Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support

Biomedical Informatics Insights
Volume 11: 1–9

DOI: 10.1177/1178222619829083
SAGE

Kira Kretzschmar[1], Holly Tyroll[1], Gabriela Pavarini[2], Arianna Manzini[2] and Ilina Singh[2]; NeurOx Young People's Advisory Group

[1]Oxford Neuroscience, Ethics and Society Young People's Advisory Group, Department of Pychiatry, University of Oxford, Oxford, UK. [2]Department of Psychiatry and Wellcome Centre for Ethics & Humanities, University of Oxford, Oxford, UK.

ABSTRACT: Over the last decade, there has been an explosion of digital interventions that aim to either supplement or replace face-to-face mental health services. More recently, a number of automated conversational agents have also been made available, which respond to users in ways that mirror a real-life interaction. What are the social and ethical concerns that arise from these advances? In this article, we discuss, from a young person's perspective, the strengths and limitations of using chatbots in mental health support. We also outline what we consider to be minimum ethical standards for these platforms, including issues surrounding privacy and confidentiality, efficacy, and safety, and review three existing platforms (Woebot, Joy, and Wysa) according to our proposed framework. It is our hope that this article will stimulate ethical debate among app developers, practitioners, young people, and other stakeholders, and inspire ethically responsible practice in digital mental health.

KEYWORDS: Chatbots, apps, therapy, mental health, artificial intelligence, human-computer interaction, conversational agent, young people, digital mental health, youth mental health

RECEIVED: February 7, 2018. **ACCEPTED:** December 17, 2018.

TYPE: Proceedings from the digital mental health conference, London, 2017 - Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Wellcome Trust (104825/Z/14/Z). A.M. is, in addition, supported by a Wellcome Trust studentship (203329/Z/16/Z) and I.S. is, in addition, supported by the NIHR Oxford Health Biomedical Research Centre (IS-BRC-1215-20005) and the Wellcome Centre for Ethics and Humanities, which is supported by core funding from the Wellcome Trust (203132/Z/16/Z).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Gabriela Pavarini, Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford OX3 7JX, UK.
Email: gabriela.pavarini@psych.ox.ac.uk

Introduction

The last two decades have seen a proliferation of Internet-delivered and mobile mental health interventions both within research settings[1,2] and on the market (eg, 7 Cups).[3] Unsupported or unguided interventions offer fully automated self-help services, whereas supported or guided interventions provide additional human support by a remote coach or clinician.[4] Research studies on the clinical benefits of these interventions for the treatment of anxiety and depression have shown mixed results,[5–7] but most have denounced the poor adherence that characterises digital mental health interventions.[8] More recently, a number of engaging, fully automated conversational agents, or chatbots, have been made available (eg, Woebot).[9] These text-based platforms, easily accessible via a mobile app or Facebook Messenger, are designed in a way that increases adherence, because they engage with users in ways that resemble human real-life interactions.[10,11]

What are the advantages and disadvantages of using fully automated conversational agents for mental health support? What ethical standards should guide this new resource? How can we ensure therapy bots are safe and effective? As a group of researchers and young people involved in ethical debates that arise from advances in technology, neuroscience, and psychiatry,[12] in this article, we hope to shed light on these questions from a young person's perspective. We also offer a detailed analysis of 3 existing platforms that are currently available in English language: Woebot,[13] Joy,[14] and Wysa.[15]

Our Approach to Mental Health Chatbots: The Importance of the Youth Perspective

The ideas presented here are the result of group discussions concerning the pros and cons of mental health chatbots, in which the Oxford Neuroscience, Ethics and Society Young People's Advisory Group (NeurOx YPAG) participated. The NeurOx YPAG is a multi-ethnic group of 14 to 18 year olds, who have been selected from several schools across Oxfordshire for their interest in ethics and mental health. The group meets periodically to discuss ethical issues in research at the intersection between technology, neuroscience, and psychiatry. The authors of this article include a sub-set of 24 members of the NeurOx YPAG who participated in the session on mental health chatbots, as well as 3 co-leaders of the NeurOx YPAG who facilitated the session. The event was organised by the BBC in partnership with the Wellcome Trust and resulted in a BBC Tomorrow's World video clip titled 'Would you trust a chatbot therapist?'.[16]

We believe that the perspectives of young people are essential to such discussion, because mental illness accounts for a

large portion of the disease burden in younger populations,[17] yet young people's mental health needs are largely unmet, due to under-investment in child and adolescents' mental health, lack of evidence-based treatments, and poorly targeted or badly implemented interventions.[18,19] At the same time, young people are the largest consumer of the digital world, and as such they constitute one of the key target groups for digital interventions.[20] For instance, Facebook Messenger, within which 2 of the 3 chatbots reviewed here are integrated, is used by 81% of 18- to 24-year-old Internet users and by 89% of 25- to 34-year-old users in the United States, compared with 67% of users between 55 and 64 years old.[21] There is also evidence that, although teenagers still consider their parents, doctors, or nurses as their main source of health information, the Internet has far overtaken other media, such as television or newspapers, as an important supplement for such sources.[22]

Mental health issues such as stress, anxiety, and depression are among the health information most searched for by teenagers online, and online content often motivates young people to change their health behaviours.[22] These data suggest that digital mental health interventions are likely to affect young people in a very important dimension of their lives. At a time where young people are seen as simultaneously in need of protection from the risks of the digital world, and as 'media-savvy' individuals,[23] who are much more confident and competent in the use of digital devices than adults,[24] we believe that young people's first-hand experience of mental health chatbots should play a role in shaping the normative debate around e-Mental Health.

What is the Current Landscape of e-Mental Health?

Online and mobile-delivered mental health interventions are proliferating and hold promise to overcome important barriers in the delivery of and access to traditional mental health support. Digital interventions are accessible to anyone with a smartphone and Internet connection, and can therefore be delivered to young people in regions that lack mental health professionals and where mental illness often goes untreated.[25] Even in contexts where specialised, publicly funded treatments are available, waiting lists are common in child and adolescent services.[26,27] Mental health apps are readily accessible and easy to use, whenever users feel sad, anxious, stressed, or just want a distraction. They are also significantly less costly than face-to-face interventions such as cognitive behavioural therapy.[28]

Beyond structural barriers, young people are often reluctant to access mental health treatment due to social and self-stigmatising attitudes to mental health interventions.[29] Many report a desire for self-reliance when coping with emotional difficulties, or a preference for obtaining support from close others such as family members or friends rather than professional support.[30,31] There is also evidence that younger teenagers tend to feel more in control of managing difficult situations in online conversations via text rather than in in-person interactions.[23] For example, the UK suicide-prevention charity Samaritans has recorded that users aged less than 25 years had the highest use of their text messaging service and were less likely to phone or visit a branch than older users.[32] Thus, as technology and digital culture become increasingly more present in young people's lives, young people may also prefer to look for support online rather than face-to-face. An online, mobile-based intervention is less likely to carry the stigma attached to formal mental health services[33] and provides a self-reliant intervention platform for those who would otherwise be reluctant to seek support.

Another common barrier to seeking other people's support concerns issues of trust and confidentiality. For example, many young people feel that their problems are too personal to be discussed with anyone or they fear that their sensitive information will be shared with others,[34,35] and particularly younger adolescents consider it easier to maintain their privacy in online conversations.[24] These individuals may find a fully automated digital intervention that could be used anonymously to be a suitable alternative outlet for disclosure of their mental health difficulties. Similarly, a recent study found that adults were more willing to disclose personal information when they thought they were talking to a nonhuman 'virtual therapist' than a human operating platform.[36]

Finally, if effective, e-Mental Health services may represent the less daunting bridge to getting professional help.[37] It is possible that, as individuals use these resources and learn about cognitive behavioural and other clinical approaches, they might develop skills to recognise when they need additional support and become progressively more open to talking to a real human. For patients who seek face-to-face help first, mental health apps may still be recommended by mental health professionals as a supplement to therapy,[3] or as a form of intermediate support while on the waiting list.

There is some evidence that the first generation of mental health digital interventions can be effective for conditions such as anxiety and depression[38]. However, as previously discussed, such interventions are also characterised by poor adherence, which may be due to the lack of the quality of human interaction that a therapist-patient relationship offers.[10] This is particularly the case for unguided interventions,[4,11] which suggests that complementing digital help with external human support may be a way to increase adherence to e-Mental Health services. However, such option is likely to be difficult to implement on a large scale.[11] Therefore, fully automated conversational agents, or chatbots, could offer a promising alternative.

Powered by artificial intelligence, the new generation of mental health platforms is meant to increase adherence by providing an engaging tool that, although cannot offer a proper therapeutic interaction, is designed to feel like users are speaking to a real human or to 'mimic human dialogues'.[11] Like other conversational agents (such as Apple's Siri), chatbots provide a

'more natural medium' for communication than the older mental health apps.[10] Chatbots process all text (and emojis) that a user might enter, and offer responsive, guided conversations and advice to help users cope with challenges to mental health. The bots offer daily check-ins on users' emotions, thoughts, and behaviours ('What is your energy like today?') and some, such as Wysa, also passively track users' movements via the phone's accelerometer.

Users receive reports that might help them gain insight into their own patterns, as well as targeted therapy exercises, including reframing one's thoughts, mindful breathing, and motivational interviewing, in the form of text, games, or video clips. According to their developers, chatbots can be used in settings with slow Internet connection, because they are designed to use limited phone data.[39] Finally, because they are commercially accessible rather than available in academic research settings only, chatbots are likely to be more sustainable than other e-Mental Health services.[10]

We have tried 3 chatbots for mental health support: Wysa, Woebot, and Joy, which have been selected because they are widely used and available directly to users, in English language. Woebot is currently available both via Facebook Messenger (16 k likes) and more recently also as a standalone mobile application (50 k+ downloads); at least in the United Kingdom Joy is currently only available via Facebook Messenger (1.3 k likes); Wysa was previously available on Messenger (3.3 k likes), but the company has discontinued this service and now the chatbot is only available as an app (500 k+ downloads). Wysa was created by Jo Aggarwal and Ramakant Vempati, initially as a side study of a larger project meant to build machine learning algorithms to detect depression from sensor feeds in mobile phones. It was launched in October 2016 and has now more than 40 million conversations and is available in 30 countries. Woebot was founded by Dr Alison Darcy and launched in June 2017 and now has more than 2 million conversations per week, across more than 120 countries.[13] Finally, Joy, which is used in more than 130 countries, was created and launched by Danny Freed in 2016 and has more than 1 million messages. It is important to note that Joy has recently undergone changes to their website and services, and is currently described as a 'mental health monitoring device', but in this paper we only consider their chatbot available on Facebook Messenger.

According to the information on their websites, these 3 chatbots were developed based on cognitive behavioural techniques.[13–15] Wysa and Joy, in addition, mention adopting therapeutic methods drawn from other third-wave cognitive behavioural therapy approaches such as dialectical behaviour therapy and mindfulness-based methods.[14,15]

Beyond the chatbots we have tested, we are aware of a few other platforms that operate under similar principles. Those include Shim, which is currently available in Swedish language (English version in Beta) as a mobile application for iOS,[40] as well as a number of mental health bots developed by tech startup X2AI,[41] including Karim (Arabic), Emma (Dutch), and Sara (English). Sara is available on Facebook Messenger @ chatwithsara (850+ likes), and is a demo version of Tess, a mental health chatbot available to hospitals and other organisations for a fee.

Similarly to other mental health mobile apps,[7] there is an apparent paucity of published evidence of the effectiveness of chatbots. From the 3 chatbots we have tested, only Woebot has published evidence from a randomised control trial. After talking to Woebot for 2 weeks, a sample of US college students (who self-identified as experiencing symptoms of depression and anxiety) showed a reduction in depression symptoms in comparison to a randomised control who only received information about depression via an e-book.[10] No difference between groups was observed for anxiety symptoms (both groups showed a reduction in anxiety) or frequency of positive and negative emotions. Eighty-five per cent of participants used the bot daily or almost daily, which is high compared with other web-based interventions.[42] A qualitative assessment revealed that participants felt generally positive about the experience, but acknowledged technical limitations; for example, many felt they were not able to have a 'natural' conversation. In terms of the positive features, they particularly appreciated daily check-ins, the bots' empathic and caring 'personality', and the learning it facilitated. More recently, Wysa has also published evidence indicating that a sample of active Wysa users showed a higher reduction in depression symptoms in comparison to less engaged app users, and that overall users found the platform helpful and encouraging.[43]

The Swedish chatbot Shim was also shown to produce increases in psychological well-being and reductions in perceived stress in a nonclinical sample, in comparison to a waiting-list control group.[11] However, these effects only applied to participants who had actively adhered to the intervention.

Taken together, these initial findings are encouraging, but of course warrant replication with larger and more diverse samples, including clinical populations. It is also important to compare these interventions to face-to-face treatments to ascertain the magnitude of the effect in comparison to traditional methods.

Current Limitations of Fully Automated Therapy in Mental Health

Automated bots are still far from recreating the richness of a face-to-face encounter with a mental health professional, despite their efforts to mirror real-life interactions.[10] Even though a minimal level of personalisation exists (eg, different tips/strategies are given for users presenting symptoms of depression vs anxiety), the support provided is still generic and perhaps more akin to a self-help book. That is, as of yet, chatbots cannot grasp the nuances of users' life history and current circumstances that may be at the root of mental health difficulties. As Woebot warns its users: 'As smart as I may seem, I'm not capable of really understanding what you need'.

For example, when the user sends longer or more complex messages, chatbots often reply not having understood or provide an off-topic, inappropriate response which, although might be comical and entertaining at times, could undermine the user's sense that the chatbot is 'listening' carefully.

Some of the automated platforms provide users with set responses to click on (eg, 'high', 'medium', or 'low', referring to energy levels), in addition to free text, which facilitates comprehension. Even though these set phrases might help users label their subjective experiences, they could also leave users feeling limited and unable to express themselves properly. All in all, there is a long way to travel before chatbots will be perceived as highly responsive and empathetic; it does not yet feel as though the technology is able to fully tailor responses to users' specific needs and circumstances.

Results of Our Discussion: Minimum Ethical Standards

Despite the above-mentioned limitations, automated chatbots may offer great potential for providing people with useful help for mental health difficulties. There are, however, several ethical issues associated with this potential. While testing the chatbots, we were particularly concerned about matters related to who has access to users' personal information and conversations; whether the digital support provided is evidence-based; and how automated bots protect users' safety in emergency situations. For their potential not to be compromised, we believe that automated bots should meet a set of minimum ethical standards concerning privacy and confidentiality, efficacy, and safety. In the remainder of this article, we will outline the ethical recommendations for chatbot creators that we developed during our group discussion about mental health chatbots. Such recommendations come, therefore, from the first-hand experience of young users.

Privacy and Confidentiality

1. Personal information, if collected, should be kept confidential;
2. Content of conversations, if shared, should be de-identified;
3. Privacy arrangements and limitations should be made transparent to users;
4. Users should have the option of being reminded of privacy arrangements and limitations at any stage.

In the informational age, privacy concerns are relevant to a variety of platforms that have become part of our everyday lives.[44,45] Addressing them is particularly compelling in the context of the chatbots discussed here, due to the sensitivity of information about users' mental well-being. Within traditional mental health settings, patients find it essential that their clinicians protect and keep their information confidential.[46] As previously mentioned, young people in particular consider the availability of trusted relationships to be a key motivational factor for seeking professional support.[29,47,48] Because therapy bots mirror real-life interactions with mental health support providers, we believe that chatbot developers should keep users' data private as far as possible. If anything is shared—assuming explicit consent is provided—it is essential that it does not include any personally identifiable information (eg, names, email addresses, and phone numbers), which should be kept strictly confidential.

When we tried the platforms, we felt more comfortable with disclosing information about our mental well-being when we had the chance to chat anonymously. On this point, the most important difference that we identified is that Wysa is only available as an app, which gives users the possibility to chat anonymously. On the other hand, Woebot and Joy are (also) available through Facebook Messenger, where all conversations are linked to users' real names (unless users create a false Messenger account).

We find having an independent mobile application very important, because data collected on Facebook Messenger are subject to Facebook's privacy policy and can be shared with third parties.[49] Having the chatbot run through an external platform means a lower degree of control over the information collected and greater vulnerability to the potential release of secure or private/confidential information to an untrusted environment. In fact, since we wrote the first draft of this article, Facebook was drawn into the Cambridge Analytica data scandal, where data from millions of users were shared for political purposes without their explicit consent. This incident highlights the need for caution when it comes to using external platforms, and the importance of ongoing oversight and monitoring to ensure compliance with privacy standards.

Wysa and Woebot indicate in their privacy policy that they do not share user content with other companies or services. They only use anonymised, aggregated data to improve and/or optimise their services. On the other hand, Joy's privacy policy states that by using their platform, users grant the company a perpetual and transferable license to 'use, edit, modify, truncate, aggregate, reproduce, distribute, prepare derivative works of, display, perform, and otherwise fully exploit the User Content', which we found very concerning.

With regards to anonymisation, it is important to note that even if platforms do not explicitly collect personal information, given the nature of these services users may type identifiable information in conversation with the bot. We find it important that any potential identifiers are hidden or removed as far as possible when data are used to optimise services, for research, or if any information is shared with third parties (assuming, of course, that users' have explicitly consented to this sharing).

If data are used for research, we find it important that this is explicitly stated and consent is sought before we begin using the platform. We would also prefer to give specific consent for specific studies, especially when data about our mental well-being are used. Critically, details concerning privacy

arrangements should be made transparent to users. We suggest that the best way to inform users about privacy arrangements is to outline these *within* the chat in an easy language and format (eg, graphically or in bullet points). Although we think that users should be given the option to click on a link to read full privacy policies on the bots' websites, we find it unlikely that many people would take the time to do so. It is, therefore, important that basic information about privacy is easily provided via the text platform *during users' first interaction* with the bot.

All the bots we tried included their privacy policy on their websites, but only Wysa and Woebot offered privacy information during our first conversation. For example, Wysa mentioned it would 'not share anything we discuss with anyone else' and Woebot told us that it would not share 'any information that you provide to Woebot.' However, we suggest that when used on Messenger, Woebot should provide further information about Facebook's privacy and data sharing policy in addition to providing a link to Facebook's privacy policy page.

Finally, users should have the option of being reminded of confidentiality arrangements at any point. We believe that the most user-friendly way to do so is to programme chatbots so that, if words such as 'privacy' or 'confidentiality' are typed into the conversation, an automated and up-to-date reminder of privacy policies is generated. Similar algorithms are already used by existing therapy bots. For instance, by messaging 'report' to Joy, its users receive an emoji-based report of their mood and mental wellness over the last 2 weeks.

In sum, we believe that privacy and transparency are of utmost importance; a lack of transparency may deter some people from using automated chatbots or undermine their trust in the platform. This may apply particularly to older teenagers, who tend to consider private in-person conversations more secure than online communication.[23] Moreover, lack of transparency can change the balance of risk and benefits for the user. We also suspect that adherence to our ethical recommendations regarding information privacy, which come from the first-hand experience of young users, will increase the number of people who are willing to share information about their mood and well-being, as well as the quantity of information shared by single users, which, in turn, will make artificial intelligence–based tools more powerful in addressing people's mental health needs.

Efficacy

1. The support provided should be evidence-based;
2. The platforms should be tested empirically;
3. Users should be informed about the extent to which the service is backed up by evidence;
4. Users should be informed about what the chatbot targets and what effects to expect.

As explained in the introductory sections, to the best of our knowledge, (preliminary) empirical evidence on the efficacy of automated chatbots has been published for Woebot and Wysa, but not Joy. As a general rule, we consider it important that the support offered by mental health chatbots is based on clinical approaches that have been empirically supported. We also find it relevant that the specific platforms are tested for their psychological or clinical effects via randomised controlled trials, and with clinical samples, especially if chatbots start to be recommended by mental health professionals as a supplement or intermediate support to therapy.[7]

Besides using evidence-based techniques and testing the efficacy of the apps, we find it important that the platform informs users about: (1) the theoretical approach that guides the service, be it cognitive behavioural, humanistic, psychodynamic, or others; (2) whether the bot has been empirically tested; (3) what population/difficulties it targets and what psychological or clinical effects users may expect from using the platform.

In terms of the apps we tested, all of them contain information about theoretical approach on the website; Woebot and Wysa additionally inform users about the published research that backs it up. At the moment, the statement under Woebot's Frequently Asked Questions is not completely accurate, as it obscures the lack of difference between experimental and control group for anxiety ('In a recent study conducted at Stanford University, using Woebot *led to significant reductions in anxiety* and depression among people aged 18-28 years old, *compared to an information-only control group*', italics added).[13] A link to the article is provided, however, as well as further information in other sections of the site. Joy's website does not provide specific information about empirical support for their specific platform.

All of the bots inform users about what effects to expect on their websites or on the app store, albeit in relatively vague terms. Wysa says it will 'help you build mental resilience skills and feel better' and 'help you manage your emotions and thoughts'[15]; Joy claims to 'help you feel like a stronger, more confident, and more fulfilled version of yourself'[14]; and Woebot is described as a 'choose-your-own-adventure self-help book' who helps you 'learn about yourself'.[13]

We would also like to see more information about the target audience/user of the chatbot, so that young people can better assess if the platform will meet their needs. At the moment, Wysa's website informs users that it is used 'around the clock and trusted by 400,000 people' and from all age groups'[15]; Woebot's website says it was originally developed 'for young adults in college and graduate school. However, we encourage anyone to try it and see if Woebot is a fit for you'.[13] Joy does not include specific information about its target audience on the website. Overall, we consider it important that chatbot websites make it clear that the service is not designed to help individuals who are experiencing severe mental health difficulties. Moreover, for safety reasons that we discuss in the next section, chatbot websites should provide information about resources that may support users with more serious difficulties.

Even though these are relevant points, young people may not explicitly search for information to find out whether platforms are empirically validated; therefore, it is important that adults inform them – and that young people inform each other – about platforms that are evidence-based and therefore safer to use. Young people may also not take the time to read a list of Frequently Asked Questions on the bot's website. Therefore, we recommend that chatbots are programmed to provide automated responses to these questions, should the user raise them. For example, if the user types in 'Who can you help?', 'How can I be sure you will help me?', 'Have you helped others in the past?', 'Are you evidence-based?', 'Do you help for real?', etc, automated answers could be generated providing up-to-date clarifications to the questions above at any point.

Safety

1. Users should be informed that they are talking to a robot;
2. Automated chatbots should encourage people to seek human support;
3. Automated chatbots should have systems in place to prevent over-reliance;
4. Automated chatbots should have systems in place to deal with emergency situations.

Not only should chatbots provide users with evidence to demonstrate whether and how they are effective; we believe that chatbot developers should also aim to reduce to a minimum the risk that online support for mental health may pose to users' safety and well-being. Clearly, at the moment, automated bots still have limited systems in place to respond to situations in which users' safety may be at risk. As we cautioned at the start of this article, therapy bots are not particularly responsive to spontaneous texts generated by users. This is one of the reasons why we think that all fully automated conversational agents for mental health support should inform users that they are talking to a robot with limited capacity to understand what a user types. Among the 3 chatbots that we have tried, both Woebot and Wysa explained early in the conversation that they are a 'robot' and an 'artificially intelligent "pocket penguin"', respectively. Woebot adds that it is 'not for everyone and some people may be better served by seeing a human therapist'. Joy did not spontaneously reveal to be an automated chatbot, but when we typed in messages it could not understand, it told us 'I am just a robot, remember?'.

We think that therapy bots should also encourage users to seek human support from close others or from mental health professionals, either face-to-face or online. This is very important, especially given that, as explained in the previous sections, empirical research on mental health chatbots is still in its infancy, and it is not clear whether the outcome would be equivalent to that obtained through any face-to-face support. It is worth noting that Wysa offers the option of getting support in a more traditional sense in conjunction with or as an alternative to the automated service. Its users can receive support from real-life mental health professional ('Wysa Coach') for a fee. This can prove particularly helpful in contexts where there is a lack of trained mental health professionals. However, this service is unavailable for users under the age of 18 years. Users are also required to pay a fee for the service, which may be limiting for those living in low-resource settings.

In addition, although we are aware that the positive and negative effects that social media might have on people's mental health still needs further exploration,[50] online platforms do sometimes contribute to the development of mental health difficulties in the first place.[51] Furthermore, we also find it possible that online platforms could cause further isolation of people who are struggling and so they might represent a step backwards in their mental health journey. These concerns, which are still speculative in nature, should be tested empirically, and specifically in young people with diverse mental health challenges. This would meet two needs: it would allow chatbot information about the target user group to become more specific and it could lead to better technical specification of the algorithm to generate tailored support to users.

We also worry that users could become over-reliant on chatbots, because they are available with the tap of an icon, 24/7, which might worsen addictive behaviours that have been observed particularly among young people in the informational age.[52,53] All the platforms we have tested are private startups and currently available free of charge, except from additional coach services, but this might change as the platforms become more sophisticated and companies move towards a more sustainable business model. They would arguably benefit from having large numbers of users or from designing a product that encourages constant use.

Similar to what has been recommended for wearable and mobile health technologies,[54] we find it essential that these platforms are specifically programmed to discourage over-reliance. For example, chatbots may provide users with tasks to complete 'in the real-world', to encourage human interaction. When trying the different bots, we were pleased to note that some systems that attempt to limit addiction seem to be already in place. For example, some of us found it helpful when, after checking in and chatting for a short time, the bots tried to bring conversations to an end – by suggesting us to click on the automatically generated message 'bye'. This is not surprising for Woebot because it was developed to integrate the evidence-based recommendations for mental health app development,[10] which include 'encouraging non-technology-based activities'.[55] We also appreciated that by texting 'stop' or 'settings' within a conversation, users could update the frequency of notifications they receive from Joy.

Finally, we consider it essential that users are encouraged to seek human support in the case of an emergency, given that chatbots are not (yet) powerful enough to deal with mental health crises. All bots we tested seemed to have a function to recognise emergencies. For example, if users type 'SOS' or 'suicide', they are sent a series of resources, however, this varied from a single US Suicide Prevention number to a more comprehensive list of helplines, from which they can get further help. We consider it essential that the resources provided are effective and tailored to the users' location. In these situations, chatbots should also send users a message that reminds them of the importance of talking to real human mental health professionals or a trusted adult or peer if they are feeling particularly unwell.

For chatbots that are not fully anonymous, as a potential additional cautionary procedure, developers should test the acceptability to the user of retaining trusted adult/peer contact information. If such information were provided, two further actions should be undertaken. In an emergency situation (what the bot deems as such), the user could be asked to choose if the bot should contact the trusted person. Another (and potentially additional) option would be that the bot itself activates contact without asking for user consent, in response to a specific pattern of user use. However, it is important to note that young people are particularly concerned with protecting their privacy from their parents' and other relatives' intrusion, and there is evidence that they use online communication as a way to do so.[23,24] Therefore, in the latter and more controversial option, the balance of trust and acceptability versus avoidance of harm must be empirically tested.

Conclusions

We started our article by arguing that bots like Woebot, Wysa, and Joy might have great potential to provide help to people struggling with mental health problems. Chatbots are widely available and accessible to anyone with a smartphone and Internet access, and they may be perceived to be less stigmatising than formal mental health support. For these reasons, they might represent the first step towards getting help.

However, chatbots' limited capacity to re-create human interactions and to offer tailored treatment, combined with the currently lack of access to mental health services in real time, raise the question of whether chatbots could do harm to users. These harms would go largely unseen unless specifically tracked. We were concerned to hear that many digital platforms and apps for mental health did not involve ongoing evaluation for harms and benefits. Such continuous assessment is essential to allow a timely response to unpredicted issues or concerns, and ethically responsible practice more generally.

We have argued that automated bots for mental health support should meet at least three minimum standards: they should respect users' privacy, be evidence-based, and ensure users' safety. Chatbots should also be as transparent as possible about what they are currently able to offer. Our recommendations are general and geared towards chatbot developers worldwide, rather than specific to particular regions. We do encourage developers to use our recommendations in combination with industry, legislative, industry, and professional standards surrounding mobile health applications that apply to the regions where the app is to be used. Overall, our recommendations around privacy are consistent with key policies produced by governments and non-government organisations that provide oversight of health apps.[56] For example, according to the newly released EU General Data Protection Regulation,[57] apps should save as little personal data as possible and strictly inform users about any data sharing with third parties.

Even though there is much to be optimistic about in this new form of mental health support, it is also important to explore the reasons underpinning the public's interest in Internet and mobile-based mental health support. Does the optimism around therapy bots reflect a concerning picture of the state of mental health services across the world? Mental health problems are among the leading causes of the global burden of disease[18,58]; yet in 2013, the US National Institutes of Health invested only US$2.2 billion in mental health research, compared with about US$5.3 billion into cancer research.[59] Moreover, although most mental health issues (75%) start by the mid-20s[60] and addressing them at a young age has been shown to reduce the personal and economic impact of mental difficulties, in the United Kingdom, only 1 in 4 children, and young people who need mental health support, does receive it.[18,19] As already mentioned, among those who do get support, much of the precious time that could be invested in recovery is spent waiting in the queue to receive support services.[26,27]

The increasing number of chatbots may indeed signal that there is a demand for mental health support that is not being met by traditional services. This means that people might rely on digital resources more and more as a substitute for mental health professional support, which highlights the importance of working hard to make these platforms effective and ethically responsible. Most importantly, however, alongside efforts to improve digital mental health resources, it is extremely important that we continue advocating for funding for research and professional services, and work to combat the stigma associated with mental health difficulties. This is critical if interventions are to be offered in a timely fashion, providing effective help to those who may be most vulnerable and most in need.

Acknowledgements

The authors thank Christopher Chapman, Dan O'Connor, and Jonathan Beamish for the invitation to join this project. The ideas presented here were based on group discussions held as part of a project on the pros and cons of mental health chatbots, led by BBC Tomorrow's World in partnership with the Wellcome Trust (available at http://www.bbc.co.uk/guides/zt8h2nb). The following NeurOx Young People's Advisory

Group members contributed to this project: Aysha Sharudin, Boris Pavlov, Charlie Davis, Daniel Mooney, Eleyna Kibble, George Tuckwell, Grace Lewis, Jasmine Heelas, James Dixon, Jessica Bransby-Meehan, Jessica Katz, Laura Seeney, Angela Lee, Martino Allegri, Maud Beard, Nav Aithani, Nellie Lumbis, Niahm Walker, Poppy Macfarlane, Samantha Bonnett, Sophie Martin, and Sophie Speakman.

Author Contributions

All authors contributed to the conceptual analysis. KK, HT, GP, and AM wrote the paper, and IS critically revised it.

ORCID iDs

Gabriela Pavarini https://orcid.org/0000-0001-5574-4021
Ilina Singh https://orcid.org/0000-0003-4497-3587

REFERENCES

1. Kessler D, Lewis G, Kaur S, et al. Therapist-delivered Internet psychotherapy for depression in primary care: a randomised controlled trial. *Lancet*. 2009;374:628–634. doi:10.1016/S0140-6736(09)61257-5.
2. Arnberg FK, Linton SJ, Hultcrantz M, Heintz E, Jonsson U. Internet-delivered psychological treatments for mood and anxiety disorders : a systematic review of their efficacy, safety, and cost-effectiveness. *PLoS ONE*. 2014;9:e98118. doi:10.1371/journal.pone.0098118.
3. Baumel A, Schueller SM. Adjusting an available online peer support platform in a program to supplement the treatment of perinatal depression and anxiety. *JMIR Ment Health*. 2016;3:e11. doi:10.2196/mental.5335.
4. Leykin Y, Muñoz RF, Contreras O, Latham MD. Results from a trial of an unsupported internet intervention for depressive symptoms. *Internet Interv*. 2014;1:175–181. doi:10.1016/j.invent.2014.09.002.Results.
5. Adelman CB, Panza KE, Bartley CA, Bontempo A, Bloch MH. A meta-analysis of computerized cognitive-behavioral therapy for the treatment of DSM-5 anxiety disorders. *J Clin Psychiatry*.2014;75:e695–e704. doi:10.4088/JCP.13r08894.
6. Andersson G, Cuijpers P. Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. *Cogn Behav Ther*. 2009;38:196–205. doi:10.1080/16506070903318960.
7. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. *J Med Internet Res*. 2017;19:e176. doi:10.2196/jmir.7332.
8. Christensen H, Griffiths KM, Farrer L. Adherence in Internet interventions for anxiety and depression. *J Med Internet Res*. 2009;11:e13. doi:10.2196/jmir.1194.
9. Sachan D. Self-help robots drive blues away. *Lancet Psychiatry*. 2018;5:547. doi:10.1016/S2215-0366(18)30230-X.
10. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment Heal*. 2017;4:e19. doi:10.2196/mental.7785.
11. Ly KH, Ly A-M, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. *Internet Interv*. 2017;10:39–46. doi:10.1016/j.invent.2017.10.002.
12. BeGOOD website. https://begoodeie.com/ypag/. Accessed February 4, 2019.
13. Woebot website. https://woebot.io/. Accessed February 4, 2019.
14. Joy website. https://www.hellojoy.ai/support. Accessed December 4, 2019.
15. Wysa website. www.wysa.io/. Accessed February 4, 2019.
16. Would you trust a chatbot therapist? http://www.bbc.co.uk/guides/zt8h2nb. Accessed February 1, 2018.
17. Patel V, Flisher AJ, Hetrick S, McGorry P. Mental health of young people: a global public-health challenge. *Lancet*. 2007;369:1302–1313. doi:10.1016/S0140-.6736(07)60368-7
18. Department of Health. *A Framework for Mental Health Research*. London: Department of Health; 2017.
19. Khan L, Parsonage M, Stubbs J. Investing in children's mental health: a review of evidence on the costs and benefits of increased service provision. https://www.crisiscareconcordat.org.uk/wp-content/uploads/2015/02/investing_in_children_mental_health.pdf. Up-dated 2015.
20. Burns JM, Davenport TA, Durkin LA, Luscombe GM, Hickie IB. The Internet as a setting for mental health service utilisation by young people. *Med J Aust*. 2010;192:S22–S26. doi:bur11029_fm [pii].
21. Percentage of U.S. internet users who use Facebook Messenger as of January 2018, by age group. https://www.statista.com/statistics/814100/share-of-us-internet-users-who-use-facebook-messenger-by-age/. Accessed February 1, 2018.
22. Wartella E, Beaudoin-Ryan L, Blackwell CK, Cingel DP, Hurwitz LB, Lauricella AR. What kind of adults will our children become? the impact of growing up in a media-saturated world. *J Child Media*. 2016;10:13–20. doi:10.1080/17482798.2015.1124796.
23. Livingstone S. Children's privacy online experimenting with boundaries within and beyond the family. In: Kraut RE, Brynin M, Kiesler S, eds. *Computers, Phones, and the Internet: Domesticating Information Technology*. New York, NY: Oxford University Press; 2006:128–145.
24. Livingstone S, Bober M. *UK Children Go Online: Surveying the Experiences of Young People and Their Parents*. London, England: London School of Economics and Political Science; 2004.
25. Demyttenaere K, Bruffaerts R, Posada-Villa J, et al. Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *JAMA*. 2004;291:2581–2590. doi:10.1001/jama.291.21.2581.
26. Smith DH, Hadorn DC; Steering Committee of the Western Canada Waiting List Project. Lining up for children's mental health services: a tool for prioritizing waiting lists. *J Am Acad Child Adolesc Psychiatry*. 2002;41:367–376. doi:10.1097/00004583-200204000-00007.
27. Anderson JK, Howarth E, Vainre M, Jones PB, Humphrey A. A scoping literature review of service-level barriers for access and engagement with mental health services for children and young people. *Child Youth Serv Rev*. 2017;77:164–176. doi:10.1016/j.childyouth.2017.04.017.
28. McCrone P, Knapp M, Proudfoot J, et al. Cost-effectiveness of computerised cognitive-behavioural therapy for anxiety and depression in primary care: randomised controlled trial. *Br J Psychiatry*. 2004;185:55–62. doi:10.1192/bjp.185.1.55.
29. Gulliver A, Griffiths KM, Christensen H, et al. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. *BMC Psychiatry*. 2010;10:113. doi:10.1186/1471-244X-10-113.
30. Rickwood DJ, Braithwaite VA. Social-psychological factors affecting help-seeking for emotional problems. *Soc Sci Med*. 1994;39:563–572. doi:10.1016/0277-9536(94)90099-X.
31. Rickwood D, Deane FP, Wilson CJ, Ciarrochi J. Young people's help-seeking for mental health problems. *Aust e-Journal Adv Ment Heal*. 2005;4:218–251. doi:10.5172/jamh.4.3.218.
32. Pollock K, Armstrong S, Coveney C, Moore J. *An Evaluation of Samaritans Telephone and Email Emotional Support Service*. Nottingham, UK: University of Nottingham; 2010.
33. Berger M, Wagner TH, Baker LC. Internet use and stigmatized illness. *Soc Sci Med*. 2005;61:1821–1827. doi:10.1016/j.socscimed.2005.03.025.
34. Dubow EF, Lovko KR Jr, Kausch DF. Demographic differences in adolescents' health concerns and perceptions of helping agents. *J Clin Child Psychol*. 1990;19:44–54. doi:10.1207/s15374424jccp1901_6.
35. West JS, Kayser L, Overton P, Saltmarsh R. Student perceptions that inhibit the initiation of counseling. *Sch Couns*. 1991;39:77–83.
36. Lucas GM, Gratch J, King A, Morency L-P. It's only a computer: virtual humans increase willingness to disclose. *Comput Human Behav*. 2014;37:94–100. doi:10.1016/j.chb.2014.04.043.
37. Christensen H, Reynolds J, Griffiths KM. Original Article: the use of e-health applications for anxiety and depression in young people : challenges and solutions. *Early Interv Psychiatry*. 2011;5:58–62. doi:10.1111/j.1751-7893.2010.00242.x.
38. Ebert DD, Zarski A, Christensen H. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth : a meta-analysis of randomized controlled outcome trials. *PLoS ONE*. 2015;72:1–15. doi:10.1371/journal.pone.0119895.
39. Wallach E. An interview with Jo Aggarwal, Co-inventor of Wysa. *The Politic*. http://thepolitic.org/an-interview-with-jo-aggarwal-co-inventor-of-wysa/. Accessed March 28, 2018.
40. Shim website. https://www.helloshim.com. Accessed July 2, 2018.
41. X2AI website. http://x2ai.com/. Accessed July 2, 2018.
42. Ludden GD, van Rompay TJ, Kelders SM, van Gemert-Pijnen JE. How to increase reach and adherence of web-based interventions: a design research viewpoint. *J Med Internet Res*. 2015;17:e172. doi:10.2196/jmir.4201.
43. Inkster B, Sarda S, Subramanian V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth*. 2018;6:e12106. doi:10.2196/12106.
44. Debatin B, Lovejoy JP, Horn AK, Hughes BN. Facebook and online privacy: attitudes, behaviors, and unintended consequences. *J Comput Commun*. 2009;15:83–108. doi:10.1111/j.1083-6101.2009.01494.x.
45. Sunyaev A, Dehling T, Taylor PL, Mandl KD. Availability and quality of mobile health app privacy policies. *J Am Med Informatics Assoc*.2015;22:e28–e33. doi:10.1136/amiajnl-2013-002605.

46. Laugharne R, Priebe S. Trust, choice and power in mental health: a literature review. *Soc Psychiatry Psychiatr Epidemiol*. 2006;41:843–852. doi:10.1007/s00127-006-0123-6.
47. Wilson CJ, Deane FP. Adolescent opinions about reducing help-seeking barriers and increasing appropriate help engagement. *J Educ Psychol Consult*. 2001;12:345–364. doi:10.1207/S1532768XJEPC1204_03.
48. Rickwood DJ, Deane FP, Wilson CJ. When and how do young people seek professional help for mental health problems? *Med J Aust*. 2007;187:S35–S39. doi:ric10279_fm [pii].
49. Facebook data policy. https://www.facebook.com/full_data_use_policy. Accessed February 1, 2018.
50. Baker DA, Algorta GP. The relationship between online social networking and depression: a systematic review of quantitative studies. *Cyberpsychol Behav Soc Netw*. 2016;19:638–648. doi:10.1089/cyber.2016.0206.
51. Hamm MP, Newton AS, Chisholm A, et al. Prevalence and effect of cyberbullying on children and young people: a scoping review of social media studies. *JAMA Pediatr*. 2015;169:770–777. doi:10.1001/jamapediatrics.2015.0944.
52. Demirci K, Akgönül M, Akpinar A. Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students. *J Behav Addict*. 2015;4:85–92. doi:10.1556/2006.4.2015.010.
53. De-Sola Gutiérrez J, de Fonseca FR, Rubio G. Cell-phone addiction: a review. *Front Psychiatry*. 2016;7:175. doi:10.3389/fpsyt.2016.00175.
54. Kreitmair KV, Cho MK, Magnus DC. Consent and engagement, security, and authentic living using wearable and mobile health technology. *Nat Biotechnol*. 2017;35:617–620. doi:10.1038/nbt.3887.
55. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps : review and evidence-based recommendations for future developments. *JMIR Ment Heal*. 2016;3: e7. doi:10.2196/mental.4984.
56. Parker L, Karliychuk T, Gillies D, Mintzes B, Raven M, Grundy Q. A health app developer's guide to law and policy: a multi-sector policy analysis. *BMC Med Inform Decis Mak*. 2017;17:141. doi:10.1186/s12911-017-0535-0.
57. 2018 reform of EU data protection regulation. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en. Accessed July 1, 2018.
58. Vos T, Barber RM, Bell B, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2015;386:743–800.
59. Ledford H. Medical research: if depression were cancer. *Nature*. 2014;515:182–184. doi:10.1038/515182a.
60. Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Üstün TB. Age of onset of mental disorders: a review of recent literature. *Curr Opin Psychiatry*. 2007;20:359–364. doi:10.1097/YCO.0b013e32816ebc8c.