Responsible innovation in online therapy

A report on technical opportunities, ethical issues, and recommendations for design

Report Date:

22.07.2019

Authors:
Dorian Peters
Diana Robinson
Karina Vold
Rafael A. Calvo

Imperial Consultants, Ltd.

This report is the independent expert opinion of the authors.

An initial set of evidence-based ethical considerations and design recommendations to guide the responsible development of digital mental health technologies.

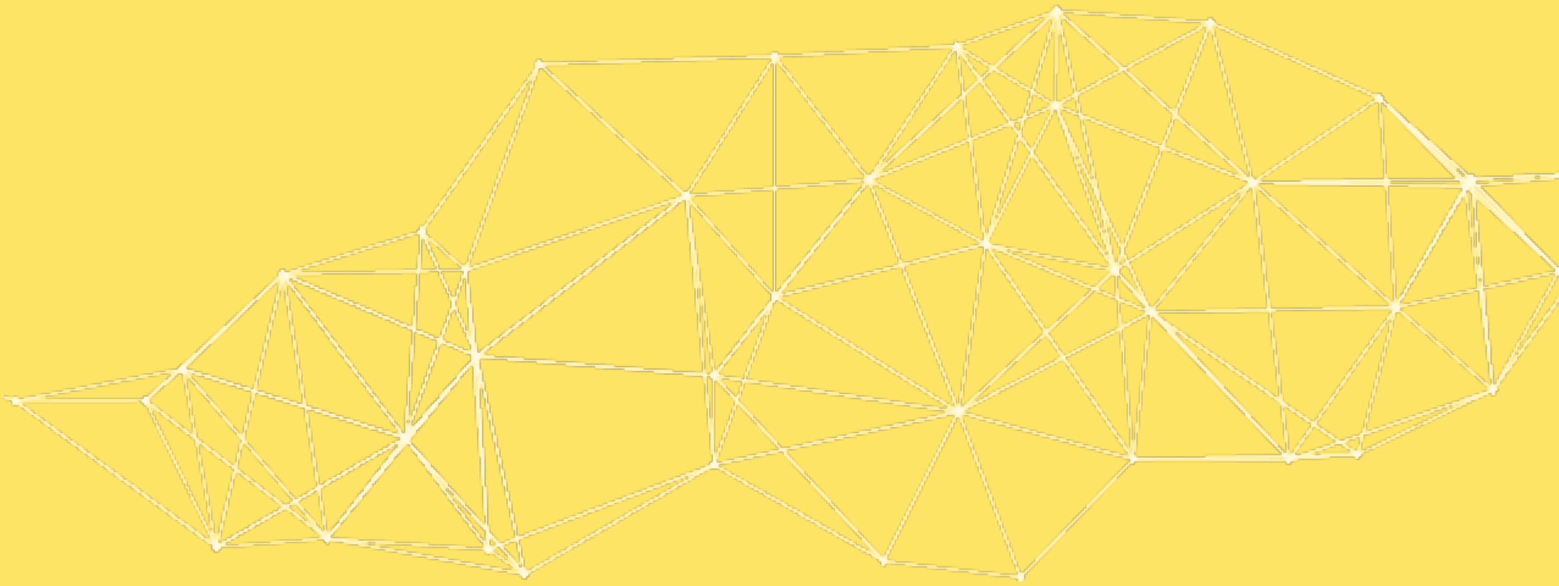

Table of Contents

Executive Summary

Many of the challenges of providing mental health care to the large number of people who need it can be addressed with technologies. Today, technology-based support for mental health treatment represents a rapidly evolving area of research and industry. But rapid change, coupled with the introduction of new technologies to such a delicate area, brings additional challenges. In this report we review some of the recent changes and trends in relation to internet-delivered therapy for depression and anxiety. We summarise some of the key ethical questions surrounding online therapy and present initial best practice recommendations for more responsible design and development of these technologies and services.

This report is based on a review of the commercial and academic literature on mental health technologies and on applied knowledge relating to the co-design and development of e-health services with clinicians and end-users in the non-profit sector. The report touches on fully- and partially-automated systems but focuses on one-on-one therapist-led services. **The objective is to contribute an initial set of evidence-based ethical considerations and design recommendations for more responsible mental health technology development**.

Introduction

By the year 2020, depression is expected to be the highest-ranking cause of disease in the developed world[1]. New technologies have the potential to increase access to support, reduce disparities, reduce costs, and even revolutionize mental health care through new technology-enabled approaches to treatment, prevention and promotion. However, in order for these potentials to manifest, it will be essential to properly address the ethical and logistical obstacles that have emerged.

Even within the relatively traditional landscape of apps and websites, recent advances in Natural Language Processing and AI are being used to improve quality and reduce the cost of treatment, but the speed of technological innovation outpaces regulation, and this has led to a number of problems: 1) There has been an explosion of new consumer technologies not evaluated for safety or efficacy, 2) End-users and mental health experts are too seldom included in design and 3) insufficient consideration is given to the ethical implications of novel technological approaches.

The valuable academic reviews addressing some of these issues can be difficult for commercial organisations to apply and build on directly. In order to help bridge research to practice for organisations, researchers, and regulators, herein we summarise key trends, challenges, and opportunities for mental health technologies, with implications for applied ethics and design practice.

Scope

While the recommendations presented herein will be applicable to most, if not all, mental health technologies, the specifics of this report are focused on online text-based one-to-one professional therapy for depression and anxiety.

Examples of technologies that sit outside of scope for this report include support for severe and other mental illnesses (e.g. suicide ideation, schizophrenia), technologies for group interaction, video/audio counseling, fully automated non-human (e.g. chatbot) or un-certified human-provided therapy (i.e. peer-based or crowd-sourced) and self-help programs. Although we touch on many of these, the full implications and available research relating to each of these varied approaches are not included.

Technology trends in mental health

Advances in Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques, combined with increased pressures to reduce the cost of healthcare, are driving companies toward creating new forms of self-help and online therapy. This trend is supported by research showing that technology can be used to support psychological wellbeing in many different ways, from treatment to prevention and resilience-building[2] via a diversity of approaches from self-help tools, to social support and access to professional counselling. While most applied advances to date take the approach of transferring traditional therapies into technology-based environments (e.g. online CBT, VR-based exposure therapy, digital symptom tracking, etc.), there has also been work exploring new forms of therapy that might be enabled by the unique affordances of new technological capabilities.

AI has been used and studied in medicine for over 30 years, and there are a number of journals, including *Artificial Intelligence in Medicine* and the *Journal of Medical Internet Research,* dedicated to it[3]. Even within the narrower domain of mental health care, there is so much research literature available that for health technology organisations it can be helpful to focus on literature reviews for an overview of advances in mental health technologies. We summarise a number of these below.

Leveraging research, consumer information and best practice guidelines

A number of literature reviews have demonstrated the capacity for properly designed digital technologies to improve mental health outcomes. For example, Hoerman et al. (2017) reviewed studies exploring the feasibility and effectiveness of online one-on-one mental health interventions employing text-based synchronous chat[4]. The study described 24 interventions covering a variety of mental health issues (e.g. anxiety, distress, depression, eating disorders, and addiction) and intervention designs. Results demonstrated that, overall, "compared with the waitlist (WL) condition, studies showed significant and sustained improvements in mental health outcomes following synchronous text-based intervention, and post treatment improvement equivalent but not superior to treatment as usual (TAU) (e.g. face-to-face and telephone counselling)".

Sanches et al. (2019) reviewed the landscape of human-computer interaction research in relation to affective disorders and found that most innovation has occurred in the areas of automated diagnosis and self-tracking, with some work on tangible interfaces.[5]

Other reviews have focused on specific technological affordances, for example, NLP techniques used in non-clinical contexts[6], or recent growth in the use of chatbots[7]. We discuss some of these academic reviews in more detail in the next section.

In addition to academic reviews, consumer product reviews can offer insights into market trends. As one example, a *Healthline* consumer information review of chatbots describes four leading mental health chatbot products (current as of July 2018): Woebot, Wisa, Joyable and Talkspace[8]. Although written for end-users, the comparison provides insights into current consumer trends and concerns, as well as the opportunities and challenges of automating mental-health chat support services.

In addition to leveraging technology research and consumer trends, companies developing mental health technologies are advised to consider available best practice guidelines, such as those provided by the American Psychiatric Association[9]. Although these guidelines are often intended for helping end-users in selecting services, in doing so, they provide a wealth of information on optimal practice.

Business drivers for innovation

Identifying business drivers is important to understanding how progress in different areas might be useful to organisations in this area. The drivers can be broadly construed within the following three categories:

- **Improving efficacy**. To improve the efficacy of existing interventions, organisations are innovating ways to integrate new sources of data now available through sensors (e.g. wearable sleep tracking, heart rate monitoring, etc.). Improved efficacy can also be achieved by improving support for counsellors through better workplace tools.
- **Reducing Cost**. Reductions in cost are frequently achieved by automating straightforward information-gathering processes that would otherwise require expensive human time. For example, for a chat service, the introductions, data gathering, and compliance requirements--all tasks that can be done by filling out a form--could be supported by an automated system such as a chatbot.
- **Extending services**. Many organisations that began by providing phone or chat counselling are increasingly offering automated or self-help tools to provide support that is accessible 24/7 and from any location (better serving remote and less mobile patients). Extended services can be used between counselling sessions or provided as part of a stepped-care approach where low cost, automated services are provided as a first line of treatment, while other more expensive treatment options are offered only if needed.

Regulatory frameworks

The proliferation of online health services has demanded that new regulatory frameworks be developed to protect patient privacy. There are at least two sets of policies that mental health chat services should consider. A number of countries are now using The Health Insurance Portability and Accountability Act (HIPAA) that sets the standard for sensitive patient data protection. Organisations with users in Europe also need to consider the General Data Protection Regulation (GDPR). Although regulatory compliance is not addressed within this report, its essential that it be considered early on in any development project as many of the software architectures available are not designed to comply. For example, Intercom, one of the most popular platforms used to build customer support chat services, does not comply with HIPPA (as of July 2019), and would therefore not be suitable for counselling.

Three categories of technology innovation

The business drivers mentioned above can motivate a variety of different approaches to technology integration including **automation** (in which technology carries out tasks previously done by humans) or **augmentation** (in which technology enhances human activity). The process of **heteromation** (in which humans enhance technology activity) can be used to gather the data that is often required to train algorithms. Each of these approaches has its own challenges and opportunities and is discussed below.

For organisations providing professional (or peer-support) counselling, natural language technologies in particular, offer opportunities for improving quality and reducing cost through software-based automation and augmentation of human services. But these also raise serious ethical questions which are discussed in further detail in section 4. Sections 5 and 6 describes approaches to mitigating ethical risks and optimising the user experience through user-centred methods.[10] Below we review the three categories of technology integration.

Automation

The last 3 years have seen a rapid growth in chatbot use within every industry.[7] The history of mental health chatbots actually began in 1964 with Joseph Weizenbaum's "Eliza", a simple expert system impersonating a psychologist. Today, a leading chatbot platform (Pandorabots), claims to serve 250,000 developers who have together built over 300,000 chatbots.

The technology behind chatbots has a lot in common with conversational systems like Apple's Siri and Amazon's Alexa. Building an automatic chatbot for a specific purpose can provide a first step towards voice-based conversational agents, or even embodied relational agents such as those developed for health by Bickmore and colleagues[11]. Many of these systems imitate empathy and have been developed to support long term human-machine relationships. [12]

Although it is the more futuristic notion of human-like robot companions that tend to capture the imagination, most automation is used in far more subtle and routine ways, such as to automate checks on safety protocols. For example, one common safety protocol is for counsellors and patients not to meet "outside the system". Therefore, it is against policy for counsellors and clients to exchange any personal contact details. An automated system can be used to detect infringements to this protocol.

Similarly, within peer-support groups, algorithms can monitor for critical situations that signal a serious risk for a patient. There have been reports that some voice assistants can automatically detect suicide ideation statements and refer a user to emergency mental health services.[13]

However, some leading mental health researchers like Dame Tyl Wykes[14] have raised concerns about the overpromise of technology for mental health and the trend towards automation. In particular, the fact that algorithms can and do make mistakes, and that they lack accountability all pose critical challenges. Moreover, the optimisation of metrics can increase bias and inequality. In considering these and other concerns, it's essential to understand user experiences and perspectives.

A report commissioned by the Academy of Medical Sciences[15] describes the various ways mental health patients in the UK would like their data to be used. The report highlights that patients prefer to have the option to talk to a human, rather than a computer, and they don't trust software to make a mental health diagnosis. However, patients do report willingness to use online therapies in conjunction with the support of a human therapist.

Augmentation

Rather than replacing humans, augmentation processes focus on increasing human capabilities and making human activity more effective, efficient and satisfying. This can be done in many different ways.[16] Examples include:

- Automated textual and visual summaries provided to therapists can assist in their decision-making.[17]
- Feedback can be provided to counsellors to clarify the impact of their interventions and help them improve their skills and techniques.

- Automated prompts can remind counsellors to perform certain tasks, ask questions, comply with therapeutic protocol or ‘check-in’ with a patient.
- Guides or “templates”, based on the context and state of the conversation, can help improve quality and support the growth of new counsellors. These guides can also help maintain consistency across providers and services.

It’s important to note that augmentation can either enhance or hinder human users, depending on how it’s designed, and in particular, it requires deep involvement of users in the design process. These issues are discussed in Section 5 on User Experience.

Heteromation (when machines outsource to humans)

Heteromation, a term coined by technology anthropologists, Bonnie Nardi and Hamid Ekbia describes “computer-mediated labour currently performed by human beings in support of technological systems and economic enterprises.”[18] The term is often used to point to the fact that human effort goes unacknowledged and poorly compensated within these arrangements. Some common examples include crowdsourcing, crowdwork, design competitions, customer reviews and self-service arrangements in airports, post offices and grocery stores.

In some contexts, humans contribute data for some purpose other than heteromation but the data is also used for the benefit of the technological system. This mass collection, appropriation and intelligent synthesis from various sources (aka. “big data”) provides many challenges but much promise for healthcare. Sanches et al. summarise the benefits: “These types of systems have been proposed as having the potential to change how health care is to be provided, not only by providing immediate support to a user, through improving adherence to a treatment or predicting episodes…but also by aggregating different health data streams across patients (big data) and helping see population-wide trends, providing the possibility of advancing theoretical frameworks for mental health and providing evidence for effectiveness of different therapies by making use of the multivariate nature of available data from different users.” [5]

In fact, data will often become the most important asset of an organisation without users being aware, beyond the value it has to them. While data use is arguably included in end-user agreements, users (both patients and counsellors) are seldom aware of the ways in which their data contributions are being used and how the consequences of that use might affect them in the future (for example, by training an automated replacement).

Even after removing personal data, Natural Language Processing algorithms can be trained and optimised to improve diagnosis and treatment. Indeed, possible uses of recorded medical consultations include the development of algorithms that detect emotional states[19], that diagnose illness[6], that recognise empathic and mirroring behaviours[20] and many more. In each of these cases, both users and therapists (not always employees) are contributing extra value to the intellectual property of an organisation, often without compensation or awareness.

The benefits and challenges of data-driven approaches have been discussed in the research literature (managing and using data, digital phenotyping[21], the ethical opportunities and challenges posed by new data sources,[10] etc.) and are further discussed in section 5.

But how much data do we really need to collect and when is it worth it? A common default mentality among technologists is “the more data you can collect, the better” but emerging cases of data misuse have shown the risks of this mentality. Data collection will need to be better weighed against increasing concerns over privacy and security moving forward. The Sanches review found “a predominance of data-driven systems, that both produce and depend on digital mental health data streams for decision-support and self-monitoring goals." (43.8% of the papers). The authors also noted an "overemphasis on data production without consideration of how it leads to fruitful interventions" and "a narrow range of therapeutic methods". This dominance of apparently technology-driven (rather than needs-driven) solutions are common in many industries where innovation is frequently initiated by technologists rather than those in the areas of application.

With the increasing salience of serious trade-offs to be negotiated between the great promises and risks posed by intelligent health technology, the greatest technical challenge has become an ethical one. How do we decide in which direction to innovate, how to proceed responsibly, and how to apply what we create for the benefit of humankind? Technologists will need to engage more deeply than ever before with experts in ethics as well as with their users to find the answers.

Ethical considerations

Designing healthcare systems supporting users with mental health conditions is by its very nature a delicate endeavor, addressing a vulnerable user group, requiring that ethical considerations are taken into account. - Sanches, et al. 2019[5]

Context

In what follows we outline some of the ethical considerations involved in the development and use of AI in mental health care. Research on these issues is especially urgent as there is an apparent dearth of literature on this topic. In the Sanches review mentioned above, for example, fewer than half (48 of 139) of papers in the human-computer interaction literature studying technologies for affective disorders explicitly mention and deal with ethical issues. In this extensive review of the research literature in Human-Computer Interaction for Affective Health, they identified a key research area to "promote ethical practices for involvement of people living with affective disorders".[5]

That being said, a few recent papers have made important contributions to this space. A 2015 survey of 226 licensed Marriage and Family Counsellors, students, and supervisors, were asked to identify ethical concerns and drawbacks of online therapy. They found that five themes emerged: (a) confidentiality, (b) impact to the therapeutic relationship, (c) licensing and liability issues, (d) issues related to crises and risky clinical situations, and (e) training and education.[22] Mittelstadt and Floridi have recently written about the ethical risks of big data in the biomedical context, identifying five main areas of concern: "(1) informed consent, (2) privacy (including anonymisation and data protection), (3) ownership, (4) epistemology and objectivity, and (5) 'Big Data Divides' created between those who have or lack the necessary resources to analyse increasingly large datasets."[23] Finally, most recently, Burr and Morley (2019)[24] have written on ethical concerns around the use of digital health technologies for mental healthcare, with a particular focus on the issue of empowerment.

In our discussion we focus on seven areas that we feel are particularly relevant for online mental health therapies--**Autonomy**, **Justice, Privacy**, **Impact**, **Atrophy**, **Authenticity**, and **Transparency**. Each of these are affected by mental health therapy in general and we will focus here on opportunities and challenges brought by technologies specifically. Within these discussions, we also touch on aspects of beneficence, non-maleficence, explicability, and responsibility. A lengthier discussion would also include the topics of informed consent, ownership of data, and group-level harms, among others. By their nature, many ethical concerns are not simply incorporated into design once but must be continuously and iteratively monitored and considered.

Autonomy

What is autonomy?

While philosophers still debate definitions of autonomy, as well as why and to what extent we should value it, many agree with John Stuart Mill's view that autonomy is "one of the central elements of well-being".[25] On this view autonomy is an instrumental good because it contributes to one's well-being. This suggests that there is an ethical imperative for healthcare providers, who aim to improve patient well-being, to support patient autonomy. And, as mentioned above, there has also been general agreement around the idea that the use of AI needs to respect and support human autonomy as well.

Capturing a definition of autonomy raises many philosophical questions that we must skim past here, but in medical ethics, the principle of autonomy includes respect for both an individual's *right to decide* and for the *freedom of whether to decide.*[26] Together, these are meant to protect both our right to make choices and our freedom to choose how and when we want to exercise that right.[24] Because of the medical context of this report we will work with this definition.

Opportunities and Concerns around autonomy

Mental health support is costly. Depending on the country, costs are covered directly by patients, government (e.g. NHS), Community Interest Companies, and insurance companies. Often the cost makes it impossible for individuals to access it in a timely fashion if at all. Many people live in areas where there are no human therapists so access to services is also limited by distance. Even people who have the means for private care, and are at short distance, find it hard to access health services due to time or stigma. Mental health in particular is highly stigmatised and many struggle with seeking help.

Technologies can reduce cost, improve access and allow more people to reach for human help when they feel it's needed. It can also provide a form of low-barrier care, self-help and psychoeducation that can have a very positive impact on health.

The nature of mental illnesses needs to be taken into account when building online therapies. This is because the illness can affect one's capacity to reason, one's perception of oneself and of others, one's ability to make decisions, and other cognitive capacities that are core to one's ability to self-govern. In a recent paper discussing the concept of empowerment in the context of digital healthcare technologies, Burr and Morley[24] note that certain psychiatric disorders impact the individual's 'decisional capacity', which is "typically 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". They discuss how this may in turn affect a patient's choice to engage with a mental health service and even restrict their ability to make healthcare decisions. In extreme cases, it may be that respecting a patient's autonomy (i.e. non-intervention) may even put the patient at risk of self-harm or harm to others.

What this suggests is (a) that respect for patient autonomy may not always contribute to well-being and may sometimes have to be traded off against other goods, such as safety, and (b) that in some cases health care providers may need to go beyond mere respect for a patient's current ability to self-govern, and actually help build and support the user's autonomy. We will provide some recommendations for how autonomy can be supported below, in our section on User Experience and Design of Online Therapy.

One example of a risk to autonomy is in the sharing of self-tracking data. Self-tracking data can serve as a means of patient empowerment --offering data which can be used in self-reflection and deliberations about personal actions and choices. The sharing of self-tracking data with family and close friends, however, especially in non-emergency situations, can have a negative effect on the empowerment and autonomy of the patient. Sanches et al. (2019) describe this as an example of "autonomy [of vulnerable populations being] claimed by their social support network, collectivized by healthcare services, or both." We'll discuss this link between privacy and autonomy in more detail in the next section, on 'Privacy'.

Finally, the use of digitally-delivered therapies presents new opportunities to assist patients' decision-making and behaviours through the use of digital interventions, but these interventions may also present new risks to patient autonomy. As members of social groups, we are constantly influenced by various external sources and actors. And while not all of these influences undermine our autonomy, some do. Hence, what distinguishes 'acceptable' and 'unacceptable' (autonomy-depleting) influences is a central question with which all accounts of autonomy must grapple.

On one standard view, attempts to influence us that appeal to our rationality do not undermine our autonomy, while **attempts to influence us that are hidden and try to subvert an individual's ability to act on their own reasons risk being *manipulative***.[27] Accordingly, several authors have argued that the use of data-driven personalised interventions as 'nudges' to change behaviour could risk becoming manipulative if they are done covertly or with the explicit intention of bypassing a patient's rationality and consent.[28] Mental health apps that merge health and commercial content are at particular risk of being manipulative because of how they attract users interested in improving their health and then, having captured their attention, target advertisements intended to serve commercial interests.[29]

Much research has been done on the ethics and public acceptability of nudging,[30] though not within the specific context of mental health apps and not regarding 'hypernudges', or online nudges that make use of personalised targeting. Nonetheless there are a few ethical considerations that can help shape best practices in this context:[31]

1. Nudges or targeted interventions should be transparent, at least to the extent possible.
2. The intended outcome should align with the interests and well-being of the person being nudged.
3. It should be possible to opt-out of these interventions, and opting-out should not be burdensome (e.g. not take more than a few clicks).

To sum up, there is an ethical imperative for healthcare providers to try to support client/user autonomy. While thus far we have focused on the autonomy of the patient, the autonomy of other users, such as therapists, should also be taken into consideration when designing and implementing new forms of digitally-delivered therapies—we will discuss this in more detail in the section on 'understanding impact'.

> ***Recommendation 1: Mental health interventions should seek to protect and support user autonomy, giving particular consideration to the use of self-tracking data, nudges, and achieving informed consent.***

Privacy

Why value privacy?

There are many different dimensions of privacy, such as privacy of thought, privacy of the body, privacy of behaviour, informational privacy, and decisional privacy. Nowadays there is also increasing discussion about the need for online privacy and data privacy. Yet it is often left unstated why we should value these different forms of privacy. Indeed, there are a wide range of potential harms that can come from a lack of privacy, including harms to autonomy, dignity, fairness, reputation, self-development, intimacy, and bodily integrity, to name a few.[32] Hence, being clear about which kinds of privacy one is concerned about and why can be useful in understanding what kinds of measures to take.

Following Lanzing (2018)[33], who has recently argued that both informational privacy (or 'data privacy') and decisional privacy are threatened by the collection and use of big data, we will focus on these two notions of privacy. 'Informational privacy' is a right that entails the ability to control who has access to one's personal information, to what extent, and for what use. 'Decisional privacy' is the right not to be accessed or interfered with in our decisions and actions, such that third parties may not access our decisions and behaviours or attempt to influence them, unless this influence was otherwise consented to.[33] Decisional privacy is instrumental for protecting autonomy, and hence some of the concerns raised in the above section will be echoed here.

Concerns around privacy

Online mental health therapy applications that collect, store, and make use of personal data raise several important concerns around privacy.

Some concerns are similar to those in other forms of therapy. Because mental health is a stigmatized topic, those that suffer from mental health conditions face the risk of stereotypes, prejudice, and discrimination, from both themselves (self-stigma) and others. This means that **if digital health records of mental health status were leaked, it could threaten one's dignity and reputation, and even put one at risk of forms of discrimination**. Furthermore, the therapeutic relationship fostered between a counsellor and a patient is an intimate one, depending on detailed knowledge of the patient's life, which crucially depends on an assurance of privacy. Certainly, any breach in the privacy of what one shares during therapy sessions might threaten one's relationships with others, as well as one's dignity and reputation.

These concerns around privacy are true in traditional (face-to-face) therapy sessions as well, of course, but **relying on digital online platforms, from electronic medical records, to online therapies, poses new threats to both informational and decisional privacy**. [33] In Hertlein et al.'s 2015 survey[22], participants expressed concerns about the authenticity of the user (such as "who has access to the computer" and "the [chance] of loss of control of who has the device at the other end"), about who else might be physically present in the same room as the counsellor ("How can the therapist or client be sure no one else is in the vicinity of the computer—that is, how can you assure confidentiality?"), and about the possibility of hackers ("security online is not guaranteed.") Hence, in the case of online therapy, patients not only have to trust their counsellor's good intentions, they also have to trust that counsellors will protect their computer screen from onlookers (or other device, e.g. tablet or mobile), protect their passwords, use secure network connections, and not use shared computers.[34] Patients furthermore have to trust the provider of the technology itself not to use the data for any unconsented purpose.

For this reason, it's essential that all users (i.e. therapists and patients) are given clear and accurate explanations about how they should conduct themselves online to ensure informational privacy is protected, about how information collected from therapy sessions will be used, and about the benefits and risks associated with online therapy. **It is important that the utmost care is taken by companies to protect the storage of this data.**

Finally, it is worth emphasizing the link between privacy and autonomy. Historically, one reason privacy of thought and decisional privacy have been valued is because these forms of privacy can carve out a 'protective space' to allow individuals the opportunity to reflect and act freely.[35] But more recently, several authors have argued that recent technological advances have strengthened this link, such that **threats to privacy are increasingly also threats to autonomy.**[36] In particular, because of the kind of personal data that is now available to be collected (such as information about online behaviour) coupled with advances in machine learning that make it possible to infer personal attributes from collected data[37], companies are increasingly able to tailor messages and services to specific individuals or groups. The takeaway is that the more personal information a company has about you, the more effectively they can target interventions in attempts to unwittingly influence you.

> ***Recommendation 2: To protect both the informational and decisional privacy of users, make transparent the use of mental health data and ensure secure storage.***

Understanding Impact (Beyond patients)

Who is impacted?

Thus far our discussions on the ethical issues around digitally-delivered mental health care therapies have focused on the patients. However, the use of online text-based therapies also involves other people who may be impacted, including therapists, developers, family members, other patients, and the wider mental health care community. Hence, there is a need to adopt a holistic approach to design and implementation that ensures that all parties affected are considered.

To illustrate how other actors may be impacted, we'll return to the topics of autonomy and privacy and consider what concerns might have been raised for counsellors, as well as consider some potential risks for those involved in the development process.

Opportunities and Risks to counsellors

Online text-based professional therapies will involve (at least) two kinds of users: patients (or clients) and therapists (or counsellors). It's important to also think of the therapist as a user of this technology, as their role as counsellor will be changed and augmented by these new tools.

Autonomy: Just as for users, digital technologies can increase autonomy by allowing counsellors the ability to work remotely. In principle, a counsellor who is approved to work in the UK, could be providing the services from anywhere in the world. Of course, there are also challenges. For example, even the partial automation of counsellor's judgement may risk disengaging them.[10] It should remain possible for counsellors to opt-out of such automation features. Without room for therapists to exercise their human judgement we may risk disengagement and skill atrophy. Additionally, for therapists to be meaningfully involved in decisions, there should be an option for them to gain further explanation of a suggestion given by an automated system. Without a clear mechanism to understand and interact meaningfully with the automated system, the autonomy of the therapist could be undermined.

Privacy: Since therapy sessions involve two interlocutors, both sides have reasonable claims to privacy. As mentioned above, it's important that all users are given clear and accurate explanations about how the information collected from therapy sessions is being used. This is especially true since users, including counsellors, may not be aware of the value of their data.

One way of using the data is in job performance evaluations. If therapy sessions are being evaluated in terms of effectiveness and successful outputs, these evaluations could foreseeably be used as a metric to evaluate the job performance of therapists. This may not be problematic, so long as therapists have given consent for their information to be used in this way and the process of evaluation is transparent, and possibly not continual so as to feel like surveillance.

Finally, in a time where industries are increasingly moving from augmentation to automation, therapists may be concerned about contributing to the automation of their own profession. While re-skilling is possible in some cases, in industries that require a vast amount of specialised training such as mental healthcare, field switching may be both less practical and less attractive to workers. We will discuss the automation of human tasks in more detail in the section on Justice below.

Risks to developers

Potential risks for the developers of new technologies should also be considered. With supervised learning, for example, a human has to assign labels to the data used to train predictive algorithms. In the case of mental health therapies, this means that an employee must read and tag private and sensitive conversations between doctors and patients. One concern is that this could have **a potentially harmful psychological impact on developers**. Therapy sessions are likely to contain sensitive content that could be disturbing, distressing, or even triggering, depending on one's own life experiences and conditions. Such a labelling task might require training on mental health, so that the developer has the necessary context for what they might read as well as training on how to cope.

Relatedly, Sanches et al. express worry about 'burnout' for HCI researchers working in the challenging area of mental health and mention the need for greater peer and institutional support. They also suggest a rethink of "how such support can be explicitly factored in in the institutional ethics or research funds". We would advocate for something similar for developers of mental health software.

> ***Recommendation 3: Consider the impact, in both opportunities and risks, for all stakeholders involved in the development and use of mental health technologies.***

Justice

What is justice?

Justice is a complex ethical principle that is closely linked to fairness and equality, though is not quite the same as either.[38] Sanches et al. describe the principle as requiring the "fair distribution of benefits, risks and costs to all people irrespectively of social class, race, gender or other forms of discrimination." In medical ethics, the principle is often subdivided into three categories: (1) distributive justice, (2) rights-based justice, and (3) legal justice. Distributive justice requires the fair distribution of resources and is particularly concerned with scarce resources. Rights-based justice requires that people's basic human rights be respected.[39] Privacy and autonomy, for example, are widely recognized as human rights and hence the concerns raised thus far tend to fall under rights-based justice. Finally, legal justice requires that people's legal rights be respected. The development and implementation of new digital technologies in mental health care raises particular concerns about distributive and rights-based forms of justice.

Concerns around justice

There are two main areas in which to analyse distributive and rights-based justice within AI-driven mental health technologies: in the design process as well as in the distribution of the final product or service. In the first, compensation and credit for the human labour involved in algorithmic design must be considered; and in the second, questions about who is able to access and benefit from the service need to be considered.

The design process: heteromation and the value of human labour

In the design process, one type of ethical challenge arises from "heteromation": the extraction of economic value from low-cost (or free) labour.[40] This includes Amazon Mechanical Turk workers who are paid very low wages to annotate data or complete tasks that are difficult for an algorithm to do. It also includes the work of completing a Captcha, or other forms of reverse Turing tests, where a person must prove that they are human by completing a task like identifying and selecting all images of crosswalks in a series of 12 photos. These tasks automatically build training sets for algorithms that will eventually be able to accomplish these tasks. Hence, these incidences represent a transfer of intellectual property to the company for which the human labourers are not credited, as well as work for which they may not be adequately compensated. These issues can be addressed in some projects by disclosing the uses or seeking approval to use the data for research and development purposes. This has been done, for example, in EQClinic, a project in which a telehealth platform is used to help medical students improve their communication skills.[41]

A related concern in the development and prototyping of products is piloting on low-income or high-need/vulnerable populations. There are trade-offs, on-the-one-hand providing a service to a population that has a critical need for it and may be willing to try an earlier developed prototype, but on the other hand putting these vulnerable populations at risk by deploying or testing unfinished solutions. One area to potentially draw upon in considering these issues is the cost-benefit considerations at play in the treatment of rare diseases for which there are not known and tested cures.[42] When it comes to experimental medical treatments there is an absolute need to obtain informed consent, so that when patients agree to testing they do so with full understanding of the potential benefits and harms. It is important to make sure that any vulnerable population is informed about other options for care, so that they may decline new (especially experimental) treatments without feeling compelled to accept them if they are posed as their only opportunity to get care.

However, there may also be positive social justice outcomes that encourage early users to act as 'data altruists.' For example, early advances in algorithmic solutions can reduce costs of these services for future generations and expand access to less advantaged segments of the populations in the coming years. There is evidence that some people may be willing to share their data, even without direct compensation, if these benefits are communicated to them.[43]

Distribution: access and effectiveness

We now turn to questions around who is able to access and benefit from online text-based mental health services. A pressing issue of distributive justice in the context of mental health technologies is the lack of users from diverse socioeconomic and ethnic groups. There is a risk that this inequality will be exacerbated with the onset of new technologies for mental health, if efforts are not made to include and design for these populations. There are (at least) two distinct concerns here: one is about *access* to treatment, and the other is about the *effectiveness* of treatment.

Access: One positive feature of online therapies is that they can increase access for rural populations, who might otherwise have to commute long distances for therapy, and to working populations, who might otherwise have to take time off in order to attend a therapy session. In these ways, online therapy, reduces the barrier to entry and could increase uptake. Conversely, however, it is unfair to assume that low-income populations all have access to the necessary computing devices and stable internet connections. Burr and Morley (2019) have recently argued that genuine empowerment of the patient crucially depends on “the prior removal of certain barriers to engagement, which patients suffering from a variety of mental health conditions face.” As national health care services move increasingly toward online therapies, there needs to be research done on which populations are equipped for uptake, so that vulnerable communities are not left out. Beyond initial uptake, there is further evidence that minority populations tend to have lower levels of attendance and retention in mental health care.[44] Thus, there is a critical need for more research into the root causes, as well as novel interventions for increasing engagement of minority populations with the use of online therapies.

Effectiveness: A further concern is about the *effectiveness* of treatments. For example, research suggests that both first- and second-generation immigrants are at increased risk for psychotic disorders, such as schizophrenia,[45] while refugees settled in western countries could be ten times more likely to have post-traumatic stress disorder.[46] Findings such as these highlight the need for research on diverse populations, in particular to understand their accompanying risk factors for mental health conditions, as well as how they might respond to treatments differently. If online therapies are developed using a data set that only includes non-immigrants, or that lacks other forms of representation, e.g. diverse socioeconomic and ethnic groups, then the therapy will be optimized to treat only that homogeneous group. Hence, it is important that the training set for the algorithm really represents the diversity of the target population that will use the therapy. This comes with its own challenges, for example, understanding the wide variation in groups affected by mental illness, the potential challenges to reaching out to particular subsets of the population (e.g. “hard to reach populations”, such as the homeless, drug-addicted, illegal immigrants, etc.[47]) and how measures can be taken to include a diversity of individuals that will yield development of inclusive and beneficial mental health interventions.

> - ***Recommendation 4: Make known the value of human labour and intellectual property in the development of algorithms to all parties, and potentially compensate for it.***
>
> - ***Recommendation 5: Research the access requirements and unique mental health situations of diverse populations in order to ensure mental health technologies are effective and inclusive.***

Atrophy

What is atrophy?

Skill atrophy is the decline in abilities that comes from underuse or neglect to perform the behaviours and tasks that keep skills up to date. Over-reliance on technology has been cited as a contributor to atrophy of skills in many different contexts. Neuroscientist Manfred Spitzer coined the term ‘digital dementia’ to capture the various forms of cognitive atrophy that result from over-reliance on technology.[48] For example, researchers found that for older adults, relying on GPS may decrease their natural ability to spatially navigate.[49]

As more tasks are automated in the context of mental health, this could result in atrophy of previously used skills of both patients and therapists. Though there is a case to be made for the replacement of particular types of skills or activities with more worthwhile utilisation of human capacities (e.g. replacing repetitive calculations or data entry with creative or empathic pursuits) there are also risks to be managed, as atrophy can lead to dependence and even safety issues, e.g. if your spatial navigation skills atrophy and GPS fails, then you could be left in a dangerous situation.[50] These risks can necessitate the need to create fail safes, in case technology fails and people need to rely on past skills, or it might mean not introducing technology into realms where humans should remain critically vigilant or engaged, such as areas that require moral or ethical decision-making--and some areas of mental health care may be among these.

Concerns around atrophy

For patients, there is a risk of losing good-decision making skills and the ability to check-in with themselves, to self-reflect, to understand and troubleshoot symptoms and emotions. Technology can be a tool to prompt analysis of mood or symptom data, provide encouragement or trigger an alert for when to get help but if someone is entirely dependent on an app on their phone for self-reflection, things could spiral quickly in the case that they are decoupled from the device (e.g. due to a loss of network connection or battery power). Additionally, dependence on an app to manage care may result in lower feelings of self-efficacy, empowerment and control.[51]

For therapists, the introduction of technology into the diagnostic and therapeutic process could result in atrophy of critical professional skills. In cognitive behavioural therapy sessions, therapists interact closely with patients through structured discussion sessions in an attempt to break down problems into separate parts (thoughts, behaviours, actions) and then to suggest strategies that patients can use to change their thinking and behaviour. The success of these sessions depends on the therapist's ability to home in on problems, deconstruct them, engage patients, and suggest strategies to adopt. All of these steps are skills that therapists develop over time, and they are also all skills that can be augmented through AI and digital technologies. This in turn makes them susceptible to atrophy. For each of these skills, the concern is that if a therapist becomes over reliant on an app that aids her sessions, over time she may lose them and struggle to be as effective in face-to-face sessions with patients. As such, technologists will need to work closely with therapists to determine the most appropriate areas for automation/augmentation.

> ***Recommendation 6: Augmentation can be highly beneficial, but take care to ensure that over-reliance on technology does not lead to atrophy of critical skills.***

Authenticity

Why value authenticity?

Above we mentioned Weizenbaum's Eliza chatbot program -- a simple expert system pretending to be a psychologist. Despite designing Eliza, Weizenbaum himself maintained that machines will always lack certain 'human' qualities, including empathy and compassion.[52] Indeed, even if a computer chatbot were sophisticated enough to effectively demonstrate human empathy and compassion, such outward behaviour merely mimics human behaviour and is quite unlikely to reflect any true inner feelings. For this reason, **Weizenbaum warned that AI technologies should not be used in contexts that require human respect, dignity, and care, as without *authentic* empathy humans could be left feeling alienated and devalued.** The context of mental health care, of course, requires all of these--respect, dignity, and care. And, as we will discuss below, even partial automation, such as online text-based one-to-one therapy, if not implemented cautiously, can threaten the 'relational authenticity' between a therapist and patient.[53]

Opportunities and concerns around authenticity

In Hertlein et al.'s (2015) study of family and marriage counsellors' ethical concerns around online therapy, one theme that emerged was the impact to the therapeutic relationship. One participant expressed concern that there may be "missed information, lost feelings/understanding, lack of intimacy and disclosure." Another therapist worried that online therapy "lacks the opportunity for physical human interaction, such as offering a crying client a tissue or engaging in therapeutic touch, which could possibly act as a barrier to joining effectively with clients." These statements capture the kinds of concerns that Weizenbaum described: that the use of AI could lead to feelings of alienation and devaluation.

Participants in Hertlien et al.'s study also worried about the loss of quality in communication that may result from the lack of nonverbal cues and body language in online therapy. One participant wrote that **"key factors of the human experience" might be missing in online therapies**, including "social relationships and nonverbal communication." These non-verbal cues, including eye contact and social touch (e.g. handshakes), have been found to significantly influence patient perceptions of clinician empathy.[54] Hence, the loss of such nonverbal cues can make it more difficult for therapists to demonstrate empathy and to build authentic relationships with clients. In addition to concerns about alienation and devaluation, some evidence suggests that relational authenticity also encourages patient engagement and trust.[55]

On the positive side, **technological interventions in mental health may also provide novel opportunities that are not available in a strictly human-to-human context.** For example, the USC Institute for Creative Technology designed a 3D avatar that functioned like a virtual therapist but was not trying to perfectly emulate a human being.[56] The result was (somewhat surprisingly) positive: "Patients admit that they feel less judged by the virtual therapist and more open to her, especially if they were told that she was operated automatically rather than by a remote person."[57] This suggests that humans might be able to have differently authentic interactions with technologically mediated systems, if they are well designed. In their recent report, Sanches et al. (2019) express a desire to see "more novel designs of systems that foster and support beneficial human interactions, beyond the design of autonomous agents imitating empathy and aimed at replacing human contact." Designs such as these may be able to explore new ways of connecting with humans and eliciting beneficial relationships and experiences that are authentic in their own way, though not authentically human.

> ***Recommendation 7: Aim to support authentic human interactions and connectivity.***

Transparency

Why value transparency?

Transparency around the collection, use and storage of data is fundamental to ensure privacy rights, and other rights, such as informed consent, are upheld. There are many areas in which transparency must be integrated and addressed within an online text-based mental health platform, but there is also an added layer of complexity when considering transparency in this context. Much of this arises from the fact that the use of text-based counselling involves a mediating platform, which introduces other parties and intermediaries into what was traditionally a strictly confidential conversation between counsellor and patient. For example, tech developers need to be involved to design and support the platform, conversations will be recorded and analysed for potential introduction of AI capabilities, then these capabilities will need to be audited in order to ensure they will be functioning correctly. All of these new layers will require some degree of transparency.

Concerns around transparency

At a high level, **there should be some basic transparency around business models** since for-profit advertising or payments from insurance providers or employer health programs may come with incentives that conflict with the best interests of the patients. Funding sources and revenue models may create conflicts of interests in data sharing and breach the trust of patients.

Relatedly, **transparency in the collection, storage, and use of data is paramount to earning patient trust.** When signing up for a platform and consenting to therapy conducted in online formats, patients should have an understanding of who will have access to what parts of their data and why. As more data is collected and recorded, the parties who have oversight and access to patient notes and therapist-patient conversational records should be clear. Text-based therapy introduces the possibility for more and different interactions for patients with the data from the session, but this access comes with both benefits and risks which need to be carefully considered.[58]

Hence, **another transparency consideration is the communication of health information to patients.** Patient understanding of their personal health risks plays a critical role in their understanding of treatment options and enables shared decision-making in which patients and clinicians collaborate to incorporate information and patient values in treatment plans. In online mental health interventions, it will be important to think about how patient understanding of their condition and treatment plans can be best communicated. Evidence suggests that visual aids, such as icon arrays or bar graphs, can be useful in improving patient comprehension of the risks they are facing, and these modes of communication could be facilitated through mobile technology.[59] However, communication through mobile technology also risks missing important social cues that can indicate the level of patient-understanding.

Finally, **it should be clear how and where AI versus humans are used.** When people are asked to share personal and sometimes sensitive details about their lives, it is essential they know who they are speaking to: an AI, a human therapist, or a hybrid care team made up of both. Knowing this will engender trust as well as understanding and context around responses, perhaps with allowances made for strange responses or lack of empathy, should the AI go awry in ways that would be unacceptable from a human therapist respondent.

> ***Recommendation 8: Ensure transparency in all aspects of the use of mental health technologies as it is critical to safe and beneficial care.***

Summary of ethical concerns

We have focused on seven areas that raise particular ethical concerns for the design, development, and use of online one-to-one mental health therapies: (1) autonomy, (2) privacy, (3) impact, (4) justice, (5) atrophy, (6) authenticity and (7) transparency. This is not an exhaustive review of concerns but is instead meant to provide an overview of the key ethical considerations in the new and emerging application of AI and data-driven technologies to mental health care. A few other important concerns were raised throughout, such as informed consent, responsibility, safety, and beneficence. In general, we advocate that clients should be made aware of the potential benefits and risks of any online mental health therapy and that informed consent should be obtained before use.

User Experience and Design of Online Therapy

Arguably, the technology experience of people living with mental health issues can only be deeply understood by engaging directly with people with lived experience as part of collaborative design and evaluation processes.

While every design decision will have impact on users, design decisions can have stronger impacts on users living with mental health problems owing to the cognitive and affective load imposed by states of distress and illness. For example, in previous studies users have reported difficulty processing large amounts of text when depressed, or choosing to avoid the use of their phones.[60] Furthermore, feedback provided by mental health tracking technologies is likely to reflect negative moods or behaviour patterns which can exacerbate symptoms. In one study, participants reported feeling guilt, disappointment, and embarrassment about their tracked data.[5] It's easy to see how well-intentioned technology-led approaches, without the oversight of experienced mental health professionals, and deep involvement from people with lived experience, could very easily inadvertently cause harm. Many researchers have asserted that deep user involvement in technology development, such as through participatory and co-design methods, is an essential part of the solution.[61]

Human-centred and participatory design in health care

> *"The design of mental health technologies has been largely top down...We have typically not done a good job of getting input from patients about their goals, needs, or preferences. Trials often bear little resemblance to clinical settings, having largely emphasized internal validity over real-world issues, such as the technological environment and implementation and sustainment."*
>
> *- Mohr et al. 2017*[62]

Increasingly, leading researchers have expressed a need for more involvement of people living with mental health issues from the earliest stages of design, and doing so in ethical ways.[63] The lack of user involvement to date can be attributed, in part, to a traditional view of "expertise" being held exclusively by clinicians/ technologists/researchers. Within the frame of what is called "human-centred design", it's acknowledged that users are also experts--not of clinical practice or technology--but of their own experience, goals and contexts. This expertise is essential to effective design outcomes but is not available to technology makers unless they engage with users.

Deep user involvement is not only necessary in order for a technology to be genuinely useful and engaging to its audience, but is also arguably, a matter of design justice, in that it represents a more democratic and consultative approach. Early and ongoing user involvement is also foundational to autonomy-support (discussed below) and provides an input channel for users to share their concerns around privacy and transparency, as highlighted above and in Hertlein et al.'s study.

One approach to effective user involvement is to employ methods for "participatory design"[64]. These methods involve users as collaborators from the earliest exploratory phases of development. Technology designers play the role of facilitators enabling users to share experiences, create design ideas, and build and experiment with prototypes along with other team members. Orlowski et al. provide specific examples of practical applications of participatory design and design thinking methods for mental health technology.[65] Sanches et al. encourage the use of participatory design as a pathway to greater ethical sensitivity in mental health technology:

> *We would like to see more ethically sensitive design practices being applied to this area. For example, more participatory design methods including the voices of people living with affective disorders, as what we saw in literature often exposed limited understandings of their realities.*[5]

Despite the optimal benefits of participatory approaches and deep user involvement, the reality of some mental health contexts (including work with children or people with severe illness) can make access difficult or impossible for non-clinicians. In response, some organisations have established advisory groups of volunteers with lived experience who participate in an ongoing fashion as representatives for the broader group. Working with carers as facilitators for collecting input from users may be another pathway to enabling user involvement. In other cases, deep engagement with health providers or carers may be the only available proxy.

It's worth noting that involving even a very small number of users is still better than none and can help with interpretation of other outcome data (i.e. usage data). Doherty et al. provide specific recommendations for adapting the design process to varying levels of user access[60]. They also highlight the importance of clearly defined outcome measures beyond symptomatology, which might include measures of improved client self-efficacy, increased levels of self-reflection, or improved therapeutic relationships (e.g. a simple measure of increased conversation, or more sophisticated text analyses).

> **Recommendation 9: Employ a human-centred approach and design with, not just for, people.**

Access and Inclusivity

User involvement that adequately represents the diversity of the potential users of a service will help to prevent blindness to the reality of the wide spectrum of audience needs within mental health service provision. This includes differing requirements relating to low income, low literacy levels, limited access to computers, mobile phones and internet connections, as well as low technology literacy (even among young people.[66]

Moreover, users will prefer different modes of technology use at different times. For example, an insomnia therapy that doesn't require keeping a phone by the bed may be far more effective, while users may not feel comfortable using an audio or video-based program within public spaces. As such, designers should consider providing clients with multiple ways of accessing materials and consider how flexibility can be provided in the delivery of services.

International guidelines for digital accessibility and 'universal design'[67] provide essential starting points for ensuring a technology does not exclude users with older devices, limited internet access, physical disabilities, or other varying requirements.

It's also worth bearing in mind (as discussed above under ethical concerns around 'impact') that patients aren't the only "users" with specific needs. Users can also include any combination of clinicians, administrators, carers, family members, and others, depending upon the service in question. Moreover, these varied user experiences are interrelated. The user experience of a patient receiving therapy via an online technology will heavily depend on the user experience of the therapist delivering that therapy from the other end. Likewise, where the use of a technology will require the involvement of carers, parents or providers, their unique needs and expectations must also be accounted for.

Finally, it's worth noting the term "user" itself, while useful for its specificity within the technology context, can be inadvertently de-humanising, and in many cases, words like "human", clients", "patients", "people", or even “lives” may be far more appropriate.

> ***Recommendation 10: Follow guidelines for universal accessibility and tailor the level and mode of content to the spectrum of audience needs.***

Autonomy-support

> *“Essentially, clinical researchers have designed tools to try to get people to do what we want them to do and how we want them to do it”* - Mohr et al.[62]

The ethical imperative to support client/user autonomy (elaborated above under “Ethical Issues”) is taken from principle into practice within the context of user experience. In fact, work applying Self-determination Theory (a leading theory of motivation and wellbeing) to technology experience has identified support for autonomy as necessary for user satisfaction and sustained engagement[68]. The literature in SDT also provides guidance with regard to what characteristics constitute “autonomy-supportive” (v. controlling) interactions. According to this work, autonomy-supportive interactions:

- Understand the other’s perspective (frame of reference)
- Seek the other’s input and ideas
- Offer meaningful choices
- Empathize with resistance and obstacles
- Minimize use of controlling language or rewards
- Provide a rationale for requested or required behaviour.

In this way, autonomy-support translates into specific design decisions. Design for privacy as a target for design is also often considered a subset of autonomy-support[60]. Design implications may involve, for example, giving the client control over when data is sent, not playing audio or video without warning, or allowing for discreet use (i.e. studies have revealed problems with app titles that include stigmatised words like “mood”, or “mental health” as users felt others may notice them on their phone). Discreet design may also involve avoiding client-identifying data on the interface whenever possible (e.g. data graph screens that do not need to include the user’s personal details). Doherty et al. caution, however, that it may be necessary (depending on the context) to include a non-identifying username to reduce the risk of confusing different people’s data.

Motivational Interviewing is another example of the principle of autonomy-support translated into specific practice, in this case, in the context of conversational support for behaviour change which might be applied within online therapeutic contexts[69].

> ***Recommendation 11: Design to support user autonomy.***

Mental health technology as service design

> *“Reconceptualization of mental health technologies as Technology Enabled Services would highlight these interventions as services that are supported by technologies rather than as human-supported technologies. The implications of this reconceptualization are that the goals and strategies of the service, the role of the provider, and the technology must all be designed*

> *and evaluated simultaneously as an integrated service."* [62]

Mental health leaders have emphasised that technologies have largely been developed without an understanding for how they fit into the larger context of a user's social support systems and mental health services (despite the finding that most positive outcomes rely on these). [62,70] Mental health technologies are not stand-alone products. In order for technologies to be successful in the real world, it is necessary for technology makers to gain an understanding of the variety of relationships, social support systems and mental health services available to users, as well as how the introduction of a technology may impact and be impacted by these elements in the larger system. This includes providing clear pathways for users to seek other forms of qualified help, as well as ways for technology data to be shared with, or kept private from, other stakeholders.

Furthermore, for mental health technologies to be successful as part of a larger healthcare context, they must fit into the lives and workflows of the people who will make use of them and provide meaningful value rather than just adding another task to their workload. A human service system view also entails the early consideration of implementation and sustainment, as there's little point in investing in the prototyping and trialling of a system if there is no way for it to be implemented and sustained in the real world. Employing human-centred design from the beginning will help inform these considerations, as well as the related ethical concerns we raised around privacy, impact, and justice.

> ▸ ***Recommendation 12: Consider support structures and the larger service system in design.***

Evaluating impact

In addition to user research and involvement, a successful user experience relies on iterative improvement based on ongoing evaluation. Health technologies additionally require clinically-relevant efficacy trials. Owing to the potentially drastic consequences of ineffective (i.e. potentially harmful) mental health technology, evaluation of both user experience and health outcomes is an essential criterion for a responsible approach.

Evaluation might initially include expert review, heuristic evaluations and internal prototype testing, and be followed by pilot studies evaluating technologies with users until there is sufficient evidence of feasibility and benefit so as to rationalize a more formal clinical evaluation. Further evaluation after release of the product can inform improvements and upgrades and is necessary for determining impact and appropriation within the complex real-world context (which is often different to the controlled environment of clinical trials). A staged approach to the evaluation of mental health technologies is described in Doherty et al. 2010[60].

> ▸ ***Recommendation 13: Evaluate impact throughout development and after release***

Multidisciplinarity

When it comes to mental health technologies, technologists should *not* attempt to "go it alone". Ensuring that users, their contexts, the healthcare system, medical research, safety, ethical implications and many other critical considerations are given expert attention requires a multidisciplinary team.

Moreover, traditional approaches to "failing fast and failing often" are potentially disastrous in a health context in which people can't always safely be used as guinea pigs for a/b testing. As such, mental health professionals must be part of the design and development team. They can help ensure more rigorous, evidence-based and appropriately safe-guarded approaches are taken.

Experts in ethics should also contribute in order to effectively assess ethical considerations from multiple standpoints. It may also be helpful for them to work directly with user experience specialists to allow broad stakeholder input into ethical issues and concerns. Depending on the nature of the project other disciplines required may include social workers, sociologists, nurses, statisticians, anthropologists, among others.

> ***Recommendation 14: Ensure multidisciplinary collaboration and oversight***

Rigorous therapeutic and research methods

In addition to recruiting multidisciplinary teams and undertaking ongoing programs of evaluation, in order to prevent harm, technology approaches need to be grounded in research. Topham et al.[71] argue that it is an ethical responsibility "to ensure that mental health technologies are grounded in solid and valid principles to maximize the benefits and limit harm". Doherty et al. [60] similarly recommend that systems be based on accepted theoretical approaches for clinical validity.

Furthermore, a need for rigorous approaches should apply, not only to the therapeutic program employed, but also to the user research and evaluation practices. A human-centred focus on lived experience suggests the importance of mixed methods approaches that employ qualitative methods for uncovering insights into subjective experience, motivation, and the causes of both engagement and disengagement. These can complement and explain results from quantitative approaches, for example, the clinical measure of symptoms, behavioural analytics or surveys, which can also be used to test the generalizability of findings from qualitative work.

> ***Recommendation 15: Employ mixed-methods and research-based approaches for design and evaluation.***

Existing quality frameworks and guidelines

A number of quality frameworks, guidelines and ratings have been developed by multidisciplinary groups of researchers and these can be applied as a basic foundation for more responsible design. For example, The Transparency for Trust Principles[72] include questions around privacy and data security, development characteristics, feasibility and health benefits, and their creators advocate that all apps should be required to provide information relating to these four principles at minimum. More specific to mental health, the Psyberguide, developed by a non-profit network of mental health professionals, bases its ratings on criteria for credibility, user experience, and transparency[73]. The American Psychiatric Association has also come out with their own mental health app evaluation model to guide psychiatrists in navigating and guiding their patients through the vast landscape of available apps[74]. Technology-specific guidelines have also been developed, including the guidelines for the design of interventions for mental health on social media.[75] Development teams should consider available standards and guidelines relevant to their project context and consider how they can comply with and be guided by these.

> ***Recommendation 16: Apply health technology quality frameworks.***

Conclusion

Applying the recommendations

One way to apply the recommendations described herein is as a simple gauge to reflect on where a team is situated with respect to each guideline. For example, a technology in development might be considered against each recommendation and then assigned a status (e.g. “Not addressed”, “In the plans”, “Actively working on it”, “Addressed & under evaluation”, “Addressed & evaluated”.) Teams can choose to keep this self-evaluation private or share it with the public. An entire technology, or specific features, can be moved from one phase to another for each recommendation. This implementation example highlights the point that “thinking about an issue” isn’t the same as “implementing a solution” which isn’t the same as having gone so far as to evaluate actual impact. Different organisations will be at different stages for different services (and for different components of each service) and a method for tracking these can help ensure a more robust process.

Moving forward

We have focused the recommendations and analysis above on the context of online text-based one-on-one mental health therapy targeting the conditions of anxiety and depression. Though most of these recommendations would apply more broadly, further work is needed on ethical and design considerations for other contexts such as self-led or group-based online therapies and those designed to address other conditions.

There are many additional research questions related to the ethical design and application of AI in mental healthcare that require further study, some of which we have touched upon or introduced in the above discussions. One of these is in access to mental health technologies to better understand how to design mental health services tailored to the needs of diverse populations. This will involve in-depth understanding of these populations and their mental health concerns as well as how to design interventions that are accessible, empowering and effective. This data will also feed into algorithmic design to make mental health platforms more broadly beneficial and inclusive.

Much work is still to be done, especially as available technologies change at an ever-increasing tempo. However, coming back to core ethical and user experience principles can help provide some stability and consistency in an otherwise continually regenerating environment.

About the authors

Rafael A. Calvo. Professor at the Dyson School of Design Engineering at Imperial College London. Rafael's research focuses on the design of systems that support wellbeing in areas of mental health, medicine and education. Until 2019 he was Professor at the University of Sydney, and director of Wellbeing Technology Lab. In 2015 he was appointed a Future Fellow of the Australian Research Council to study the design of wellbeing-supportive technology. He is the recipient of five teaching awards and has published four books and over 200 articles in the fields of HCI, wellbeing-supportive design, affective computing, and computational intelligence. His books include *Positive Computing: Technology for Wellbeing* (MIT Press) and the *Oxford Handbook of Affective Computing*. He has consulted for organizations in the US, Europe, South America and Australia and is Associate Editor for the *IEEE Transactions on Technology and Society*, *Frontiers in Psychology-Human Media Interaction* and the *Journal of Medical Internet Research–Human Factors*.

Dorian Peters. Author, researcher and technology designer, Dorian has published two books on design: *Interface Design for Learning* (New Riders) and *Positive Computing: Technology for Wellbeing* (MIT Press). She specialises in design for health and digital wellbeing and is currently a visiting fellow at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. She creates human-centered technologies, facilitates design workshops and consults for non-profit and industry organizations such as the Movember Foundation, Stanford, Carnegie Mellon, Atlassian, and Google.

Diana Robinson. PhD candidate in computer science at the University of Cambridge and Student Fellow at the Leverhulme Centre for the Future of Intelligence, Diana specialises in human-computer interaction, philosophy and business. She was a Visiting Scholar at the MIT Media Lab in the Opera of the Future group. Prior to that, she worked as a Commodity Risk Analyst in BP's Integrated Supply and Trading business. She was a Princeton Project 55 Fellow in 2012/13. Diana holds an MBA from the Cambridge Judge Business School and a BA in philosophy from Princeton University.

Karina Vold. Postdoctoral Research Associate at the Leverhulme Centre for the Future of Intelligence and Research Fellow at the Faculty of Philosophy at the University of Cambridge. Karina specialises in philosophy of mind and artificial intelligence, and her recent work focuses on theories of cognitive extension, intelligence augmentation, AI ethics, and neuroethics. She received her BA in philosophy and political science from the University of Toronto and her PhD in Philosophy from McGill University. Karina is currently a Canada-UK Fellow for Innovation and Entrepreneurship and a Digital Charter Fellow at the Alan Turing Institute. She will begin as Assistant Professor of Philosophy, Humanities, and Technology at Carleton College in September 2020.

References

[1] According to a 2005 Green Paper on Improving the mental health of the population: Towards a strategy on mental health for the European Union, by the European Commission: https://ec.europa.eu/health/ph_determinants/life_style/mental/green_paper/mental_gp_en.pdf

[2] Calvo, R., & Peters, D. (2014). *Positive Computing: Technology for Wellbeing and Human Potential*. Cambridge: The MIT Press.

[3] Peek, N., Combi, C., Marin, R., & Bellazzi, R. (2015). Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. *Artificial intelligence in medicine*, 65(1), 61-73.

[4] Hoermann, S., McCabe, K. L., Milne, D. N., & Calvo, R. A. (2017). Application of synchronous text-based dialogue systems in mental health interventions: systematic review. *Journal of medical Internet research*, 19(8), e267.

[5] Sanches, P., Janson, A., Karpashevich, P., Nadal, C., Qu, C., Daudén Roquet, C., ... & Sas, C. (2019). HCI and Affective Health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[6] Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5), 649-685.

[7] Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, 22(5), 811-817.

[8] https://www.healthline.com/health/mental-health/chatbots-reviews#1

[9] https://www.psychiatry.org/psychiatrists/practice/mental-health-apps

[10] Naqshbandi, K., Hoermann, S., Milne, D., Peters, D., Davies, B., Potter, S., & Calvo, R. A. (2019). Codesigning technology for a voluntary-sector organization. *Human Technology*, *15*(1).

[11] Cauell, J., Bickmore, T., Campbell, L., & Vilhjálmsson, H. (2000). Designing embodied conversational agents. *Embodied conversational agents*, 29-63.

[12] Bickmore, T. and R. Picard. (2005). Establishing and Maintaining Long-Term Human-Computer Relationships. *Journal ACM Transactions on Computer-Human Interaction* (TOCHI). Vol. 12(2): 293-327.

[13] For an example, see the CNBC report from June 6, 2018: https://www.cnbc.com/2018/06/06/siri-alexa-google-assistant-responses-to-suicidal-tendencies.html

[14] Wykes, T., & Brown, M. (2016). Over promised, over-sold and underperforming? – e-health in mental health, *Journal of Mental Health*, 25:1, 1-4,

[15] https://www.ipsos.com/ipsos-mori/en-uk/future-data-driven-technologies-and-implications-use-patient-data

[16] Some discussed in Naqshbandi et al., 2019 above.

[17] Kim, J. Y., Calvo, R. A., Yacef, K., & Enfield, N. J. (2019). A Review on Dyadic Conversation Visualizations-Purposes, Data, Lens of Analysis. arXiv preprint arXiv:1905.00653.

[18] http://blog.castac.org/2017/05/automation-and-heteromation/

[19] Calvo, R. A., D'Mello, S., Gratch, J., & Kappas, A. (Eds.). (2015). *The Oxford handbook of affective computing*. Oxford Library of Psychology.

[20] Wu, K., Liu, C., Taylor, S., Atkins, P. W., & Calvo, R. A. (2017). Automatic mimicry detection in medical consultations. In 2017 IEEE Life Sciences Conference (LSC) (pp. 55-58). IEEE.

[21] Torous, J., Gershon, A., Hays, R., Onnela, J. P., & Baker, J. T. (2019). Digital Phenotyping for the Busy Psychiatrist: Clinical Implications and Relevance. Psychiatric Annals, 49(5), 196-201.

[22] Hertlein, K. M., Blumer, M. L. C., & Mihaloliakos, J. H. (2015). Marriage and Family Counselors' Perceived Ethical Issues Related to Online Therapy. The Family Journal, 23(1), 5–12.

[23] Mittelstadt, B. and Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Sci Eng Ethics* 22:303–341.

[24] Burr, C., & Morley, J. (2019). Empowerment or Engagement? Digital Health Technologies for Mental Healthcare. Digital Health Technologies for Mental Healthcare (May 24, 2019).

[25] Mill, J.S. (1859/1975). *On Liberty*, David Spitz, ed. New York: Norton.

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

[27] Susser, D., Roessler, B., & Nissenbaum, H. (2018). Online Manipulation: Hidden Influences in a Digital World. Available at SSRN 3306006.

[28] For example, Susser et al., 2018; Vold, K. and Whittlestone, J. (Forthcoming). Privacy, Autonomy, and Personalised Targeting: Rethinking How Personal Data is Used. Report on Data, Privacy, and the Individual in the Digital Age, by IE University's Center for the Governance of Change; and Lanzing, M. (2018). "Strongly Recommended" Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies. *Philosophy and Technology*, 1-20.

[29] Sax, M., Helberger, N. & Bol, N. (2018). Health as a Means Towards Profitable Ends: mHealth Apps, User Autonomy, and Unfair Commercial Practices. Journal of Consumer Policy 41: 103.

[30] For example, see: Selinger, E., & Whyte, K. (2011). Is there a right way to nudge? The practice and ethics of choice architecture. *Sociology Compass* 5(10), 923–935; and Petrescu, D. C., Hollands, G. J., Couturier, D. L., Ng, Y. L., & Marteau, T. (2016). Public acceptability in the UK and USA of nudging to reduce obesity: The example of reducing sugar-sweetened beverages consumption. *PLOS ONE* 11(6): e0155995.

[31] Sunstein, C. R. (2015). The ethics of nudging. *Yale Journal on Regulation* 32, 413–450.

[32] For in depth discussions see Solove, D. (2008). *Understanding privacy.* Cambridge, MA: Harvard University Press, and Roessler, B. (2005). *The value of privacy*. Cambridge: Polity Press.

[33] Lanzing, M. (2018). "Strongly recommended" revisiting decisional privacy to judge hypernudging in self-tracking technologies. Philosophy & Technology, 1-20.

[34] Further discussion in Hertlein et al., 2015, and Nelson, W.A. (2010). The ethics of telemedicine. *Healthcare Executive*, 25, 50-53.

[35] Westin, A. (1967). Privacy and freedom, New York: Atheneum.

[36] See: Vold and Whittlestone, forthcoming; Lanzing, 2018; and Susser, Daniel and Roessler, Beate and Nissenbaum, Helen F., Online Manipulation: Hidden Influences in a Digital World (December 23, 2018). Available at SSRN: https://ssrn.com/abstract=3306006

[37] See: Hu, H.-J. Zeng, H. Li, C. Niu, & Chen, Z., (2007). Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web, 2007* Banff, AB, ACM: pp. 151-160; and Kosinski, M., Kohli, P., Stillwell, D. J., Bachrach, Y., & Graepel, T. (2012). Personality and website choice. *ACM Web Science Conference*, Evanston, Illinois, USA, 251–254.

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

[39] Following Gillon R. (1994). Medical ethics: four principles plus attention to scope. *BMJ (Clinical research ed.)*, *309*(6948), 184–188.

[40] Ekbia, H. R., & Nardi, B. A. (2017). Heteromation, and other stories of computing and capitalism. MIT Press.

[41] Liu, C., Scott, K. M., Lim, R. L., Taylor, S., & Calvo, R. A. (2016). EQClinic: a platform for learning communication skills in clinical consultations. *Medical education online*, *21*, 31801.

[42] Morel, T et al. "Quantifying benefit-risk preferences for new medicines in rare disease patients and caregivers." *Orphanet journal of rare diseases* vol. 11,1 70. 26 May. 2016,

[43] See, for example, Chelsea L. R., Kaphingst, K. A., and J. D. Jensen. (2018). When Personal Feels Invasive: Foreseeing Challenges in Precision Medicine Communication. *Journal of Health Communication* 23:2, pages 144-152; and Halvorson, G., and B. Novelli. 2014. Data altruism: Honoring patients' expectations for continuous learning. Commentary, Institute of Medicine, Washington, DC.

[44] Alegría, M., Polo, A., Gao, S., Santana, L., Rothstein, D., Jimenez, A., … Normand, S. L. (2008). Evaluation of a patient activation and empowerment intervention in mental health care. *Medical care*, *46*(3), 247–256.

[45] Borque, F. van der Ven, E. and A. Malla. (2011). A meta-analysis of the risk for psychotic disorders among first- and second-generation immigrants. Psychol Med. 41(5): 897-910.

[46] Fazel, M., Wheeler, J., & Danesh, J. (2005). Prevalence of serious mental disorder in 7000 refugees resettled in western countries: a systematic review. The Lancet, 365(9467), 1309–1314.

[47] Bonevski B, Randell M, Paul C, et al. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Med Res Methodol*. 2014;14:42. Published 2014 Mar 25.

[48] Spitzer M. (2012). "Digital Dementia: What We and Our Children are Doing to our Minds." München: Droemer, 7.

[49] Konishi, K. and Bohbot, V. D., (2010). Grey matter in the hippocampus correlates with spatial memory strategies in human older adults tested on a virtual navigation task. Abstract Society for Neuroscience's Annual meeting (2010).

[50] Hernández-Orallo, J. & Vold, K. (2019). AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI. AAAI /ACM Conference on Artificial Intelligence. Ethics, and Society (AIES 2018), Honolulu, Hawaii, USA. January 27-28, 2019.

[51] Hernández-Orallo and Vold, 2019.

[52] Weizenbaum, J. (1976). Computer Power and Human Reason: From Judgment to Calculation. San Francisco: WH Freeman.

[53] Turkle, S. (2011). Authenticity in the age of digital companions. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 62–76).

[54] Montague E, Chen P, Xu J, Chewning B, Barrett B. (2013). Nonverbal interpersonal interactions in clinical encounters and patient perceptions of empathy. J Participat Med. 5:e33.

[55] Laugharne, R., Priebe, S., McCabe, R., Garland, N., & Clifford, D. (2012). Trust, choice and power in mental health care: Experiences of patients with psychosis. *International Journal of Social Psychiatry*, *58*(5), 496–504.

[56] Brigida, A. (2013). A Virtual Therapist. USC Viterbi School of Engineering, Online Blog. October 18, 2013. Available online: https://viterbi.usc.edu/news/news/2013/a-virtual-therapist.htm

[57] Tieu, A. (2015). We now have an AI Therapist, and she's doing her job better than humans can. Futurism: The Byte. July 16, 2015. Available online: https://futurism.com/uscs-new-ai-ellie-has-more-success-than-actual-therapists

[58] Kahn, M.W., Bell, S.K., Walker, J., Delbanco T. (2014). Let's Show Patients Their Mental Health Records. *JAMA.*2014;311(13):1291–1292.

[59] Zipkin, D. A., Umscheid, C. A., Keating, N. L., Allen, E., Aung, K., Beyth, R., … Feldstein, D. A. (2014). Evidence-based risk communication: a systematic review. Annals of Internal Medicine, 161(4), 270–80.

[60] Doherty, G., Coyle, D., & Matthews, M. (2010). Design and evaluation guidelines for mental health technologies. Interacting with Computers, 22(4), 243–252.

[61] See: Sanches, 2019; Orlowski, 2018; and Mohr et al., 2017

[62] Mohr, D. C., Weingardt, K. R., Reddy, M., & Schueller, S. M. (2017). Three Problems with Current Digital Mental Health Research . . . and Three Things We Can Do About Them. Psychiatric Services, appi.ps.2016005.

[63] See: Sanches, 2019; Orlowski, 2016; and Mohr, 2017

[64] Jesper Simonsen and Toni Robertson (Eds.). 2012. *Routledge international handbook of participatory design.* Routledge.

[65] Orlowski, S., Matthews, B., Bidargaddi, N., Jones, G., Lawn, S., Venning, A., & Collin, P. (2016). Mental Health Technologies: Designing With Consumers. JMIR Human Factors, 3(1), e4.

[66] Robotham, D., Satkunanathan, S., Doughty, L., & Wykes, T. (2016). Do we still have a digital divide in mental health? A five-year survey follow-up. *Journal of medical Internet research*, *18*(11), e309.

[67] See web accessibility standards (https://www.w3.org/standards/webdesign/accessibility), the universal design guide (http://universaldesign.ie) and the Inclusive Design Toolkit (http://www.inclusivedesigntoolkit.com)

[68] Robotham, D., Satkunanathan, S., Doughty, L., & Wykes, T. (2016). Do we still have a digital divide in mental health? A five-year survey follow-up. Journal of Medical Internet Research, 18(11).

[69] See: Shingleton, R. M., & Palfai, T. P. (2016). Technology-delivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research. *Patient education and counseling*, *99*(1), 17-35.

[70] Brugha, T. S., Bebbington, P. E., MacCarthy, B., Sturt, E., Wykes, T., & Potter, J. (1990). Gender, social support and recovery from depressive disorders: a prospective clinical study. *Psychological Medicine*, *20*(1), 147-156.

[71] Phil Topham, Praminda Caleb-Solly, Paul Matthews, Andy Farmer, and Chris Mash. 2015. Mental Health App Design: A Journey From Concept to Completion. In Proceedings of the 17th International

Conference on Human-ComputerInteraction withMobile Devices and Services Adjunct (MobileHCI'15). ACM, New York, NY, USA, 582–591.

[72] Wykes, T., & Schueller, S. (2019). Why Reviewing Apps Is Not Enough: Transparency for Trust (T4T) Principles of Responsible Health App Marketplaces. *Journal of Medical Internet Research*, *21*(5), e12390.

[73] See: http://www.Psyberguide.org

[74] See: https://www.psychiatry.org/psychiatrists/practice/mental-health-apps/app-evaluation-model

[75] Manikonda, L., & De Choudhury, M. (2017, May). Modelling and understanding visual attributes of mental health disclosures in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 170-181). ACM.