Responsible AI—Two Frameworks for Ethical Design Practice

Dorian Peters, Karina Vold, Diana Robinson, and Rafael A. Calvo, *Senior Member, IEEE*

***Abstract*—In 2019, the IEEE launched the P7000 standards projects intended to address ethical issues in the design of autonomous and intelligent systems. This move came amidst a growing public concern over the unintended consequences of artificial intelligence (AI), compounded by the lack of an anticipatory process for attending to ethical impact within professional practice. However, the difficulty in moving from principles to practice presents a significant challenge to the implementation of ethical guidelines. Herein, we describe two complementary frameworks for integrating ethical analysis into engineering practice to help address this challenge. We then provide the outcomes of an ethical analysis informed by these frameworks, conducted within the specific context of Internet-delivered therapy in digital mental health. We hope both the frameworks and analysis can provide tools and insights, not only for the context of digital healthcare but also for data-enabled and intelligent technology development more broadly.**

***Index Terms*—Digital health, ethics, value-sensitive design (VSD).**

I. Introduction

In the last year, the Association for Computing Machinery (ACM) released new ethical standards for professional conduct [1] and the IEEE released guidelines for the ethical design of autonomous and intelligent systems [2] demonstrating a shift among professional technology organizations toward prioritizing ethical impact. In parallel, thousands of technology professionals and social scientists have formed multidisciplinary committees to devise ethical principles for the design, development, and use of artificial intelligence (AI) technologies [3]. Moreover, many governments and international organizations have released sets of ethical principles, including the OECD Principles in 2019 [4], the Montreal Declaration in 2017 [5], the U.K. House of Lords report "AI in the U.K.: ready willing and able?" in 2018 [6], the European Commission High-Level Expert Group (HLEG) on AI in 2018 [7], and the Beijing AI Principles in 2019 [8]. Indeed, recent reports indicate that there are currently more than 70 publicly available sets of ethical principles or frameworks for AI, most of which have been released within the last five years [3], [9], [10].

The recent focus on ethical AI has arisen from increasing concern over its unintended negative impacts, coupled with a traditional exclusion of ethical analysis from engineering practice. While engineers have always met basic ethical standards concerning safety, security, and functionality, issues to do with justice, bias, addiction, and indirect societal harms were traditionally considered out of scope. However, expectations are changing. While engineers are not, and we believe should not, be expected to do the work of philosophers, psychologists, and sociologists, they do need to work with experts in these disciplines to anticipate and mitigate ethical risks as a standard of practice. It is no longer acceptable for technology to be released into the world blindly, leaving others to deal with the consequences. Engineering educators have already responded to this change in sentiment by evolving curricula to help ensure the next generation of technology makers is better equipped to engineer more responsibly [11], [12].

Yet, moving effectively from ethical theory and principles into context specific, actionable practice is proving a significant barrier for the widespread uptake of systematic ethical impact analysis in software engineering [13], [14]. In this article, we hope to contribute to resolving some of this translational difficulty by presenting two frameworks (the *Responsible Design Process* and the *Spheres of Technology Experience*) together with the outcomes of an example ethical analysis in the context of digital mental health. We hope that both the frameworks and the case study will serve as resources for those looking for guidance in translating ethical principles into technology practice.

II. Ethics Imperative in Health and Intelligent Systems

Verbeek explains that "When technologies co-shape human actions, they give material answers to the ethical question of how to act" [15]. He also highlights how technologies

Manuscript received November 11, 2019; revised February 6, 2020; accepted February 6, 2020. Date of publication February 17, 2020; date of current version March 11, 2020. This article was recommended for publication by Associate Editor M. Michael. The work of Karina Vold was supported in part by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust under Grant RC-2015-067, and in part by the Digital Charter Fellowship Programme at the Alan Turing Institute, U.K. *(Corresponding author: Dorian Peters.)*

Dorian Peters is with the Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge CB2 1SB, U.K., and also with the Dyson School of Design Engineering, Imperial College London, London SW7 1AL, U.K. (e-mail: dp605@cam.ac.uk).

Karina Vold is with the Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge CB2 1SB, U.K., and also with Alan Turing Institute, London NW1 2DB, U.K. (e-mail: kvv22@cam.ac.uk).

Diana Robinson is with the Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge CB2 1SB, U.K., and also with the Department of Computer Science, University of Cambridge, Cambridge CB3 0FD, U.K. (e-mail: dmpr3@cam.ac.uk).

Rafael A. Calvo is with the Dyson School of Design Engineering, Imperial College London, London SW7 2DBR, U.K. (e-mail: r.calvo@imperial.ac.uk).

Digital Object Identifier 10.1109/TTS.2020.2974991

inscribe the values of the designers, engineers, and businesses who make them [16], [17]. As a result, responsibility must go beyond the narrow definition of safety which, until recently, has largely constituted professional norms for technologists.

Furthermore, ethical implications should be considered early on and throughout the design, development, and implementation phases since value-laden tradeoffs are often made even during the earliest stages of design. Achieving ethically desirable outcomes will neither be easy nor straightforward. For instance, ethical design cannot be implemented as a single, one-off, review process since technologies (especially, intelligent ones) are continuously changing, as are the ways users appropriate them, and the socio-technical contexts within which they exist. Therefore, ethical impact evaluation must be an ongoing, iterative process—one that involves various stakeholders at every step, and can be re-evaluated over time, and as new issues emerge.

While society-wide ethical considerations are a relatively new focus within technology engineering [18] ethical enquiry has a long history within healthcare, perhaps because health practitioners work directly with the people they serve and often within sensitive and high-risk contexts. *Principles of Biomedical Ethics* [19] have been taught for 40 years. Hence, those working on the engineering of health technologies will need to adhere to both technology-related and biomedical ethical principles.

Fortunately, at the level of basic principles, the AI communities and biomedical ethicists might already be largely in agreement. A recent analysis suggests that the plethora of ethical AI frameworks can be consolidated into just five meta-principles, four of which also constitute the principles for biomedical ethics. These are: Respect for Autonomy, Beneficence, Nonmaleficence, and Justice, with the addition of "explicability" for AI [3].

Other recent systematic reviews of AI ethics principles have produced somewhat different taxonomies (see [14], [20], [21]). For example, Jobin *et al.* [9] reviewed 84 ethical guidelines and proposed 11 principles: 1) transparency; 2) justice and fairness; 3) nonmaleficence; 4) responsibility; 5) privacy; 6) beneficence; 7) freedom and autonomy; 8) trust; 9) dignity; 10) sustainability; and 11) solidarity. They found a wide divergence in how these principles were interpreted and in how the recommendations suggested they be applied. They do, however, note some evidence of convergence (by number of mentions) around five principles: 1) transparency; 2) justice and fairness; 3) nonmaleficence; 4) responsibility; and 5) privacy.

Between this set of principles and the set we will be using from the meta-analysis by Floridi *et al.* [3], there is a significant overlap. We have found the latter set to be particularly practical in the health domain, owing to its overlapping with biomedical ethics, however, we understand that other principles are also important. However, broad principles fall short of dictating specific actions in practice. Indeed, it has been acknowledged that these ethical frameworks do not provide enough contextual guidance for engineers to make use of them (e.g., [10], [14], and [22]). For example, the principle of fairness could lead to affirmative action, providing extra support for a group, or not, depending on the context.

In order for abstract principles to translate into actionable practice, the engineering discipline will need a variety of solutions. For example, the emerging development of the IEEE's PS7000 specifications stands to contribute on this point. Additionally, work by Gebru *et al.* [13] on "datasheets for datasets" in which they advocate for clear documentation for datasets that records "motivation, composition, collection process, recommended uses, and so on" is one example of a suggestion that would operationalize more abstract principles, such as transparency and accountability.

However, there will always be ethical decisions and tradeoffs that are not amenable to universally applicable specifications, and that need to be made with sensitivity to specific context and stakeholders. In these cases, we will need methods for conducting this kind of decision-making rigorously. Responsible innovation requires these methods to be anticipatory, reflexive, inclusive, and responsive [23].

The very fact that there has been some convergence around a set of principles (rather than a single principle) seems to indicate a kind of value pluralism—the view that there are multiple values that are equally fundamental, and yet may sometimes conflict with each other. How to navigate tradeoffs between equally important values when conflicts arise is where constructive reflection and discourse may be most needed, and where it will be important to acknowledge that cultural and contextual differences may affect what the right outcome is within practice (further discussions in [24] and [25]). Here, methods for value-sensitive design (VSD) [16] can help development teams and stakeholders articulate and align with explicit values.

Moreover, data-enabled technologies employ a wide-range of techniques that are used differently in different contexts, and these diverse contexts raise unique concerns that will require different ethical tradeoffs. It is unlikely that the same priorities and solutions applicable to one domain, context, or project, will translate across to others [22], [25], [26]. This means that technology development teams will need to conduct bespoke ethical evaluations for each project, in the same way, user research and specifications analyses are unique to each project. This need for ethical impact assessment for technology is akin to the need for environmental impact assessment in other types of engineering [27].

It is impossible to provide ethical principles that will be specific enough to provide answers in practice, and yet broad enough to apply universally. But it is possible to provide a *process*. While every team and organization may devise their own answers to ethical dilemmas, they should have systematic processes by which to do so consciously and rigorously, leaving a record of these processes along with the values and rationale they employed to make decisions. This record can provide the public with transparency regarding the rationale for a decision after the fact, as well as give the design team confidence that such a decision was made in a systematic and professional way. Such process will not guarantee a product has no negative consequences, but it will help mitigate the

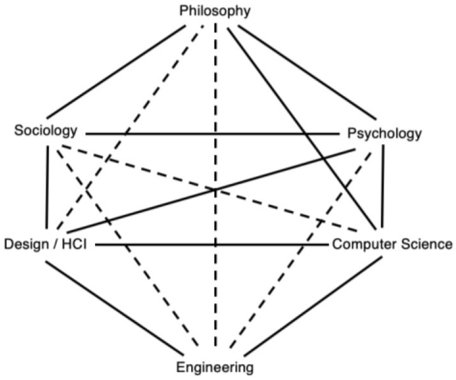

Fig. 1. Connections among six fields centrally involved in the ethical design and development of AI and data-enabled systems (based on Sloan foundation's hexagonal mapping of "connections among the cognitive sciences" 1978, reproduced in Gardner 1985, p. 37) key: unbroken lines = strong interdisciplinary ties, broken lines = weak interdisciplinary ties.

risks, and provide professionals with the reassurance of having acted responsibly.

In the next section, we present two frameworks that can help provide structure for such a process for ethical impact analysis.

III. Moving From Principles to Practice: Frameworks for Responsible Technology Development

A. *Multidisciplinarity in AI Ethics*

We begin our discussion of frameworks with a diagrammatic representation of disciplines central to the development of ethical AI and how they interconnect (see Fig. 1). The primary intention is to emphasize the importance of grounding all ethical impact assessments in multidisciplinary expertise. It is likely that new requirements in an increasingly AI-enhanced world will lead to the development of new specializations which blur traditional disciplinary boundaries. Nevertheless, there is no single discipline capable of handling the task of ethical analysis single-handedly. Given the complexity of the problems, the best outcomes are likely to come from the richest diversity.

For the digital health case study we present later, we leveraged expertise from four different disciplines, including design, engineering, human–computer interaction, and philosophy. An ideal project would include even more disciplines, such as psychology and sociology (see Fig. 1), as well as, end users, domain experts, and other stakeholders. In the mental health context, for example, users may include patients, therapists, and family, while domain experts would include therapists, mental health researchers, and others working within the healthcare system.

The importance of this multivocal approach cannot be overstated as there can be a tendency for agile teams to consist of just programmers, designers, and managers. Taking digital mental health as a cautionary example, a failure to involve psychologists, health practitioners, end users, and other domain experts has led to an exploding industry of mental health tools that lack evidence, inclusivity, and effectiveness (and at worst, cause harm) [28]–[30]. This has been possible because, while traditional channels for healthcare are highly regulated, technology regulation lags behind and these technologies are unusually quick to implement and disseminate. As Nebeker *et al.* [31] have cautioned: "it is critical that the minimal requirements used to make a digital health technology available to the public are not mistaken for a product that has passed rigorous testing or demonstrated real-world therapeutic value."

B. *Framework 1—The Responsible Design Process*

Sometimes technology designers, like policymakers, are forced to make values-based tradeoffs. For example, they might be able to increase privacy at the expense of security or increase accuracy at the expense of privacy. Moreover, some technologies may increase the wellbeing of some at the expense of others. Value-laden decisions arise as part of engineering and can either be addressed in a cursory way by one or a few individuals, or in a systematic and robust way by teams. Only the latter approach can hold up to scrutiny should negative consequences emerge after the fact. As such, we need a technology development process that makes room for this sort of robust decision making and for the ethical impact analysis on which it must stand.

Innovators are often asked to address the consequences of their technologies, but this *post-hoc* approach is increasingly seen as limited. A number of nonregulatory approaches have been developed to take into account the broader social impact of new technologies including anticipatory governance, technology assessment, and VSD [23].

They can all be included within "responsible innovation," a growing field of research exploring ways of "taking care of the future through collective stewardship of science and innovation in the present" [23]. As with many design practices, the goal is to embed deliberation within the design and innovation process.

An important aspect of responsible innovation is the concept of human wellbeing, which is also at the center of many current ethical frameworks. For example, the IEEE centers its ethics specifications on human wellbeing. So too do several government frameworks [4], [5].

As such, we argue that a responsible technology development process will need to incorporate evidence-based methods for evaluating the impact on, and designing for, human wellbeing, by drawing on psychology (see [32], [33]).

However, the promotion of human wellbeing is not a complete solution. After all, decisions must be made as to whose wellbeing is being considered. When technology makers are forced to make tradeoffs that increase the wellbeing of some but at the expense of others, at the cost of long-term ecological impacts, or in spite of other negative side effects, then issues to do with justice, equality, and other values arise. This is where ethical analysis, drawing on philosophy and other disciplines, must come in.

As such, our conception of a responsible development process involves taking existing design processes, particularly those that are anticipatory, reflexive, inclusive, and responsive,

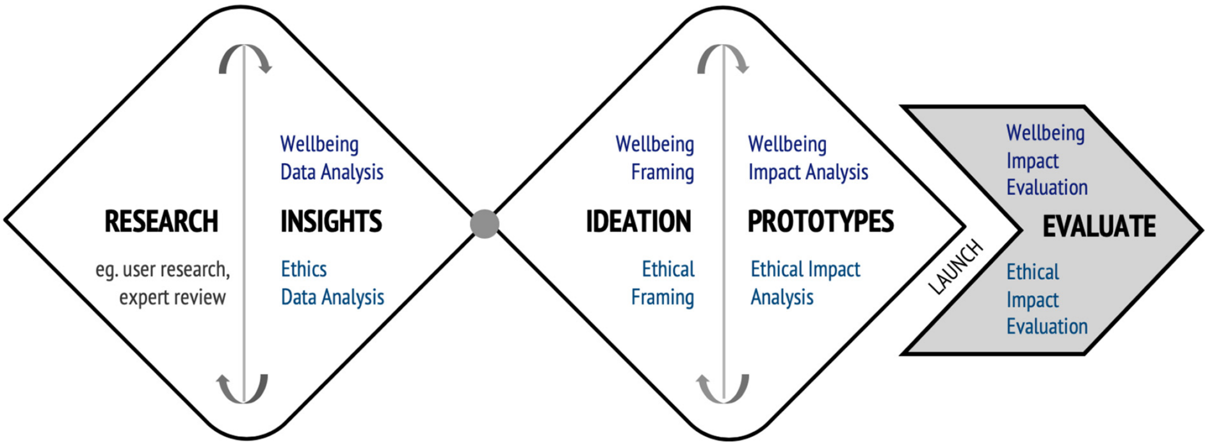

Fig. 2. Responsible design process framework. A process for technology development in which wellbeing support and ethical impact analysis are incorporated at each phase. A post-launch evaluation phase is also added.

and augmenting them with methods for ethical analysis and wellbeing-supportive design. The approach described here could include activities such as those used in VSD [16].

While development processes are as varied as technologists themselves, there are a series of developmental phases that find their way into most, if not all, approaches, and these include: research, ideation, prototyping, and testing. The U.K. design council consolidated these commonalities and created a popular "double diamond" diagram to illustrate them [34]. The broadest and narrowest points of the diamonds represent points of divergence and convergence. We began with this standard process and integrated stages for wellbeing support and ethical decision making to create the resulting responsible design process framework presented in Fig. 2.

Wellbeing in the diagram refers to human psychological wellbeing and it is included separately to other ethical issues because evidence-based design methods grounded in psychological research already exist for it and allow it to be attended to empirically. Analysis of other ethical dimensions, such as fairness, data governance, ecosystem wellbeing, or democratic participation cannot rely predominantly on psychological research and will require different methods. We describe each phase of the process in further detail as follows.

Research: The research phase involves investigating the needs, preferences, contexts, and lives of the people who will be served or, otherwise, impacted by a technology. This phase may include standard approaches to user research (e.g., design thinking methods, ethnographies, participatory workshops, etc.) as well as expert review and secondary research in relation to the specific domain. Standard approaches can surface wellbeing and ethical issues, however, tailoring these methods to focus participants on ethical or psychological reflection may be helpful.

Insights: This phase involves the analysis of the data from the research phase, and synthesis into specific insights for design. Data analysis can be done through the lens of wellbeing theory with a view to anticipating harms and opportunities for supporting healthy psychological experience. Ethics data analysis can be done through the lens of an ethical framework, and with a view to identifying potential biases, ethical risks, and tensions.

Ideation: The ideation phase involves the divergent generation of ideas for design solutions. Ethical reflection can be integrated into the ideation phase through framing. For example, introducing wellbeing psychology concepts into the ideation phase can help the team focus on root psychological causes of user needs while introducing ethics concepts into ideation can sensitize the team to ethical tensions that may arise so that brainstorming can involve resolutions to these.

Prototypes: In this phase, the team converges on and builds various design solutions. Responsible impact analysis involves collaborative speculation on the wellbeing and ethical impacts (good and bad) to which a particular design concept may lead. This will ideally involve a wide range of stakeholders including end users.

Evaluation (in Use): The real-life ethical impacts that a technology will have on people, their communities and the planet, can only be fully understood once the product or service is in real-world use. Teams must speculate and test in advance, but unintended use patterns are realities in our complex socio-technical systems. Wellbeing impact evaluation involves evaluating the impact of technology use on a user's psychological experience during and after use. Ethical impact evaluation involves evaluating the ethical impacts of a technology's use, not just on its users, but often, also on those indirectly affected, such as their friends and families, communities, society as a whole, and the planet.

In the framework described above, we have taken a familiar process and incorporated phases for the integration of ethics and wellbeing in an attempt to provide a map for a more responsible development process. The map also provides a landscape within which to research, develop, and situate new methods and tools to support each of these phases. For example, research can be directed at identifying effective methods for "ethical data analysis" within the insights phase, "ethical framing" within ideation, and "ethical impact evaluation" during use. Moreover, ethics-based methods and tools

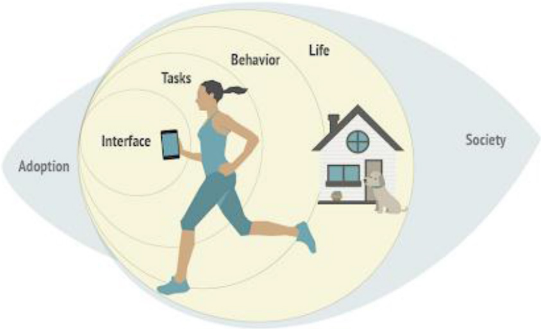

Fig. 3. Six spheres of technology experience (adapted from Peters, Calvo, and Ryan [26]).

that already exist can be more easily integrated within the standard development process in this way. It is worth noting that, while it is true an ideal project would start with responsible methods from the beginning, we are aware that few projects represent the ideal. Integrating wellbeing and ethics from any point is likely better than not at all and will contribute to more responsible outcomes.

C. Framework 2—The Spheres of Technology Experience

There has been a lack of appreciation for the different resolutions at which technologies can have an impact on the experience. A technology can make an impact through the design of its interface, based on the tasks it is designed to support, through the behaviors it promotes, or as a collective result of widespread societal use.

For instance, consider the way games impact wellbeing and autonomy (autonomy is both a constituent of wellbeing [35] and a central principle of AI ethics frameworks in its own right). A user may experience a strong sense of autonomy and wellbeing during game play, but because the game is designed to increase compulsive engagement, a resulting addiction may diminish the same user's experience of autonomy at a life level (as overuse crowds out time for taking care of work, family, and other things of greater importance to her). Therefore, does the game support or hinder autonomy? The answer in this context is probably "both," and therefore, a fair assessment of ethical impact relies on an evaluation that takes into account the impact at different resolutions.

Therefore, it is clear that greater precision is required in order to effectively identify impacts at different granularities within the technology experience. Calvo *et al.* [36] first highlighted this need with respect to autonomy and presented a framework distinguishing four "spheres of autonomy." Peters *et al.* [32] expanded on this substantially, developing, as part of a larger model, a framework of technology experience which identifies six distinct spheres within which wellbeing can be influenced. It is this framework that we believe can be usefully applied to provide a structure to ethical impact analysis conducted during a responsible development process. We provide an illustration of the six spheres in Fig. 3.

This "Spheres of Technology Experience" framework is described in detail by Peters *et al.* [32] wherein they also provided methods for wellbeing-supportive design. An application of the framework to human autonomy in AI is described in [37]. Below we provide just a brief description of each sphere to show how each can also help to structure ethical analysis.

Adoption: It refers to the experience of a technology prior to use, including the marketing and socio-cultural forces leading a person to use it. A development team may want to consider the ethical impacts of the forces leading to uptake, and to what extent users are choosing to use a product freely or being coerced or pressured to do so. As a simple example, a new upgrade can be forced upon existing users—an approach that is notoriously un-user-friendly—or it can be introduced in a way that better respects autonomy, as when developers provide an option to "try the upgrade" first.

Interface: The interface sphere is the first sphere of the "user experience" and involves interacting with the product itself, including the use of navigation, buttons, and controls. At this level, an ethical analysis might explore issues to do with autonomy and inclusion, for example, to what extent does the interface support autonomy by providing meaningful options and controls? and, are users of different abilities or cultures being excluded?

Task: Broadening the lens of analysis beyond the interface, the task sphere refers to discrete activities enabled, or enhanced, by the technology. For example, in the case of a fitness app, "tracking steps" or "adding a meal to the diary" constitute tasks. Ethical impacts arising from an analysis of these tasks might include the risk of inadvertently contributing to eating disorders or anxiety. Awareness of these risks can help designers structure tasks in ways that respect the diverse needs of users and provide safeguards against negative outcomes.

Behavior: Combinations of tasks contribute to an overall behavior. For example, the task "step-counting" might contribute to the overall behavior: "exercise." For technologies intending to positively impact on a particular behavior, it is important to consider the psychology literature on the effects of various approaches to supporting that behavior (and/or work with a psychologist).

Life: The final and broadest sphere within the user's experience is life, which captures the impacts of a technology at a life level. Not all technologies will have impacts significant enough to yield measurable effects on a person's quality of life overall. While a self-driving car may impact measures of autonomy and wellbeing at the life sphere for someone who is vision impaired, the extent to which a cooking timer is customizable probably will not. While many technologies have only narrow application, and there is little reason to expect them to impact the life sphere, others, such as those that target wellbeing directly (i.e., meditation and fitness apps) or those used daily (workplace technologies, entertainment products, and social media) do need to consider and anticipate life-level impacts.

Society: Expanding beyond the user experience into the broadest sphere, society involves the direct and collateral impact on nonusers, nonhuman life, and the environment. The self-driving car mentioned above may promote wellbeing for some users but decrease it for those whose livelihoods depend on driving. This can only be revealed at a societal level of analysis which entails the exploration of emergent and complex systems. This sphere presents the greatest challenges to impact analysis. Identifying and anticipating ethical impacts at this level will not only require multidisciplinary expertise but also ongoing evaluation after a technology is released into use. Nevertheless, some specific methods already exist to assist developers in anticipating societal impact. “Consequence scanning” is a method developed by the nonprofit organization, Doteveryone, dedicated to responsible innovation [38]. The method provides a step-by-step process for collaboratively identifying ethical risks and tensions associated with a new or planned product or service.

It is important to qualify that the boundaries between the six spheres of technology experience are merely conceptual and should not be seen as concrete. Instead, they are intended to provide a way of organizing thinking and evaluation that allows for the identification of contradictory parallel effects.

The ways in which the Spheres of Technology Experience framework allows us to identify and target wellbeing and ethical impact helps to ensure analysis is both more thoroughly and clearly articulated. The fact that empirical measures already exist that can be applied to these spheres also makes their application practical and actionable. Existing measures (described in [32]) can help developers to quantitatively compare different technologies and designs with regard to their different impacts on a range of ethical and wellbeing-related attributes and within different spheres. Some examples of how measures have been used to evaluate the ethical impact already exist. For example, Kerner and Goodyear [39] conducted a study investigating the psychological impact of wearable fitness trackers. Additionally, a series of studies comparing how different game designs impact wellbeing and autonomy have been conducted using psychological measures [40], [41].

While the above examples focus on autonomy and wellbeing, the spheres can be used to articulate impact in relation to any ethical values, and at any stage within the responsible development process.

IV. Case Study—Responsible Digital Mental Health Technologies

A. Project Background

During the process of identifying methods for responsible design practice, we were commissioned by a health technology company to explore the ethical tensions arising in the health domain and recommendations for how these could be addressed. The company wanted to follow more responsible practices, and sought to anticipate unintended consequences. We viewed this an opportunity to enrich our experience with ethical analysis and to apply the frameworks described above.

The product in question was a text-based online therapy program for depression and anxiety. As part of the research phase (see Fig. 2) of our responsible design framework, the expert review we provided was later combined with commercial user research data to help create insights and inform ideation.

We believe an expert-led analysis within the research phase is a valuable way to: 1) involve disciplinary experts; 2) draw on existing knowledge; 3) sensitize the development team to ethical issues early on; 4) inform the design of user studies; and 5) assist the interpretation of user data.

The analysis below is the outcome of a multidisciplinary literature review and analysis involving a team of researchers with significant professional experience in the co-design and development of digital health technologies. Specifically, and consistent with the claim that ethical impact analysis must be a multidisciplinary endeavor, the team consisted of combined expertise in engineering, design, human–computer interaction, psychology, and philosophy. It also employed the Spheres of Technology Experience framework described above to guide the identification of key ethical considerations within digital mental health.

The outcomes pertain to a narrow genre of technologies, rather than a specific product, but the same approach could be taken for a product-specific context.

The analysis is structured according to the five ethical principles described in Section II. As such, recommendations are grouped into these five categories: 1) Respect for autonomy (Section IV-C); 2) Beneficence (Section IV-D); 3) Nonmaleficence (Section IV-E); 4) Justice (Section IV-F); and 5) Explicability (Section IV-G). While there are many sets of principles, we chose these five for this analysis because they align with the principles of medical ethics.

The analysis is presented in the form of a series of recommendations intended for design and development teams of mental health technologies. Each is presented with: elaboration that connects it to real-world context; a brief overview of how it is addressed in practical ethics; how it may be considered in the context of the application; and specific practical strategies for development which draw on design and engineering literature.

B. Online Therapy Within Digital Mental Health

Depression is the leading cause of disability worldwide [42] making data-enabled mental health technologies a critical area of research and industry. These technologies have the potential to increase access to therapy, reduce disparities, reduce costs, and improve the effectiveness of mental health care. But rapid change, coupled with the rapid introduction of new technologies into such a sensitive area, has brought new ethical challenges involving transparency, patient involvement, and human autonomy.

A number of authors within human–computer interaction have articulated some of the ethically loaded socio-technical challenges facing digital mental health developers [28]–[31], [43], [44]. For example, Orlowski *et al.* [44] stated: “Design solutions not generated with end users themselves are more likely to fail... Moreover, from an ethical and moral perspective, egalitarian ways of working, such as those exemplified by participatory design, represent a promising opportunity to redress the legacy of consumer disempowerment in mental health.”

Another important criticism of traditional practice in digital mental health revolves around respect for autonomy, a core ethical principle. As Mohr *et al.* [45] explained: "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." The literature has also highlighted the ethical issue of transparency as critical to this area. For example, the Psyberguide, developed by a nonprofit network of mental health professionals, includes transparency as one of three criteria for quality ratings of mental health technologies (the other two being credibility and user experience) [46].

The specific analysis herein focuses on online text-based one-to-one professional therapy for depression and anxiety. This is an augmentative approach to online therapy which, rather than replacing humans, aims to use data to increase human capabilities and to make human activity and interaction more effective, efficient, and satisfying. The analysis is presented as a series of recommendations followed by philosophical justification and practical strategies for implementation.

C. Respect for Autonomy

Recommendation 1: Mental health technologies should be designed to protect and support user autonomy. 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 [19]. 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 [47]. Respect for autonomy is essential for the development of any digital health technology, but in the case of mental health technologies, it is particularly challenging. This is because certain mental illnesses 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.

Burr and Morley [47] discussed how the presence of a mental illness might also affect a patient's choice to engage with a mental health service and restrict their ability to make healthcare decisions. In extreme cases, respecting a patient's autonomy (i.e., nonintervention) may even threaten safety, if there is a risk of harm to self or others. What this suggests is that: 1) respect for patient autonomy is defeasible such that it may sometimes have to be traded off against other goods and 2) in some cases healthcare providers may need to go beyond respect for a patient's current ability to self-govern to help build and support the user's autonomy in the long term.

Online professional text-based therapies will involve (at least) two kinds of users: 1) patients and 2) therapists (counselors). It is important to also think of the therapist as a user of this technology, as their role will be changed and augmented by these new tools. This is also true for other kinds of data-enabled medical technologies which will affect not only the patient but also healthcare professionals.

Recommendation 2: To protect the privacy and autonomy of users, make transparent the use of mental health data and ensure secure storage. Online mental health therapy applications that collect, store, and make use of personal data raise several important concerns around privacy, which in turn can pose risks to user autonomy [48], [49]. In particular, because of the kind of personal data that is now available to be collected (e.g., biometrics, location, and online behavior) combined with advances in machine learning that make it possible to infer personal attributes from collected data (e.g., [50] and [51]) companies are increasingly able to tailor messages and services to specific individuals or groups. This means that the more personal information a company has about someone, the more effectively they can target interventions in an attempt to influence them which may present new risks to patient autonomy.

Even features that can serve as a means of patient empowerment, such as self-tracking, which can be used to boost self-reflection, can pose risks to autonomy. But the sharing or use of this data, be it with family, friends, or even healthcare professionals—especially in nonemergency situations—can negatively affect patient autonomy. Sanches *et al.* [43] described this as an example of "autonomy [of patients being] claimed by their social support network, collectivized by healthcare services, or both." This explains why designing for privacy as a target can also be considered a subset of autonomy support [52].

Furthermore, since therapy sessions involve two interlocutors, both sides have reasonable claims to privacy. It is 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 counselors, may not be aware of the value of their data. One way of using the data, for example, is for analyzing the counselors' conversations. This may be problematic, but also presents an opportunity to provide feedback if handled in a manner that does not feel intrusive.

1) Practical Strategies for Respecting Autonomy: The literature on self-determination theory, a robustly evidence-based psychological theory of wellbeing and motivation [35] provides guidance with respect to what characteristics constitute "autonomy-supportive" (versus controlling) environments and interactions. According to this article, autonomy-supportive interactions as follows.

1) Understand the other's perspective (frame of reference).
2) Seek the other's input and ideas.
3) Offer meaningful choices.
4) Empathize with resistance and obstacles.
5) Minimize the use of controlling language or rewards.
6) Provide a rationale for requested or required behavior.

These can be translated into design guidance for digital technologies. For example, applying empathy is the cornerstone of human-centered design so employing human-centered methods is likely to increase autonomy-supportive outcomes. Seeking the user's input and ideas is also achieved through human-centered and participatory processes. In the mental health context, this requires engagement with people with have suffered mental illness as only they have direct expertise around frames of reference, threats to autonomy and privacy within their contexts, and insights into the kinds of obstacles that are most salient for them. Moreover, insights from these processes can inform what meaningful choices can be added to the technology. With respect to the protection of privacy specifically (recommendation 2), meaningful choices are likely to involve giving the client control over when and with whom data is shared.

Security experts should be consulted in ensuring that the data collected and analyzed in the course of mental health therapy is safely and securely stored, and that it is only shared through properly encrypted channels. In many countries, this is required by the Health Insurance Portability and Accountability Act (HIPAA) and similar legislation.

D. Beneficence

Recommendation 3: Consider the wider impact of both opportunities and risks for all stakeholders involved in the development and use of mental health technologies. In biomedical ethics, the principle of beneficence is typically thought of as a commitment to "do good." In this context, this will require that the potential benefits of a design or development choice be balanced against potential risks, both for an individual user and for society more broadly [19].

First, it is important to identify all those who stand to benefit from a particular technology. In addition to patients, the use of online text-based therapies impacts 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.

Moreover, the collateral impact might involve other, less obvious, stakeholders, such as developers themselves. For example, with supervised learning, a human has to assign labels to data used to train predictive algorithms. In the case of mental health therapies, this means that an employee must read and tag sensitive conversations between doctors and patients. This could have a harmful psychological impact on developers since therapy sessions are likely to contain content that could be distressing or triggering, depending on one's own life experiences. Such a labeling task might require training, 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.* [43] expressed 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 rethinking how such support can be explicitly factored into institutional guidelines and budgets.

Recommendation 4: Research the access requirements and unique mental health situations of diverse populations in order to ensure mental health technologies are effective for all relevant groups. In many cases, the risks of a new technology are not evenly distributed. In the context of data-enabled digital mental health therapies, the relative dearth of research and understanding on the needs of people from diverse socioeconomic and ethnic groups may put members of those groups at greater risk.

If online therapies are developed using a data set that only includes relatively affluent university students, or that lacks other forms of representation, then the therapy will only be optimized for a homogeneous group. Hence, it is important that the training set for the algorithm genuinely represents the diversity of the target population that will use it.

In practice what this means is that in some cases sets used to evaluate algorithms might need to come from a different statistical distribution than the training set. This comes with its own challenges, for example, understanding the wide variation in groups affected by mental illness, reaching out to "hard to reach populations" (e.g., the homeless, refugees, those addicted to drugs, etc.) and how measures can be used to yield inclusive and broadly beneficial interventions.

Recommendation 5: Aim to support authentic human interactions, connectivity, and engagement. Another example of the kind of balancing that needs to be done to ensure beneficence, involves the opportunities and risks that digital health technologies pose for authentic relationships. The context of mental healthcare requires respect, dignity, and empathy. However, even highly sophisticated AI systems lack human empathy and are at best able to mimic these traits. Thus, even partial automation in mental healthcare, if not implemented cautiously, could threaten "relational authenticity" [53].

In Hertlein *et al.*'s [54] study of family and marriage counselors' 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 concerns that the use of AI could lead to feelings of alienation and devaluation.

A related concern is a reduction in the quality of communication that may result from the lack of nonverbal cues and body language. This is true for online therapy, but also more broadly for other forms of data-enabled digital health interventions. There is some evidence, for example, that the data entry required for electronic medical records (EMRs) disrupts the nonverbal relationship between health-care providers and patients (e.g., [55]). Other research has shown that nonverbal cues, including eye contact and social touch (e.g., handshakes), have been found to significantly influence patient perceptions of clinician empathy [56]. Hence, the loss of such nonverbal cues can make it more difficult for health care providers to demonstrate empathy and to build authentic relationships with clients.

In addition to concerns about alienation and reduced quality of communication, some evidence suggests that relational authenticity also encourages patient engagement and trust [57]. Hence, any reduction in relational autonomy might, in turn, diminish engagement and trust. In their recent report, Sanches *et al.* [43] expressed 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."

On the other hand, technological interventions in mental health may provide new opportunities for engagement that are not available in a strictly human-to-human context. For example, a 3-D avatar that functions like a virtual therapist but was not trying to perfectly emulate a human being [58]. 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" [59]. This suggests that patients might be able to have differently authentic interactions with technologically mediated systems, if they are well designed. 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.

1) Practical Strategies for Beneficence: Arguably, the technology experience of people living with mental health issues can only be well understood by engaging directly with them as part of a collaborative design and evaluation process. This experience will be shaped by socio-economic and cultural circumstances and will, therefore, differ among individuals, yet meaningful patterns will still exist. User involvement that adequately represents the diversity of potential users of a service is therefore critical to bringing about genuine benefit. This inclusive process will also help to prevent blindness to the reality of the wide spectrum of audience needs within mental health service provision. This includes differing requirements due to low income, disability, low literacy, limited access to computers, mobile phones, and Internet connections, as well as low technology literacy (even among young people) [60].

In addition, users will prefer different modes of technology use at different times. For example, an insomnia therapy that does not 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.

E. Nonmaleficence

Within medical ethics, nonmaleficence is an obligation not to cause harm. This also applies to the design and development of data-enabled digital mental health therapies. The difficulty with this principle is avoiding the "known unknowns"—that is, harms that one foresees, though with some uncertainty—as well as the "unknown unknowns"—that is, harms that one does not foresee. The latter requires evaluating (and re-evaluating) impact both during development and after release.

Recommendation 6: While augmentation can be beneficial, ensure that over-reliance on technology does not lead to atrophy of critical skills or diminish competence. One example of a foreseeable, though uncertain, the harm is atrophy. Skill atrophy is the decline in abilities that comes from underuse or neglect to perform the behaviors 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 (e.g., [61] and [62])—a concern that dates back at least as far as Plato's discussion of the diminishing effects that writing would have on memory [63]. 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 for more worthwhile use 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. These risks can necessitate the need to create fail-safes (procedures for cases in which technology malfunctions and people need to rely on past skills), or they might necessitate not introducing technology into realms where humans should remain critically vigilant or engaged, such as areas that require value judgments [64]. Some areas of mental health-care may be among these.

For patients, there may be a risk of losing good decision-making skills and the ability to check-in with themselves, to self-reflect, as well as 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 a device for self-reflection they may lose competence at self-management when they are decoupled from the device (e.g., due to a loss of network connection, a damaged device, or no battery power). Additionally, dependence on a technology to manage care may result in lower feelings of self-efficacy, empowerment, and control [64].

For therapists, the introduction of technology into the diagnostic and therapeutic process could result in atrophy of critical professional skills. In cognitive-behavioral therapy sessions, therapists interact closely with patients through structured discussion sessions to break down problems into separate parts (thoughts, behaviors, and actions) and then to suggest strategies that patients can use to change their thinking and behavior. 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. If a therapist becomes over-reliant on an app that aids in these skills, over time she may lose them and struggle to be as effective in face-to-face sessions with patients.

Technologists will need to work closely with therapists and patients to determine appropriate areas for automation and augmentation and then evaluate outcomes after release.

Recommendation 7: To avoid risks arising from stigma, design to protect the privacy of users and always ensure secure storage of mental health data. Another foreseeable though uncertain harm is privacy. Because mental health is a stigmatized topic, those that suffer from mental health conditions face the risk of bias and discrimination, from both themselves (self-stigma) and others. This means that if digital health records of mental health status are leaked, hacked, or accessed by unconsented third parties, a user's dignity and reputation could be threatened, and they could be put at risk of discrimination.

These concerns are true in traditional (face-to-face) therapy as well, but relying on digital online platforms, from EMRs, to online therapies, poses new risks to both informational and decisional privacy [65]. In Hertlein *et al.*'s [54] survey, 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 counselor ("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 counselor's good intentions, they also have to trust that counselors will protect their computer screen or other devices from onlookers, protect their passwords, use secure network connections, and not use shared computers [54], [66]. Patients furthermore have to trust the provider of the technology not to use the data for any unconsented purpose.

For this reason, it is important that the utmost care is taken by companies to protect and anonymize the use and storage of sensitive data. Doherty *et al.* [52] suggested additional design implications to protect against the risks of stigma.

1) Practical Strategies for Nonmaleficence: There are a number of practical strategies that help ensure the principle of "do no harm" is followed. First, in addition to user research and involvement, a safe user experience design depends upon iterative improvement based on the ongoing evaluation. Health technologies also 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 consist of 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 to justify a more formal clinical evaluation. Further evaluation after the release of the product can inform improvements and upgrades and is necessary for determining impact and appropriation within complex real-world contexts (which are often very different to the controlled environments of clinical trials). Our framework for a responsible design process calls for just this kind of staged approach to evaluation (see Doherty *et al.* [52] for further discussion of a staged approach to the evaluation of mental health technologies more specifically).

Moreover, as alluded to earlier, 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. Traditional approaches to "failing fast and often" are potentially disastrous in a health context in which people cannot 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 safety-conscious approaches are taken.

Experts in ethics should also contribute in order to effectively assess ethical considerations from multiple standpoints. It may be helpful for them to work directly with user experience specialists to allow broad stakeholder input into ethical concerns.

In addition to involving multidisciplinary teams and undertaking ongoing evaluation, technology approaches need to be grounded in research to prevent harm. Topham *et al.* [67] argued 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.* [52] similarly recommended 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-centered focus on lived experience suggests the importance of mixed methods including qualitative methods for uncovering insights into subjective experience, motivation, and the causes of engagement and disengagement. These can complement and explain results from quantitative approaches, such as symptom scores, behavioral analytics, or surveys.

Finally, a simple safeguard for avoiding nonmaleficence is to apply existing quality frameworks. A number of quality frameworks and guidelines 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 [68] includes 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 mental health professionals, bases its ratings on criteria for credibility, user experience, and transparency [46] while the American Psychiatric Association has an app evaluation model for psychiatrists [69]. Technology-specific guidelines have also been developed, including the guidelines for the design of interventions for mental health on social media [70].

With respect to ensuring anonymity to prevent harms from stigma (recommendation 7), design implications may involve allowing for discreet use. For example, studies have revealed problems with app titles that include stigmatized words like "mood" or "mental health" because users worry others will see them [52]. The 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 personal details).

F. Justice

Justice is a complex ethical principle that is closely linked to fairness and equality, though is not quite the same as either [71]. Sanches *et al.* [43] described 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 [72]. Privacy and autonomy, for example, are widely recognized as human rights and hence some of the concerns raised thus far would fall under rights-based justice. Finally, legal justice requires that people's legal rights be respected. The development and implementation of data-enabled digital mental health technologies raises particular concerns about distributive and rights-based forms of justice. Because the law differs by jurisdiction, we will not discuss legal justice.

There are two main areas in which to analyze distributive and rights-based justice within data-enabled mental health technologies: 1) in the design process and 2) in the distribution of the final product or service. In the first, compensation and credit for the human labor involved in algorithmic design must be considered; and in the second, questions about who is able to access and make use of the service need to be considered.

Recommendation 8: Make known the value of human labor and intellectual property in the development of algorithms to all parties, and potentially compensate for it. With regard to the design process, one type of ethical challenge arises from "heteromation": the extraction of economic value from low-cost (or free) labor [73]. This includes all sorts of labor, from Amazon Mechanical Turk workers, who are paid very low wages to complete tasks that are difficult for an algorithm to do, to the work of completing a Captcha, or other forms of reverse Turing tests, where a person must prove they are human by completing a task (e.g., identifying and selecting all images of crosswalks in a series of nine photographs). These tasks automatically build training sets for algorithms that will eventually be able to accomplish these tasks themselves. Hence, there may be a transfer of intellectual property to the company for which the human laborers 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 of data 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 [74].

A related concern in the development and prototyping of products is piloting on low income, high need, or otherwise vulnerable populations. 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 seems sensible. On the other hand, it may involve 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 no known and tested cures [75]. When it comes to new or experimental medical technologies 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 reasonably decline new (especially, experimental) treatments without feeling compelled to accept them.

There may, of course, 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 for future generations and expand access to less advantaged segments of the population. There is evidence that some people may be willing to share their data, even without direct compensation, if these benefits are communicated to them [76], [77].

Recommendation 9: Follow guidelines for universal accessibility and tailor the level and mode of content to the spectrum of audience needs. Certainly, one positive feature of online therapies is that they can increase access for remote and working populations. In these ways, online therapy, reduces the barrier to entry and could increase uptake. It is unfair, however, to assume that low-income populations all have access to the necessary computing devices and stable Internet connections. Burr and Morley [47] have recently argued that genuine empowerment of a 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 healthcare services move increasingly toward online therapies, research must be 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 retention in mental healthcare [78]. Thus, there is a critical need for more research into the root causes of this finding, as well as ways to better tailor to these populations with the use of online therapies. This includes designing for differing requirements relating to income, literacy levels, and technology access [60].

1) Practical Strategies for Supporting Justice: International guidelines for digital accessibility and "universal design" provide essential starting points for ensuring a technology does not exclude users with older devices, limited Internet access, physical disabilities, or other varying requirements.

Furthermore, as mentioned, researchers have expressed a need for more involvement of people living with mental health issues in technology design [43]–[45]. 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.

One popular approach to user involvement is "participatory design" [79] which involves including users as collaborators from the earliest exploratory phases of development. Orlowski *et al.* [44] provided specific examples of practical applications of participatory design and design thinking methods for mental health technology. Likewise, where the use of a technology will require the involvement of carers, parents, or providers, their unique needs should also be included.

Finally, it is worth noting that the term "user" itself, while useful for its specificity within the technology context, can be inadvertently de-humanizing, obscuring ethical responsibilities. Therefore, in many cases, words like "human," "clients," "patients," "people," or even "lives" may be far more appropriate.

G. Explicability

In addition to the four traditional bioethical principles, Floridi *et al.* [3] included explicability for the AI context, which they describe as enabling the other principles through both intelligibility and accountability. Other terms, such as "transparency," are also frequently used in AI ethics frameworks to capture a similar duty [e.g., the IEEE (2019) [2] uses both transparency and "accountability"]. In general, the idea is that we (i.e., designers, users, and society more generally) need to be able to understand data-enabled systems enough to somehow hold to account their functions (both in terms of their input data and their outputs). We will focus our discussion on the concepts of transparency and accountability as aligned with the IEEE guidelines [2].

Recommendation 10: Ensure transparency and accountability in all aspects of the use of mental health technologies as it is critical to safe and beneficial care. Transparency to do with the collection, use, and storage of data is fundamental to ensuring privacy and other rights, such as informed consent. There are many areas in which transparency must be integrated within an online text-based mental health platform, and many of these arise from the use of a mediating platform which introduces other parties into what was traditionally a confidential conversation between counselor and patient. For example, developers need to be involved in order to design and support the platform; conversations may be recorded and analyzed for potential introduction of AI capabilities; and then these capabilities will need to be audited in order to ensure they function correctly. All of these new layers will require some degree of transparency and accountability.

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, it should be made clear which parties have access to patient notes and therapist-patient conversational records. Additionally, text-based therapy introduces the possibility for different interactions with the data from a session, but this access comes with both benefits and risks which need to be carefully considered [80].

At a high level, there should also be basic transparency and accountability 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 patients. Funding sources and revenue models may create conflicts of interest in data sharing and breach the trust of patients.

1) Practical Strategies for Explicability: Quality frameworks for digital health provide a valuable starting point for applying principles of transparency and accountability. For example, The Transparency for Trust Principles [68] require standard information to be communicated to users in understandable ways, including information around privacy, data security, development characteristics, feasibility, and health benefits. The Psyberguide [46] bases ratings on transparency as well, so examples of technologies that meet the transparency criteria provide models for practical approaches to implementation.

V. Conclusion

The recommendations described above present the result of an ethical analysis conducted by a particular team of professionals in the context of a particular technology type within a specific domain. Analyses by other teams would yield different outcomes although it is reasonable to assume that, for a given context, patterns of concerns will emerge that overlap.

As digital ethics continues to grow in importance, the ethical principles for AI can help us to structure ethical impact analyses. However, the translation of ethical principles into actionable strategies in practice is challenging. We have presented a number of frameworks, including one for a responsible design process and another for providing greater resolution to technology experience as a contribution toward helping address the difficulty in moving from principle to practice within ethical impact analysis.

We have also provided a description of the outcomes of an expert-led ethical analysis, conducted in the context of digital health, in order to show one way such an analysis might be contribute to early stages in a development process.

Of course, our contribution goes only a very small way toward the full integration of ethical impact assessment needed within engineering practice. More research and experimentation with various tools and methods, as well as focused research on ethical implications pertinent to specific application areas and technologies, is still very much needed. We hope the frameworks presented can provide some help toward shaping that path.

Acknowledgment

The authors would like to thank IESO Health for the support and expertise provided.

References

[1] ACM. (2018). *ACM Code of Ethics and Professional Conduct*. [Online]. Available: https://www.acm.org/code-of-ethics

[2] IEEE. (2019). *Ethically Aligned Design*. [Online]. Available: https://ethicsinaction.ieee.org

[3] L. Floridi *et al.*, "Ai4People—An ethical framework for a good Ai society: Opportunities, risks, principles, and recommendations," *Minds Mach.*, vol. 28, no. 4, pp. 689–707, 2018.

[4] OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/

[5] Universite de Montreal. (2017). *Montreal Declaration for Responsible Ai*. [Online]. Available: https://www.montrealdeclaration-responsibleai.com/the-declaration

[6] U.K. House of Lords Select Committee on Artificial Intelligence. (2018). *Ai in the U.K.: Ready, Willing and Able? Report of Session 2017–19*. [Online]. Available: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

[7] High-Level Expert Group on Artificial Intelligence (AI HLEG). (2018). *Ethical Guidelines for Trustworthy AI*. [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1

[8] Beijing of Artificial Intelligence (BAAI). (2019). *Beijing AI Principles*. [Online]. Available: https://baip.baai.ac.cn/en?fbclid=IwAR2HtIRKJxxy9Q1Y953H-2pMHlbIr8pcsIxho93BtZY-FPH39vV9v9B2eY

[9] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, pp. 389–399, Sep. 2019.

[10] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how. An overview of Ai ethics tools, methods and research to translate principles into practices," 2019. [Online]. Available: arXiv:1905.06876.

[11] A. Narayanan and S. Vallor, "Why software engineering courses should include ethics coverage," *Commun. ACM*, vol. 57, no. 3, pp. 23–25, 2014.

[12] M. E. Sunderland, B. Taebi, C. Carson, and W. Kastenberg, "Teaching global perspectives: Engineering ethics across international and academic borders," *J. Res. Innov.*, vol. 1, no. 2, pp. 228–239, 2014.

[13] T. Gebru *et al.*, "Datasheets for datasets," 2020. [Online]. Available: arXiv:1803.09010v6.

[14] T. Hagendorff, "The ethics of Ai ethics: An evaluation of guidelines," 2020. [Online]. Available: arXiv:1903.03425.

[15] P. Verbeek, "Materializing morality: Design ethics and technological mediation," *Sci. Technol. Human Values*, vol. 31, no. 3, pp. 361–380, 2006.

[16] B. Friedman and D. G. Hendry, *Value Sensitive Design: Shaping Technology With Moral Imagination*. Cambridge, MA, USA: MIT Press, 2019.

[17] P. Verbeek, *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago, IL, USA: Univ. Chicago Press, 2011.

[18] J. A. Hamad, M. Hasanain, M. Abdulwahed, and R. Al-Ammari, "Ethics in engineering education: A literature review," in *Proc. IEEE Front. Educ. Conf. (FIE)*, 2013, pp. 1554–1560.

[19] T. L. Beauchamp and J. F. Childress, *Principles of Biomedical Ethics*. Oxford, U.K.: Oxford Univ. Press, 2001.

[20] D. Greene, A. L. Hoffmann, and L. Stark, "Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning," in *Proc. Hawaii Int. Conf. Syst. Sci.*, 2019, pp. 1–10.

[21] Y. Zeng, E. Lu, and C. Huangfu, "Linking artificial intelligence principles," in *Proc. AAAI Workshop Artif. Intell. Safety (AAAI-Safe AI)*, 2019, pp. 1–15.

[22] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The role and limits of principles in Ai ethics: Towards a focus on tensions," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, 2019, pp. 195–200.

[23] J. Stilgoe, R. Owen, and P. Macnaghten, "Developing a framework for responsible innovation," *Res. Policy*, vol. 42, no. 9, pp. 1568–1580, 2013.

[24] B. Mittelstadt, "Ai ethics—Too principled to fail?" in *Proc. Soc. Sci. Res. Netw.*, 2019, Art. no. 3391293.

[25] J. Whittlestone, R. Nyrup, A. Alexandrova, K. Dihal, and S. Cave, *Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research*. London, U.K.: Nuffield Found., 2019.

[26] A. Zimmermann and B. Zevenbergen, *Ai Ethics: Seven Traps*, Tinker, Oklahoma City, OK, USA 2019.

[27] R. A. Calvo, D. Peters, and S. Cave, "Advancing impact assessment for intelligent systems," *Nat. Mach. Intell.*, vol. 2, pp. 89–91, Feb. 2020.

[28] M. E. Larsen *et al.*, "Using science to sell apps: Evaluation of mental health app store quality claims," *NPJ Digit. Med.*, vol. 2, no. 1, p. 18, 2019.

[29] E. Anthes, "Mental health: There's an app for that," *Nat. News*, vol. 532, no. 7597, p. 20, 2016.

[30] L. Ennis, D. Rose, M. Denis, N. Pandit, and T. Wykes, "Can't surf, won't surf: The digital divide in mental health," *J. Mental Health*, vol. 21, no. 4, pp. 395–403, 2012.

[31] C. Nebeker, J. Torous, and R. J. B. Ellis, "Building the case for actionable ethics in digital health research supported by artificial intelligence," *BMC Med.*, vol. 17, no. 1, p. 137, 2019.

[32] D. Peters, R. A. Calvo, and R. M. Ryan, "Designing for motivation, engagement and wellbeing in digital experience," *Front. Psychol.*, vol. 9, p. 797, May 2018.

[33] R. A. Calvo and D. Peters, *Positive Computing: Technology for Wellbeing and Human Potential*. Cambridge, MA, USA: MIT Press, 2014.

[34] UD Council. (2004). *Double Diamond Diagram*. [Online]. Available: https://www.designcouncil.org.uk/news-opinion/what-framework-innovation-design-councils-evolved-double-diamond

[35] R. M. Ryan and E. L. Deci, *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York, NY, USA: Guilford, 2017.

[36] R. A. Calvo, D. Peters, D. M. Johnson, and Y. Rogers, "Autonomy in technology design," in *Proc. Conf. Human Factors Comput. Syst.*, Apr. 2014, pp. 37–40.

[37] R. A. Calvo, D. Peters, K. Vold, and R. M. Ryan, "Supporting human autonomy in AI systems: A framework for ethical enquiry," in *Ethics of Digital Well-Being: A Multidisciplinary Approach*, C. Burr and L. Floridi, Eds., Switzerland: Springer, 2020.

[38] S. Brown *Consequence Scanning Manual, Version 1*. London, U.K.: Doteveryone, 2019.

[39] C. Kerner and V. A. Goodyear, "The motivational impact of wearable healthy lifestyle technologies: A self-determination perspective on fitbits with adolescents," *Amer. J. Health Educ.*, vol. 48, no. 5, pp. 287–297, 2017.

[40] R. Ryan, S. Rigby, and A. Przybylski, "The motivational pull of video games: A self-determination theory approach," *Motivation Emotion*, vol. 30, pp. 344–360, Nov. 2006.

[41] W. Peng, J.-H. Lin, K. A. Pfeiffer, and B. Winn, "Need satisfaction supportive game features as motivational determinants: An experimental study of a self-determination theory guided exergame," *Media Psychol.*, vol. 15, no. 2, pp. 175–196, 2012.

[42] WHO. (2018). *Fact Sheet: Depression*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[43] P. Sanches *et al.*, "HCI and affective health: Taking stock of a decade of studies and charting future research directions," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Mar. 2019, p. 245.

[44] S. Orlowski *et al.*, "Mental health technologies: Designing with consumers," *JMIR Human Factors*, vol. 3, no. 1, p. e4, 2016.

[45] D. C. Mohr, K. R. Weingardt, M. Reddy, and S. M. Schueller, "Three problems with current digital mental health research...and three things we can do about them," *Psychiatric Services*, vol. 68, no. 5, pp. 427–429, 2017.

[46] (2019). *Psyberguide*. [Online]. Available: http://www.Psyberguide.org

[47] C. Burr and J. Morley, "Empowerment or engagement? Digital health technologies for mental healthcare," in *Proc. Digit. Health Technol. Mental Healthcare*, May 2019, pp. 1–29.

[48] K. Vold and J. Whittlestone, "Privacy, autonomy, and personalised targeting: Rethinking how personal data is used," in *Report on Data, Privacy, and the Individual in the Digital Age, by IE University's Center for the Governance of Change*, Madrid, Spain: IE Univeristy, 2019.

[49] D. Susser, B. Roessler, and H. Nissenbaum, "Online manipulation: Hidden influences in a digital world," in *Proc. Soc. Sci. Res. Netw.*, 2018, Art. no. 3306006.

[50] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on user's browsing behavior," in *Proc. ACM 16th Int. Conf. World Wide Web*, 2007, pp. 151–160.

[51] M. Kosinski, D. Stillwell, P. Kohli, Y. Bachrach, and T. Graepel, "Personality and website choice," in *Proc. Web Sci. Conf.*, Jun. 2012, pp. 251–254.

[52] G. Doherty, D. Coyle, and M. Matthews, "Design and evaluation guidelines for mental health technologies," *Interact. Comput.*, vol. 22, no. 4, pp. 243–252, 2010.

[53] S. Turkle, "Authenticity in the age of digital companions," *Soc. Behav. Commun. Biol. Artif. Syst.*, vol. 8, pp. 501–517, Jan. 2007.

[54] K. M. Hertlein, M. L. Blumer, and J. H. Mihaloliakos, "Marriage and family counselors' perceived ethical issues related to online therapy," *Family J.*, vol. 23, no. 1, pp. 5–12, 2015.

[55] D. A. Ludwick and J. Doucette, "Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries," *Int. J. Med. Informat.*, vol. 78, no. 1, pp. 22–31, 2009.

[56] E. Montague, P.-Y. Chen, J. Xu, B. Chewning, and B. Barrett, "Nonverbal interpersonal interactions in clinical encounters and patient perceptions of empathy," *J. Participatory Med.*, vol. 5, p. e33, Aug. 2013.

[57] R. Laugharne, S. Priebe, R. McCabe, N. Garland, and D. Clifford, "Trust, choice and power in mental health care: Experiences of patients with psychosis," *Int. J. Soc. Psychiatry*, vol. 58, no. 5, pp. 496–504, 2012.

[58] A.-C. Brigida. (2013). *A Virtual Therapist*. [Online]. Available: https://viterbi.usc.edu/news/news/2013/a-virtual-therapist.htm

[59] A. Tieu. (2015). *We Now Have an Ai Therapist, and She's Doing Her Job Better Than Humans Can*. [Online]. Available: https://futurism.com/uscs-new-ai-ellie-has-more-success-than-actual-therapists

[60] D. Robotham, S. Satkunanathan, L. Doughty, and T. Wykes, "Do we still have a digital divide in mental health? A five-year survey follow-up," *J. Med. Internet Res.*, vol. 18, no. 11, p. e309, 2016.

[61] B. Sparrow, J. Liu, and D. M. Wegner, "Google effects on memory: Cognitive consequences of having information at our fingertips," *Science*, vol. 333, no. 6043, pp. 776–778, 2011.

[62] E. Z. Patai *et al.*, "Hippocampal and retrosplenial goal distance coding after long-term consolidation of a real-world environment," *Cerebr. Cortex*, vol. 29, no. 6, pp. 2748–2758, 2019.

[63] Plato, *Plato's Phaedrus*. Cambridge, U.K.: Cambridge Univ. Press, 1952.

[64] J. Hernandez-Orallo and K. Vold, "Ai extenders: The ethical and societal implications of humans cognitively extended by Ai," in *Proc. 2nd AAAI/ACM Annu. Conf. AI Ethics Soc.*, 2019, pp. 507–513.

[65] M. Lanzing, "Strongly recommended: Revisiting decisional privacy to judge hypernudging in self-tracking technologies," *Philosophy Technol.*, vol. 32, pp. 549–568, Jun. 2018.

[66] W. A. Nelson, "The ethics of telemedicine. Unique nature of virtual encounters calls for special sensitivities," *Healthcare Executive*, vol. 25, no. 6, pp. 50–53, 2010.

[67] P. Topham, P. Caleb-Solly, P. Matthews, A. Farmer, and C. Mash, "Mental health app design: A journey from concept to completion," in *Proc. 17th Int. Conf. Human–Comput. Interact. Mobile Devices Services Adjunct (MobileHCI)*, 2015, pp. 582–591.

[68] T. Wykes and S. Schueller, "Why reviewing apps is not enough: Transparency for trust (T4T) principles of responsible health app marketplaces," *J. Med. Internet Res.*, vol. 21, no. 5, 2019, Art. no. e12390.

[69] AP Association. (2019). *App Evaluation Model*. [Online]. Available: https://www.psychiatry.org/psychiatrists/practice/mental-health-apps/app-evaluation-model

[70] L. Manikonda and M. Choudhury, "Modeling and understanding visual attributes of mental health disclosures in social media," in *Proc. Int. Conf. Human Factors Comput. Syst.*, Jan. 2017, pp. 170–181.

[71] A. Sen, *The Idea of Justice*. London, U.K.: Penguin, 2010.

[72] R. Gillon, "Medical ethics: Four principles plus attention to scope," *BMJ*, vol. 309, no. 6948, p. 184, 1994.

[73] H. R. Ekbia and B. A. Nardi, *Heteromation, and Other Stories of Computing and Capitalism*. Cambridge, MA, USA: MIT Press, 2017.

[74] C. Liu, K. M. Scott, R. L. Lim, S. Taylor, and R. A. Calvo, "EQ clinic: A platform for learning communication skills in clinical consultations," *Med. Educ. Online*, vol. 21, no. 1, 2016, Art. no. 31801.

[75] T. Morel, S. Aymé, D. Cassiman, S. Simoens, M. Morgan, and M. Vandebroek, "Quantifying benefit-risk preferences for new medicines in rare disease patients and caregivers," *Orphanet J. Rare Diseases*, vol. 11, no. 1, p. 70, 2016.

[76] C. L. Ratcliff, K. A. Kaphingst, and J. D. Jensen, "When personal feels invasive: Foreseeing challenges in precision medicine communication," *J. Health Commun.*, vol. 23, no. 2, pp. 144–152, Jan. 2018.

[77] G. Halvorson and W. Novelli, *Data Altruism: Honoring Patients' Expectations for Continuous Learning, NAM Perspectives*, Inst. Med., Washington, DC, USA, Jan. 2014.

[78] M. Alegri'a *et al.*, "Evaluation of a patient activation and empowerment intervention in mental health care," *Med. Care*, vol. 46, no. 3, p. 247, 2008.

[79] J. Simonsen and T. Robertson, *International Handbook of Participatory Design*. London, U.K.: Routledge, 2012.

[80] M. W. Kahn, S. K. Bell, J. Walker, and T. Delbanco, "Let's show patients their mental health records," *J. Amer. Med. Assoc.*, vol. 311, no. 13, pp. 1291–1292, 2014.

Dorian Peters received the M.M.Des degree from the Univeristy of Sydney, Sydney, NSW, Australia and the B.A. degree from the Carnegie Mellon University, Pittsburgh, PA, USA.

She is an Author, a Researcher, and a Technology Designer. She has published two books on design *Interface Design for Learning* (New Riders) and *Positive Computing: Technology for Wellbeing* (MIT Press). She specialist in design for health and digital wellbeing. She currently works with the Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, U.K., and Imperial College London, London, U.K. She also facilitates design workshops and consults for nonprofit and industry organizations.

Karina Vold received the B.A. degree in philosophy and political science from the University of Toronto, Toronto, ON, Canada, and the Ph.D. degree in philosophy from McGill University, Montreal, QC, Canada, in 2017.

She is a Postdoctoral Research Associate with the Leverhulme Centre for the Future of Intelligence and a Research Fellow with the Faculty of Philosophy, University of Cambridge, Cambridge, U.K. She specializes in philosophy of cognitive science and artificial intelligence. She is also currently a Digital Charter Fellow with the Alan Turing Institute, London, U.K. Her recent work focuses on theories of situated cognition, cognitive enhancement, AI ethics, and neuroethics.

Dr. Vold serves as an Associate Editor for the IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY.

Diana Robinson received the M.B.A. degree from the Cambridge Judge Business School, Cambridge, U.K., and the B.A. degree in philosophy from Princeton University, Princeton, NJ, USA. She is currently pursuing the Ph.D. degree in computer science with the University of Cambridge, Cambridge.

She is a Student Fellow with the Leverhulme Centre for the Future of Intelligence, University of Cambridge. She specializes in human–computer interaction, philosophy, and business. She worked as a Commodity Risk Analyst in BP's Integrated Supply and Trading Business. She was a Visiting Scholar with the MIT Media Lab, Opera of the Future Group, Cambridge, MA, USA. She was a Princeton Project 55 Fellow in 2012/2013.

Rafael A. Calvo (Senior Member, IEEE) received the Ph.D. degree from the Universidad de Rosario, Rosario, Argentina, in 2000.

He is a Professor and the Director of Researcher with the Dyson School of Design Engineering, Imperial College London, London, U.K., and the Co-Lead with the Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, U.K. Until 2019, he was a Professor with the University of Sydney, Sydney, NSW, Australia, and a Future Fellow with the Australian Research Council. He has coauthored *Positive Computing* (MIT Press). He has authored two books and many publications in the fields of learning technologies, affective computing, and computational intelligence.

Prof. Calvo is a recipient of five teaching awards for his work on learning technologies. He has the Co-Editor of the IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY. He was also the Co-Editor of the *Oxford Handbook of Affective Computing*, and an Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, the IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, and the *Journal of Medical Internet Research Human Factors*.