Journal of Responsible Technology 11 (2022) 100044

Contents lists available at ScienceDirect

Journal of Responsible Technology

journal homepage: www.sciencedirect.com/journal/journal-of-responsible-technology

Involving psychological therapy stakeholders in responsible research to develop an automated feedback tool: Learnings from the ExTRAPPOLATE project

Jacob A Andrews [a,b,*], Mat Rawsthorne [c], Cosmin Manolescu [l], Matthew Burton McFaul [j], Blandine French [a,e], Elizabeth Rye [e], Rebecca McNaughton [e], Michael Baliousis [d,e], Sharron Smith [e,h], Sanchia Biswas [d,e], Erin Baker [e], Dean Repper [e,i], Yunfei Long [f], Tahseen Jilani [g], Jeremie Clos [k], Fred Higton [a], Nima Moghaddam [h], Sam Malins [b,d]

[a] NIHR Mindtech Medtech Co-operative, University of Nottingham, Nottingham, UK
[b] Mental Health and Clinical Neurosciences, University of Nottingham, Nottingham, UK
[c] Institute of Mental Health, Nottingham, UK
[d] Specialist Services, Nottinghamshire Healthcare NHS Foundation Trust, Nottingham, UK
[e] Patient and Practitioner Reference Group, Nottingham, UK
[f] School of Computer Science and Electronic Engineering, University of Essex, UK
[g] HDR UK, Digital Research Service, University of Nottingham, Nottingham, UK
[h] School of Psychology, University of Lincoln, Lincoln, UK
[i] Trent Psychological Therapies Service, Nottingham, UK
[j] Virtual Health Labs Ltd., London, UK
[k] School of Computer Science, University of Nottingham, Nottingham, UK
[l] Mental Health Services for Older People (MHSOP), Nottinghamshire Healthcare NHS Foundation Trust, Nottingham, UK

ARTICLE INFO

Keywords:
Patient and public involvement
Responsible research and innovation
Autonomous systems
Psychological therapy

ABSTRACT

Understanding stakeholders' views on novel autonomous systems in healthcare is essential to ensure these are not abandoned after substantial investment has been made. The ExTRAPPOLATE project applied the principles of Responsible Research and Innovation (RRI) in the development of an automated feedback system for psychological therapists, 'AutoCICS'. A Patient and Practitioner Reference Group (PPRG) was convened over three online workshops to inform the system's development. Iterative workshops allowed proposed changes to the system (based on stakeholder comments) to be scrutinized. The PPRG reference group provided valuable insights, differentiated by role, including concerns and suggestions related to the applicability and acceptability of the system to different patients, as well as ethical considerations. The RRI approach enabled the *anticipation* of barriers to use, *reflection* on stakeholders' views, effective *engagement* with stakeholders, and *action* to revise the design and proposed use of the system prior to testing in future planned feasibility and effectiveness studies. Many best practices and learnings can be taken from the application of RRI in the development of the AutoCICS system.

Abbreviations

ACT Matrix Acceptance and Commitment Therapy Matrix
ARC process Awareness, Relationships, Cultural agreements
AREA framework Anticipate, Reflect, Engage, Act
CICS Consultation Interaction Coding Scheme
EPSRC Engineering and Physical Sciences Research Council
ExTRAPPOLATE Explainable Therapy Related Annotations: Patient and Practitioner Oriented Learning Assisting Trust and Engagement
HDI Human Data Interaction
NIHR National Institute for Health Research

NB. All figures to be presented in colour.

* Corresponding author at: NIHR Mindtech Medtech Co-operative, Institute of Mental Health Triumph Road, University of Nottingham, Nottingham, NG7 2TU.
E-mail address: Jacob.andrews@nottingham.ac.uk (J.A. Andrews).

https://doi.org/10.1016/j.jrt.2022.100044
Received 17 December 2021; Received in revised form 1 June 2022; Accepted 3 August 2022
Available online 6 August 2022

NLP	Natural Language Processing
PPI	Patient and Public Involvement
PPRG	Patient and Practitioner Reference Group
RRI	Responsible Research and Innovation
SUS	Systems Usability Scale
UK	United Kingdom

1. Introduction

Machine learning applications have potential to greatly improve healthcare, but there have already been many failed attempts to implement these where implications for clinicians, patients and carers have not been adequately considered (Panch et al., 2019). Machine learning has been defined as techniques "commonly used to solve a variety of real-world problems with the help of computer systems which can learn to solve a problem instead of being explicitly programmed" (Kühl et al., 2020). Machine learning is itself considered a subfield of artificial intelligence, which describes "a broad discipline that aims to understand and design systems that display properties of intelligence" (Panch et al., 2018). Automated systems developed using machine learning are often abandoned when healthcare professionals, patients, or carers do not understand their algorithms and, therefore, do not trust such systems (Markus et al., 2021). Furthermore, the use of systems developed using machine learning may have unforeseen social or ethical implications, for example, bias from limited representation of minority groups in the data used to train algorithms (Vokinger et al., 2021).

Responsible Research and Innovation (RRI) has the potential to ensure that these issues are addressed. RRI is an initiative to promote early exploration of the potential consequences of research for society and to ensure that research programmes are sustainable and inclusive (Burget, Bardone, & Pedaste, 2017). Stilgoe and colleagues developed a framework for Responsible Innovation, which encompasses four dimensions: anticipation, reflexivity, inclusion and responsiveness (Stilgoe et al., 2013). The United Kingdom's Engineering and Physical Sciences Research Council have adapted these dimensions to produce their own Framework for Responsible Innovation, using the acronym AREA: Anticipate, Reflect, Engage and Act. The framework states that responsible innovation should *anticipate* potential unintended impacts of research, *reflect* on purposes, motivations and ambiguities of the approach, *engage* to discuss these issues broadly and inclusively, and *act* to use these considerations to influence the research trajectory (Engineering and Physical Sciences Research Council, 2021). Sharing methods for applying RRI to the very earliest stages of developing trustworthy autonomous systems is important to enable others to understand how these concepts can be put into practice in a meaningful, practical and impactful way (Rivard & Lehoux, 2020).

In this paper, we present learnings from the ExTRAPPOLATE project, which sought to apply the principles of responsible research and innovation in the early development of a novel automated feedback system for clinical psychologists (AutoCICS). The system was designed to help improve their clinical practice. The application of the EPSRC and Stilgoe et al.'s (2013) AREA framework of Responsible Research and Innovation is discussed, reflecting on the best practices, shortcomings and learnings from the process undertaken. Findings are related to the field of Human Data Interaction, to understand how stakeholder views related to factors of legibility, agency and negotiability (Victorelli et al., 2020).

1.1. ExTRAPPOLATE: Background to the AutoCICS system

The effectiveness of psychological therapy varies significantly between individual therapists (Baldwin & Imel, 2013; Barkham et al., 2017). Furthermore, experience does not seem to significantly improve a clinical psychologist's success rate (Goldberg et al., 2016). One factor that may contribute to this is the lack of readily available feedback for clinical psychologists. Traditional methods of quality control rely on expert observation and rating, which is costly and time-consuming. Furthermore, the outputs of these assessments do not necessarily identify how effective a therapist is (Branson et al., 2015). There is an obvious need for systematic, objective and routine means of appraising the quality of psychological therapy and feeding back to the clinical psychologist to help them improve (Perepletchikova, 2011; Waller & Turner, 2016).

The Consultation Interaction Coding Scheme (CICS) is a means of identifying interaction types within psychological therapy session transcripts, for example when a client is describing their problem, when they are evaluating themselves or their therapy, or when they are noticing changes in the experience of their condition (Malins et al., 2020). This has the benefit of allowing analysis of a particular session by examining its constituent parts. In addition, the coding scheme provides a way to rate patient activation, that is, the degree to which the patient is actively engaged with therapeutic processes (Hibbard et al., 2004).

Recent work has shown that CICS coding scheme ratings can give an indication of clients' outcome prognosis from interactions at initial sessions and is superior to client and therapist predictions of outcome based on client confidence and ability to achieve therapeutic change (Malins et al., 2021a, 2021b). Thus, coding session transcripts according to this scheme could permit recognition of where a therapist could improve, by adjusting the way they interact in specific targeted ways for particular clients.

However, manually coding session transcripts with the CICS coding scheme is time consuming, preventing routine use. The ExTRAPPOLATE project explored the possibility of using Natural Language Processing (NLP) to automate the process of coding sessions, with the possibility that an automated system (AutoCICS) could analyse a therapy session and provide guidance to a therapist on how to improve their practice in a single, computer-based tool.

Automated text classification, a popular topic in NLP, can unlock information embedded in clinical text by extracting particular features of language (e.g. mention of symptoms, sentence length and emotion words) from patient-clinician conversations (Ewbank et al., 2020). Prior work has shown that candidate language features can be selected to automatically identify interaction-types and levels of patient activation (Rawsthorne et al., 2020). Co-produced linguistic feature selections can be extracted from psychological therapy transcripts and machine learning applied using Support Vector Machines (SVMs) to assess whether in-session patient activation and interaction-types may be accurately predicted (Rawsthorne et al., 2020). Building on such prior work, we created the initial proof of concept for a tool, AutoCICS, that could analyse a therapy session using the CICS coding scheme and give feedback to a therapist to improve their practice.

1.2. ExTRAPPOLATE: Responsible Research Through Stakeholder Involvement

Alongside a technical workstream, the ExTRAPPOLATE project sought to put into practice the principles of RRI through patient and public involvement (PPI), by bringing together a group of key stakeholders (the Patient and Practitioner Reference Group; PPRG). This group engaged with the research team in a dialogue to explore the implications of the AutoCICS system and discuss practical factors associated with its design and proposed use to inform its development.

PPI is well suited to the RRI principles because it aims to enable patients and members of the public to contribute to research to ensure that it addresses issues of importance to them. PPI is envisaged as a way for research to be conducted 'with' or 'by' patients and members of the public who may ultimately be affected by the research, rather than having research done 'to' or 'on' them (Jackson et al., 2020). One main benefit of involvement is that it enables research projects to be informed by the knowledge gained by patients and members of the public through their relevant experiences (Mockford et al., 2012). Boivin and colleagues (2018) describe PPI contributors as "experience-based experts who contribute knowledge that is complementary to that of scientists

and professionals".

In the context of RRI applied to novel healthcare technologies, combining PPI with consultation of healthcare professionals can enable fulfilment of all four constituent parts of the EPSRC's AREA framework for Responsible Innovation. Discussions with stakeholders (in this context: patients, carers and healthcare professionals) helps anticipate the impacts new technologies might have, including unintended negative consequences of innovations developed in good faith (*anticipate*). Well conducted stakeholder engagement requires researchers to describe their research in ways which are accessible to non-specialists, and which place the work in a wider societal context, providing a space for stakeholders to raise questions and challenge assumptions (*reflect*). Stakeholder engagement also has an emphasis on inclusivity, wherein people with different backgrounds, different life experiences and different levels of professional status are treated as valued equals in discussions about the research, enabling effective debate (*engage*). Research projects benefit from these discussions, with research teams taking on board what they have learnt from stakeholders to inform the project direction (*act*). In this paper, we describe how we applied RRI principles through the AREA framework to inform research and development plans for the AutoCICS tool, in the context of psychological therapy.

2. Materials and Methods

The methods described in this section are summarised in a table in Appendix A for quick reference.

2.1. People involved

The core project team consisted of: three technical experts (YL, TJ, JC); two clinical psychologists (SM, NM); an assistant psychologist (CM); a service-user researcher (MR); an interdisciplinary researcher in digital mental health (JA); an expert in human-computer interaction design from the digital health industry (MBM); and two patient and public involvement advisors (FH and one wishing to remain anonymous). The latter were involved as core team members to oversee the process and offer guidance on workshop design and preparation of materials for lay audiences.

Separate from this core project team, an expert advisory panel was assembled: the Patient and Practitioner Reference Group (PPRG). This group consisted of individuals from five different roles/perspectives, all of whom were stakeholders in the project by virtue of their knowledge and experience of psychological therapy. The five perspectives were: patient (2 people), carer (1 person); clinical psychologist (3 people); psychological therapy trainer (3 people); and psychological therapy service manager (3 people). Some group members were able to provide a perspective from multiple roles, for example where they had experience training clinical psychologists but also worked with patients themselves.

The PPRG reference group were recruited using multiple methods. Psychological therapists, trainers and managers were recruited through advertisements circulated in local mental health organisations and by personal contacts of the research team. The patients and carers were known to the service user researcher through previous research projects, and two of the group had past experience of undertaking research projects themselves. Some literature highlights a professionalization paradox, whereby patients and members of the public lose their lay perspective as they become more trained in research (Ives, Damery, & Redwod, 2013). This was not considered to be an issue in the present study because, similar to the position taken by Staley (2013) the memories and perspectives of the recipient of psychotherapy (or carer thereof) were not considered to be lost or tarnished by virtue of prior involvement in research. The research team did consider whether the views expressed by the patient and carer group would be unduly positive about the system due to past affiliation with the research team. However, in the workshops, participants provided open and honest critique about the research, which may not have been forthcoming if they had had no prior relationship with the research team, thus the relationship was judged to be more beneficial than detrimental to the research process.

After consultation with the University of Nottingham School of Medicine Research Ethics Committee, it was determined that ethical approval was not required for the workshops, as they were considered an involvement activity with experienced group members to inform the development of a novel system. Patients and carers were reimbursed for their time at standard involvement rates.

2.2. Aims and use of the AREA Framework

With the methods described below, we aimed to *anticipate* future consequences from the roll out of the AutoCICS system. Three online workshops were conducted with a multi-stakeholder group (the PPRG reference group). Workshop 1 used the Trustscapes research tool, a civic engagement process tool developed as part of the UnBias Fairness Toolkit (Lane et al., 2018), to elicit concerns and considerations of importance from four key stakeholder groups. These group members were selected for having relevant tacit knowledge of the context in which the system would be used. The Trustscapes tool provides a framework to consider stakeholders' concerns and considerations, and elicits stakeholder views on how the approach under discussion should be conducted to ensure it works well. In workshop 2, we aimed to *reflect* by inviting stakeholders to critique the design and proposed use of the AutoCICS system and the research trajectory, with an opportunity to provide an early rating of the system design using the System Usability Scale (SUS) (Brooke, 1996). Workshops were led by researchers who were independent of the development of the CICS coding scheme, to enable participants to speak relatively more openly and freely about their views on the proposed AutoCICS system. We sought to *inclusively engage* throughout all workshops by including representatives from five different, relevant perspectives on the development of the system. Further engagement was made possible by the involvement of a service user researcher and two patient and public involvement volunteers on the core research team. The research team sought to *act in response* to stakeholder views in a number of ways, with initial responses to stakeholder concerns proposed to the reference group at Workshop 3. This final workshop provided a space to consider together with stakeholders whether the proposed actions were suitable, whether they addressed the concerns raised by the PPRG reference group at earlier workshops and what could additionally be done to address their concerns and anticipated problems with the system. Details of each workshop and its outcomes are provided below.

2.3. Stakeholder Involvement Design

The design of the three workshops is summarised in Fig. 1. The three workshops were originally designed to take place in person, but after lockdown restrictions were imposed as part of the COVID-19 pandemic, the involvement design was reconfigured to enable online participation. Each member of the PPRG reference group was provided with a one-to-one induction session to learn to use the online platform Miro (Khusid, 2011).

2.3.1. Workshop 1 Design

Workshop 1 was designed to allow all members of the PPRG reference group to meet for the first time, to discuss ground rules for the project, and to begin eliciting concerns about the AutoCICS system and suggestions for its future development. In the first part of Workshop 1, each PPRG sub-group met on a separate Miro board, facilitated either by a member of the research team or by a nominated member of the perspective group. A video call in each Miro board enabled discussion. Where possible, these were recorded (with attendees' consent). After each attendee introduced themselves, the first activity involved a

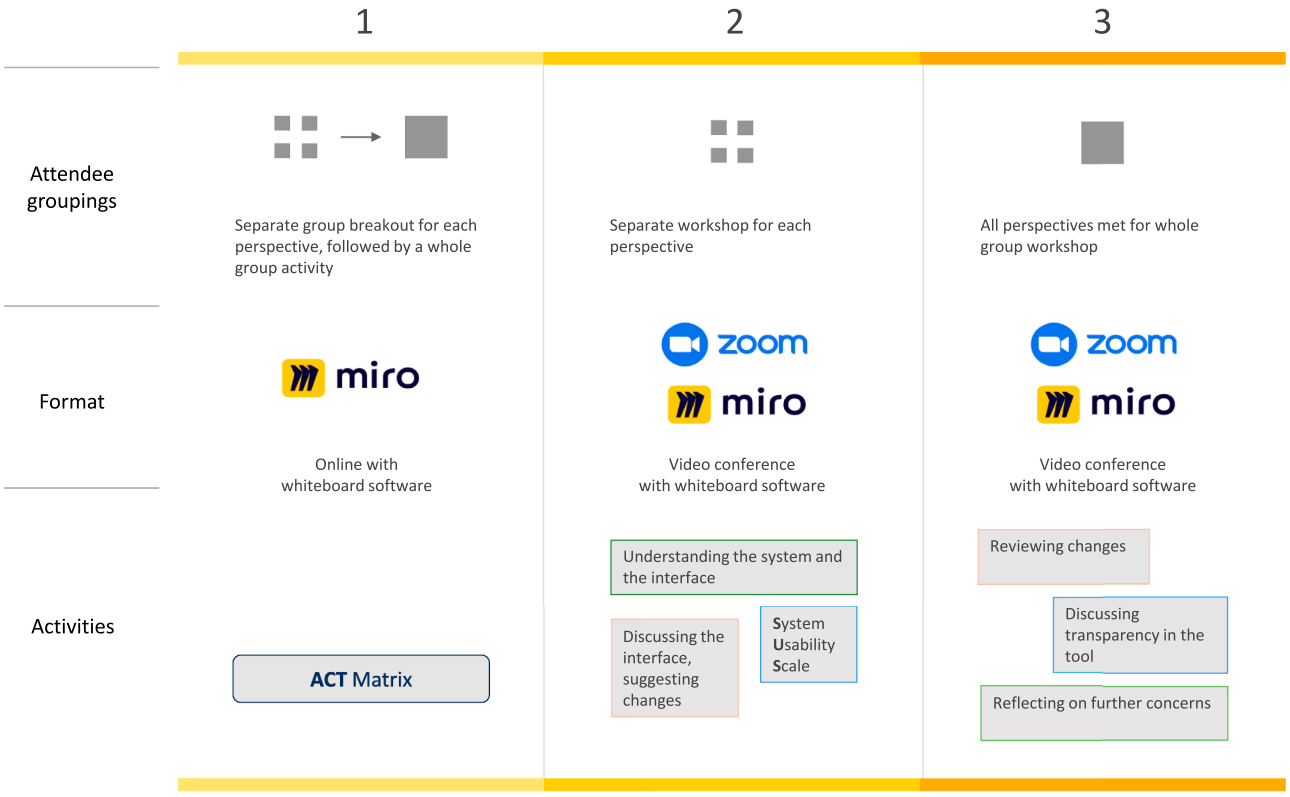

Fig. 1. Design of the three workshops conducted in the research project.

facilitated discussion using four panels adapted from the Trustscapes tool from the Unbias Fairness Toolkit (Rovatsos et al., 2019). The four Trustscapes panels asked attendees to comment in these areas:

- Describe your concerns relating to automated rating and feedback on quality of psychological therapy
- What is important to you about how therapy is assessed and the impact of it being digitally rated?
- How do you think these issues (raised above) are being addressed by companies and authorities (such as the NHS, service providers, regulatory bodies, etc)?
- Ideally, what would you like to see done to make sure digital rating of therapy works well?

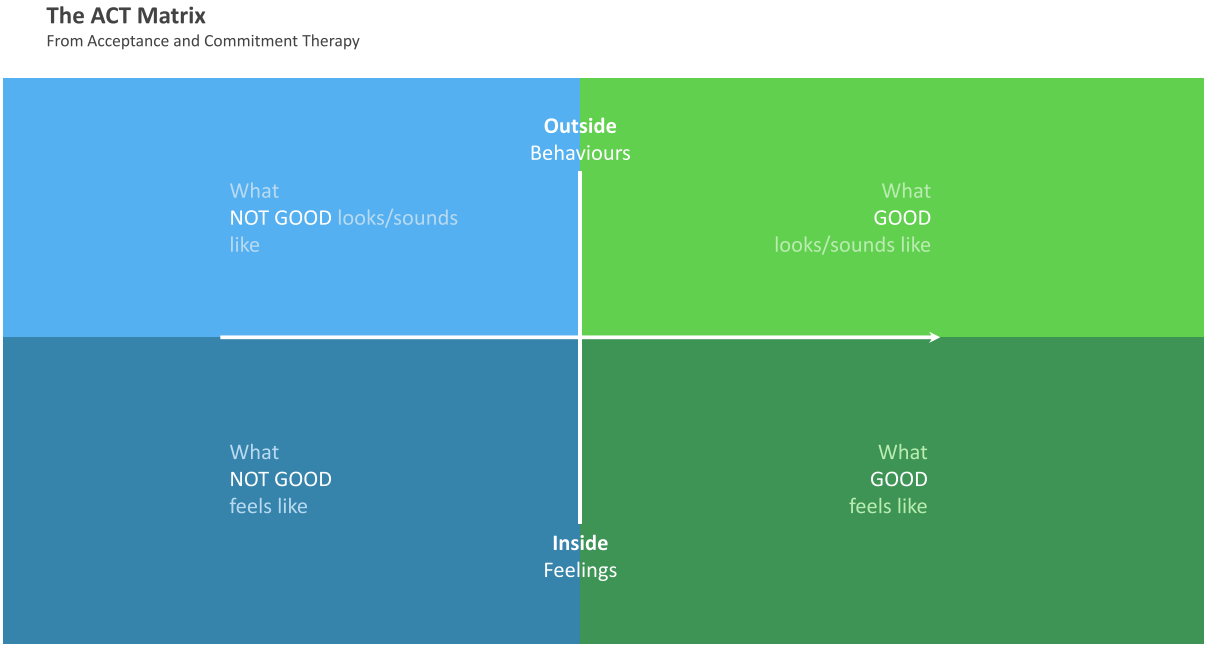

Fig. 2. The ACT (Acceptance and Commitment Therapy) Matrix

The online whiteboard functionality allowed attendees to write digital sticky notes and place these in the relevant panel. Facilitators noted key points to be shared with the whole group in part 2, after a break.

Involving a diverse range of stakeholders can be challenging, and to set intentions for effective collaborative working, we adapted Polk and colleagues' Acceptance and Commitment Therapy Matrix (Polk et al., 2016). While this is traditionally used as a therapeutic tool, in this project we used it as per Atkins' ARC process (Awareness, Relationships, Cultural agreements; (Atkins et al., 2019)), to help the group reflect on how they would like to work together. The matrix (figure 2), was presented to attendees on Miro to collectively set ground rules and intentions. Attendees typed into sticky notes and placed these in the relevant quadrant.

2.3.2. Workshop 2 design

The purpose of the second workshop was to demonstrate the technology under development and invite PPRG reference group members to provide ideas, thoughts and suggestions to inform its development, as well as collecting an initial usability score. Originally, we envisaged showing a minimum viable product to the PPRG. However, given time constraints and multiple delays relating to the COVID-19 pandemic, it was not possible for the technical research team to develop a functioning product. Instead, digital illustrations showed demonstration screens that could form part of the final interface (Figs. 3a and 3b). These illustrations were developed by the research team with the principle aim of enabling workshop attendees to imagine how the system would function and thus be able to conceptualise its use and articulate their response to it. The illustrations also provided a starting point for discussions on the interface design. The priority in the development of the illustrations was clarity on the purpose and functionality of the system.

In addition to commenting on the proposed functionality and design of the system, stakeholders were encouraged to reflect on the understandability and meaningfulness of the 'explanations' detailing how the algorithm worked (Fig. 3a), as these explanations were conceptualised as a way to enable end users to understand how the algorithm had reached its conclusions, and thereby provide transparency in the system. The research team also invited the PPRG reference group to score the system using the System Usability Scale (SUS) (Brooke, 1996). Facilitators made notes within the Miro platform to capture attendees' comments, and attendees wrote their thoughts in sticky notes on the platform.

2.3.3. Workshop 3 design

The design for Workshop 3 was adapted in response to the analysis of workshops 1 and 2 and through discussions with the research team. Five key changes to the use and design of the system, and around 20 other changes, were chosen to present back to the PPRG reference group to understand whether these addressed their concerns. It was still not possible at this point to present a working prototype, so instead, proposed changes to the use and design of the system were discussed with the PPRG reference group. Revised illustrations of the proposed user interface were produced and shown to the PPRG for comments. Discussion in the research team had shown that transparency was still an issue of concern, though not enough clarity had been obtained in workshop 2 to be able to suggest related changes. Thus, transparency was included as a 'talking point' in the final design of workshop 3, with examples of algorithm explanations from Google and Amazon presented within Miro to provide a comparator with the technology under development and stimulate discussion. As there had been reluctance to complete the SUS usability scale by some attendees in workshop 2, limited data on the SUS had been collected and it was decided not to follow this up in workshop 3.

2.3.4. Post hoc debrief

Following workshop 3, a further session was requested by some members of the PPRG reference group to discuss what had worked well in the session and what could have been improved. This session was facilitated by a member of university staff who had not been involved in the workshops.

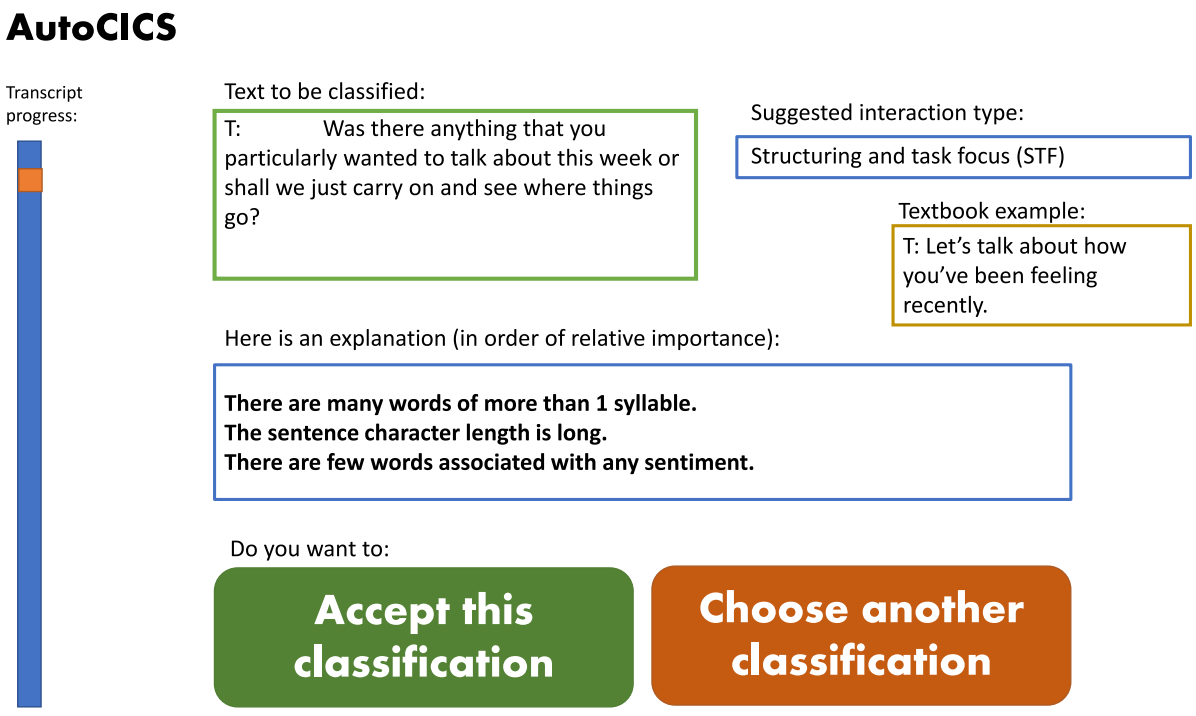

Fig. 3a. Illustration demonstrating a proposed interface for the AutoCICS therapist feedback tool, which was presented to the stakeholder group for discussion. This screen would enable a therapist to review how their session had been analysed and coded by the automated tool, and to correct any codings/classifications that the therapist considered not to be accurate. The explanation box was designed to provide users with insight into how the coding algorithm took its decisions, providing transparency and explainability to users.

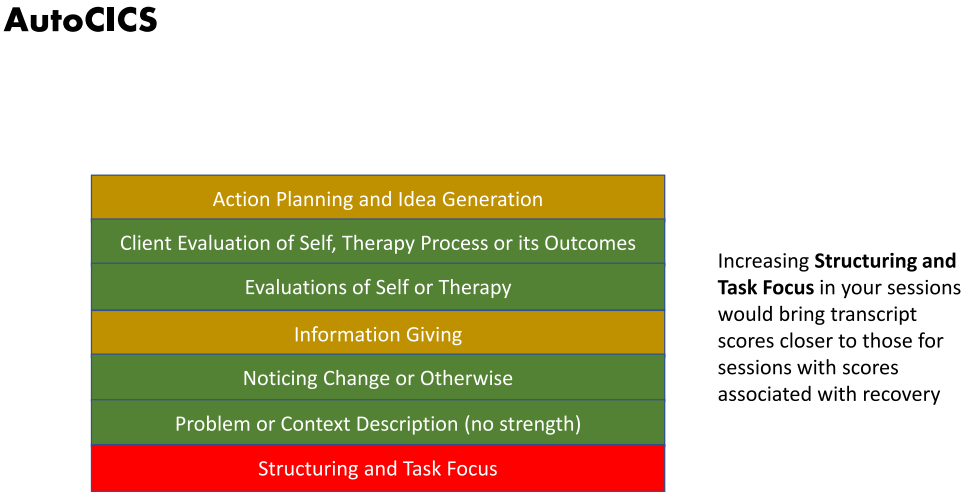

Fig. 3b. Illustration demonstrating a proposed interface for the AutoCICS therapist feedback tool, which was presented to the stakeholder group for discussion. This screen uses a traffic light system to feedback to therapists on how to improve their next session with the current client.

2.3.5. Analysis

The three workshops were designed to obtain PPI input from key stakeholders to inform future technical research to develop the AutoCICS system. A pragmatic, thorough, purpose-driven approach was used to analyse the data from the workshops, with multiple team members reviewing the data.

Data from workshops 1 and 2 consisted of written contributions on the Miro board and transcribed utterances from audio-recorded workshops. These were analysed together to inform workshop 3. The data were separated into individual comments, and these were listed in Microsoft Excel. Some contributions were identifiable by perspective but were otherwise anonymised at the point of transcription. All Miro sticky notes were also added to the Excel sheet in individual cells. Where transcribed notes were the same as any Miro sticky note entry, without adding further detail, these duplicate entries were removed.

The first and last authors read through this dataset consecutively and considered how each entry might inform the research programme trajectory. Notes were written by either the first or last author alongside each data cell to either: set out how the research trajectory would change in light of the suggestion, explain why the suggestion could not be enacted, or respond to points framed as questions. This exhaustive approach taking account of every suggestion from the PPRG reference group ensured accountability of the process. The first and last author met to discuss and agree the complete set of responses.

For the purposes of presenting the changes back to the PPRG reference group in Workshop 3, a reduction phase was conducted, as there was not sufficient time in workshop 3 to discuss every comment from workshop 1 and 2. In the reduction phase, the first author identified overarching areas of concern or suggested change, leading to the generation of five key changes, and 20 other comments and responses, for discussion in Workshop 3. This prioritisation was then validated with two other members of the research team, including a PPI volunteer. Between workshops 2 and 3, the core project team met and reflected on any other areas that could usefully be discussed with the stakeholder reference group (PPRG).

Attendee contributions to the ACT matrix from Workshop 1 were not analysed, as the matrix activity served only as a tool for workshop attendees to reflect on how they would wish to work together. The completed ACT matrix was briefly reviewed at the start of workshops 2 and 3 to encourage a positive environment for a successful collaboration. Some examples of entries have been included in the results section below. System Usability Scale scores from workshop 2 were analysed in line with the creator's guidance (Brooke, 2013).

The data from Workshop 3 were analysed using Template Analysis (King, 2012) to produce descriptive categories summarising the data. First, all written comments were collated along with a transcription of the workshop audio recording and notes taken by the research team during the workshop. The data were then reviewed line by line to identify commonalities, with themes iteratively renamed and reorganised to produce a final set of descriptive categories.

2.3.6. Project outputs

The PPRG reference group were involved in creating project outputs, including the design of a conference poster and in making contributions to this paper (reading and commenting on the draft manuscript).

3. Findings

3.1. Findings from Workshop 1

The first workshop was attended by 3 members of the core project team and 9 members of the PPRG reference group: 2 patients, 1 carer, 2 clinical psychologists, 2 therapy trainers and 2 service managers. Attendance at each workshop is set out in a table in Appendix B. The first and second parts of the session were successfully coordinated and all attendees were able to use the Miro software effectively. The research team made notes at each session on the Miro boards to capture PPRG comments, and the reference group themselves also made use of this functionality to contribute their own written comments. The ACT Matrix activity collected a wide variety of thoughts from the PPRG reference group on desirable and undesirable behaviours for the project (Fig. 4).

The Trustscapes activity enabled collection of a multitude of thoughts and ideas across the four panels on the Miro boards for each sub-group. Key points captured by facilitators are presented in Table 1.

Multiple groups raised concerns about the reductionist nature of the approach, mentioning that the system would lack the ability to address nuance or to consider relational aspects of therapy. Transparency was another key issue, with PPRG members highlighting the need to check the conclusions drawn by the system with the therapist and with patients alike. Concerns were also raised about the accuracy and applicability of the system to different therapy types and different mental health conditions.

3.2. Findings from Workshop 2

Due to limited availability of PPRG reference group members, workshop 2 was run as a separate, asynchronous session for each perspective, except service managers, who were unavailable to arrange

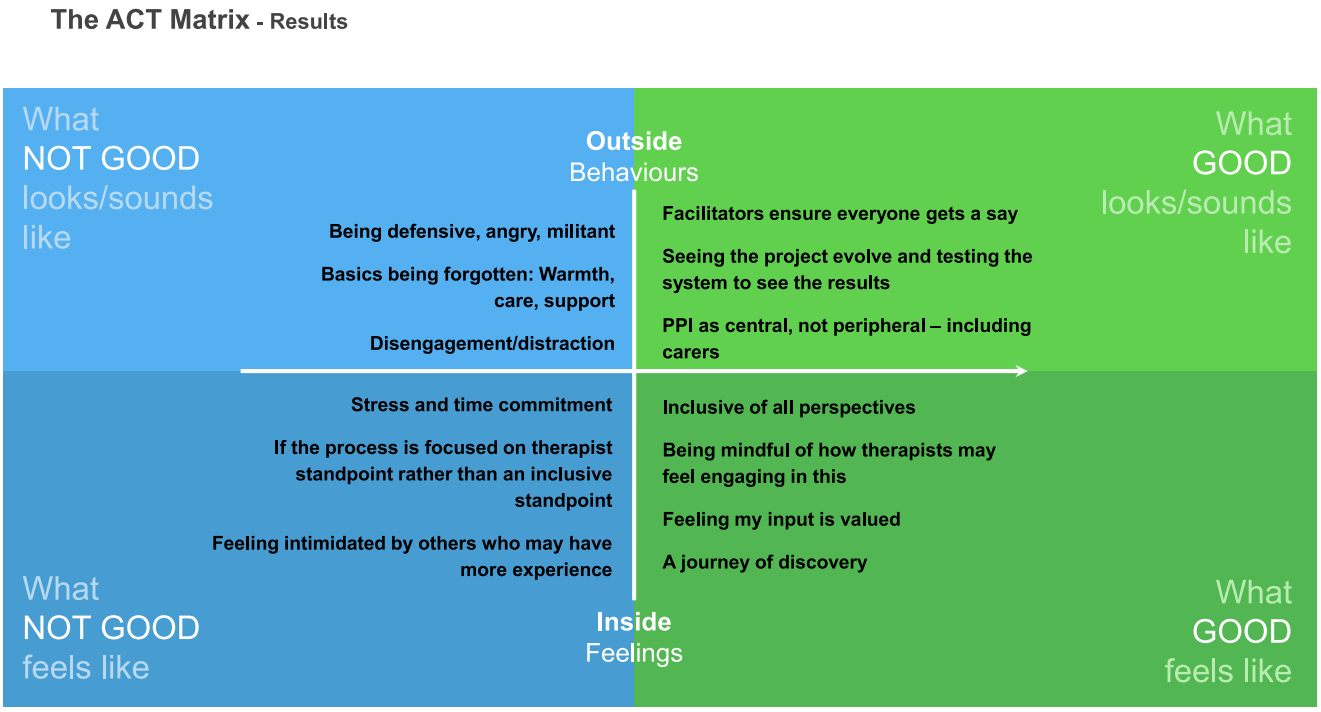

Fig. 4. ACT Matrix - Example comments in each section.

Table 1
Trustscapes: Key points by perspective

Perspective	Comments
Patient and carer	Authenticity of interactions, does [use of the system] hijack the therapy? Transparency, does the client or carer routinely see coding results, can they feedback/input on the process? Reductionist, i.e. missing nuances, leading to misinterpretation, fear of being misunderstood and fear of being assessed according to a reductionist paradigm.
Therapist	[Concerns about] ethics and consent, etc. [Will it have] validity with different models/approaches, e.g. CBT vs. psychodynamic, etc. [The research needs to] clarify purpose and limitations [of the tool]. [What is its] accuracy and applicability? [Consideration should be given to] cultural sensitivity, e.g. differences in age, ethnicity, demographics, etc.
Therapy trainer	Need to check findings [of the system] with the user. Concern that the computer will not pick up the nuances of therapy underpinning. For a model such as CBT it is very tempting to imagine you can rate the therapy process easily using an algorithm - however emotion and therapeutic relationship are just as important as technical intervention and these are idiosyncratic – matched to the patient in the moment – there is more to CBT than adherence, especially with more complex clinical presentations.
Therapy service manager	Need to see it in practice. Governance frameworks [need to be considered] in advance. [Will it have] accuracy across a range of real-life variety in therapy?

to meet. Separate sessions were successfully convened with: a) patients (2 attendees) and a carer (1 attendee), b) clinical psychologists (3 attendees) and c) therapy trainers (2 attendees).

Scores on the System Usability Scale (SUS) were entered on the Miro board by members of the clinical psychologist and psychotherapy trainer sub-groups, but not by patients and carers, who indicated they understood the system to be more for the use of therapists, and that patient ratings of usability of the system would be irrelevant in their view.

Table 2 presents median average SUS scores for each item from the clinical psychologist and therapy trainer sub-groups. We used these median scores as the basis to calculate overall scores per perspective, following the guidance from the developers of the scale (Brooke, 2013). This generated an overall score of 65/100 from the clinical psychologist

Table 2
Median scores on the SUS from the clinical psychologist and psychotherapy trainer sub-groups.

SUS scale item (each scored from 1 = Strongly disagree to 5 = Strongly agree)	Clinical psychologist median score (n=3)	Psychotherapy trainer median scores (n=2)
I think that I would like to use this system frequently.	2	1.5
I found the system unnecessarily complex.	3	3
I thought the system was easy to use.	3	2.5
I think that I would need the support of a technical person to be able to use this.	2	2.5
I found the various functions in this system were well integrated.	4	2.5
I thought there was too much inconsistency in this system.	1	2.5
I would imagine that most people would learn to use this system fairly quickly.	4	3
I found the system very cumbersome to use.	2	2.5
I felt very confident using the system.	3	2
I needed to learn a lot of things before I could get going with this system.	2	2.5
Overall I would rate the user friendliness of this product as:	Good*	[unrated]

* Question 11 is rated from 7 options, 'worst imaginable', 'awful', 'poor', 'ok', 'good', 'excellent', best imaginable'. All those who completed this item rated it 'good'.

group, and 46.25/100 for the psychotherapy trainer group, both of which are placed in the lower 50th percentile for scores on this scale (Brooke, 2013), suggesting improvements can be made.

3.2.1. Themes from Workshop 2

Fig. 5 presents sticky notes entered by PPRG group members, demonstrating how data appeared on the Miro board. Each theme is discussed in detail below.

3.2.1.1. Concerns relating to the algorithm. Concerns relating to the algorithm were raised by patients and carers, therapy trainers and therapists themselves. Patient and carer concerns included that the algorithm design did not incorporate silence, tears, upset and emotion, as it only analysed transcribed language, and that measures of the therapeutic relationship had not been considered. They were also concerned that a system based on analysis of in-therapy language would depend on patients expressing their thoughts clearly, while in their experience, this was not always possible, for example if the patient was over-tired or very depressed.

Clinical psychologists thought that knowledge of the algorithm analysing their words might make them overly self-conscious about the language they used, while therapy trainers considered the algorithm would miss a lot of meaning by analysing individual sentences alone, particularly as this may miss aspects of context. Therapy trainers also commented that sudden improvements sometimes occur, while for other patients their condition may get worse before getting better during a course of therapy, and that this may be difficult to account for using machine learning analyses of sessions.

3.2.1.2. Ethical Concerns. Ethical concerns about the tool were raised by patients and carers. These included aspects of consent, whether the patient would be made aware of how the tool worked, what would be done with the data, where it would be stored, who would have access to it and whether patients and carers would be able to see its use. Concerns were also raised about how the use of the tool would affect culture in terms of therapist time, energy and morale.

3.2.1.3. Comments on the user interface. Suggestions on the design of the interface included concern about some of the language used – the CICS coding system enables rating of patient activation in a session as 'strong' or 'weak', and patients and carers felt this language was unhelpful in the context of mental health. Therapists saw some positives of the system interface design, including that they were able to recode automated classifications that may not be correct, although the consensus was that this process would be too time consuming to use on a routine basis. Therapists found the explanation box to be redundant as the explanations were not meaningful enough for clinicians to take action. The interface design presented in Workshop 2 included a final screen which provided a prediction of whether patients would reach recovery after 10 sessions given the analysis of the present session. Therapy trainers considered this likely to be inaccurate, and to be unhelpful irrespective of accuracy, as a negative output may discourage therapists rather than motivating them to change their practice.

3.2.1.4. Concerns about applicability. Concerns about applicability of the tool from patients and carers were that the tool may not work effectively for people with autism, or for whom English was not their first language, or for people with other communication needs. Concerns were also raised that the system was based on the analysis of language, when other, non-verbal mechanisms were also at play in many therapeutic modalities. Therapy trainers pointed out that a traffic light system may not work well for colour blind users.

3.2.1.5. Use in clinical practice. Concerns were expressed regarding the practicality of the system. Expectations that a therapist would check through the coding after each session with each of their patients were derided by all groups as taking too much time to be practical, unless a large amount of resource was dedicated to its use. Therapy trainers commented that it would be useful to have just a final analysis of the session from the system without the need to check the coding

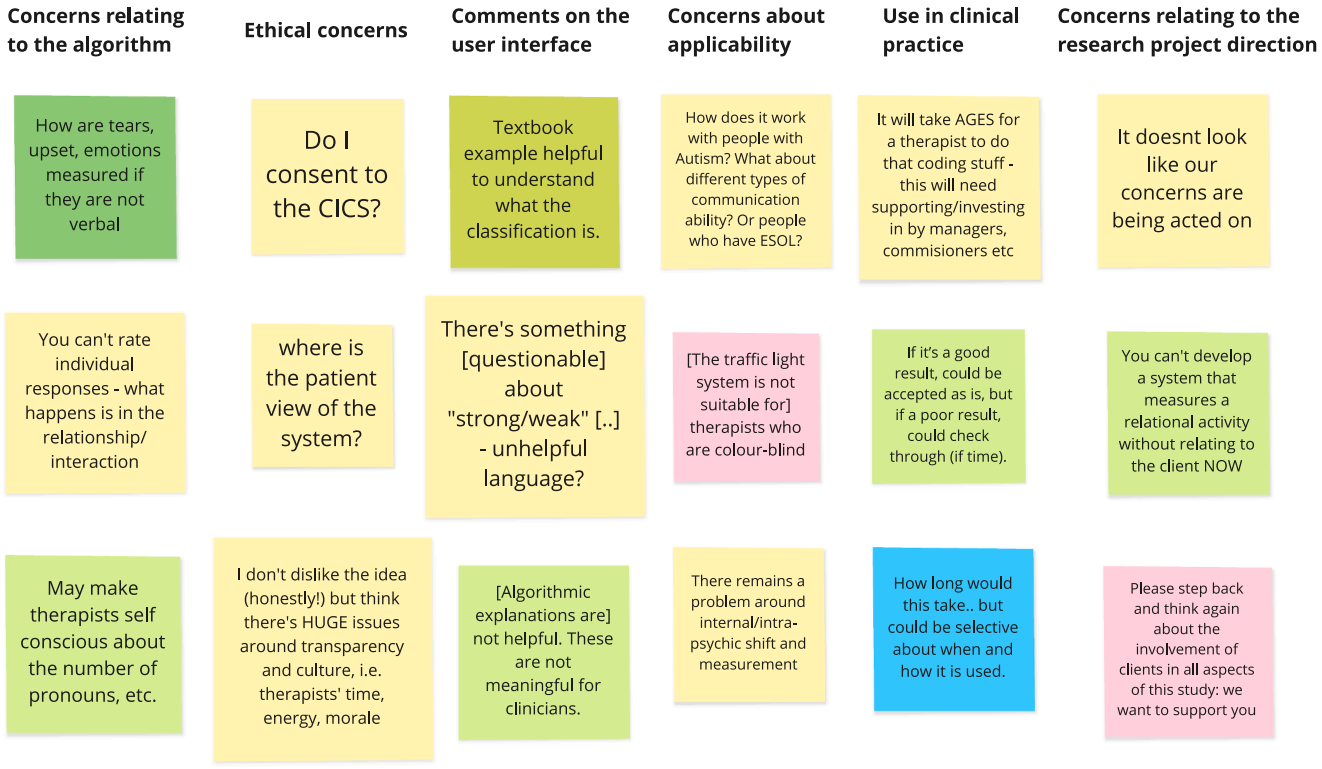

Fig. 5. Example sticky notes from Workshop 2 discussions based on the interface design illustrations.

throughout, or with the possibility to check through if the prediction was negative or highlighted something of concern. Patients mentioned that they would be concerned if the coding was not checked, however, in case of errors with the algorithm. The timing of the roll-out of the system was a concern for one patient, who commented that currently psychological therapy staff are under a lot of pressure. Therapists thought the system could be informative in clinical practice, while therapy trainers saw benefits of the system in producing a report that could be discussed in supervision sessions. Therapy trainers also thought the tool would have a strong application in research.

3.1.2.6. Concerns relating to the research project direction. A key part of PPI is inviting comments and suggestions from stakeholders on the direction of the research. In workshop 2, patients and carers raised concerns that things they had said at the first workshop had not yet been instigated in the design of the system and emphasised the value and importance of involving patients throughout the process. They suggested that greater thought needed to be put into how they should be involved in the process.

3.3. Action planning

The action plan devised by the project team in light of the findings from workshops 1 and 2 included five key changes to be highlighted and discussed with the PPRG reference group in workshop 3. These were:

- The intention for the system would now be simply as a 'feedback tool', rather than a 'coding tool', meaning that therapists would not have to check the coding of every line of every session.
- The indication of likelihood of patient 'recovery' on the feedback screen would be revised to indicate likelihood of patient 'improvement', to reduce the pressure on the therapist.
- Outputs would indicate more clearly how the therapist should improve (i.e. which interaction types they should aim for more of, and which less).
- An intention was set out to explore how larger sections of text could be analysed at once, rather than relying on line-by-line analysis, which may otherwise lack context.
- An intention was set out to consider the applicability of the tool more carefully, and design ways of increasing the applicability of the tool.

In the third workshop, the research team presented the five main proposed changes back to the reference group and invited their input on these, to see if the suggested changes needed to be modified, or if there was anything the group deemed to be of greater importance. This pragmatic, purpose-driven process aimed to build trustworthiness, credibility and accountability, because the research team reported back to the reference group the decisions that were being made, and asked for their validation of, and further input on, these decisions. Further changes to the system were also derived based on PPRG input, including improvements to the interface's accessibility, the usage approach, and proposed updates to the algorithm which could be explored in future work. These were also presented in the Miro board for workshop 3.

3.4. Findings from Workshop 3

The third workshop was attended by patients (2 attendees), a carer (1 attendee), a clinical psychologist (1 attendee), a therapy trainer (1 attendee) and a service manager (1 attendee). The workshop took the form of an active discussion between all attendees about the research team's proposed changes to the use and design of the system, described above. Textbox 1 presents a list of themes generated through the Template Analysis of the transcript, Miro sticky notes and workshop notes.

Textbox 1
Themes derived from discussion in Workshop 3.

- Exclusion of patients from the feedback process
  - Oversimplification/Reductionism
  - Patient involvement in the validation of the tool
  - Patient consent and understanding of the tool
  - Potential benefits of the system compared to existing situation in psychotherapy
  - Practical application of the system
  - Transparency

In the session, patients and carers raised concerns that use of the system excluded patients and did not permit patients to influence the feedback given to therapists. Furthermore, it was raised that the patient may have difficulty understanding how the system worked and therefore would not be able to provide true informed consent. Without understanding how the system had made its decisions, patients may be fearful of being subject to bias or mistakes in the analysis of their data. As the work was at a developmental stage, it was suggested that the work was starting in a position where the patient was excluded and that there could be harmful effects of the system for the patient, for example in increased anxiety about their speech being recorded and analysed. Discussing readings from the tool with the patient at a subsequent therapy session was suggested as means to help mitigate this anxiety and increase involvement of the patient in the process. The patient and carer group also suggested that patient and carer representatives should be involved in the validation of the system.

All group members reiterated that they considered the system to be reductionist and may not be accurate. However, therapists added that current methods of feedback to therapists are also problematic, for example patients feel interpersonal pressure to state their therapy was good/useful even where this is not the case. In relation to the practical application of the system, all group members agreed that the system should only be one of a number of approaches that clinical psychologists use to review their own practice, and that they should not make decisions about patient care exclusively based on the suggestions provided by the system.

The topic of transparency was also discussed in the third workshop. When considering examples of how Google and Amazon attempted to make their advertising systems transparent to customers, PPRG reference group members stated that these were not comparable to a system for use in healthcare, since inaccurate suggestions in a commercial environment would have limited repercussions, while decisions relating to mental health care may have more serious consequences.

3.5. Post hoc debrief meeting

After the final workshop, an additional meeting was held in which patients and carers from the PPRG reference group voiced concerns about the research process and provided constructive criticism of the organisation of the project. This related to: patient and carer expectations of their role being greater than it turned out to be; a perceived disconnect between research team PPI advisors and PPRG members; and the research team's failure to provide details of why certain suggestions made by the PPRG were not included in the action plan presented in workshop 3. In the discussion, it was agreed that having a clinical psychologist from the research team present at stakeholder events would have been useful to address concerns – the research team's hope of allowing openness and transparency by their absence was overridden by participants' need for explanations relating the use of the tool in clinical practice. Actions arising from this meeting included that further documentation of how the patients' and carer's views had been reflected on and why their suggestions had not been presented for discussion as a main topic in the third workshop would be circulated after the meeting. The debrief session was not used to generate input on the AutoCICS design process itself.

4. Discussion

4.1. Summary of findings

The ExTRAPPOLATE project has taken an RRI approach to involve key stakeholders from five perspectives to elicit concerns, suggestions and feedback about an automated system to inform clinical psychology practice at an early stage in the design process. Overall, this has provided essential information for the AutoCICS system development that could not have been identified through other methods. Concerns expressed by stakeholders ranged from practical usage issues and interface design factors to consideration of whether the algorithm in development would pick up on all relevant details of the patient and therapist's discussion, and how the excess time burden required by the system could be reduced. The application of an RRI framework has ensured that the research *anticipated* future consequences of the introduction of the system, *reflected* on stakeholders' critique of the system, *inclusively engaged* to discuss matters openly and honestly, and enabled the research team to *act in response* to stakeholder views.

Incorporating multiple stakeholder groups enabled representation of multiple viewpoints in the findings. The patient and carer group were keen to emphasise that patients should be involved in the feedback process and to highlight the ethical issues around informed consent. The patient and carer group also emphasised the importance of considering the applicability of the algorithms to different patient subgroups. Both therapists and patients raised the need to address the reductionist nature of the approach, for example by incorporating measures of emotion, and providing more detailed, quantified feedback to enable action. Therapy trainers largely focused on practical concerns relating to using the system and suggested ways its use could be made more efficient, as well as identifying specific interface improvements to be implemented. Service managers were most concerned about ensuring governance requirements were met, and that the system would have applicability across different therapy types. We thus demonstrate how concerns about AI and machine learning vary across different stakeholder types in the field of clinical psychology, and respond to calls for the involvement of psychologists, patients and organisational leaders in research exploring the application of such tools (Carr, 2020; Kyratsis et al., 2012; Tahan & Zygoulis, 2019).

While many of the PPRG's comments were specific to the case of the AutoCICS system development, certain themes emerged from the analysis which were more general and add to prior work exploring patient and healthcare professional concerns about the use of algorithms and artificially intelligent systems (Richardson et al., 2021). Many of the issues raised relate to the field of Human Data Interaction, which has a focus on agency, legibility and negotiability (Victorelli et al., 2020): patients and carers emphasised that the initial iteration of the system did not permit the patient to engage with the results of the system, or to reflect together with the therapist on the guidance it provided, effectively limiting their agency and negotiability in matters pertaining to their own data. Durán and Jongsma (2021) similarly indicate that AI-based decision-making can reduce patient autonomy, which may be problematic where decisions informed by AI-based systems do not align with patient values and preferences. Additionally, AI hesitancy has been highlighted as a barrier to patient trust in novel applications of technology in healthcare in previous work (Nadarzynski et al., 2019). Researchers have highlighted how lack of trust often relates to a lack of legibility, that is, little transparency in the workings of algorithm-based systems. Although the design of the AutoCICS system included 'explanations' (Fig. 3a), clinical psychologists considered these to be of limited use as they were not actionable. Wider research recognises pitfalls to explainability, which may not result in true legibility for users (Ehsan & Riedl, 2021). The PPRG also highlighted the dangers of algorithms introducing bias into decision-making where these have not been trained on representative populations, limiting applicability (Musbahi et al., 2021). Our findings provide clear ways in which the AutoCICS system and its proposed usage could be improved, which will inform the next steps of the algorithmic development as well as informing protocols for its use in research trials and later implementation.

4.2. Best practices

The development methods used in our workshops drew from the RRI framework by Stilgoe and colleagues (2013). Their framework highlights the "power of technology to produce both benefit and harm" and emphasises the importance of researchers taking responsibility for "possible hazards their research might unleash" (2013). The ***anticipate*** dimension of the AREA framework seeks to address the possibility of undesirable outcomes from new technologies directly. The Trustscapes tool was effective in facilitating the task of anticipating future potential problems with the system. Stakeholders were able to provide a wide range of worst-case scenarios and stated clearly how the system (as proposed) could lead to increased patient anxiety, increased therapist burden, and inaccurate predictions if algorithms were trained on a homogenous patient cohort.

In relation to ***reflexivity***, Stilgoe and colleagues comment that "Responsibility demands reflexivity on the part of actors and institutions [..] holding a mirror up to one's own activities, commitments and assumptions, being aware of the limits of knowledge and being mindful that a particular framing of an issue may not be universally held" (Stilgoe et al., 2013). ExTRAPPOLATE employed a two-step process to ensure adequate reflection on the issues at hand. Through use of an action plan, reviewed with the PPRG in Workshop 3, the research team were able to look past their own potentially biased views of the future potential of the system and construct a development plan informed by the real-world experiences of patients, carers, and psychological therapy professionals at different levels. This process was further enhanced by the involvement of representatives from all stakeholder groups in the drafting of this paper (reading and providing commentary on drafts), by which understandings could be verified and learnings shared.

By its nature, the ExTRAPPOLATE project was an inclusive, interdisciplinary venture which included among the research team membership two PPI volunteers and one service user researcher. In addition, the workshops enabled the research team to ***engage*** with key stakeholders from outside the core team, enhancing the development of the system. Von Schomberg (2012) states that RRI is "A transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products". The ExTRAPPOLATE project iterated through three workshops, developing and updating an action plan to improve the acceptability and usability of the AutoCICS system.

The system as presented was highly criticised by patient and carer stakeholders for not being patient-centred, and for introducing uses of their utterances which were perhaps not wholly understandable to lay patients and carers, emphasising the need to make changes. The use of a convenience sampling method meant that patient and carer members of the PPRG reference group were known to, and familiar with, the research team, meaning they were comfortable to voice their concerns, potentially resulting in feedback which was more honest and more direct than may have been the case with an unfamiliar group. This further enabled the research team to successfully *engage* with relevant stakeholders and ensure the system design would meet their needs.

Van Ouheusden (2011) reminds us that participation mechanisms used in responsible research initiatives also need to be openly reflected upon, with stakeholders able to question these. In Workshop 1, the PPRG and the research team together completed the ACT matrix (Figs. 2 and 4) to set out positive intentions for working together. In addition, a debrief meeting subsequent to the main workshop series was held at the request of the patient and carer subgroup. Responding to the needs of those involved in the research project in these ways brought the project in line with the final dimension of the EPSRC's AREA framework, ***act***. RRI

requires researchers and innovators "to change shape or direction in response to stakeholder and public values and changing circumstances" (Stilgoe et al., 2013), and the ExTRAPPOLATE project demonstrated how this could be done, even where views are held differently between differing stakeholder groups.

In relation to the tools used to facilitate this work, the use of an online whiteboard software package [Miro, (Khusid, 2011)] was successful in permitting written as well as oral contributions of ideas. Some feedback from the PPRG reference group suggested that a discussion without the use of the Miro board would have been more personable and therefore preferable. The research team found that using the Miro boards facilitated the presentation of materials which would otherwise have been more difficult. While an in-person event may have allowed greater interaction between the research team and the PPRG, the use of the Miro software was a viable alternative during a period of national restrictions on face-to-face contact.

4.3. Shortcomings and learnings relating to the RRI approach

There were many learnings from the workshops, including some relating to national restrictions on face-to-face meeting. Firstly, the PPRG were all recruited from a single county, meaning that the assembled group were not necessarily representative of psychological therapy patient and therapist populations across the United Kingdom. The use of more extensive outreach initiatives to publicise the project and advertise an opportunity to contribute could have enabled greater representativeness within the group. However, the use of online methods to facilitate the workshops (as a workaround for restrictions on personal contact at the time) had benefits, as it enabled stakeholders who may not have been able to attend in-person sessions due to other responsibilities or time commitments, to still be involved.

As a result of the reorganisation of clinical work during the COVID-19 pandemic in 2020 and 2021, multiple delays occurred during the project, which meant we were not able to conduct certain activities that had been envisaged. This included demonstrating a functioning prototype of the AutoCICS system to workshop attendees. As a replacement, we developed illustrations showing how the system's user interface could look, to enable attendees to reflect on the proposed functioning of the system. The input we were able to collect from participants in the project demonstrates that the visualisations were sufficient for the PPRG reference group to gain an understanding of the purpose and aims of the software and be able to express their reactions and concerns about its potential use. This input has been extremely important in producing plans for the next stages of the research and development of the AutoCICS system.

The involvement process would have benefitted from the use of a quantitative method to rank the importance of different action points discussed by the PPRG reference group. As it was, not all ideas from the second workshop were presented for discussion in the third workshop, and some members of the group were frustrated that their ideas had not been highlighted for discussion. Had these been ranked by importance to the PPRG during the second workshop, with further detail provided by the research team on the practical limitations of the project scope, there would have been a strong argument for taking forward particular ideas, which may have been more satisfying for all parties involved. Rankings and prioritisation exercises have been conducted in other projects in related areas (Hollis et al., 2018; Musbahi et al., 2021), offering models for the adaptation of our approach in future work.

The rating of the system using the System Usability Scale (SUS) was not completed by the patients or carers. They considered it irrelevant to them as they were not the intended users of the interface. The research team had envisaged that collecting patient and carer views using the standardised SUS scale would provide insights into the overall complexity of the system, however the fact that this was not completed limited the knowledge that could be generated in this area. The overall development plan for the AutoCICS system includes further usability testing at later stages, when more knowledge can be gained to account for this limitation.

The patient and carer members of the PPRG reference group suggested that the research team should have been clearer about their role from the outset. Prior to the first workshop, potential members were invited to the group by email and were provided with a verbal introduction to the project and their role at a one-to-one induction session. The learning taken from this was that the PPRG (and the research team) would have benefitted from a more comprehensive *written* document, setting out the terms of reference for the group, including a detailed role profile. The UK Standards for Public Involvement (National Institute for Health and Care Research 2022) highlights the importance of having a shared understanding of roles, responsibilities and expectations in public involvement. Similarly, Pollard et al (2015) developed guidelines for PPI in research, including the suggestion that 'person specifications' may be used to enhance clarity around role expectations. Learnings from this project have illustrated the difficulties that may be experienced in the absence of such shared understandings.

The facilitated debrief session with patients and carers was organised *post hoc*, on request of PPRG members. Prior scheduling of such a feedback session as part of the process (rather than something organised afterward) may have helped ensure that workshop attendees felt confident during the workshops, knowing they would have an opportunity to feed back on the process afterward. This was also perhaps more important given the lack of face-to-face contact, and the absence of informal, unplanned meetings between researchers and reference group members that may have taken place if the workshops had been in-person. Leese et al (2022) report similar experiences of disconnect between researchers and PPI volunteers when using purely online communication methods, and the need to organise extra online meetings to facilitate relationship building during times of restrictions on face-to-face contact. These reflections will remain valid for future endeavours to include more hard-to-reach groups via online methods.

5. Conclusions

This project was successful in its aim to involve a wide-ranging stakeholder group, the PPRG, in a responsible research and innovation approach to the early design of the AutoCICS system. This approach enabled the research team to gain important insights from patients, carers, therapists, service managers, and therapy trainers on the promises and risks of the system early in its development. This enabled corrective action to be taken in relation to the initial design and planned use of the system, before it was fully developed for feasibility testing. The next steps will involve further development and testing of the system for both feasibility and effectiveness, which will be essential to understand whether the proposed benefits of using the system in clinical practice are valid and convincing.

This paper has presented an approach for engaging a range of key stakeholders, eliciting their views and validating learning taken from their input. By involving practitioners with different roles as well as patients and carers in the research, including in single-perspective and cross-perspective sessions, we were able to transform our understanding of how the AutoCICS tool could usefully function, and to refine plans for further research and development of the system for the benefit of all stakeholders. Reflections on the process raised: the importance of clear role expectations, methods for addressing concerns, considerations for online-only PPI, and the importance of post-involvement debriefing. Future research using RRI initiatives with multi-stakeholder groups could benefit from attending to these points, perhaps by involving experienced PPI coordinators who would be able to provide relevant training and guidance based on their experience.

CRediT authorship contribution statement

Jacob A Andrews: Conceptualization, Methodology, Investigation,

Formal analysis, Resources, Data curation, Visualization, Supervision, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing. **Mat Rawsthorne:** Conceptualization, Investigation, Resources, Writing – review & editing, Funding acquisition. **Cosmin Manolescu:** Methodology, Investigation, Resources, Writing – original draft, Writing – review & editing. **Matthew Burton McFaul:** Visualization, Investigation. **Blandine French:** Investigation, Writing – review & editing. **Elizabeth Rye:** Writing – review & editing. **Rebecca McNaughton:** Writing – review & editing. **Michael Baliousis:** Writing – review & editing. **Sharron Smith:** Writing – review & editing. **Sanchia Biswas:** Writing – review & editing. **Erin Baker:** Writing – review & editing. **Dean Repper:** Writing – review & editing. **Yunfei Long:** Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing. **Tahseen Jilani:** Conceptualization. **Jeremie Clos:** Conceptualization. **Fred Higton:** Supervision. **Nima Moghaddam:** Conceptualization. **Sam Malins:** Conceptualization, Resources, Funding acquisition, Formal analysis, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is funded by an Engineering and Physical Sciences Research Council's (EPSRC) Human Data Interaction (HDI) Network Plus, Grant Award EP/R045178/1, project reference 301671, and with support from Health Data Research UK. This work has been supported by the National Institute for Health Research (NIHR Development and Skills Enhancement Award, Dr Sam Malins, NIHR300822). This project also received support from the NIHR MindTech MedTech Co-operative and the NIHR Nottingham Biomedical Research Centre. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care. Funders had no role in the study design; in the collection, analysis or interpretation of data; or in the writing of the report; or in the decision to submit the article for publication.

Appendix A: Summary of Online Workshop Activities

Workshop/ session	Activity	Materials/Tools	AREA framework dimension(s)
Individual on-boarding session	Connect with group member via video call. Introduction to the project. Introduction to the Miro online collaboration platform, helping each group member to try navigating around boards and creating sticky notes.	One 'board' in the Miro online collaboration software (Miro.com). Microsoft Teams video calling.	N/A
Workshop 1	**Trustscapes** – Four question headers were presented on Miro, relating to trust in automated tools. Adapted from: https://unbias.wp.horizon.ac.uk/fairness-toolkit/#:~:text=3.-,TrustScapes Using Trustscapes involved discussion and typed entry of sticky notes on the Miro board under question headers. Conducted in separate perspective groups first (ie. Patients and carers, clinical psychologists, service managers, trainers), then facilitators of each group fed back key points to the whole group on a separate, whole group Miro board. **ACT Matrix activity** – together as a whole group determining behaviours and thinking that would support collaborative working (see https://contextualscience.org/act_matrix).	Four separate Miro boards, one per perspective. One further board for the whole group. Trustscapes toolkit (see link left), copied onto each board. ACT Matrix presented on the whole group board. Zoom/MS Teams to facilitate recording of discussion.	Anticipate, Engage
Workshop 2	**ACT Matrix** – reminder of intentions expressed. **System interface** - presentation of visuals showing planned interface. Group discussion of different aspects of importance, for example: functionality, appearance, use. **System Usability Scale** (SUS) – used to rate the system in its initial iteration (see https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html). **Alternatives -** Presentation of alternative interfaces for comparison and discussion.	Miro board presenting the interface design, SUS scale, and small coloured dots for voting on the SUS. Zoom/MS Teams recommended to facilitate recording of discussion.	Engage
Workshop 1+2 analysis, preparation for workshop 3	Produce an action plan for updating the interface and the planned use of the system based on group input in prior workshops. Produce lay-friendly summary for workshop 3.	Prepare materials on a Miro board for easy access and to prompt discussion.	Act
Workshop 3	**Action plan** – present the proposed action plan to the group. Show any revised visuals based on earlier feedback. Listen to feedback on the plan and revised visuals. **Transparency and trust** – Discussion around transparency, explainability and trust in the tool. Comparison with other online tools making use of algorithms. **Final poll** – Prepare a separate online form with any final questions for the group. This is useful when it is important to get all stakeholder's views on an issue but where it may not be practical to go round each member in turn within the workshop. (Not in fact used due to lack of time).	Miro board showing action plan and revised visuals. Zoom/MS Teams recommended to facilitate recording of discussion MS Forms/Google Forms recommended for final poll.	Engage, Act
Debrief session	Discussion of what went well in the process and what could be improved. Particularly of use for patient/carer groups.	Zoom/MS Teams.	Engage

Appendix B: Table of attendees at each workshop

	Workshop 1	Workshop 2(a)	Workshop 2(b)	Workshop 2(c)	Workshop 3	Debrief session
Patients	2	2	-	-	2	2
Carers	1	1	-	-	1	1
Clinical psychologists	2	-	3	-	1	-
Therapy trainers	2	-	-	2	1	-
Service managers	2	-	-	-	1	-
Research team	3	2	1	1	1	2

References

Atkins, P. W. B., Wilson, D. S., & Hayes, S. C. (2019). *Prosocial: using evolutionary science to build productive, equitable, and collaborative groups*. Oakland, CA: New Harbinger Publications.

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 258–297). Wiley.

Barkham, M., Lutz, W., Lambert, M. J., & Saxon, D. (2017). Therapist effects, effective therapists, and the law of variability. In L. Castonguay, & C. Hill (Eds.), *How and why are some therapists better than others? Understanding therapist effects* (pp. 13–36). American Psychological Association.

Boivin, A., Richards, T., Forsythe, L., Gregoire, A., L'Esperance, A., Abelson, J., & Carman, K. L. (2018). Evaluating patient and public involvement in research. *BMJ, 363*. https://doi.org/10.1136/bmj.k5147. k5147.

Branson, A., Shafran, R., & Myles, P. (2015). Investigating the relationship between competence and patient outcome with CBT. *Behav Res Ther, 68*, 19–26. https://doi.org/10.1016/j.brat.2015.03.002

Brooke, J. (2013). SUS: a retrospective. *Journal of Usability Studies, 8*, 29–40.

Brooke, John (1996). SUS - A quick and dirty usability scale. In Patrick Jordan, et al. (Eds.), *Usability Evaluation In Industry* (1st). London: Taylor & Francis.

Burget, M, Bardone, E, & Pedaste, M (2017). Definitions and conceptual dimensions of responsible research and innovation: A literature review. *Science and engineering ethics, 23*(1), 1–19. https://doi.org/10.1007/s11948-016-9782-1

Carr, S. (2020). 'AI gone mental': engagement and ethics in data-driven technology for mental health. *Journal of Mental Health, 29*(2), 125–130. https://doi.org/10.1080/09638237.2020.1714011

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*(5), 329. https://doi.org/10.1136/medethics-2020-106820

Ehsan, U., & Riedl, M. O. (2021). Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. *arXiv preprint arXiv:2109.12480*.

Engineering and Physical Sciences Research Council. (2021). *Framework for Responsible Innovation*. Retrieved 17/09/2021 from https://epsrc.ukri.org/research/framework/.

Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2020). Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. *JAMA Psychiatry, 77*(1), 35–43. https://doi.org/10.1001/jamapsychiatry.2019.2664

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., & Wampold, B. E. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology, 63*(1), 1.

Hibbard, J., Stockard, J., Mahoney, E. R., & Tusler, M. (2004). Development of the Patient Activation Measure (PAM): Conceptualizing and Measuring Activation in Patients and Consumers. *Health Services Research, 39*(4 Pt 1), 1005–1026. https://doi.org/10.1111/j.1475-6773.2004.00269.x

Hollis, C., Sampson, S., Simons, L., Davies, E. B., Churchill, R., Betton, V., Butler, D., Chapman, K., Easton, K., Gronlund, T. A., Kabir, T., Rawsthorne, M., Rye, E., & Tomlin, A. (2018). Identifying research priorities for digital technology in mental health care: results of the James Lind Alliance Priority Setting Partnership. *Lancet Psychiatry, 5*(10), 845–854. https://doi.org/10.1016/S2215-0366(18)30296-7

Ives, J., Damery, S., & Redwod, S. (2013). PPI, paradoxes and Plato: who's sailing the ship? *Journal of medical ethics, 39*(3), 181–185.

Jackson, T., Pinnock, H., Liew, S. M., Horne, E., Ehrlich, E., Fulton, O., Worth, A., Sheikh, A., & De Simoni, A. (2020). Patient and public involvement in research: from tokenistic box ticking to valued team members. *BMC Med, 18*(1), 79. https://doi.org/10.1186/s12916-020-01544-7

Khusid, A. (2011). *Miro*. In (Version 0.7.0.0) https://miro.com/.

Kühl, N., Goutier, M., Hirt, R., & Satzger, G. (2020). Machine learning in artificial intelligence: Towards a common understanding. *arXiv preprint arXiv:2004.04686*.

King, N. (2012). Doing template analysis. In G. Symon, & C. Cassell (Eds.), *Qualitative organizational research: Core methods and current challenges*. London: Sage.

Kyratsis, Y., Ahmad, R., & Holmes, A. (2012). Technology adoption and implementation in organisations: comparative case studies of 12 English NHS Trusts. *BMJ Open, 2*(2), Article e000872. https://doi.org/10.1136/bmjopen-2012-000872

Lane, G., Angus, A., & Murdoch, A. (2018). UnBias Fairness Toolkit (Version 1). *Zenodo*. https://doi.org/10.5281/zenodo.2667808. Retrieved 28/02/2020 from.

Leese, J., Garraway, L., Li, L., Oelke, N., & MacLeod, M. (2022). Adapting patient and public involvement in patient-oriented methods research: Reflections in a Canadian setting during COVID-19. *Health Expectations, 25*(2), 477–481. https://doi.org/10.1111/hex.13387

Malins, S., Moghaddam, N., Morriss, R., Schroder, T., Brown, P., & Boycott, N. (2021a). Predicting outcomes and sudden gains from initial in-session interactions during remote cognitive-behavioural therapy for severe health anxiety. *Clin Psychol Psychother, 28*(4), 891–906. https://doi.org/10.1002/cpp.2543

Malins, S., Moghaddam, N., Morriss, R., Schroder, T., Brown, P., & Boycott, N. (2021b). The predictive value of patient, therapist, and in-session ratings of motivational factors early in remote cognitive behavioural therapy for severe health anxiety. *Br J Clin Psychol*. https://doi.org/10.1111/bjc.12328

Malins, S., Moghaddam, N., Morriss, R., Schroder, T., Brown, P., Boycott, N., & Atha, C. (2020). Patient activation in psychotherapy interactions: Developing and validating the consultation interactions coding scheme. *J Clin Psychol, 76*(4), 646–658. https://doi.org/10.1002/jclp.22910

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform, 113*, Article 103655. https://doi.org/10.1016/j.jbi.2020.103655

Mockford, C., Staniszewska, S., Griffiths, F., & Herron-Marx, S. (2012). The impact of patient and public involvement on UK NHS health care: a systematic review. *Int J Qual Health Care, 24*(1), 28–38. https://doi.org/10.1093/intqhc/mzr066

Musbahi, O., Syed, L., Le Feuvre, P., Cobb, J., & Jones, G. (2021). Public patient views of artificial intelligence in healthcare: A nominal group technique study. *DIGITAL HEALTH, 7*. https://doi.org/10.1177/20552076211063682, 20552076211063682.

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digit Health, 5*. https://doi.org/10.1177/2055207619871808, 2055207619871808.

National Institute for Health and Care Research, Chief Scientist Office, Health and Care Research Wales, & HSC Public Health Agency. (no date). *UK Standards for Public Involvement: Working together*. Retrieved 27/05/ 2022 from https://sites.google.com/nihr.ac.uk/pi-standards/standards/working-together.

Panch, T., Mattie, H., & Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. *NPJ Digit Med, 2*, 77. https://doi.org/10.1038/s41746-019-0155-4

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of global health, 8*(2).

Perepletchikova, F. (2011). On the topic of treatment integrity. *Clinical Psychology: Science and Practice, 18*(2), 148–153.

Polk, K. L., Schoendorff, B., Webster, M., & Olaz, F. O. (2016). The Essential Guide to the ACT Matrix: A Step-by-Step Approach to Using the ACT Matrix Model in Clinical Practice. *New Harbinger Publications*. https://books.google.co.uk/books?id=QHotDAAAQBAJ.

Pollard, K., Donskoy, A.-L., Moule, P., Donald, C., Lima, M., & Rice, C. (2015). Developing and evaluating guidelines for patient and public involvement (PPI) in research. *International Journal of Health Care Quality Assurance, 28*(2), 141–155. https://doi.org/10.1108/IJHCQA-01-2014-0001

Rawsthorne, M., Jilani, T., Andrews, J., Long, Y., Clos, J., Malins, S., & Hunt, D. (2020). *ExTRA: Explainable Therapy-Related Annotations. 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*. nov. Dublin, Ireland.

Richardson, J. P., Smith, C., Curtis, S., Watson, S., Zhu, X., Barry, B., & Sharp, R. R. (2021). Patient apprehensions about the use of artificial intelligence in healthcare. *npj Digital Medicine, 4*(1), 140. https://doi.org/10.1038/s41746-021-00509-1

Rivard, L., & Lehoux, P. (2020). When desirability and feasibility go hand in hand: innovators' perspectives on what is and is not responsible innovation in health. *Journal of Responsible Innovation, 7*(1), 76–95. https://doi.org/10.1080/23299460.2019.1622952

Rovatsos, M., Mittelstadt, B., & Koene, A. (2019). *Landscape Summary: Bias In Algorithmic Decision-Making: what is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?* .

Staley, K. (2013). There is no paradox with PPI in research. *Journal of medical ethics, 39*(3), 186–187.

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

Tahan, M., & Zygoulis, F. (2019). Artificial Intelligence and Clinical Psychology - Current Trends. *Journal of Clinical & Developmental Psychology*. https://ssrn.com/abstract=3433670.

van Oudheusden, M. (2011). Questioning 'participation': a critical appraisal of its conceptualization in a Flemish participatory technology assessment. *Sci Eng Ethics, 17*(4), 673–690. https://doi.org/10.1007/s11948-011-9313-z

Victorelli, E. Z., Dos Reis, J. C., Hornung, H., & Prado, A. B (2020). Understanding human-data interaction: Literature review and recommendations for design. *International journal of human-computer studies, 134*, 13–32. https://doi.org/10.1016/j.ijhcs.2019.09.004

Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. *Commun Med, 1*, 25. https://doi.org/10.1038/s43856-021-00028-w

von Schomberg, R (2012). Prospects for Technology Assessment in a framework of responsible research and innovation. In M. Dusseldorp, & R. Beecroft (Eds.), *Technikfolgen abschätzen lehren: Bildungspotenziale transdisziplinärer Methode* (pp. 39–61). Springer VS.

Waller, G., & Turner, H. (2016). Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour Research and Therapy, 77*, 129–137. https://doi.org/10.1016/j.brat.2015.12.005