers that exist in the case of specific mental health conditions, which may prevent DHTs from increasing patient empowerment.

However, it is not sufficient to restrict our focus in this paper to the issues of decisional capacity alone—understanding and critically evaluating the concept of empowerment requires a broader focus. A key concern is that so-called 'empowering technologies' focus too narrowly on monitoring and providing information to an individual, on the assumption that a more informed process of deliberation is sufficient for empowerment (Morley & Floridi, 2019). However, there are many problems with this assumption.

First, and foremost, it is not clear *exactly how* digitally-mediated access to information will empower people. This is primarily because, despite its common use, empowerment is a term that is used both loosely and inconsistently (Roberts, 1999) and is, consequently, embedded in a range of competing discourses that have highly variable aims: from the need to give people choice to the importance of providing people with an opportunity to change their position in society (Starkey, 2003). All these variable conceptualizations are in use in the wider health promotion discourse (Sheehan, 2014) but, as has been highlighted elsewhere (Morley & Floridi, 2019), the narrative that is used in the context of digitising healthcare services (including mental health services) positions empowerment as a self-reflexive and transformative process (Garcia et al*.,* 2014).

At first, this view of empowerment might not appear to be problematic. Indeed, there have been some early findings that this process can result in, at least moderate, positive impacts on the mental health of adolescents (Kenny, Dooley, & Fitzgerald, 2015) if appropriate evidence-based design recommendations are followed (Bakker, Kazantzis, Rickwood, & Rickard, 2016). This means that mental health-focused DHTs that aim to 'empower' individuals by taking action to actively improve their mental health through a process of self-reflection are likely to play an important part in the future of mental health care, especially in terms of making mental health support more accessible and reducing barriers to seeking help (Bakker et al., 2016). These opportunities should not be ignored. However, this conception of empowerment raises unique ethical issues, such as how it can be leveraged in ways that overlook socioeconomic factors that determine whether an individual can benefit from the use of a mental health DHT in the manner described. Moreover, the self-reflexive process presumes that an individual actively *wants* and *feels able to* engage with the process in the first place (e.g. to download an app, open it and register a user name)—this presumption is far from guaranteed in a wide variety of mental health disorders.

Thus, the overarching argument tends to ignore the fact that there are many factors that moderate an individual's ability and motivation to even *engage* with this active process of self-reflection.

Such moderating factors (or variables) are well articulated by the Engagement Capacity Model (ECM) (Sieck, Walker, Retchin, & McAlearney, 2019), which stresses that an individual may fail to engage with healthcare services if they feel unable to do so due to (a) low resources, (b) low self-efficacy (competence), or (c) low willingness. These variables are themselves the result of a dynamic interplay between an individual, their environment, and the corresponding behaviours creating a complex feedback loop where each of these factors constantly influence each other. For example, a change in an individual's environment, such as a reduction in income and consequential decision that paying for a smartphone contract is no longer affordable, might reduce the amount of resources they feel they have available to them to improve their mental health, in turn reducing their level of willingness to engage with the mental health services that are available (e.g. those accessible via a desktop computer at the library), making them less likely to consider engaging with the self-reflexive process of empowerment, and as a consequence lowering the confidence they have in their capacity (self-efficacy) to take the steps necessary to improve their mental health.

Genuine empowerment, therefore, requires attending to the wider psychosocial factors that could constrain an individual's ability to engage with healthcare services, both online and offline. For instance, far greater attention needs to be paid to the unequal distribution of mHealth resources throughout society and the existence of considerable perverse incentives within the system that will discourage the lowering of barriers to adoption. For example, while it may be better for the system and for the individual themselves to 'self-treat' at home through the use of a mindfulness app there are likely to still be incentives in the system for health practitioners to want to see the person in a clinical setting so that it generates a payment.

Our intention, in highlighting these complex sociotechnical and (later) bioethical issues, is not to present the future of digitally-enhanced mental health service provision as dystopian or impossible to achieve. We believe that it is possible to capitalise on the opportunities presented by DHTs in a responsible manner, but this requires making it clear that DHTs are not neutral technologies.[5] As such, there is a responsibility on all parts of the system to encourage the design of DHTs that, in complete awareness of the complex space within which they operate, actively re-ontologise the way that mental health care services are delivered, with the goal of genuine

[5] In (Morley & Floridi, 2019), one of the authors defends a view of framing DHTs as 'digital companions', which can have significant (positive and negative) effects on relationships key to maintaining positive mental health, such as those between: (a) clinical advice and behaviour change; (b) perception of self and behaviour change; (c) need for social interaction and desire to socialise.

improvement to the patient experience, as determined within the bounds of long-established bioethical principles that we will now discuss.

2 Engagement and DHTs: Five Principles to Guide Critical Evaluation

In this section we present several conceptual and ethical concerns that need to be addressed if we are to achieve the goal of increasing patient engagement and, in turn, empowerment. These concerns are structured according to the principles outlined in (Floridi et al., 2018), which comprise the four traditional principles of biomedical ethics [i.e. beneficence, non-maleficence, autonomy, and justice (Beauchamp & Childress, 2013)] as well as an additional principle (i.e. explicability) that is included to capture specific ethical issues that arise with the use of AI. These five principles were found to be well-represented in several significant policy documents that address the ethical issues with AI (see Floridi et al., 2018), and are well-suited to the present article because of their grounding in biomedical ethics.

2.1. Autonomy

In biomedical ethics, the principle of autonomy incorporates respect for both an individual's *right to decide* and *freedom of whether to decide* (Beauchamp & Childress, 2013). The motivation behind the latter component, as Sen (2010, p. 18) notes, is that "[t]he freedom to choose our lives can make a significant contribution to our well-being, but going beyond the perspective of well-being, the freedom itself may be seen as important […] we are under no obligation to seek only our own well-being, and it is for us to decide what we have good reason to pursue." In short, although humans have a right to decide, we also have the freedom to choose how and whether to exercise that right. However, freedom alone is insufficient for autonomy—agency is also required and provides the basis for social recognition of one's right to decide, including the capacity to express informed consent.

Contemporary theories of relational autonomy maintain that an individual's agency, or capacity for intentional action, is in large part determined by their sociocultural environment.[6] These approaches contrast with procedural accounts of autonomy, which view autonomous decision-making in more cognitivist terms and may downplay the significance of the wider environmental dynamics that contribute to overt choice behaviour (see Owens & Cribb, 2013 for a discussion).

[6] A related idea is captured in the well-known *capability approach*, which focuses on the real opportunities for action that different sociocultural environments afford, the individual differences in people's abilities (or capacity) to transform resources in ways conducive to their well-being, and the unequal distribution of such opportunities throughout society (Sen, 2010).

Relational theories of autonomy help bring into stark relief the need to design and evaluate DHTs at a level of abstraction that articulates the socially embedded (or situated) nature of the user, in order to fully appreciate the interpersonal differences in capacity for engagement (e.g. time demands, literacy levels, finances, social support). For example, Lucas et al. (2017, p. 2, emphasis added) explored whether virtual human interviewers could "increase *willingness* of service members to report PTSD symptoms", by reducing barriers to engagement that may result from the perceived stigma that comes from reporting symptoms to a human interviewer. They show how such a technology has the potential to increase an individual's relational autonomy, by creating a wider set of opportunities for seeking treatment and respecting the barriers to engagement that certain mental health conditions present—in this case the barrier was low willingness caused by a concern regarding perceived stigma. In this manner, retaining an emphasis on relational autonomy may help ensure that DHTs do not end up embodying overly-individualistic values of what it means to 'live well' but rather help demonstrate the prudential value of social relatedness.

Thus, although DHTs can create new opportunities available to individuals by altering the landscape of affordances that a user perceives (Bruineberg & Rietveld, 2014), respect for individual autonomy also requires acknowledgement of the different values that individuals bring to bear when choosing whether to engage. This is often embodied in the idea that the *right* to choose is not a *duty* to choose (Beauchamp & Childress, 2013). This aspect of autonomy can cause difficulties for technology designers and developers, as well as the healthcare professionals that use their products. Two concerns are significant.

Firstly, there is a concern that can arise when insufficient consideration is given to the scope of autonomy. For example, a patient may autonomously decide to *disempower* themselves, in order to have someone else (e.g. their doctor or caregiver) make decisions on their behalf. Alternatively, an individual experiencing depression may be fully informed about their mental health and the options available to them in terms of recovery, but nevertheless autonomously decide not to engage with the variety of DHTs available to them—their mental health may be an important part of their self-identity and how they make sense of the world.[7] Examples such as these pose challenges for determining the efficacy of a DHT. As White et al. (White, Imperiale, & Perera, 2016, p. 2) note, delivering mental health services is problematised by the challenge of specifying what constitutes a "'good outcome' for people in the particular contexts in which they are living their lives". An individual may autonomously decide that their own journey of recovery

[7] This perspective is captured by the *recovery approach* (Anthony, 1993, p. 527), which maintains that recovery "is 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."

requires learning to live within acceptable limitations that are intrinsically chosen and governed, rather than based on extrinsic and optimal standards represented by the outcome measure chosen by the healthcare provider.

Secondly, there is a concern that can arise when too much consideration is given to the scope of autonomy. For example, if a technology developer focuses too narrowly on the day-to-day decisions of a user (e.g. whether to adhere to a self-directed course of therapy delivered via an mHealth technology), they may fail to appreciate how a single decision fits within a patient's broader healthcare regime. As (Kukla, 2005, p. 37) states: "The bulk of our health care activities take the form, not of crisis management and punctate decision-making, but of ongoing practices, including large amounts of self-management and surveillance, wherein we are inducted into standards set by medical institutions with which we have prescribed forms of direct contact." The point here is that the individual decision of whether to adhere to a course of treatment on any day, typically made multiple times during a course of treatment, may be the wrong level of abstraction to focus on when determining whether a user's autonomy is respected. The meaningful choice that requires consideration is the initial choice of whether to engage in a course of treatment and how to integrate the treatment into ongoing practices, as opposed to the subsequent choices that may result from individual prompts (or nudges)—perhaps delivered via smartphone notifications and serving to remind a user to continue with a self-determine course of therapy (e.g. CBT).

2.2. Beneficence

The principle of beneficence typically emphasises the promotion of patient welfare but can also be extended to include the welfare of the caregivers. Consideration of how to 'do good' in the context of healthcare and DHTs, therefore, need not, and perhaps should not, be limited to the individual patient—deploying a new DHT in a healthcare pathway can be highly beneficial for patients, but could prove to be overly-demanding for clinical staff.

Novel technologies are creating new opportunities to 'do good', by unlocking possible treatment options that did not exist previously (Fernández-Caballero et al., 2017). However, ensuring that the principle of beneficence is upheld when designing, implementing, and using DHTs requires that we have some way of measuring a wide range of outcomes and that the measures used are suitable for the context in which they are deployed (e.g. clinical, epidemiological, or allocational decisions)[8]. This can prove to be challenging for a number of reasons.

[8] See (Hausman, 2015) for an argument that claims that no single measure can adequately capture the value of health outcomes across all three contexts.

In order to determine whether DHTs 'do good' it is important to consider how effective they are in bringing about their stated goals—this includes a *comparative evaluation* against relevant existing services. However, depending on the type of comparative analysis being conducted, certain measures may prove to be limited. For instance, alongside other key performance indicators that commissioners use to assess the overall quality of care, patient-reported outcome measures (PROMs) can provide a valuable source of information about a patient's subjective attitudes towards a procedure or treatment. As Nelson et al. (2015, p. 1) notes, "the systematic use of information from PROMs leads to better communication and decision making between doctors and patients and improves patient satisfaction with care". However, PROMs can be either specific or generic, and in the case of the former, can be specific in myriad ways (e.g. disease- or condition-specific, population-specific)[9]. This leads to certain constraints on their applicability. For instance, if the PROM has been validated for a specific population (e.g. elderly patients) this can rule out comparisons with the wider population due to differences in the dimensions being assessed (e.g. an instrument for measuring adolescent well-being will focus on different factors from well-being of elderly patients due to different expectations concerning typical levels of functioning.

DHTs, such as sensors/wearables, smartphones, and social media are enabling new forms of data collection and measurement when combined with techniques such as big data analytics and machine learning. However, DHTs are not immune to the aforementioned limitations on measurement, and technology designers must consider what to measure and how best to measure it during the design process. Furthermore, technology designers also face additional ethical challenges that go beyond the choice of measurement tool.

One such challenge, is the need to balance the evidence-standard required of health interventions, exemplified by the reproducible results of randomised-controlled trials, with the opportunity presented by DHTs to deliver far more personalised care. If too much emphasis is put on optimising the outcome for an individual 'user' during the designing, testing and evaluating phases, then it would be ethically wrong to launch that product at scale on the market—where it would be used by individuals with grossly different socioeconomic circumstances—due to the chances of it having a negative impact on those that do not match the 'profile' of the individual for which it was tailored. If, however, the opposite was true and the focus was on reproducibility of the results, we risk missing the opportunity to improve outcomes for individuals who have more specific needs that have, up until now, been unmet by the provision of generic mental health services.

[9] Examples can be found at: http://phi.uhce.ox.ac.uk/inst_types.php

Another key ethical concern is that although subjective reports are a valuable source of information about how a patient may evaluate the positive impact of some intervention, treatment or therapy, the way the information is collected, stored, and used could raise concerns among users. This is especially important in the case of mental health, where concerns over privacy can be particularly significant. For instance, a DHT designer may be aware that some biometric signals carry mutual information about an individual's mood or emotional state or that natural language processing techniques can be used to infer information about common mental health disorders such as depression and anxiety (Burr & Cristianini, 2019). Moreover, they may be aware that such techniques can be used to bypass the need for explicit user input (e.g. completion of a questionnaire), allowing them to be used at scale without high costs. Although the reliability and validity of using digital footprints or biometric signals to bypass traditional forms of psychometric assessment is currently inadequate for clinical use, this does not prevent the use of such techniques in the wider ecosystem of mHealth apps and IoT devices (Bellet & Frijters, 2019). As such, from the perspective of a designer, the decision not to utilise such techniques within a health and wellness app could be judged as a missed opportunity and a failure to "do good".

However, the use of such a technique to measure the effectiveness of a possible intervention may not necessarily be seen the same way by the user, who may have decided to present themselves in public in such a way that their mental health condition is not obvious to their friends, family or colleagues. This ability to choose the "face" we wear in public, therefore, could be undermined by a designer's attempt to use novel techniques (e.g. big data and machine learning) to measure our inner lives by bypassing the need for explicit feedback (e.g. a self-reported questionnaire) (Bellet & Frijters, 2019; Burr & Cristianini, 2019). In turn, the discovery of such techniques by a user, who may have wished to keep their mental health condition private, could lead to self-surveillance of future online interactions that end up overriding the initial desire to "do good".[10] The simple point here, well-known to bioethicists, is that consideration of how best to meet the principle of beneficence goes hand in hand with a requirement of considering the possible risks of harm.

2.3. Non-Maleficence

Avoiding harm is sometimes treated as an overriding principle in the delivery of healthcare (i.e. 'above all do no harm'), although there are many instances of where this fails to be useful in practice and sometimes morally indefensible in principle (see Beauchamp & Childress, 2013). As

[10] This is likely one of the primary motivations behind the backlash to a study by Facebook that demonstrated how user's emotional states could be manipulated (Kramer, Guillory, & Hancock, 2014).

such, it is typically agreed in bioethics that independent of context, there is no *a priori* rank ordering of the norms of beneficence and non-maleficence. Nevertheless, the bias towards the principle of non-maleficence can be seen in the NICE Evidence Standards Framework, which is used for evaluating DHTs deployed in the NHS and places a significant emphasis on demonstrating how proposed DHTs should be evaluated according to the proportional risk that their use would pose within the healthcare system (Greaves et al., 2018). The framework is founded on a *proportionate approach to risk*, which categorises DHTs according to their function so that more rigorous standards are applied to DHTs that have the potential for causing greater harm. For example, DHTs that are designed for 'active monitoring' of patients—included in the highest risk tier of the framework—should ideally be supported by a high-quality randomised controlled study that demonstrates how the DHT has comparative utility according to relevant clinical outcomes in the target population, using validated condition-specific measures. Again, here we see the need to consider the scope of measures deployed for assessing DHTs and their potential impacts on service users (see previous section).

Unfortunately, the NICE framework notes that its evaluative scope is limited and less relevant to DHTs that are "downloaded or purchased directly by users (such as through app stores)" and is "not [yet] designed for use with DHTs that incorporate artificial intelligence using adaptive algorithms" (National Institute for Health and Care Excellence, 2018). This limited scope is understandable when we consider the variation in *standards of due care* that are relied upon in order to avoid negligence. In the first instance, an app developer does not have the same professional duty of care to an individual that a doctor does to a patient. In the second, the adaptive nature of the algorithms in question may place epistemic limits on the duty of care that can be exercised due to the lack of explainability inherent in some forms of AI (Watson et al., 2019). Nevertheless, the fact that app developers are not yet beholden to the same duty of care that governs the obligations of a formal caregiver does not mean that they are exempt from giving *appropriate consideration* to possible risks and benefits that their product may cause. How we delineate and specify the concept of 'appropriate consideration' though, must instead make reference to a broader ethics of social responsibility. It is, perhaps, for these reasons that so many organisations are currently at work trying to specify codes of conduct (Department of Health and Social Care, 2019) or empirically-informed design guidelines (Calvo & Peters, 2014), which can help provide ethical support for the development and use of DHTs in wider contexts.

A central challenge for the development of such ethical frameworks is how to deal with trade-offs between maximising opportunities and minimising risks. Several specific trade-offs arise in relation to the *over-use* and *under-use* of DHTs for mental health.

Firstly, and in relation to the *over-use* of DHTs, while there is broad consensus that CBT is an effective treatment for common mental health disorders such as anxiety and depression, CBT is not harmless. As such, there are potential risks that could emerge from over-use of DHTs for CBT, such as deterioration of existing symptoms, emergence of new symptoms, and strains on family relations (Schermuly-Haupt, Linden, & Rush, 2018). Such risks may also help explain the findings of a study performed by Breedvelt et al. (2019), which analysed GP's attitudes to mHealth interventions for depression, and found that GPs thought that unguided use of such interventions (e.g. automated self-care) is likely to be less effective than guided care. In short, although the proliferation of therapy-based apps may provide greater access, and in turn reduce barriers to engagement for those who need support, there is a trade-off between improved access or scalability on the one hand, and potential decrease in efficacy and possible increase in the risk of harm on the other.

Another instance of the possible over-use of DHTs can be found in the ongoing debate around the automated monitoring of suicidal ideation on social media. Others have already raised concerns about the ethical challenges raised by mental health professionals using social media as a way of monitoring patients, including the tension between the principles of beneficence and non-maleficence (Lehavot, Ben-Zeev, & Neville, 2012). A notable concern in relation to using DHTs to automatically monitor individuals is that the reliability and validity of such techniques is currently insufficient for clinical use (Burr & Cristianini, 2019). Therefore, there is a risk that if deployed at scale, such techniques would likely lead to a high-rate of false positives, which in turn could result in the over medicalisation and stigmatisation of otherwise healthy and normal attitudes, behaviours and cognitions.[11]

Secondly, and in relation to the *under-use* of DHTs for mental health, it can be argued that an over-cautious approach to mitigating risk can stifle research and lead to harm by failing to advance treatment options. This is particularly relevant in the case of IoT devices and ubiquitous computing where there is a genuine opportunity to gather valuable environmental data (or 'ecologically valid' data) that could help researchers to understand how the environment affects the presentation of mental health disorders. For instance, one epidemiological study used Google Trends data and NHS prescription data for antidepressants to explore the distribution and prevalence of seasonal affective disorder (Lansdall-Welfare, Lightman, & Cristianini, 2019), while another used Twitter data (i.e. NLP) to discover a diurnal variation in emotions (Lansdall-Welfare,

[11] Such a concern is reminiscent of concerns raised in an open letter to the DSM-V, which noted how lowering diagnostic thresholds for certain categories (e.g. ADD) could lead to *epidemiological inflation*, and in some cases could lead to the inappropriate prescription of pharmacological substances to vulnerable populations (e.g. the use of neuroleptics in children diagnosed with disruptive mood dysregulation disorder). (Kamens, Elkins, & Robbins, 2017, p. 682)

Lightman, & Cristianini, 2019). Both of these population-level studies were done using publicly-available datasets but required extensive forms of data collection and expensive data storage. Greater collaboration between researchers and technology companies who already have access to this data, as well as additionally valuable meta-data that is not publicly available, would likely extend our scientific and medical knowledge about possible risk factors for mental health disorders. It is also possible that, as more is understood about how environmental factors co-determine mental health, DHTs and the information that we gather from them could contribute to raising the standards of due care—as the evidence base grows the number of unintended consequences from lack of knowledge shrinks. This potential for DHTs, particularly those involving the use of artificial intelligence to spark human curiosity that can lead to better outcomes (Holm, 2019), is one reason why governments should take a proportionate risk-based approach to the ways in which such DHTs are regulated.

2.4. Justice

Although DHTs could be used to optimise back-end operational processes for efficiency purposes [e.g. to release time for clinicians to ensure the right care is delivered in the right place at the right time and to improve equity of care (Nelson, Herron, Rees, & Nachev, 2019)], it is also likely that their use will have impacts on society in ways that are unequally distributed. For instance, there could be economic inequalities [e.g. those who have to rely on free-apps are far more likely to experience privacy harms due to the exploitative monetisation of their data (Polykalas & Prezerakos, 2019)] or epistemic inequalities (e.g. those with higher levels of health and media literacy who are better placed to make use of developments). However, there are also more specific concerns that can be discussed.

In addition to DHTs that are employed and embedded within formal healthcare systems, there are many more DHTs that can be accessed through third-party services (e.g. app stores). The quality and variety of these DHTs is vast, including apps that teach mindfulness-based stress reduction, online community support forums, and services that connect users with chatbots or human wellness coaches. We here focus on the latter.

Wellness coaching often has similar goals to formal healthcare services and can include NLP-based chatbots that deploy some form of CBT or paid-for online services that connect users to another human. It has been reported that some of these services are reliant on unlicensed "coaches" who deliver simple forms of motivational therapy or emotional-health coaching (Barras, 2019), rather than a clinically recognised form of mental health therapy. While improving access

on one level, and perhaps even extending the set of opportunities for engagement, these services also raise several ethical concerns.

Primarily these concerns stem from the fact that the logic underpinning many of these ideas is overly technologically deterministic, presenting the 'problem' of emotional wellbeing as something with a well-defined causal chain that can be 'solved' algorithmically (Janssen & Kuk, 2016). This approach assumes that a DHT is a neutral collection of code and data rather than a node in much broader social *system* composed of human and artificial agents (Ananny & Crawford, 2018), the impact of which needs to be assessed not *in silico* but *in socio.* When such a social systems approach to analysis (Crawford & Calo, 2016), instead of a product analysis, is taken it becomes much clearer that when the effects of many small, seemingly neutral, interactions (e.g. one user 'talking' to an NLP-based chatbot) are aggregated at a societal level the impact can be morally significant (Floridi, 2013). For instance, if these chatbot or video-based consultation services do little more than provide paid-for access to someone to talk to, it can be argued that they end up commercialising (and perhaps replacing) an important social function that has typically been provided by friends and families. This could result in the commodification and diminishment of social relationships by reducing the opportunity for an actual friend to cultivate virtues such as empathy, or compassionate listening. Moreover, a genuine friend or family member may also be able to offer more insightful support, due to a wider understanding of the contextual factors (e.g. lifestyle, previous experiences). Alternatively, if these services end up drawing users away from formal healthcare services, which are governed by stricter evidence standards (see previous section), they could result in harm to the user due to inadequate care.

Although aspects of these concerns may fall more naturally within the remit of the principle of non-maleficence (i.e. avoid harm), there is also a social justice concern related to the fact that these services may further increase social inequalities in access to forms of treatment by creating a market that is only available to segments of society, and perhaps more importantly, the compounding effects of isolation that result from shifting the burden of care. Here, the aggregate effect is the loss of community. People will rely less on their neighbours, friends and family to provide them with advice, which will give them less opportunities to build up trusting relationships that hold together divergent and contrasting views (Durante, 2010), and undermine the likelihood that responsibility (burden of care) for maintaining the wellbeing of each node (individual) is evenly distributed across the network (Floridi, 2016a). Instead, this responsibility is shifted solely to the individual, which can potentially be very damaging to that individual's mental wellbeing for two primary, interconnected, reasons: (1) the individual becomes increasingly isolated, unable to benefit from the cathartic social support captured by the 'a problem shared, is a problem halved'

idiom; and (2) the individual feels too much backward-facing moral responsibility (blame) for having experienced a deterioration in their mental wellbeing and feels increasingly unable to interact with other 'blameless' individuals, resulting in further isolation (Wardrope, 2015).

2.5. Explicability

Much of the current literature about explicability in the context of artificial intelligence and 'black-box' decision-making focuses on the need to make it possible for an individual to understand *how* an algorithm made a decision through the use of specific statistical or visualisation techniques such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) or SHapely Additive exPlanations (SHAP) (Lundberg & Lee, 2017)[12]. In the context of medical care, and particularly mental healthcare, this focus is necessary but not sufficient as it does not reflect the fact that explanations are social and contextual, about more than causal attribution (Miller, 2019), and reliant on meaningful dialogue between user, developer and model (Mittelstadt, Russell, & Wachter, 2019). In short, purely quantitative explanations fail to take into account the need to make a result, or specific piece of advice, *meaningfully interpretable* (or understandable to a specific end-user (Guidotti et al., 2018).

What counts as interpretable, and therefore actionable, advice is not an agreed standard (Doshi-Velez & Kim, 2017; Dosilovic, Brcic, & Hlupic, 2018). Instead, what counts for one individual may not count as such to another individual with a different set of epistemic and normative standards, experiences, and baseline knowledge (Binns, 2018). This is particularly related to variance in health literacy level, which has been shown to have a significant influence on an individual's ability to evaluate the quality, reliability and actionability of a healthcare information source (Chen et al., 2018). For example, those with lower health literacy levels are more likely to rely on social media sources of health advice, including mental health advice, than traditional online sources, such as websites providing clinically-validated information (Chen et al., 2018). This may be because, in the absence of an ability to determine the difference in credibility between the two sources, these individuals rely more heavily on bandwagon heuristics and conflate popularity (e.g. likes and shares) with credibility and reliability (Borah & Xiao, 2018). In the context of mental health, a lack of such considerations is particularly concerning as it means that those with low eHealth literacy, presented with conflicting information or recommendations about how to improve their health, are potentially more at risk than others of suffering from health anxiety (so-

[12] Both LIME and SHAP are methods that can be used to 'explain' the output of any machine learning model, typically used for 'explaining' classifiers.

called 'cyberchondria') (McMullan, Berle, Arnáez, & Starcevic, 2019) and more vulnerable to poor-quality and potentially harmful advice.

This is a concern from an equity of care perspective due to the following cycle: (a) lower levels of eHealth literacy have been found to be associated with other disadvantaging sociodemographic factors (Paige, Krieger, & Stellefson, 2017); (b) individuals with a lower income are more likely to rely on unregulated (and free) online sources of mental health care provision; (c) the poorer quality of advice delivered through these unregulated services means that individuals are unlikely to see an improvement in their mental state; (d) this lowers their self-efficacy; (e) this lessens their willingness to engage with mental health services; (f) this increases the risk of these individuals feeling unable to participate in society, both socially and economically; which (g) lessens their chances of improving their circumstances or their eHealth literacy, creating a situation of cumulative disadvantage. As such, we can acknowledge the importance of keeping the patient as a key part of the decision-making process as much as possible, in order to mitigate the worst effects that result from a lack of awareness.

The only way such nuances in design needs for mental health DHTs are going to be elicited is if the 'users' are treated as part of the solution, rather than as a problem that needs to be overcome (Aitken et al., 2019). This requires all parts of the system (e.g. designers, commissioners, policymakers, etc.), committing to the use of techniques such as those encapsulated under the headings of value sensitive design (Friedman, Hendry, & Borning, 2017) or responsible research and innovation (Jirotka, Grimpe, Stahl, Eden, & Hartswood, 2017; Stahl & Wright, 2018a; Stilgoe, Owen, & Macnaghten, 2013), which stresses the importance of considered and extensive stakeholder engagement throughout the development, deployment and use of DHTs. Not only will such a commitment improve the design of the technology and ensure it achieves positive outcomes for its users [e.g. as DeepMind Health found by developing their Streams App *with,* rather than *for,* clinicians in the Royal Free Hospital (DeepMind Health, 2019)], but also meet the requirements of perceived usefulness and ease of use, to enhance the likelihood of adoption.

Such engagement practices can, therefore, be seen as a way of improving the social responsibility of DHTs by encouraging their designers and commissioners to take into account the expectations of stakeholders with regards to the impacts of the DHT on individuals, society and the wider system (Zhao, 2018). As such they are a means of moving from principles to practice (Winfield & Jirotka, 2018) and are a key 'tool' in the governance toolbox alongside impact assessments, judicial review, model repositories (Edwards & Veale, 2018), and best practice guidelines or codes of conduct. However, in cases where the risks to end-users, in this case patients, are at their highest, it might be that these governance approaches are insufficient. For

example, Hall, Gertz, Amato, & Pagliari (2017) assessed the information for consumers' of 15 direct-to-consumer genetic testing companies available in the UK against the UK Human Genetics Commission (HGC) best practice principles and failed to find one provider compliant with all of the principles. Given that the results from these tests often include the statistical likelihood of the individual developing a specific disease, the risk posed to the individual's psychological integrity by not presenting this information in an interpretable format, is quite high (Andorno, 2004). Instances of such high risk may result in calls for a move up from ethically-aligned standards to ethically-aligned regulation (Winfield & Jirotka, 2018). While this may well be necessary to protect patient safety, it is important that the transition from "soft ethics" to governance and legislation (Floridi, 2018) is done in a way that is proportionate and capable of producing regulation that is neither too semantically strict, flexible nor overly unpredictable (Arvan, 2018).

3 Allowing for contextual flexibility

All ethical principles, including the bioethical principles that we have used as a means of guiding our critique, constrain behaviours. However, the way that they constrain behaviours may not always be interpreted consistently across different contexts (e.g. between different cultures, peoples and organisations) (Turilli, 2007). This creates a tension between the need for universal principles, such as non-maleficence, beneficence, autonomy, justice and explicability, and the need to respect differences in their implementation, application, and relative weighting of importance (Binns, 2018). For example, a clinical researcher might have a different interpretation of justice and give it a different weighting than a policy-maker. In addition, patients and clinicians are likely to interpret 'harm' (non-maleficence) differently.

If regulation is designed in a way that makes the interpretation of these principles too 'strict' it will limit society's ability to reflect on them (i.e. flexibly interpret, discuss and evaluate), making it harder to judge whether or not they have been adequately applied in different circumstances (D'Agostino & Durante, 2018). However, if regulation is designed in a way that is too open to interpretation it will fail to protect society from the risks that have been highlighted (Floridi, 2016b). There is no simple or straightforward way out of this tension. Ethics in this sense is a practice of ongoing discussion and critical engagement, and as such any set of ethical guidelines or principles should be treated as "living documents" that require continuous investment to maintain (Floridi et al., 2018).

4 Conclusion

Discussing the clinical applications of machine learning, Watson et al. note how the "opportunity costs of not using our best available tools for disease detection and treatment are substantial—12 million people a year receive misdiagnoses in the United States, with about six million facing potential harm as a result. Nearly one third of all preventable deaths in the United Kingdom are attributable to misdiagnosis" (2019, p. 2). As we have demonstrated, additional opportunity costs exist in the context of failing to use DHTs effectively for delivering mental healthcare. However, to ensure that these opportunities are pursued in an ethically responsible manner, it is vital that those responsible for delivering healthcare understand the importance of framing the challenges in the appropriate way—the concepts we use matter.

In this paper, we critically evaluated the concept of empowerment as it applies to DHTs and mental healthcare, showing how an insufficient consideration of wider psychological and socioeconomic factors runs the risk of missed opportunities for patient engagement and a misunderstanding of the role that key bioethical principles play in shaping healthcare delivery. Different mental health disorders will present different barriers to engagement and must be considered in relation to the situated nature of the individual concerned. To better articulate these concerns, we deployed five principles related to the ethical development and use of artificial intelligence, which are grounded in the literature on bioethics. These principles served as a structure to frame our discussion of some of the specific ethical issues that arise with the use of DHTs for mental healthcare—there will obviously be many more that we have not considered. It is well understood in the bioethical literature that these *prima facie* principles are general guidelines (Beauchamp & Childress, 2013), which serve to establish more specific rules that could be used to assist the design and development of relevant DHTs. As such, they can only serve as a starting point in the ethical evaluation of specific technologies with specific uses in specific contexts. However, as we have shown, their higher-order level of abstraction can be of significant value in drawing attention to relevant ethical differences between the use of DHTs in healthcare systems broadly construed and the use of DHTs in the narrower context of mental healthcare.

It is vital that we continue to scrutinise the design, development, and use of DHTs in all areas of healthcare. While traditional ethical principles will still play a valuable role, the novel features of DHTs (e.g. artificial intelligence) alter their nature and specificity when applied to these new contexts. Therefore, and to appropriate a term from computer science, if we wish to avoid creating vulnerabilities that arise from being locked-in to a *legacy system of values* we must be willing to regularly evaluate our use of normative concepts. If we fail to do this, we may be unable to determine whether DHTs are genuinely empowering all users or simply serving as a costly distraction that prevents our healthcare system from serving those who need the most support.

Statement of Contribution: CB and JM contributed equally to the design, research, and writing of this article.

Bibliography

Aitken, M., Tully, M. P., Porteous, C., Denegri, S., Cunningham-Burley, S., Banner, N., … Willison, D. J. (2019). Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research. *International Journal of Population Data Science*, *4*(1). https://doi.org/10.23889/ijpds.v4i1.586

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Andorno, R. (2004). The right not to know: an autonomy based approach. *Journal of Medical Ethics*, *30*(5), 435–439. https://doi.org/10.1136/jme.2002.001578

Anthony, W. A. (1993). Recovery from mental illness: The guiding vision of the mental health service system in the 1990s. *Psychosocial Rehabilitation Journal*, *16*(4), 11–23. https://doi.org/10.1037/h0095655

Arvan, M. (2018). Mental time-travel, semantic flexibility, and A.I. ethics. *AI & SOCIETY*. https://doi.org/10.1007/s00146-018-0848-2

Bakker, D., Kazantzis, N., Rickwood, D., & Rickard, N. (2016). Mental Health Smartphone Apps: Review and Evidence-Based Recommendations for Future Developments. *JMIR Mental Health*, *3*(1), e7. https://doi.org/10.2196/mental.4984

Barras, C. (2019). Mental health apps lean on bots and unlicensed therapists. *Nature Medicine*. https://doi.org/10.1038/d41591-019-00009-6

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed.). New York, N.Y.: Oxford University Press.

Bellet, C., & Frijters, P. (2019). Big Data and Well-being. In J. Helliwell, R. Layard, & J. Sachs (Eds.), *World Happiness Report 2019*. Retrieved from https://worldhappiness.report/ed/2019/big-data-and-well-being/

Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology*, *31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Borah, P., & Xiao, X. (2018). The Importance of 'Likes': The Interplay of Message Framing, Source, and Social Endorsement on Credibility Perceptions of Health Information on Facebook. *Journal of Health Communication*, *23*(4), 399–411. https://doi.org/10.1080/10810730.2018.1455770

Bravo, P., Barr, P. J., Scholl, I., Elwyn, G., McAllister, M., Elwyn, G., & McAllister, M. (2015). Conceptualising patient empowerment: a mixed methods study. *BMC Health Services Research*, *15*(1). https://doi.org/10.1186/s12913-015-0907-z

Breedvelt, J. J., Zamperoni, V., Kessler, D., Riper, H., Kleiboer, A. M., Elliott, I., … Bockting, C. L. (2019). GPs' attitudes towards digital technologies for depression: an online survey in primary care. *British Journal of General Practice*, *69*(680), e164–e170. https://doi.org/10.3399/bjgp18X700721

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00599

Burr, C., & Cristianini, N. (2019). Can Machines Read our Minds? *Minds and Machines*. https://doi.org/10.1007/s11023-019-09497-4

Burr, C., Taddeo, M., & Floridi, L. (2019). The Ethics of Digital Well-Being: A Thematic Review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3338441

Calvo, R. A., & Peters, D. (2014). *Positive Computing: Technology for Wellbeing and Human Potential*. United States: MIT Press.

Charland, L. C. (2015). Decision-Making Capacity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2015). Retrieved from https://plato.stanford.edu/archives/fall2015/entries/decision-capacity/

Chen, X., Hay, J. L., Waters, E. A., Kiviniemi, M. T., Biddle, C., Schofield, E., … Orom, H. (2018). Health Literacy and Use and Trust in Health Information. *Journal of Health Communication*, *23*(8), 724–734. https://doi.org/10.1080/10810730.2018.1511658

Chiauzzi, E., DasMahapatra, P., Cochin, E., Bunce, M., Khoury, R., & Dave, P. (2016). Factors in Patient Empowerment: A Survey of an Online Patient Research Network. *The Patient - Patient-Centered Outcomes Research*, *9*(6), 511–523. https://doi.org/10.1007/s40271-016-0171-2

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*(7625), 311–313. https://doi.org/10.1038/538311a

D'Agostino, M., & Durante, M. (2018). Introduction: the Governance of Algorithms. *Philosophy & Technology*, *31*(4), 499–505. https://doi.org/10.1007/s13347-018-0337-z

DeepMind Health. (2019, April 15). Retrieved from https://deepmind.com/applied/deepmind-health/working-partners/how-were-helping-today/

Department of Health and Social Care. (2019, February 19). Code of conduct for data-driven health and care technology. Retrieved April 15, 2019, from GOV.UK website: 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

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv:1702.08608 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1702.08608

Dosilovic, F. K., Brcic, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040

Durante, M. (2010). What Is the Model of Trust for Multi-agent Systems? Whether or Not E-Trust Applies to Autonomous Agents. *Knowledge, Technology & Policy*, *23*(3–4), 347–366. https://doi.org/10.1007/s12130-010-9118-4

Dzogang, F., Lightman, S., & Cristianini, N. (2018). Diurnal variations of psychometric indicators in Twitter content. *PLOS ONE*, *13*(6), e0197002. https://doi.org/10.1371/journal.pone.0197002

Edwards, L., & Veale, M. (2018). Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? *IEEE Security & Privacy*, *16*(3), 46–54. https://doi.org/10.1109/MSP.2018.2701152

Fernández-Caballero, A., Navarro, E., Fernández-Sotos, P., González, P., Ricarte, J. J., Latorre, J. M., & Rodriguez-Jimenez, R. (2017). Human-Avatar Symbiosis for the Treatment of Auditory Verbal Hallucinations in Schizophrenia through Virtual/Augmented Reality and Brain-Computer Interfaces. *Frontiers in Neuroinformatics*, *11*. https://doi.org/10.3389/fninf.2017.00064

Floridi, L. (2010). *Information: a very short introduction*. Oxford ; New York: Oxford University Press.

Floridi, L. (2013). Distributed Morality in an Information Society. *Science and Engineering Ethics*, *19*(3), 727–743. https://doi.org/10.1007/s11948-012-9413-4

Floridi, L. (2016a). Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi, L. (2016b). Tolerant Paternalism: Pro-ethical Design as a Resolution of the Dilemma of Toleration. *Science and Engineering Ethics*, *22*(6), 1669–1688. https://doi.org/10.1007/s11948-015-9733-2

Floridi, L. (2018). Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180081. https://doi.org/10.1098/rsta.2018.0081

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Foley, T., & Woollard, J. (2019). *The digital future of mental healthcare and its workforce*. Retrieved from topol.hee.nhs.uk

Friedman, B., Hendry, D. G., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction*, *11*(2), 63–125. https://doi.org/10.1561/1100000015

Garcia, J., Romero, N., Keyson, D., & Havinga, P. (n.d.). Reflective healthcare systems: Mirco-cylce of self-reflection to empower users. *Interaction Design and Architecture(s)*, *23*(1), 173–190.

Greaves, F., Joshi, I., Campbell, M., Roberts, S., Patel, N., & Powell, J. (2018). What is an appropriate level of evidence for a digital health intervention? *The Lancet*, *392*(10165), 2665–2667. https://doi.org/10.1016/S0140-6736(18)33129-5

Grisso, T., & Appelbaum, P. S. (1998). *Assessing competence to consent to treatment : a guide for physicians and other health professionals*. New York: Oxford University Press.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Hall, J. A., Gertz, R., Amato, J., & Pagliari, C. (2017). Transparency of genetic testing services for 'health, wellness and lifestyle': analysis of online prepurchase information for UK consumers. *European Journal of Human Genetics*, *25*(8), 908–917. https://doi.org/10.1038/ejhg.2017.75

Hausman, D. (2015). *Valuing Health: Well-Being, Freedom, and Suffering*. New York: Oxford University Press.

Hindmarch, T., Hotopf, M., & Owen, G. S. (2013). Depression and decision-making capacity for treatment or research: a systematic review. *BMC Medical Ethics*, *14*(1). https://doi.org/10.1186/1472-6939-14-54

Holm, E. A. (2019). In defense of the black box. *Science*, *364*(6435), 26–27. https://doi.org/10.1126/science.aax0162

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly*, *33*(3), 371–377. https://doi.org/10.1016/j.giq.2016.08.011

Jirotka, M., Grimpe, B., Stahl, B., Eden, G., & Hartswood, M. (2017). Responsible research and innovation in the digital age. *Communications of the ACM*, *60*(5), 62–68. https://doi.org/10.1145/3064940

Kamens, S. R., Elkins, D. N., & Robbins, B. D. (2017). Open Letter to the DSM-5. *Journal of Humanistic Psychology*, *57*(6), 675–687. https://doi.org/10.1177/0022167817698261

Kenny, R., Dooley, B., & Fitzgerald, A. (2015). Feasibility of "CopeSmart": A Telemental Heath App for Adolescents. *JMIR Mental Health*, *2*(3), e22. https://doi.org/10.2196/mental.4370

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kukla, R. (2005). Conscientious Autonomy: Displacing Decisions in Health Care. *Hastings Center Report*, *35*(2), 34–44. https://doi.org/10.1353/hcr.2005.0025

Lansdall-Welfare, T., Lightman, S., & Cristianini, N. (2019). Seasonal variation in antidepressant prescriptions, environmental light and web queries for seasonal affective disorder. *The British Journal of Psychiatry*, 1–4. https://doi.org/10.1192/bjp.2019.40

Lehavot, K., Ben-Zeev, D., & Neville, R. E. (2012). Ethical Considerations and Social Media: A Case of Suicidal Postings on Facebook. *Journal of Dual Diagnosis*, *8*(4), 341–346. https://doi.org/10.1080/15504263.2012.718928

Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., & Morency, L.-P. (2017). Reporting Mental Health Symptoms: Breaking Down Barriers to Care with Virtual Human Interviewers. *Frontiers in Robotics and AI*, *4*. https://doi.org/10.3389/frobt.2017.00051

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *ArXiv:1705.07874 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1705.07874

Lupton, D. (2016). *The Quantified Self*. USA: Polity Press.

McMullan, R. D., Berle, D., Arnáez, S., & Starcevic, V. (2019). The relationships between health anxiety, online health information seeking, and cyberchondria: Systematic review and meta-analysis. *Journal of Affective Disorders*, *245*, 270–278. https://doi.org/10.1016/j.jad.2018.11.037

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 279–288. https://doi.org/10.1145/3287560.3287574

Morley, J., & Floridi, L. (2019). *Against Empowerment: How to Reframe the Role of mHealth Tools in the Healthcare Ecosystem (Draft)*.

Nagy, P., & Neff, G. (2015). Imagined Affordance: Reconstructing a Keyword for Communication Theory. *Social Media + Society*, *1*(2), 205630511560338. https://doi.org/10.1177/2056305115603385

National Institute for Health and Care Excellence. (2018). *Evidence standards framework for digital health technologies*. Retrieved from https://www.nice.org.uk/Media/Default/About/what-

we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf

Nelson, A., Herron, D., Rees, G., & Nachev, P. (2019). Predicting scheduled hospital attendance with artificial intelligence. *Npj Digital Medicine*, *2*(1), 26. https://doi.org/10.1038/s41746-019-0103-3

Nelson, E. C., Eftimovska, E., Lind, C., Hager, A., Wasson, J. H., & Lindblad, S. (2015). Patient reported outcome measures in practice. *BMJ*, *350*(feb10 14), g7818–g7818. https://doi.org/10.1136/bmj.g7818

NHS England. (2019). *The NHS Long Term Plan*. Retrieved from NHS website: https://www.longtermplan.nhs.uk/wp-content/uploads/2019/01/nhs-long-term-plan.pdf

Owens, J., & Cribb, A. (2013). Beyond Choice and Individualism: Understanding Autonomy for Public Health Ethics. *Public Health Ethics*, *6*(3), 262–271. https://doi.org/10.1093/phe/pht038

Paige, S. R., Krieger, J. L., & Stellefson, M. L. (2017). The Influence of eHealth Literacy on Perceived Trust in Online Health Communication Channels and Sources. *Journal of Health Communication*, *22*(1), 53–65. https://doi.org/10.1080/10810730.2016.1250846

Polykalas, S. E., & Prezerakos, G. N. (2019). When the mobile app is free, the product is your personal data. *Digital Policy, Regulation and Governance*, *21*(2), 89–101. https://doi.org/10.1108/DPRG-11-2018-0068

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *ArXiv:1602.04938 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1602.04938

Roberts, K. J. (1999). Patient empowerment in the United States: a critical commentary. *Health Expectations*, *2*(2), 82–92. https://doi.org/10.1046/j.1369-6513.1999.00048.x

Schermuly-Haupt, M.-L., Linden, M., & Rush, A. J. (2018). Unwanted Events and Side Effects in Cognitive Behavior Therapy. *Cognitive Therapy and Research*, *42*(3), 219–229. https://doi.org/10.1007/s10608-018-9904-y

Sen, A. (2010). *The Idea of Justice*. London: Penguin.

Sheehan, M. (2014). Reining in patient and individual choice. *Journal of Medical Ethics*, *40*(5), 291–292. https://doi.org/10.1136/medethics-2014-102161

Sieck, C., Walker, D., Retchin, S., & McAlearney, A. (2019). The Patient Engagement Capacity Model: What Factors Determine a Patient's Ability to Engage? Retrieved March 30, 2019, from Catalyst website: https://catalyst.nejm.org/patient-engagement-capacity-model/?utm_campaign=Connect%20Weekly&utm_source=hs_email&utm_medium=email&utm_content=70937477&_hsenc=p2ANqtz-9iyYCA7cZ07BERqjc6bZfyUmsoykOeFDRfMu9OAAxkEwMcmOIxeQ6s7AjzvfxHDfuTuPrEeL3FZwMVEVDa8DRGkFSPAw&_hsmi=70937477

Spencer, G. (2015). 'Troubling' moments in health promotion: unpacking the ethics of empowerment: G. Spencer. *Health Promotion Journal of Australia*, *26*(3), 205–209. https://doi.org/10.1071/HE15049

Stahl, B. C., & Wright, D. (2018a). Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. *IEEE Security & Privacy*, *16*(3), 26–33. https://doi.org/10.1109/MSP.2018.2701164

Stahl, B. C., & Wright, D. (2018b). Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. *IEEE Security & Privacy*, *16*(3), 26–33. https://doi.org/10.1109/MSP.2018.2701164

Starkey, F. (2003). The 'Empowerment Debate': Consumerist, Professional and Liberational Perspectives in Health and Social Care. *Social Policy and Society*, *2*(4), 273–284. https://doi.org/10.1017/S1474746403001404

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

The Topol Review Board. (2019). *The Topol Review: Preparing the healthcare workforce to deliver the digital future*. Retrieved from topol.hee.nhs.uk

Turilli, M. (2007). Ethical protocols design. *Ethics and Information Technology*, *9*(1), 49–62. https://doi.org/10.1007/s10676-006-9128-9

Wardrope, A. (2015). Relational Autonomy and the Ethics of Health Promotion. *Public Health Ethics*, *8*(1), 50–62. https://doi.org/10.1093/phe/phu025

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: beyond the black box. *BMJ*, l886. https://doi.org/10.1136/bmj.l886

White, R. G., Imperiale, M. G., & Perera, E. (2016). The Capabilities Approach: Fostering contexts for enhancing mental health and wellbeing across the globe. *Globalization and Health*, *12*. https://doi.org/10.1186/s12992-016-0150-3

Winfield, A. F. T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180085. https://doi.org/10.1098/rsta.2018.0085

Zhao, W.-W. (2018). Improving Social Responsibility of Artificial Intelligence by Using ISO 26000. *IOP Conference Series: Materials Science and Engineering*, *428*, 012049. https://doi.org/10.1088/1757-899x/428/1/012049