Koa Health Apps Ethics Audit

Ethics Audit of Foundations and Mindset

April 2021

Foreword 2
1. Introduction 3
2. Audit of Foundations 4
2.1 How the app and its model work 4
2.1.1 Recommender model 4
2.1.2 Heart rate and breathing rate model 6
2.1.3 Data points 7
2.1.4 Dashboard 7
2.2. Mental health apps and ethics 8
2.2.1 Summary of ethical implications 9
2.3. Social context for the model 9
2.3.1 Summary of societal analysis 11
2.4. Ethical assessment 12
2.4.1 Compliance of Foundations with Koa ethics commitments 12
2.4.2 Analysis of societal factors and ethical issues 14
2.4.3 Ethics in product design 16
2.4.4 Summary of ethical issues and recommendations 22
2.5. Algorithmic impact assessment 25
2.5.1 Algorithmic bias 25
2.5.2 Recommender model 27
2.5.3 Heart and breath rates model 30
2.6. Data Management assessment 32
2.6.1 Data governance, lifecycle and risk management 32
2.6.2 Privacy Policy assessment 34
2.6.3 Security of personal information 38
2.6.4 Summary of data management issues and recommendations 39
2.7. Desirability and Acceptability assessment 41
2.7.1 Desirability and acceptability grounds and issues 41
2.7.2 Usability and concept testing 43
2.7.3 Summary of desirability and acceptability issues and recommendations 44
2.8. Summary of conclusions and recommendations for Foundations 45
3. Audit of Mindset 50
3.1. Theoretical framework and social context for the model 51
3.1.1 The Cognitive Behaviour model 51
3.2. How Mindset works 52
3.2.1 The system overview 52
3.2.2 Mindset validation and testing 54
3.3. Ethics assessment 55
3.3.1 Koa 10 commitments in Mindset 55
3.3.2 Mindset ethics analysis 57
3.3.3 Summary of ethics recommendations 60
3.4. Data management assessment 61
3.4.1 Privacy Policy readability 63
3.4.2 Variables for future data management assessments 65
3.5. Desirability and acceptability assessment 65
3.5.1 Desirability analysis 65
3.5.2 Acceptability analysis 67
3.5.3 Summary of desirability and acceptability recommendations 68
3.6. Summary of conclusions and recommendations for Mindset 68
4. Overall Conclusions 71
5. References 72
5.1 Foundations Audit 72
5.2 Mindset Audit 77

Foreword

This is the second audit of Koa Health's ethics strategy, albeit the first as Koa, since we span out of Telefónica Alpha in October 2020. As with the first audit, this one has been carried out by Eticas Research and Consulting.

Our first audit focused on a prototype wellbeing product called REMIX! This audit examines the successor to that prototype, Foundations, our commercial product to tackle stress and build resilience. The audit also assesses Mindset, Koa's app to support symptom management in patients with depression. Mindset was being readied for its commercial launch during this audit, and as such its evaluation is not as extensive as that of Foundations.

In terms of how we have progressed since our first audit, I am pleased to see that we've maintained strong performances with respect to data management and security, supporting the health and wellbeing of our users, and avoiding discrimination and bias. On the latter point, the processes that we have established for our design and content development, appear to be paying dividends as we grow Foundations, although Eticas quite rightly note some areas for improvement.

Whilst we have made further efforts on understandability over the last year, we also raised the bar for ourselves, in particular by introducing a target reading age of 11 for all of our content, including our terms and conditions. I am pleased that we've managed to get to reading ages of 12-14 for Foundations content, and 16 for our terms and conditions, and we'll continue to push to meet our goal, alongside working to add more details into the terms and conditions and to improve understandability for vulnerable groups.

A big focus of our efforts during 2021 is on further embedding ethics into our R&D process. We will be introducing external ethics reviews for Koa-led research projects - we already have these when we undertake research with our university and hospital partners. In addition, we shall work with Eticas to ensure that we consider trade-offs between our ethical principles much earlier in the development of research projects and product features.

I would like to thank the Eticas team for their diligence and professionalism. What follows is the views of the auditors, but on behalf of Koa I wholeheartedly welcome this audit and its recommendations.

Ollie Smith
Strategy Director and Head of Ethics

Koa Health

1. Introduction

This Ethics and Algorithmic Audit produced by Eticas Research and Consulting delivers an assessment of desirability, acceptability, algorithmic fairness, and data management for two of **Koa Health's** products: **Foundations and Mindset**. The products are also reviewed in terms of ethics, taking into account the company's 10 ethical commitments in order to produce recommendations for improvement.

The audit began during December 2020, with data collection and interviews undertaken from December 2020 through March 2021. As Mindset will be commercially available from April 2021, certain aspects of the product were not finalized, which limited some aspects of the audit. In these cases, Eticas has focused on providing recommendations that will aid the ethical compliance of the commercially available version.

2. Audit of Foundations

The audit reviews the application **Foundations in its version 3.2.0**. At the time of this audit, Foundations is defined as a mental wellbeing app designed to help organisations support their teams, enabling people to take care of their mental wellbeing on their own terms, in a cost-effective way. Employers will offer foundations as an **Employee Assistance Program (EAP)** to help employees reduce stress.

The app offers **152 different activities and programmes** (as of 1, March , 2021) to help users manage their stress and other mental problems. Employees interact with the app, its programmes and activities with the aim to build resilience. Some employers have access to a summary report or a beta dashboard to see general information about Foundations' usage, track "stressful" moments or situations and facilitate employees' resilience. The app is available on iOS and Android.

The assessment is based on four primary data collection techniques: a) a literature review concerning the use of technologies in the domain of Foundations and its ethical implications; b) the review of Koa documents describing technical specifications and self-assessing ethics compliance of the system; c) Interviews with the Koa team[1] and d) a thorough evaluation of the system functioning and data management, including a digital ethnography of its recommendations. The analysis has been oriented towards examining ethical compliance, complementing Koa self-assessment and providing guidelines for the system's ethical design, management, and testing.

2.1 How the app and its model work

The Foundations app has one embedded model and one planned model at the time of this audit. The embedded model that is already available **in production** is a **recommendation system** for the user to discover new activities**.** The planned model that has been developed but **not yet deployed** is a **heart rate and breathing rate measurement system**. Whilst the app is intended to be used by employees, an employee dashboard also exists. In this section we will summarise how the models work and describe the data points the app collects and the information the dashboard gives out.

2.1.1 Recommender model

Foundations has a list of **152 activities classified in 12 categories** at the moment of this audit. The activities are related to 6 different areas of focus. The user must choose at least one area of focus when the app is installed, and the program and its activities that are shown in the different widgets are related to this area of focus, based on rules.

Table 1. List of activities and programmes available in the app

[1] **Roles interviewed** were: Strategic Director and Head of Ethics, Project Manager, responsible Data Scientist for each model, Content and Psychology Lead, Service Design Strategist, Director of Cyber Security (Dec'20-Jan'21).

Categories of activities (145) and programmes	Areas[2] of focus:
1. Relaxation (2 programmes, 15 activities) 2. Working with thoughts (2 programmes, 10 activities) 3. Positive thinking (1 programme, 6 activities) 4. Unwind now (11 activities) 5. Boost your self-confidence (2 programmes, 9 activities) 6. Sleeping well (4 programmes, 21 activities) 7. Falling asleep (1 programme, 11 activities) 8. Relaxing sounds for sleep (17 activities) 9. Thoughtful communication (1 programme, 9 activities) 10. Mind your body (1 programme, 9 activities) 11. Mind your life (2 programme, 17 activities) 12. Covid-19 Staying resilient in times of crisis (17 activities)	1. Anxious thoughts 2. Feeling down 3. Difficult relaxing 4. Trouble sleeping 5. Low self-esteem 6. Feeling stressed

Source: Koa.

Once the user chooses an area of focus and a program, she/he will see different widgets on the main screen:

- **Next activity** inside the active program
- **Feel better now**: Activity suggestions based on how the user says she is feeling right now, for example tense, anxious or sad. The app suggests three activities for that feeling. These suggestions are based on rules, and from all the activities that relate to that feeling, the app chooses randomly three of them.
- **Today's activity for your focus**: Other two activities based on the area of focus, chosen based on rules and then two of them picked randomly.
- **Other activities for you**: Two other activities from the general pool of activities suggested by a recommender model.

The recommender model used in the "Other activities for your widget" is based on the activities' popularity. There is a planned version 2 of the model where the suggestions will be personalized for each user.

2.1.1.1 Current version

The current RecSys v1.0 model is based on the **popularity** of different activities during certain hours, thus **differentiating between day and night activities**.

The **training of the current version** of the model takes into account all the impressions and clicks of the different activities of the app, excluding tester users. It generates a ranking based on the probability to be viewed by users. This ranking is adjusted to take into consideration activities with a small number of views. In this model, each user has an equal probability of seeing a specific activity.

[2] You can only have one area of focus active at the time.

2.1.1.2 Future version

The next version of the model will be an evolution of the first version, one where the user will receive personalized suggestions. This version will give suggestions based on two algorithms. The first one is a **Contextual Bandit Algorithm**, where context, apart for a time of a day is also the user's history of app usage. It will mostly contain behavioural data like activities that were seen, viewed and liked. The second algorithm will be **a collaborative filtering algorithm**. For that similar behavioural data will be used. These two approaches will be combined by using features from collaborative filtering algorithms as part of the context for the bandit algorithm.

This model aims to optimize for click rate on activities by showing the user activities that the model suggests as relevant. The datasets and data model are not yet decided, although the team plan to use recall, precission@k k=2 and ROC-AUC metrics to measure the model's accuracy and precision.

2.1.2 Heart rate and breathing rate model

Some activities will have a model associated in order to **measure the heart and breathing rate** of the user. The current version, developed but yet to be deployed, measures the heart rate and breathing rate with the aid of the accelerometer of the mobile device. In a future version, the team plans to create a model to detect if the activity actually reduced the user's stress based on these readings.

2.1.2.1 Current version

The most accurate way to measure heart rate is by using an **ElectroCardioGram (ECG)**. It is nowadays considered ground truth, but other options also exist. Two of these options are the BallistoCardioGram (BCG) and SeismoCardioGram (BCG). Both of them can be implemented using the accelerometers available on some mobile devices.

The **goal of the HR/BR model v1.0 is to track the evolution of heart rate and breathing rate** of a user in order to provide live feedback on their ability to relax (e.g. slower heart rate and breathing rate) while performing a breathing exercise within the Foundations app while sitting or lying down. The shortest breathing exercise is an audio activity that lasts for 3 minutes, so the model has at least 3 minutes to estimate the rates. The first estimation comes after the first 5 seconds, and it gives a new estimate in buffers of 20 seconds.

The **model's input is raw accelerometer signals**, while the output is a single value for heart rate and breathing rate for a given time window. The model is composed of a set of filters and time-frequency transformations applied to the accelerometer signal (FFT) to extract the dominant frequency related to the ballistic and oscillatory components of the beating heart and expansion of the thorax while breathing. The model parameters are trained using a Bayesian optimizer on a custom dataset created internally at Koa. The main model is complemented with a logistic accuracy estimation model that predicts the accuracy of the estimation from the signal's spectral entropy. Based on this estimate, a further post-processing step is applied to remove inaccurate values from the final result.

The development and training of the algorithm are based on a custom **dataset created at Koa**. This dataset consists of approx. 24h of recordings of ECG and respiration signals from a biosignal acquisition

platform matched with smartphone accelerometer signals. The dataset includes recordings from **28 subjects (14 males, age range 20-50)** at rest in **two positions** (sitting holding the phone in the hands and laying with the phone on the chest). Subjects did **not have a history of cardiovascular diseases**. The phones used to develop the dataset cover a range of **high tier and low tier Apple and Android smartphones**.

2.1.2.2 Future version

The current primary use case is to track the evolution of a user's heart rate and breathing rate to provide live feedback on their ability to relax (e.g. slower heart rate and breathing rate) while performing a breathing exercise within the Foundations app while sitting or lying down. The biosignal extracted will be further used to assess the user stress level from stress biomarkers present in literature and inform the user regarding his/her state.

2.1.3 Data points

The list of **data collected by the application** are:

- **Name and email of user**
- **Usage data**: time of day of use, duration of use, activities seen, clicked and rated
- **Data on heart/breath rate.**
- **Information from users' interaction with the app**, like open questionnaires, closed questions, favorite activities and personal preferences.

When performing these activities, and if users' have previously consented, Foundations will capture and process information from users' smartphones and show it to them. The basis for these collections is consent, and it can be withdrawn at any time.

2.1.4 Dashboard

The **dashboard information** provided to those responsible from the companies and organizations offering the service to their workers consists of aggregated insights related to the usage of the app so that they can understand its impact. The information reported is:

- **Signups**: number of signups since launch, number of signups from the last 30 days, number of signups per week since launch, number of signups per day from the last 30 days.
- **Engagement**: total activities engaged with since launch, total activities from the last 30 days, active users in the last 30 days, total number of minutes spent since launch, total number of minutes spent from the last 30 days, cumulative user events per hour in the day.
- **Content popularity**: Top 20 activities engaged with since launch, percentage of activities marked as helpful since launch, Top 10 programmes engaged with since launch, percentage of programmes marked as helpful since launch.
- **Additional insights**: Percentage of notifications opt-in since launch, Selected user motivation from using the app.

2.2. Mental health apps and ethics

Apps offering psychological support are increasingly used to address different mental health problems (Becker et al., 2014; Anthes, 2016), often showing successful outcomes in reducing stress (Harrison et al., 2011; Economides et al., 2018). In many cases, these technologies have been useful instruments for patients' therapeutic guidance and self-assessment.

Nevertheless, ethical issues concerning these types of technologies are multiple. Firstly, the **need for further evidence-based research** on these technological developments has been stressed. Empirical assessments should also be presented to users transparently and understandably. Indeed, the literature has framed the lack of studies or scientific evidence for several technological tools as a significant ethical and legal concern (Becker et al., 2014; Aguilera and Muench, 2012; Ahthes, 2016; US Federal Trade Commission, 2016; Paganin and Simbula, 2020). Related problems identified in this framework are the **scope and quality of existing testing** (Batra et al., 2017), which often includes small and non-controlled samples examined during a short period of time. It has also been indicated that many psychological therapies delivered in person may not always remain efficacious when delivered through these software (Heffner et al., 2015). This issue does not apply to Foundations, which is not a therapeutic tool. However, it is essential to regularly assess these technologies' efficiency and effectiveness regarding various users' groups to address these issues. This may help to avoid, on this basis, those interventions showing poor results or even adverse incidence over users' health[3] or behaviour (Stratton et al., 2017).

In the case of health **apps aimed at addressing work-related mental problems**, it has been indicated that qualitative assessments should consider the following specific factors: "design, development stage, and implementation of the app; the working context in which it is being used; employees mental models; practicability; resources; and skills required of experts and users." (de Korte et al., 2018: 13). Environmental aspects and stress factors related to concrete jobs and job positions should be part of these validation studies.

Secondly, we should consider **power relations behind these technologies'** governance and administration from an organizational perspective as well. The development of apps to address mental health problems for employees, such as Foundations, has been extended in recent years to ensure sustainable productivity. Some of these apps have shown to provide robust results in terms of mental health and stress symptoms at work (Stratton et al., 2017). However, it has been pointed out that **Stress Management protocols within the work domain** should be targeted to specific individuals and not be mandatory for the whole workforce (Stratton et al., 2017). These findings have been obtained for various technologies, which call to consider both workers and work contexts characteristics (de Korte et al., 2018). Considering the specific employees' mental health status is crucial for ensuring both the ethical grounds and efficiency of such systems. This involves an unavoidable trade-off between their standardization and relative capacity to avoid unfair discrimination while providing similar results for different social groups.

[3] This is the case of an app to treat addiction by measuring blood alcohol level. It was revealed that the system might have encouraged a group of patients to drink more instead of less (Gajecki et al., 2014).

Thirdly, **socio-economic biases and issues regarding the accessibility to these technologies** have been identified, including possible digital divide problems or limitations related to devices needed to use these systems, which may hinder plural access to psychological support (Burns et al., 2011). In this regard, the development of software that can be used in different mobile phones has been recommended. Moreover, differences in **users' cultural capital** should also be regarded as a general framework for communicating these apps' goals and capabilities. Best practice concerning informed consent involves developing cultural-based strategies to address the specific needs and characteristics of social groups, such as language or identity. Simultaneously, free apps and possible advertising-funded apps should be transparent concerning data shared with third parties.

Fourthly, **in terms of privacy**, these technologies often require the acquisition and processing of large amounts of sensitive data related to users' mental health, tracking of users daily activities or other special categories of personal data such as political opinions. Besides implementing data minimization approaches, by-design mechanisms for preventing data breaches regarding sensitive data about mental health are essential for ensuring data security (Luxton et al., 2011). Moreover, security specifications and functionalities should be proportional to these risks, ensuring data protection (Njie, 2013). Text messaging or geolocation have been identified as potential privacy issues in these contexts (Ackerman, 2013; Caetano, 2013). Regarding users' requirements for data protection, a proactive approach towards explainability and informed consent has been proposed in this domain (de Korte et al., 2018).

2.2.1 Summary of ethical implications

- ❏ Need for experimental research with various populations of users. Systematic evidence-based validation must be ensured.
  - As part of these evaluations, consider possible differential factors concerning jobs sectors and job positions that may influence these technologies' effectiveness.
  - Examine the trade-off between standardization of treatment within each working sector/organization and the need to target individuals' psychological/life specifics and their working conditions.
- ❏ Consider the need to analyse and address possible socio-economic discrimination derived from the software's technical specifications (needed infrastructure, ease of use by groups of population/cultural capital, etc.).
  - The exploitation policy should be made explicit to users.
- ❏ Develop a robust data protection policy (data minimization, security, informed consent) to ensure the integrity of special categories of personal data and purpose limitation.

2.3. Social context for the model

This section aims to place Foundations in the social context of its implementation by addressing the main **societal factors** that could **facilitate or limit its efficacy and smooth adoption** according to the existing state of the art. Issues identified include the influence of workplace environments and socio-demographic factors, such as gender, stigma or workers profiles. It should be noted that the analysis

is not aimed at replacing a social impact assessment but to contribute to developing hypotheses aimed at integrating key factors in the empirical evaluation of the system usage.

The **Foundations app** is presented as a form of addressing losses in companies' profit and competitiveness due to the adverse effects of workforce stress, anxiety and depression. The system is targeted to workers of companies contracting the service and aims to support workers to be more resilient and resistant to these mental issues. Although it is not targeting any industry, Foundations has already been deployed in healthcare, education, finance, telecom and industrial sectors, and used by UK and US workers above 16 years old.

The literature has revealed that **being exposed to chronic hostile working conditions** leads to stress (Ravalier, 2018) and other mental disorders (Dewa et al., 2014; Siegrist, 2008), also favoring many other related health issues, such as diabetes or cardiovascular problems (Chandola et al., 2006; Rosengren et al., 2004; Sohail and Rehman, 2015; Berkman et al. 2014). Moreover, mental problems derived from work stress often led workers to engage in unhealthy behaviours, such as drinking, using drugs or lashing out (Nguyen et al., 2019). For these reasons, the occupational phenomenon officially classified by the WHO as workplace burnout is approached as a multivariable phenomenon. Therefore, it is also essential to consider any treatment for such mental disorders as conditioned by various structural factors at both social and organizational levels.

Currently, Foundations mainly address these mental problems **in the UK and US social contexts** by providing companies and other organizations with tools to manage workers' mental status. **In the UK**, work-related stress is the first cause of long-term sickness absence and the second reason for sickness leave shorter than four weeks in public service workers (Chartered Institute of Personnel Development, 2016). Moreover, stress accounts for 45% of all working days lost due to poor health (Health and Safety Executive, 2016).

There is a significant **variation in the incidence of mental problems across different productive sectors** and organizational responsibilities. Within the primary care sector, about 23% of workers have been shown to suffer from mental distress. The level of responsibility within their organizations and also marital and health statuses have been identified as critical drivers for stress (Calnan et al., 2011). Social workers within the country have described how workload, lack of managerial support and adequate reflective supervision are key stress causes (Ravalier, 2018). Work-life balance (WLB) has also been identified as a fundamental factor leading to mental distress among UK construction workers (Kotera et al., 2019). This confirms other studies highlighting this impact (Aazami et al., 2015) and the adverse effects of poor WLB on work productivity at an organizational level (Mendis and Weerakkody, 2014). Gender differential impact has also been underlined. In the IT sector, it has been shown how the roots of stress are more related to personal factors for females, while organizational and environmental factors are the primary stress drivers for males (Haque et al., 2016). Another disadvantaged social group significantly affected by working conditions is immigrants engaged in precarious employment. Their most common mental problems derive from these working environments and practices, and include anxiety and depression. It has also been found that workers' social exclusion and the precarious nature of employment can have adverse mental health effects (Muoka and Lhussier, 2020).

In **the US**, mental disorders are also one of the leading causes of ill health leading to work absence, affecting both high work loss and total lost workdays (Zaidel et al., 2018). Moreover, as in the UK, mental illnesses are also risk factors for injury and illness, leading to work absences (Airaksinen et al., 2017; Birnbaum, 2010). The adverse effects of occupational stress on productivity, particularly related to hostile working environments and violence, have also been stressed (Rasool et al., 2020). While these scenarios' increasing economic impact has been emphasized (Sime, 2019), specialized organizations recommend clean and well-equipped workspaces, fair wages, and work-life balance as some of the leading solutions to this problem (Nguyen et al., 2019).

Differences between sectorial and social groups also offer useful insights. Different studies have shown that the increasing perception of job insecurity in the United States correlates with work stress (Burgard, 2009; Fan et al., 2015). Moreover, research with medically healthy employed men and women between 30 and 60 years old has revealed that job insecurity and home stress are related to elevated depression and anxiety symptoms (Burgard et al., 2009). This study also shows the importance of considering work-life balance and their interrelations as part of therapeutic strategies and managerial intervention scenarios. Along these lines, a well-perceived balance between work and family life has been shown to be an essential driver of mental health in the case of hospice nurses (Bernett, 2019). A study among bisexual Latino men in the New York City Metropolitan Area revealed that this population experienced adverse mental health outcomes due to pressures in their work environments, family demands, and work-life balance, which are also crucial factors inducing mental disease (Muñoz-Laboy, 2015).

It should also be noted that **adherence to psychotherapy** has shown to be more effective than antidepressants in minimizing the risk of future work leaves in the USA (Gaspar et al., 2020). Other practices such as **mindfulness** have helped to minimize work depression, anxiety, non-severe psychiatric symptoms at the workplace (Lacerda et al., 2018). In the above framework, mobile technologies and apps for mental self-care have also been supported by the WHO, in its Mental Health Action Plan 2013–2020, and also by other public organizations such as the UK National Health Service (NHS). Among the elements supporting this approach, these systems' capacity for reaching people without access to care has also been pointed out (Anthes, 2016). This is in line with one of the Foundations means and purposes.

However, the Work Health Survey of Mental Health America shows that 69% of workers say that they prefer **to remain silent about their stress** within the workplace and 50% strongly agreed with this statement (Nguyen et al., 2019). This is particularly relevant for Foundations since it is clear from the literature that there is a strong stigma concerning how workers manage and are able to manage their mental disorders.

2.3.1 Summary of societal analysis

- ❏ There is a need for a holistic approach towards work-related mental problems that integrate aspects such as habits, work-life balance, and physical health. Lacking these analyses may affect both the efficacy and acceptability of Foundations.

- ❑ Consider variation and specificities of mental problems across productive sectors and working positions to capture users' specific needs and interests. This includes specific organizational and socioeconomic factors.
    - Under these premises, address users' autonomy by protecting and explaining their anonymity. This should be considered to minimize stigma and negative-related effects on usage.

2.4. Ethical assessment

Ethical issues will be analyzed in this section, considering both broad social aspects affecting the ethical principles guiding Foundations and also the specific commitments established by Koa for their systems. Following the Koa ethics self-assessment's conclusions (Ethics Audit Internal document, EIA), this part of the analysis's primary purpose will be to contribute to the application of ethical standards.

2.4.1 Compliance of Foundations with Koa ethics commitments

There is a complete set of ethical instruments and protocols applied to Foundations. In this regard, the audit efforts are oriented to assess and improve contributions to Koa's ethics commitments and introduce missing elements of analysis. The analysis of the 10 commitments for each of the products conducted in this section will aim to increase their scores in the future.

Table 2. Analysis of Koa ethics commitments

Commitment	Analysis
1.We aim to support users to achieve their optimal balance of health and happiness	This is not a **problematic aspect in Foundation** due to how its aims and capabilities are framed and communicated, **focusing on users' wellbeing**. This also minimizes the importance of the lack of a measure of happiness pointed out in the EAI.
2.We will ensure that our recommendations are not based on discriminatory bias	No discriminatory recommendations have been identified through our digital ethnography of the 152 activities. However, **possible biases regarding graphics** and accessibility for vulnerable groups should be assessed in the future (see section 5.3). A methodological framework for analysing algorithmic bias, focusing on gender, is provided in Section 6.
3.We follow best practice in giving users control over how we use their personal data	The main instruments in this regard are the users' consent protocol and the app Privacy Policy. Data minimization also contributes to this purpose. Communication about data processing is clear and ARCO rights are properly integrated into these instruments. However, the **explainability of algorithms** should be added/improved in the Privacy Policy. Only a reference to ARCO rights is included in its current version: *"The data*

	*protection laws give you a series of rights regarding the personal information that we manage about you. Specifically, the rights of access, rectification, erasure, limitation, objection, portability, as well as not being subject to **automated decisions** and to remove your consent at any time."*
4.We will deploy the best available techniques to prevent any user from becoming addicted to any of our services	No indirect evidence of addictive functionalities has been found. Still the role of the app in terms of **user investment (see Table 3)** should be considered in future analysis.
5.We will explain how our services work to support you in having the greatest possible health and happiness; in doing this, we will ensure that such explanations are comprehensible, aiming for a reading age of no more than 11	As shown in section 7.2.1, the Privacy Policy's **readability has a reading age of 16**. Although explanations about how the system works and recommendations seem to be clear, it is recommended to assess all the system presentations using the tools presented in 7.2.1 to ensure broad age access. Moreover, we have also examined a set of content within the system recommendations to establish Foundations' reading age and found that it reaches 12-14. Even though this is almost aligned with the 11-age established limit, either the age limit **should be reconsidered, or this issue should be assessed** in the future as content is modified.
6.We will publish the ethical approvals of our research and external audits of our work, although we may remove some commercially sensitive information	Koa is only partly complying with this commitment since, whilst external audits are published, ethical approvals of research are not as yet. However, , Koa states that it plans to put in place processes for this during 2021.
7.We use the state-of-the-art industry standards of encryption to protect your data	The technological **development process follows this commitment**. Data security protocols are developed following both GDPR and ISO27001. Privacy Enhancing Technologies (PET) include TLS secure communication and symmetric 256-bit encryption - RSA public-key SHA-2 algorithms. The methodology used for integrating these requirements into the design includes the review, every two weeks, of the system design. Both technical and legal experts are involved in this iteration. As part of this exercise, threats are measured by addressing a comprehensive list of specific risk scenarios.
8.We will create products and services that preserve as much privacy as possible, for you and your community	Given the sensitivity of data shared by users and employers' role in data governance, ensuring purpose limitation, anonymization, and secure communication is critical for the system's ethics. The above-mentioned

	system security protocols and the use of data minimization are in line with this commitment.
9.We will not generate revenue through serving adverts to end-users of our services	According to Foundations policy, **no personal data is shared with third parties with advertising or revenue goals**. Services involved, included in the PP, fulfil concrete purposes such as hosting (e.g. AWS) or analytics (e.g. G Analytics).
10. We will hold external ethics audits at least once annually to assess progress against our ethics strategy, including algorithmic audits	This document is aimed at fulfilling this goal.

Source: own elaboration based on Koa commitments.

2.4.2 Analysis of societal factors and ethical issues

It has been proposed that, given the limited **empirical validation, regulatory oversight and scientific research** of many of these technologies, apps should only help to improve the relationship between psychiatrists and patients. They should not be designed to replace experts and psychologists, so possible harmful interventions on vulnerable patients are reduced (Torous and Roberts, 2017). In this regard, **clinical and organizational safeguards must be taken in those cases of mobile apps available directly to consumers** since they may create a gap "in protection for vulnerable patients" (Torous and Roberts, 2017:6). This involves issues related to both *digital divide*, which can lead specific groups of people to disadvantage treatment concerning health-related services (Chang et al., 2004), and individuals with low health literacy (Tieu et al., 2015).

Furthermore, it has been recommended to present and use mobile health **as an "adjuvant tool"** to address the above ethical issues (Torous and Roberts, 2017; Hsin et al., 2016). In Foundations, this means providing clear information about the scope of the system **to users**, pointing out that this scope does not include **clinical treatment, and detailing third parties involved in data processing**, so they can properly understand the system mechanisms and goals and avoid "therapeutic misconception" (Torous and Roberts, 2017:7). In this regard, education has been framed as an essential part of mobile health. Therefore, gathering up to date and feasible data about Foundations' efficacy with respect to improving mental wellbeing is key to ensure that precise information is given to users.

It is essential to ensure **that the use of the app is voluntary** and to adapt **informed consent** to vulnerable groups. Given that Foundations management does not include a therapist (Torous and Roberts, 2017), the app must guarantee these factors by design. Aspects to be considered are: sharing enough information about how the system works with users, assessing their decision-making capacity (including age, disabilities, etc.), and their authenticity of choice (Roberts, 2016). In the case of Foundations, the latter involves ensuring that workers are able to reject using the app without having to offer any explanation and without any consequences for them. In this regard, **organizational culture** is key. Supervisor communication and feeling comfortable to report dishonest or unfair practices are critical drivers for workers' mental wellbeing (Nguyen et al., 2019).

Another ethical key aspect concerning eHealth services is ensuring that integrating possible **addictive mechanisms** behind their design is reduced at the minimum. The six variables considered by the literature regarding addiction to apps, variable rewards, social reciprocity, infinite scrolling, the illusion of choice, user investment, and gamification, are reviewed with this purpose (Neyman, 2017).

Table 3. Addiction primary dimensions in Foundations design

Addiction variable	Definition	Review
Variable rewards	Random and unpredictable rewards produce more of the neurotransmitter dopamine than regular rewards. In apps, they are based on notifications and other processes.	The app model is mostly oriented towards self-reflection and providing suggestions for concrete activities, which can be repeated across a longer process. This minimizes the risk of passive reception of a randomized stimulus. Therefore, the use of notifications and dopamine generators (Eyal, 2014) seems to be minimal. **Only a few notifications are used**.
Social reciprocity	These are compensations derived from social interaction and reciprocity. In software applications, chemical satisfaction is received from outcomes of these interactions, for instance in the form of likes.	Mutual user exchanges do not play a role in this app since it's only based on users-machine interactions.
Infinite scrolling	This is achieved by loading content on a single page instead of spreading it across a series of pages. It produces an interface through which consuming content is allowed by scrolling instead of moving to a different page.	Foundations content is distributed across multiple sections and layers, so this issue is not relevant.
Illusion of choice	User choices can be oriented by software design through the layout of their applications. While some applications seem to empower users with reviews or notifications about different products and services, they often provide a limited number of options.	This problem mostly applies to commercial strategies. Choices (recommendations and programs) available in Foundations are made explicit to users and fit specific functions.

User investment	Many social media applications take advantage of the human tendency to invest time in activities they feel they "construct" (the so called "Ikea effect") by giving users the power to curate their profiles.	Users' capacity to modify or reconstruct the system structure and layout is limited. In line with the CBT approach, **users must invest effort in self-assessing their mental status** and work in potential outcomes.
Gamification	"Closely tied to variable rewards, "gamification" is defined in the tech industry as the process of using game mechanics to reward the completion of tasks." (Neyman, 2017: 4).	Only one game is used by Foundations as an activity. The app also uses the concept of "earned badges", but in a very subtle way at the moment: there are no notifications when a new badge is earned, they can only be consulted in the user profile section and they are a summary of completed programmes by the user.

Source: own elaboration based on Neyman, 2017.

Lastly, **privacy** is of utmost importance to guarantee users' integrity and the acceptability of the system. Data protection breaches, sometimes based on selling profiles for users' information to third parties (Glenn and Monteith, 2014), must be avoided. Moreover, advertising and the possible sharing of personal data with third parties should be made explicit to users. In the case of Foundations, no third products or services are advertised to end-users as its revenue model.

2.4.3 Ethics in product design

The Koa team has an internal guide and process to **help avoid bias in language and design**. When building products and working in teams, biases can lead to the exclusion of perspectives, which can ultimately lead to the exclusion of users who might be most in need of such products. The Koa team has implemented a checklist to check for bias in projects, and the content of the app follows a peer-review process paying attention to these aspects. Color, graphics and avatars are designed without targeting or reflecting any gender in particular.

This is a sample of questions the Koa team ask of themselves when creating content:

- Gender: are pronouns mostly male or balanced?
- Ethnicity: are non-white experiences being inadvertently excluded?
- Class/Wealth: Is there an assumption that the user is of a certain socioeconomic status?
- Sexuality: Is the language heteronormative?
- Disability: Are all abilities taken into consideration?
- Religion: Are non-western religions included?
- Immigration status: does this need to be factored in?
- Education/Occupation: Does it assume a certain level of literacy?
- Age: Is there an assumption on the user's age?
- Parent/carer: Does it take into account dependents the user might have?
- Language: Is it written in plain English?

With that in mind, the Koa team has already identified these areas to improve:

- Review the activity “Mindful walk” for people that cannot walk
- The audio activities are not usable by deaf people

2.4.3.1 Ethics in design

Although this audit does not include a usability and user experience study, we have analysed the type of activities and how they are delivered. We have identified 9 ways of delivering or getting information inside the app: text, long text, data entry, audio, quiz, audio (sounds), audio with text and video with text. We have counted how the different activity categories (learning, relaxation, journaling, game, blog, movement, and reflection) use different ways to deliver/obtain information.

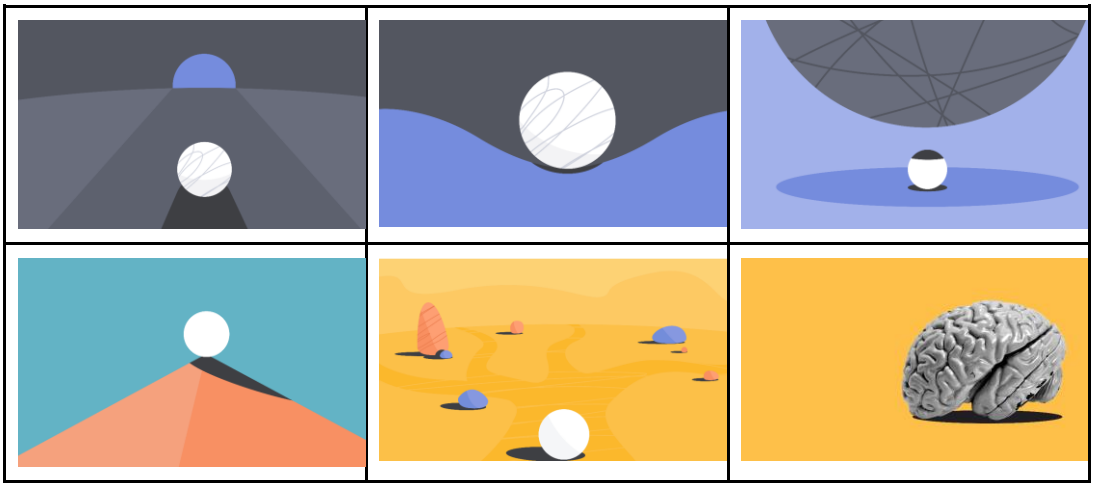
Table 4. List of activities per category and way of delivering information

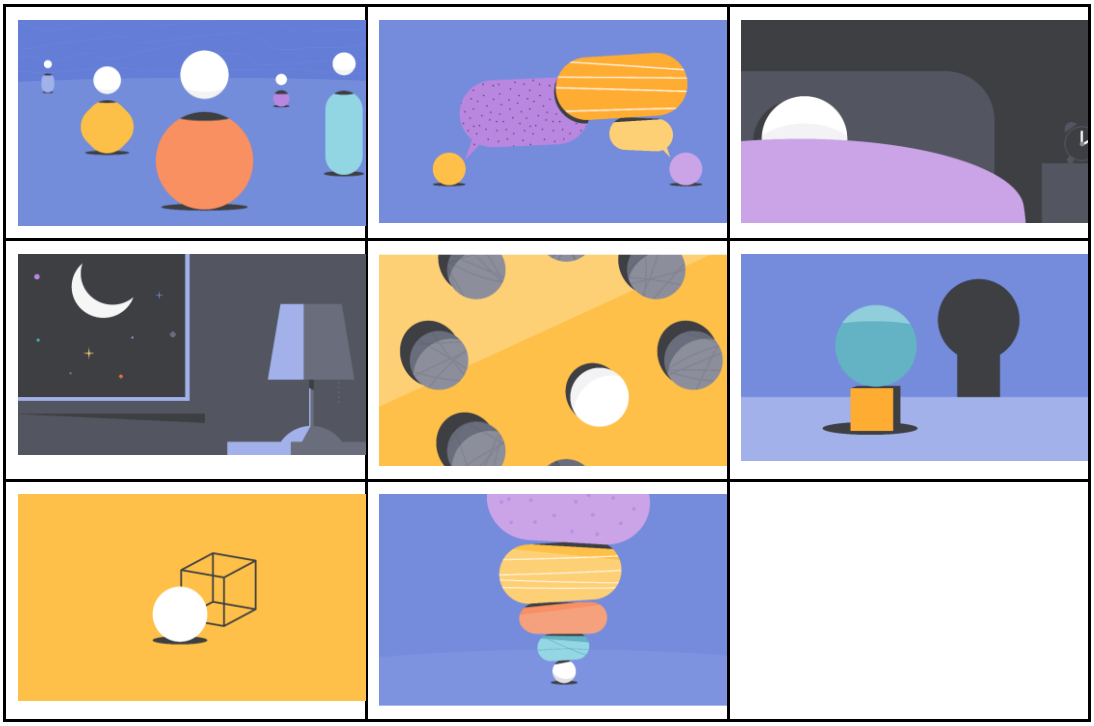
Activity	Total	Text	Long Text	Data Entry	Audio	Quiz	Audio (Sounds)	Audio with Text	Video with Text
Learning	68	42	14	0	0	9	0	3	0
Relaxation	43	0	0	0	24	0	17	0	2
Journaling	16	0	0	16	0	0	0	0	0
Game	1	0	0	0	0	1	0	0	0
Blog	12	0	12	0	0	0	0	0	0
Movement	3	0	0	0	3	0	0	0	0
Reflection	9	4	0	5	0	0	0	0	0
Total	152	46	26	21	27	10	17	3	2

Source: own elaboration.

Paying attention to the list of questions above, we have reviewed the written copy, the photos and graphical designs used in the summary of a programme, the header of a long text activity and the pictures showing totally or partially people.

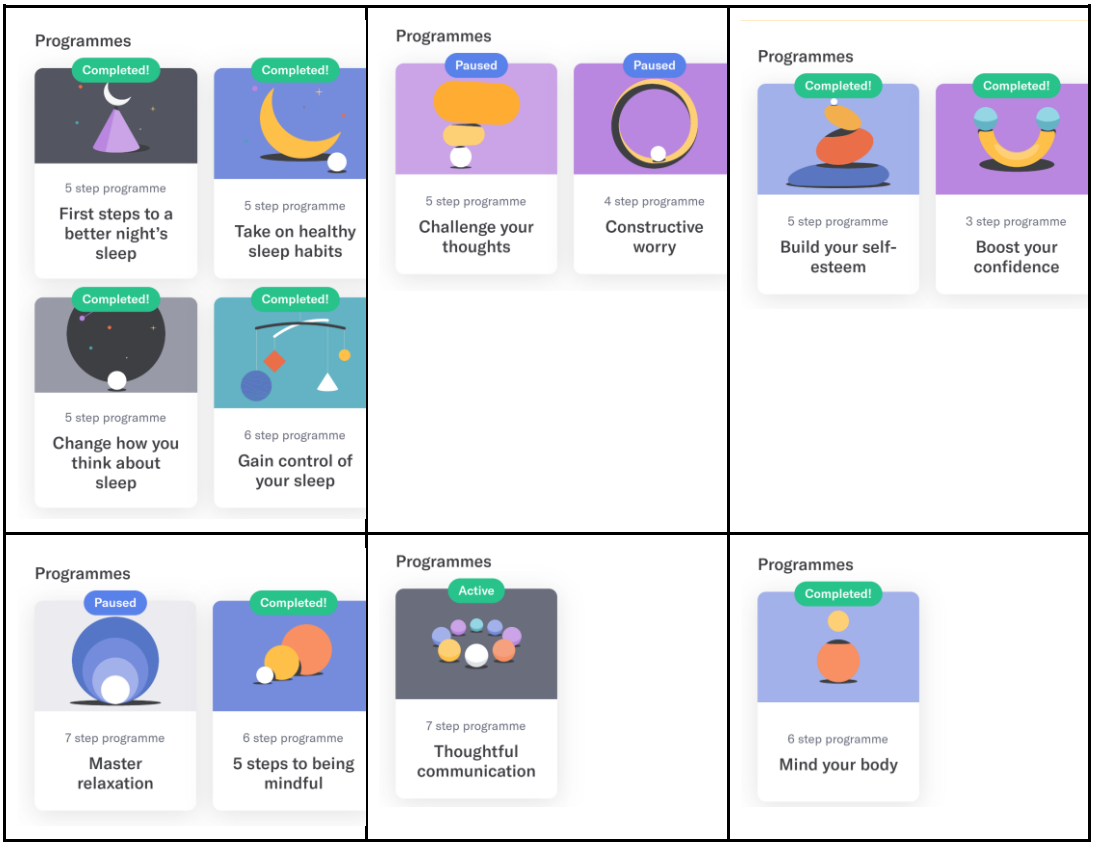
Figure 1. Abstract designs used in activities delivered as a long text

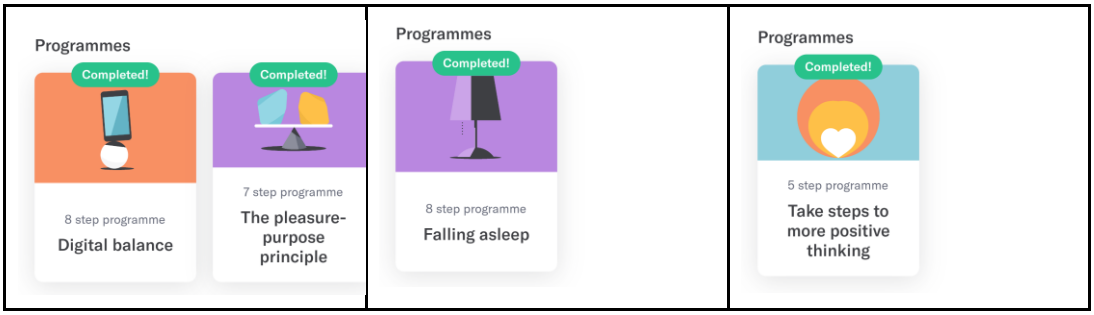




Source: Koa.

Figure 2. Abstract design to depict the different programmes



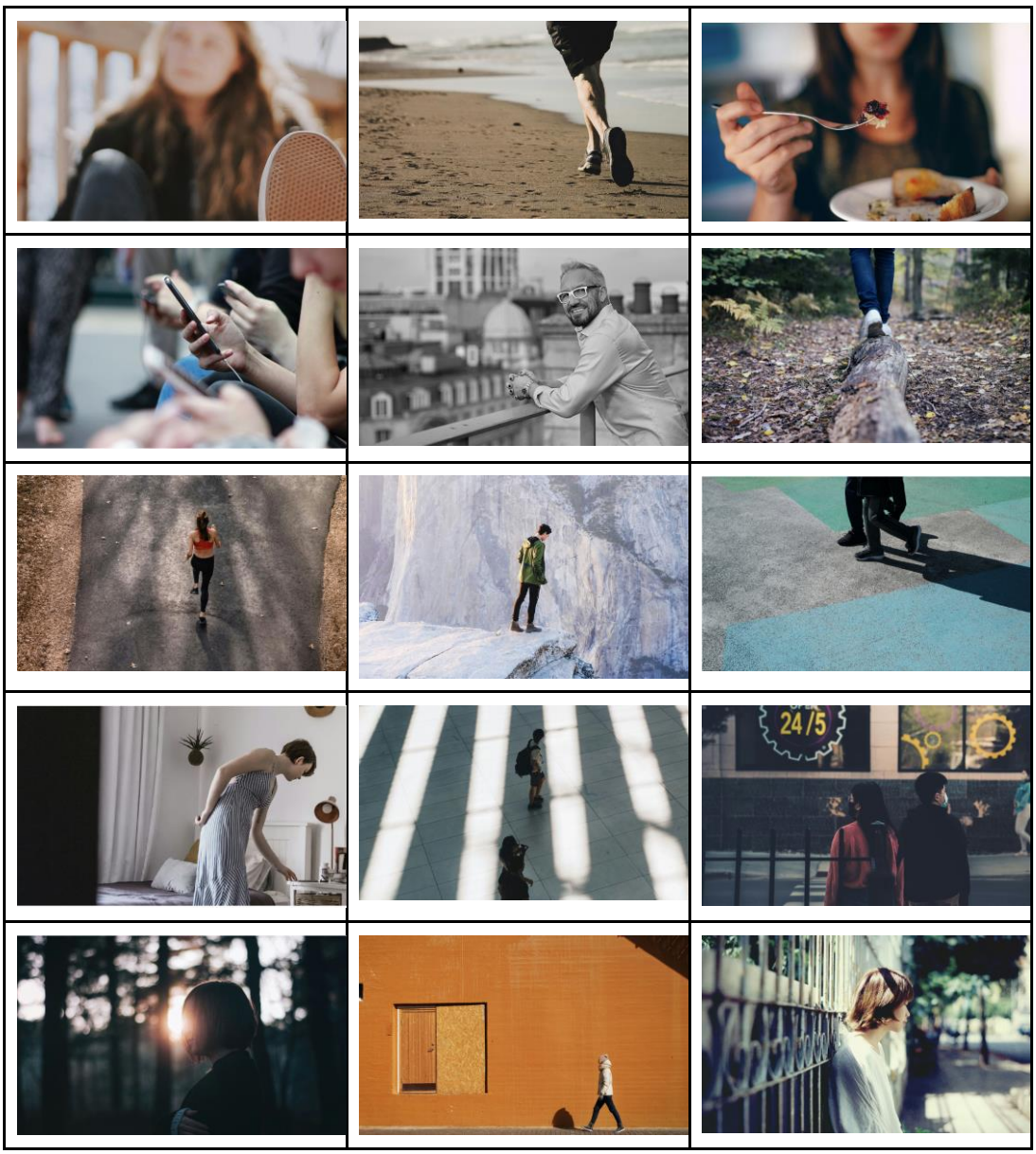


Source: Koa.

In these 30 abstract designs used in long text activities and programs we find:

- 19 designs use a white ball to denote selfness and explain the concept
- 5 designs use a ball of another colour to denote selfness and explain the concept
- 6 designs use a different pattern to explain the concept

Figure 3. Photos of people used in the (long text) activities



Source: Koa.

Issues found cover three dimensions:

- Firstly, within the long text activities, we have seen a **predominance of pictures that depict Caucasians**. It should be noted that even though Asian people are integrated into the sample, no black people have been identified. This could be considered as aligned with a prevalence of using a white item for identifying self at the abstract level. However, such an association's relative significance derives from the lack of "racial" diversity in people's illustration and not from the actual (subjective) character of the (predominantly white) items.
- Secondly, there is also a prevalence of **young people** for illustrating activities and behaviours. This should be considered when examining target groups.
- Thirdly, we have found that more than **90% of relaxation activities use audio** (24 activities out of 26, if we exclude the 17 ones that only sound), limiting access to these typologies of activities and information for certain groups of people. Visual impairment could also be a factor influencing the level of accessibility. Still, significant use of audios for Foundations activities could be well adapted for visually impaired people. In this regard, the development of a VoiceOver model for the app designed to provide users with visual impairments the same volume of data as is accessible to sighted users using the device could be considered based on existing analysis and models (Kamei-Hannan et al., 2015; Torres-Carazo et al., 2016; Berling et al., 2018).

2.4.3.2 Ethics in users' access

As part of Koa's commitment to ensure that explanations are comprehensible, we have analyzed the readability of some activities. We have chosen 3 activities that are delivered as text (T), 3 activities delivered as long text (LT) and the one activity that has video and text available (VT).

The different texts are checked against different indices using the open-source Python library "Readability" available at https://github.com/andreasvc/readability/. This library calculates text statistics as well as the Flesch Kincaid Reading Ease and several grade level indicators, which equate the readability of the text to the U.S. grade level system.

The following indexes are checked:

- The **Flesch Kincaid Reading Ease** test works by counting the number of words, syllables, and sentences in the text. It then calculates the average number of words per sentence and the average number of syllables per word. The idea is that **shorter words** (with few syllables) and shorter sentences are easier to read. The higher the score, the easier the text is to understand. The result is a number between 0 and 100. Higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read. A value between 60 and 80 should be easy for a 12- to 15-year-old to understand.

- **Flesch–Kincaid grade level**. Test used extensively in the field of education. As in Flesch Kincaid Reading Ease**,** it is based on the idea that **shorter words** (with few syllables) and shorter sentences are easier to read**.** But the "Flesch–Kincaid Grade Level Formula" instead presents a score as a U.S. grade level.

- **Gunning fog index** is a readability test for English writing that takes into account two qualities to determine readability: the average number of words in sentences and the percentage of **complex words** (words with three or more syllables).

- **SMOG grade.** SMOG is an acronym for "Simple Measure of Gobbledygook". It is used particularly for checking health messages. It also takes into account the **complex words** (words with three or more syllables), in three 10-sentence samples. Analysed texts of fewer than 30 sentences are statistically invalid, because the formula was normed on 30-sentence samples.

- **Coleman–Liau index.** It relies on **characters** instead of syllables per word. Although opinion varies on its accuracy as compared to the syllable/word and complex word indices, characters are more readily and accurately counted by computer programs than are syllables.

- **Automated readability index**. As the Coleman-Liau index, it relies on a factor of **characters** per word, instead of the usual syllables per word. This index was designed for real-time monitoring of readability on electric typewriters.

Additionally, an average grade level is calculated as the arithmetic average of Flesch–Kincaid grade level, Gunning fog index, SMOG grade, Coleman–Liau index and Automated readability index. We recommend using this average in order to account for the readability of a text as it takes into consideration different approaches to measure complex texts.

Table 5. Text Statistics

Statistic	T1	T2	T3	LT1	LT2	LT3	VT
No. of sentences	17	35	43	32	20	15	36
No. of words	278	498	586	456	423	323	547
No. of complex words	32	71	75	52	66	35	45
Percent of complex words	11.5%	14.3%	12.8%	11.4%	15.6%	10.8%	8.23%
Average words per sentence	16.35	14.23	13.63	14.25	21.15	21.53	15.19
Average syllables per word	1.33	1.44	1.47	1.36	1.5	1.34	1.35

Source: own elaboration.

Table 6. Text Readability Estimated U.S Grades and Ages [4]

Readability Index	T1 Grade (Age)	T2 Grade (Age)	T3 Grade (Age)	LT1 Grade (Age)	LT2 Grade (Age)	LT3 Grade (Age)	VT Grade (Age)
Flesch Kincaid Reading	77.6	70.6	69	77.7	58.4	71.3	77.3

[4] A table to look up ages for the different U.S. Grade Levels can be found at https://en.wikipedia.org/wiki/Education_in_the_United_States#Educational_stages

Ease[5]	(12-13)	(12-13)	(13-15)	(12-13)	(15-18)	(12-13)	(12-13)
Flesch Kincaid Grade Level	6.49 (11-12)	6.95 (11-12)	7.02 (12-13)	5.96 (10-11)	10.4 (15-16)	8.66 (13-14)	6.26 (11-12)
Gunning Fog Score	11.1 (16-17)	11.4 (16-17)	10.6 (15-16)	10.3 (15-16)	14.7 (19-20)	12.9 (17-18)	9.37 (14-15)
SMOG Index	n.a.	10.8 (15-16)	10.2 (15-16)	9.98 (14-15)	n.a.	n.a.	9.12 (14-15)
Coleman Liau Index	9.21 (14-15)	8.76 (13-14)	10.1 (15-16)	8.36 (13-14)	10.6 (15-16)	9.02 (14-15)	9.33 (14-15)
Automated Readability Index	8.23 (13-14)	7.02 (12-13)	7.84 (12-13)	6.71 (11-12)	11.4 (16-17)	10.3 (15-16)	7.86 (12-13)
Average grade level	**8.77 (13-14)**	**8.98 (13-14)**	**9.15 (14-15)**	**8.26 (13-14)**	**11.8 (16-17)**	**10.2 (15-16)**	**8.39 (13-14)**

Source: own elaboration.

The minimum estimated grade for the analyzed texts is a 7th-8th U.S. grade, which corresponds with 12-14 years old (middle school). This analysis shows that **four out of seven activities have this estimated grade**, whereas the other three have more complicated texts.

Recommendations based on these results:

- Add the **readability analysis to Koa internal process** to refine the content and make it more comprehensible.
- Review the existing activities with a readability analysis to identify the most complex texts in order to refine them and make them easier to understand.

2.4.4 Summary of ethical issues and recommendations

The following table summarizes the ethical issues identified during the audit and provides concrete recommendations for addressing them. These lines of action were reviewed and adjusted on the basis of Koa feedback.

Table 7. Foundations' ethics assessment

Ethics-related issues	Foundations strategy	Recommendations for improvement
Lack of evidence-based studies. Need to assess the system impact on specific groups of users	Empirical research concerning theoretical basis, concept and usability of Foundations	As described in Section 3, Koa's research has already used medium-large samples, but it tested the effects of Foundations on users during a short period of time.

[5] Flesch Kincaid Reading Ease does not provide the result as an estimated grade level, instead its result is a number between 0 - 100 that can be translated into U.S. grade range levels. A look up table can be found at https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests#Flesch_reading_ease

and regularly.	are being conducted by Koa. It is planned to conduct these studies on a regular basis.	These studies should also **consider the differential impact of the app concerning different job positions and working sectors**. Moreover, **accessibility regarding vulnerable groups** (including different disabilities) should be tested. Lastly, **mid- and long-term impact** should also be measured. It should be noted that Koa is designing new RCTs. The studies will address many of these aspects. The demographic analysis will include (but limited to) ethnicity, age, disability, socioeconomic, employment and methodologies used will seek to test the efficacy and its maintenance over long periods of time. However, since these RCTs are not focused on differential impact across job sectors or positions or assess accessibility, it is suggested to consider these aspects in the future.
The system does not replace clinicians' roles nor provide clinical treatment, so it should be presented as an adjuvant and self-assessment tool designed to reach well-being.	The app is designed and presented as a wellbeing instrument. This is explained to users as part of the app PP and within its content, in the intro section.	The Koa approach seems complete. In terms of improvement, **communication regarding the system's limitations** could be reinforced within the app introduction. Foundations' lack of a duty of care[6] could be explained in this context.
Ensure informed consent and informational mechanisms for vulnerable groups. Explainability of algorithms should be achieved in this framework.	Most of the consent information is offered in PP and within the app presentation, activities and recommendations.	Differential explanation and consent mechanisms should be **further developed to facilitate access to vulnerable collectives**. In particular, this includes consent for disabled people. In particular, besides mandatory information aimed at confirming voluntariness, non-coercion, information about risks and benefits, **consent materials in alternative media** (Stineman and Musick, 2001)**, including**

[6] Legal obligation and compliance with standards of reasonable care (i.e. Hippocratic oath) while carrying out activities (health treatment) that could foreseeably injure others.

		video and audio, could be integrated into the system. Almost no information about automated processing is being provided. The **how and why of algorithmic processing should** be presented in the PP. Address **readability** in the final version of the system recommendations to ensure homogeneity and easy access.
Guarantee lack of organizational coercion regarding the use of Foundations.	Foundations strategy is to keep users anonymous for employers. Data minimization, anonymization and not sharing personal data with employers organizations are some of the mechanisms used to achieve this.	Existing data protection specifications, protocols and tools are considered appropriate and proportional to this risk. **Anonymity should be reinforced in the public presentation and onboarding of the ap**p to foster trust and address employees' reluctance to share information about their mental status revealed by the literature.
Addiction	The Koa internal ethics assessment includes a measure of addiction. However, this issue has not been tested for Foundations yet.	No direct evidence of addictive features is found. However, it is recommended to **consider user investment, rewards (in particular, the final set of notifications) and possible gamification as relevant aspects to be examined in future addiction assessments**.
Discrimination	Koa conducts iterative self-assessments concerning possible discriminatory features and outcomes for the app.	No discriminatory recommendations have been found. However, the pictures used for illustrating the activities could be **more diverse in terms of ethnicity and age**. They could also further consider the **accessibility of people with disabilities**. It is recommended to take these drivers for plurality into account.

Source: own elaboration.

2.5. Algorithmic impact assessment

This section is aimed at introducing the methodological framework for the future assessment of **biases in Foundations**. The strategy focuses on contrasting gender disparity as a tool for gathering indirect evidence about differential impact by each protected group and establishing grounds for auditing the

ML system's potential for discrimination. The actual application of selected metrics, explained below, was not conducted by Eticas due to data protection requirements.

In this regard, it is important to note that access to gender identifiers, even under a robust pseudonymization framework, is not considered proportional to the existing risk of bias identified by Koa. Therefore, legal access to needed data is not possible to be achieved. According to this perspective, the reuse of these gender identifiers would be beyond the scope of original purposes for data collection, **limiting Koa's legitimate interest** in conducting the analysis.

This constitutes a relevant example of possible **tensions between rights to data privacy and transparency**, on the one hand, and the need for **establishing mechanisms for algorithmic bias prevention**, on the other. Bias risks can be assessed as low in Foundations due to its algorithmic model aimed at assigning wellness recommendations (assigning a benefit) without using gender as a specific category for classifying and assigning recommendations. However, algorithmic discrimination is still theoretically possible. Among possible sources of algorithmic discrimination not involving the collection of protected categories of personal data, we find proxy biases: ML models do not require the protected categories to be integrated into training data to discriminate since systems use anonymous information to “learn” individuals groups based on their belonging to certain categories (Yeom et al., 2017). The so-called “proxy bias” can lead to an algorithm that has not been trained in the category “race”, to “learn” it based on the aggregate processing of other directly or indirectly related attributes, such as geographic data, purchase, mobility or preference (“likes” in Foundations).

Therefore, a certain balance between this risk -and its actual impact- and hazards derived from releasing these data for an algorithmic assessment must be considered case by case. In this regard, the **Koa approach can be considered as best practice**. Still, since Koa is this data controller regarding training data, it is recommended to assess the possibility of conducting the analysis described below in-house.

2.5.1 Algorithmic bias

In order to frame algorithmic bias, we should first distinguish between different forms of discrimination. Following definitions by Lippert-Rasmussen (2013), generic discrimination occurs when someone treats a person A worse than s/he would treat another person B because A has some attribute that B does not have. **Group discrimination happens when such attribute consists of simply belonging to a socially salient group**, i.e., a group in which membership "is important to the structure of social interactions across a wide range of social contexts" (Lippert-Rasmussen, 2013:30), and requires animosity against this group, or the belief that people in this group are inferior, or the belief that they should not intermingle with others. In order to be considered discriminatory, bias should involve one or more of the so-called protected groups, which correspond to the following protected attributes, which is based on the attributes included in the Equality Act of the UK 2010, Section 4, and the European Charter of Human Rights. It should be noted that this is not an exhaustive list since it may be adapted or modified, depending on the context:

Table 8. Protected groups and attributes

Protected groups (non exhaustive)	Protected attributes
Children and Elderly	Age
Disable people (physical and mental)	Disability
Women and Transsexual	Gender and Gender reassignment
Multiple social groups (e.g. African American, etc.)	Race, color, ethnicity
Pregnant	Pregnancy
Muslim, Jewish	Religion or belief
Gay people, lesbian people, etc.	Sexual orientation
Low-income people	Property

Source: own elaboration.

Statistical discrimination is group discrimination based on a fact that is statistically relevant. A classic example of statistical discrimination is not hiring a highly-qualified woman because women have a higher probability of taking parental leave. Instead, non-statistical discrimination occurs when the highly-qualified woman is not hired because she has said that she intends to have a child and take parental leave (Lippert-Rasmussen, 2013). If we disregard animosity, inherent to humankind, and consider that the algorithm considers any feature used by a learning algorithm as statistically relevant, we can say that algorithms can discriminate (Castillo, 2018).

A more precise definition of **algorithmic bias** -or algorithmic discrimination- involves the **systematic production of disadvantageous outcomes against socially salient groups**, particularly disadvantaged groups. This bias is embedded in the mathematical properties of an algorithm.

Algorithmic bias has been divided into two different types, depending on the stage of the machine learning process at which it happens (Danks and London, 2017). Firstly, biased models **can be biased due to the collection and use of biased training data** when training or modelling algorithms during the initial stages of development -in the processing stage - (Cowgill, 2019). Secondly, post-algorithmic or processing bias relates to the modelling of the system **caused by its interactions with users**.

There are four main steps in the detection of algorithmic bias:
1. Define an assignment of elements in the data to groups,
2. Define a protected group,
3. Determine a set of metrics aimed at measuring bias, and,
4. Measure and compare across groups.

The first step simply **sorts the data items into groups,** which can be overlapping ("soft" assignment) or non-overlapping ("hard" assignment). In most cases, the data items would correspond to people, and hence, the groups will be done on individual characteristics. Any characteristic of individuals can be used to create such groups, but particular attention is placed in *protected* characteristics. Protected

characteristics correspond to attributes of people that anti-discrimination law mentions.[7] These groupings are created in the data in order to evaluate the extent to which an algorithm may treat or affect a group differently from another.

The second step determines **which of the groups that have been defined will be protected**, meaning that the algorithm's application must not further disadvantage them and that the impact of the algorithms on them will be specially monitored. In some cases, protected groups are categories that are legally protected (e.g., people with disabilities). In other cases, the definition of what constitutes a protected group is related to a commitment that may not be legally binding, such as an intention to increase the participation of women or minorities who might be underrepresented in certain positions. In Foundations, we have selected gender as a starting point for algorithmic bias assessment based on data availability and the potentiality of this category for discrimination within the recommender system.

The third step determines the set of metrics to be used. In general, these metrics quantify the extent to which an algorithm **treats** people differently (disparate treatment) and the extent to which an algorithm has a different **impact** on different people (disparate impact). There are multiple and often overlapping definitions of metrics that should be used to evaluate algorithmic bias.

In the fourth step, after selecting the above metrics, the data is analyzed to obtain values and confidence intervals for these measurements. If the data goes through several steps in a system (such as data collection and data analysis), which is a common situation, the analysis is carried out for each step separately. The **computation of metrics is done by using a combination of existing libraries**, which are general-purpose and custom code for a particular purpose.

2.5.2 Recommender model

The next version of the recommender model will be implemented as a **mixture of a collaborative filtering algorithm and a contextual bandit algorithm**. A contextual bandit model makes predictions based on the state of the *environment*. It tries to identify the most appropriate content at the best time for an individual user. Collaborative filtering shows users items based on criteria like “Customers who viewed this item also viewed” or “Because you watched...” User profiles are constructed based on explicit ratings such as likes, and implicit ratings like viewing time. To come up with recommendations, the user profile is compared to other users’ profiles to find matches. Items rated highly by users with similar profiles but that have not been seen are then recommended to the user.

Recommender systems implemented totally or partially from collaborative filtering are well studied and they have a list of known biases and issues to pay attention to.

The type of bias we may find in recommender systems are:

- **Selection bias:** The algorithms used to predict user preferences are designed to have high prediction accuracy on the assumption that the missing ratings are missing at random, i.e., that there is no bias operating over which items are rated and which are not.

[7] Article 21 of the EU Charter of Fundamental Rights.

- **Cold-start bias:** Popular but older items are hard to avoid, and new things are harder to find. Likewise, the earlier an individual user gives a positive rating to an item, the more of an effect that item will have on their future recommendations, even if their tastes change or mature.
- **Popularity bias:** Very popular items are likely to get recommended to every user (and since recommendations make ratings more likely, popular items tend to increase in popularity).
- **Over-specialization:** Occurs when a recommender algorithm offers much more narrow choices than the full range of what the user would like.
- **Homogenization:** Recommendation algorithms encourage similar users to interact with the same set of items, therefore homogenizing their behavior, relative to the same platform without recommended content.

The current **recommendation model based on popularity** addresses the cold-start bias and even the selection bias by giving an adjusted probability to items/activities with few views. It is expected to have a popularity bias as it recommends the most viewed items, as well as a high homogenization. Homogenization of users' behavior does not correspond directly with an increase in utility. If we would like to know the impact of the feedback loop on the population, a global homogenization metric could be considered, calculated as a Jaccard index. Of all the recommender types studied by Allison et al. (2018), the recommender-based on popularity increased the homogeneity the most.

2.5.2.1 Metrics

Metrics have been selected due to their capacity to measure possible gender bias in the Foundations recommender system. This type of bias can be caused by the ability of the system to amplify existing social biases. **Bias disparity**, in this case, would relate to specific differences between input and recommendations targeting gender groups. A **jupyter notebook script has been provided to Koa** with the formulas from Bias Disparity in Recommendation Systems to measure the bias disparity.

We also propose to study the popularity of items per group definitions, the most recommended items per group definitions and the Gini coefficient. This metric is used to measure the inequality of a distribution, so the higher the Gini coefficient, the more unequal are the values in the studied distribution. Traditional recommender systems are expected to make popular items become even more popular and non-popular items become even less popular because a traditional recommendation strategy always shows the most relevant items.

2.5.2.2 Group definitions

Foundations does not collect any demographic data from the user, the only available data in that regard are the name and email address of the user. Thus, a study based on **guessed gender** from names is the only one we foresee as possible to conduct.

2.5.2.3 Measurements and results

As the audit team could not have access to the data needed in order to calculate gender bias disparity, Eticas provided a **script written in python as s jupyter notebook,** so that Koa can proceed with this analysis. The script runs by default on random datasets of users, activities, recommendations and ratings.

Data could not be accessed because of **privacy reasons since the data protection strategy was guided by the principle of data minimization** and did not contemplate auditing purposes within its legal basis for data collection. This has different implications in terms of algorithmic fairness and transparency. As already mentioned, proxies have shown to "revert" systems designed to be blind to special data categories (Barocas and Selbst,2016). Concealing algorithmic models to these data categories does not always prevent unexpected biases and can lead to algorithmic discrimination under certain scenarios (Corbett-Davies, 2017). Instead, gathering personal data or sensitive attributes could be seen as a strategy for detecting and possibly curing planned and unintentional biases. Consequently, determining the existence of legitimate risk factors for collecting subsets of data integrating these protected attributes should, therefore, be part of the technological development process.

Along these lines, **Koa should revise and reinforce its protocols regarding technical mechanisms and legal protocols for secure algorithmic audits** in those cases where the system at hand does not collect/use these categories of personal data. Collecting sensitive categories of personal data for these specific purposes should be based on a **risk assessment**. In order for this assessment to be properly conducted, Koa protocols should seek to integrate methodological instruments to search for indirect evidence of bias as part of preliminary phases of system audits, which could justify further data processing or collection. This strategy should cover the whole data processing, ensuring the integration of a legal basis for such data processing purposes (auditing algorithmic biases) since the very data collection process. It may focus on certain categories of personal data, applying the principle of data minimization, and should use pseudo anonymization data in all cases.

Therefore, besides conducting the **above gender analysis to collect indirect evidence of algorithmic bias or differential impact, other measures should be taken**. In order to address algorithmic bias in Foundations and ensure algorithmic fairness in future developments, a twofold strategy is proposed:

- ➔ Firstly, best practices concerning the balance between data minimization and algorithmic models documenting should be deployed. This includes developing strategies to avoid undesired blind spots in future technological projects. While reducing data collection categories corresponding to protected groups could minimize the risks of biases, it can also reinforce and obscure partial assumptions embedded in the model (Holstein, 2019). Not having disaggregated data regarding social collectives subjected to these biases, under certain contextual and technical conditions (Turner Lee, 2018), could create barriers for monitoring relevant machine learning processes, such as those derived from proxy bias. Two measures should be taken in this regard. On the one hand, creating robustly pseudonymized and statistically relevant training sub-datasets, including protected groups categories, may help to monitor algorithmic discrimination in the future. This does not necessarily involve integrating the same data collection categories into the system to be put in production, so the data minimization principle can still be followed. On the other hand, concrete legal frameworks and governance protocols within Koa teams for this purpose should be established (Baer, 2019). Different teams should intervene since the beginning of the development process, particularly during the model's design and training. They should systematically identify potential risks of biases and establish the methodology for secure pseudonymized data collecting/storing. The legal department should establish a case-by-case legal basis for data collection and design adapted consent and privacy policy protocols.

➔ Secondly, RCT studies addressing the system's functioning, including its efficiency and usability, could indirectly address algorithmic discrimination. Indirect evidence of discrimination could be collected by measuring the system's differential impact, including Foundations outputs (i.e., recommendations, activities) and outcomes (i.e., overall impact on well-being). This includes selecting subpopulations as part of the RCT allowing statistical representation of users. Then data collecting techniques adapted to link dependent, independent variables and indicators to these groups should be applied. Therefore the study design should consider these categories to assign participants into experimental and control groups randomly. The outcome variable being studied should also cover these disparate impact examinations. However, two limitations should be considered. On the one hand, the population that participates may not be clearly representative of the whole studied subpopulations. On the other, it should be noted that the qualitative and quantitative results of these analyses may not provide direct evidence of algorithmic biases since they will be able to capture the algorithmic model source of identified biases.

2.5.3 Heart and breath rates model

The model used to measure heart and breathing rate has reported general Mean Absolute Error (MAE) values from training of the algorithm that is based on a custom **dataset created at Koa**. The MAE is calculated as the average of the absolute errors between the estimation and true heart rate and breathing rates measured with ECG and respiration signals from a biosignal acquisition platform. The accuracy (ACC) is defined as the absolute error < a given threshold. When calculating the heart rate, the chosen threshold is 5 beats per minute. When calculating the breathing rate, the chosen threshold is 3 breaths per minute.

2.5.3.1 Metrics

In order to reflect on the possible bias the following steps should be followed:

- review the demographic data of the subjects
- calculate the MAE values per gender to identify any gender bias
- calculate the MAE values per kind of device (iphone, android high tier, android low tier) as proxy of economic status

2.5.3.2 Measurement and results

Device types are measured as a proxy for economic status. Six different devices were tested that correspond to high tier devices (iPhone11, Google Pixel, Samsung S10) and low tier devices (iPhone6, BQ Aquaris, Huawei). Results are also provided per females (F) and males (M).

Table 9. Heart Rate per device type

Device	Sitting MAE	Sitting ACC	Laying MAE	Laying ACC
iPhone 11	5.14	0.72	1.73	0.98

iPhone 6	5.44	0.7	1.48	0.97
BQ Aquaris	5.72	0.7	1.71	0.96
Huawei	6.14	0.65	1.45	0.96
Google Pixel	4.62	0.73	2.57	0.93
Samsung S10	4.66	0.71	1.57	0.96

Source: own elaboration based on Koa data.

Table 10. Heart Rate per gender by operative system

Device	Sitting MAE		Sitting ACC		Laying MAE		Laying ACC	
	F	M	F	M	F	M	F	M
Android	4.73	6.11	0.76	0.63	1.51	2.04	0.97	0.94
iOS	5.02	5.29	0.72	0.70	1.54	1.62	0.98	0.97
Average	4.88	5.70	0.74	0.67	1.53	1.83	0.98	0.96

Source: own elaboration based on Koa data.

Table 11. Breathing Rate per device type

Device	Sitting MAE	Sitting ACC	Laying MAE	Laying MAE
iPhone 11	3.31	0.66	2.17	0.82
iPhone 6	2.98	0.69	2.42	0.8
BQ Aquaris	3.27	0.73	2.62	0.76
Huawei	2.9	0.69	2.54	0.76
Google Pixel	3.46	0.64	2.34	0.79
Samsung S10	2.92	0.68	1.79	0.82

Source: own elaboration based on Koa data.

Table 12. Heart Rate per gender by operative system

Device	Sitting MAE		Sitting ACC		Laying MAE		Laying ACC	
	F	M	F	M	F	M	F	M
Android	3.42	2.95	0.64	0.67	2.35	2.73	0.78	0.78
iOS	2.80	3.30	0.70	0.66	2.48	2.48	0.87	0.77
Average	3.11	3.13	0.67	0.67	2.61	2.61	0.83	0.78

Source: own elaboration based on Koa data.

Summary of results

- No significant differences are found across genders and operative systems/type of device.
- The implemented **models work better when lying, especially the heart rate model**. This should be assessed in future versions of the model. Users should be properly informed about relevant efficacy rates differences, as in this case, before doing the exercises.

2.6. Data Management assessment

Issues concerning **privacy by design in Foundations** relate to health data protection in storage, communication, and usage. Similar technologies have shown privacy-related problems such as inadequate measures to avoid gathering health-generated data without making explicit transmission to clinicians or other companies. Moreover, the collection and treatment of geographic and geospatial data and approaches to balance access and privacy have been shown to have several limitations to protect users' identities (Lane and Schur, 2010). It has been stressed that purpose limitation regarding the use of biomedical data, adapting consent to potential reuse of personal data, is not always guaranteed (Vayena et al., 2016).

Taking the above into account, this assessment's main aim is to evaluate the system's **capacity to protect users' privacy**. In the case of Foundations, this should mostly be achieved by providing by-design mechanisms to avoid data breaches' negative externalities. This section will address privacy in Foundations focusing on data management and compliance with data protection principles of purpose limitation, data minimization, and security. The analysis focuses **on existing instruments used for data processing. It takes data protection principles reflected in the GDPR as a reference for ethical analysis without assessing legal compliance**.

2.6.1 Data governance, lifecycle and risk management

Koa is the **controller** in Foundations. **Processors**, who process personal data on behalf of the controller, include those companies providing their employees with the Foundations' service. Personal data is also shared with other processors (third parties), such as services providers such as Facebook, Linkedin or Google, with purposes hosting, providing customer support, analytics or application functionality such as notifications.

Data minimization has been implemented in Foundations. No demographic data are collected. Only **first name and email** are used for onboarding and stored by the app. However, the system uses special personal data categories, particularly data concerning health, including heart rate, breathing rates and psychological status. The **complete categories of personal data** to be collected by the system includes all the types reflected below. Data collection purposes corresponding to each of these categories and the legal basis for their processing are listed in the following Table.

Table 13. Collected data and legal basis of the processing

Collected data	Concept	Legal basis of the processing

Email and first name.	App provision	Performance of a contract
Activity data, such as how often and for how long users use the App, how they navigate between screens, the activities they use, and which screens they spend more time on.	Improve App (including aspects related to performance, navigation, availability and usability)	Legitimate interest
Contact data (to send information).	Marketing	Legitimate interest
Information from user interaction with the app, including open questionnaires.	Help manage stress	Consent
Information from user interaction with the app.	Personalized notifications based on activity	Consent
Breathing rate or heart rate.	Body tracking	Consent

Source: own elaboration.

While emails and first names present a high risk of identification, reidentification using only information about navigation or personalized notifications is less likely to occur. Even body tracking such as heart data[8] information has shown to offer some possibilities of reidentification when combined with other data. Stored pseudo-anonymized data, such as location or access codes, or usernames and emails, **could lead to individuals' identification.** Processors could also link their users' identity to special categories of personal data, in particular, "health data" defined by the GDPR (Article 4, 15) as:

"'data concerning health' means personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status".

Moreover, data collected will contain the users' notes, reflecting users' thoughts and feelings, which may include personal identifiers. These journal records are reflected in the module "Working with thoughts" and the section "Keeping a thought record"[9]. Likewise, the system managers could **achieve reidentification** based on the above data by using different mechanisms. A twofold strategy should be established in this context. On the one hand, given the nature of diaries mainly oriented to self-assessment, it is recommended not to transmit this data to Koa servers and keep it only on users' devices. On the other hand, this privacy-enhancing capability of the system could be pointed out in the Privacy Policy to stress the application of security mechanisms for ensuring data security.

[8] See Wang L et al. (2017), Unlock with Your Heart: Heartbeat-based Authentication on Commercial Mobile Phones. Available at: https://dl.acm.org/doi/10.1145/3264950

[9] It should be noted that these text data is not processed using NLP. A specific algorithm audit is recommended in case an NLP algorithm is used in the future to classify these notes or further personalized treatment.

However, following Koa's instructions, service providers will have access to a limited set of personal data, which they are obliged to erase "right after their services are finished"[10]. In the case of companies contracting Foundation, access to users insights will not include personal information such as name, email addresses or text entered into the app. Still, reidentification could also be achieved under certain circumstances. The signup and usage time information presented in the dashboard could be screened to identify employers able to access the app in a certain period of time (daily rates are provided). Depending on the rollout of the app, the size of the group of employees using the app and the pace of signups, there could be a potential risk to identify early adopters, their time of consumption of the app and their main motivations to use it, ranging from difficulties relaxing to low self-esteem. Instead, once a large number of employees operate the app, employers' **potential risk of re-identification in workplace contexts can be judged very low**.

Still, special and proportional safeguards should be implemented for the treatment of personal data in Foundations. **Privacy by design measures** should be oriented towards minimizing the risk of unauthorized access to app provision and managing stress data. Measures should be taken to inform users about the need for avoiding to provide personal information through questionnaires. In the following sections, we will analyze the three main instruments designed to ensure this: The Privacy Policy, the data management within the system and its security mechanisms.

2.6.2 Privacy Policy assessment

The following Table shows the analysis of the main aspects to be addressed by the Foundations privacy policy.

Table 14. Privacy policy observations

Requirement	Definition	Observations
Data controller-DPO	Koa is the data controller and has designed a DPO.	All required information is provided. The Data Protection Office contact is *dpo@Koahealth.com*.
Data processors and third parties	Companies hiring the system are presented as data processors. Their role is explained as follows: "*Only Koa and its sub processors, following its instructions, will have access to your personal information as described in this Privacy Policy. Where the App is offered by an employer (Customer) to its employees, Koa may provide aggregated insights related to usage of the App, so that they can understand its impact.*".	Processors include companies and other organizations offering the service. Their categories and the aims of data sharing are clearly explained.

[10] See Koa Privacy Policy at: https://foundations.Koahealth.com/privacy-policy-web/

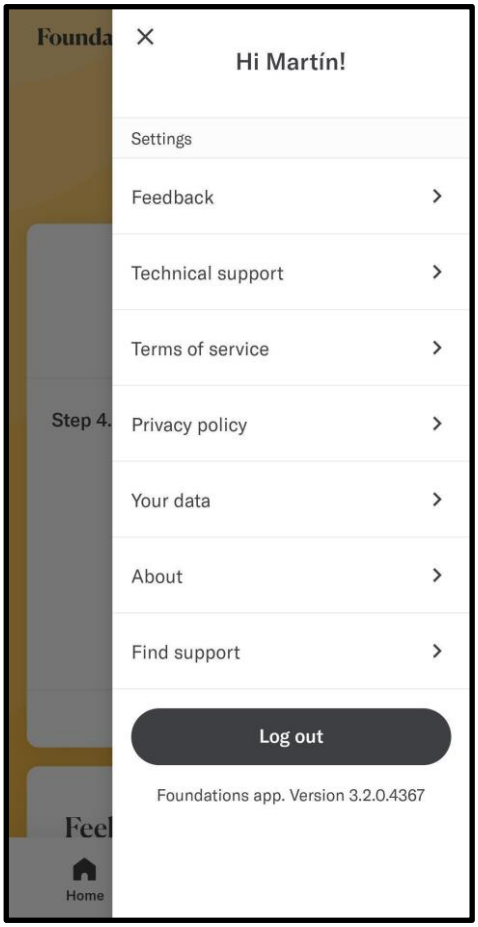
	Data sharing with third parties, also acting as processors, is presented as follows: *"We may share some of your personal data with service providers for specific activities such as hosting, providing customer support, analytics or application functionality such as notifications."*	
Purposes of data collection	Purposes of data collection are properly detailed in section 2. *"Why do we collect personal data about you and what do we do with it?"*. Purpose limitation is justified under security tools-mechanisms: *"We protect all communications between the App and the servers in line with best practice by using TLS for encryption and server authentication. We use ISO 27001 certified systems in order to protect your registration information including email and password. We store your personal data in an encrypted database."*	Data collection purposes are explained in a disaggregated manner. Data management strategy and security measures to ensure purpose limitation are also clarified. The information provided concerning **some data collection purposes could be further specified**, including: ● *"Provision of basic App services".* In this case, categories of processed personal data could be provided. ● *"Marketing"- "We process your contact data to send you information about our services or products. We may use third party services to facilitate communication*" The type of personal data shared with these third parties could be described.
Basis for the processing	As stated above, informed consent, performance of a contract and legitimate interest are used as legal basis for the processing of different categories of personal identifiers or pseudo-identifiers. Regarding informed consent, the protocol includes the following statements: a."Your consent is the basis for the collection and process of personal data to manage your stress..." b."Explicit consent is given to the main purpose of the app. The privacy policy is accepted when creating the account."	**Improvements regarding consent** may include: For a: It could be rephrased to show that technology supports users so they can manage their stress, not that the technology does it. For b: The onboarding process could be refined. Superficial presentation should be avoided, so the onboarding process could explicitly ask users to read the full PP. Since informed consent should involve informed and affirmative action, it is recommended to ask data subjects to open the PP before using the app. This could be promoted by design, including a click box at the end of the PP.

		At the end of the consent section, the right to withdraw and how to achieve it should be repeated. Lastly, the need for specific consent materials for those individuals with accessibility problems should be evaluated (See Table 7).
Data Erasure	Data retention periods for each type of personal data are clearly detailed in section 5. *"How long do we keep your data?"* In case of inactive users: *"If you are not active in our App, we will erase your data after 12 months from last access."* The right to withdraw from processing at any time is included.	Regarding the **12 months data retention period**, it should be mentioned whether this applies to all personal data. This section could also repeat how individuals' rights are protected if some **data is not erased by service providers** following the specifications reflected in Standard Contractual Clauses or Privacy Shield certificates (section 4 of the PP).

Source: own elaboration.

In terms of **transparency and control**, ARCO rights are clearly explained in the Privacy Policy so users can exercise them. Opt-out options for each of these services are included in the Cookies Policy. Koa's responsibilities as a controller over these data processors and the existence of binding contracts (Standard Contractual Clauses) are described in the PP. Moreover, information options and instruments (Privacy Policy, Terms of Service, Technical Support and Your Data) to facilitate or enforce ARCO rights available in the app, as it can be seen below are multiple and accessible.

Figure 4. Foundations settings



Source: Koa.

However, while the following is presented as the information collected for improving users experience, the PP offers insufficient detail about the actual data collected in this framework:

"We process information to improve the user experience. Based on analysis of how users use the App we can make judgements like if loading times are slow, or if information is too hard to find, and use this to improve the user experience."

"We may share some of your personal data with service providers for specific activities such as hosting, providing customer support, analytics or application functionality such as notifications. We only share the minimum information and authorize our service providers to process your information following our instructions. We make sure that our service providers erase all your personal information right after their services are finished."

The same limitation is found regarding the information collected through cookies: "*User activity in the App: Frequency of access to the App, time spent on different screens, functions used etc.*". Both policies **could include more detailed lists of categories of personal data shared with processors**.

2.6.2.1 Readability

The readability of the privacy policy available at https://foundations.Koahealth.com/privacy-policy-app/ (effective from November 5, 2020) is checked against different indices using the open source

Python library "Readability" available at https://github.com/andreasvc/readability/. As section *5.3.2 Ethics in users' access* introduced, readability is checked against different readability indices that estimates the years of education needed to understand a piece of writing. **Foundations app privacy policy has an average U.S. grade level of 11 (16-17 years old)**. Results show readability based on Flesch Reading Ease is 10th to 12th U.S. grade (15-18 years old, fairly difficult to read), and readability based on the average U.S. grade level is 11th grade (16-17 years old, high school - junior).

Table 15. Privacy policy statistics

Text Statistics	
No. of sentences	148
No. of words	2567
No. of complex words	430
Percent of complex words	16.8%
Average words per sentence	17.34
Average syllables per word	1.54

Source: own elaboration.

Table 16. Privacy policy readability

Readability Indexes	Score / Grade	Ages[11]
Flesch Kincaid Reading Ease	58.8	15-18
Flesch Kincaid Grade Level	9.37	14-15
Gunning Fog Score	13.6	18-19
SMOG Index	12.3	17-18
Coleman Liau Index	11	16-17
Automated Readability Index	10.1	15-16
Average U.S. Grade Level	11.3	16-17

Source: own elaboration.

2.6.3 Security of personal information

Securing confidentiality is key for ehealth tools supporting mental wellbeing and providing treatment in this domain (Torous and Roberts, 2017). Along these lines, the **anonymity of users** for companies and respect of purpose limitation contractual clauses by third parties are vital for ensuring high ethical standards. Foundations does not share any personal identifiers with both customers and third parties. However, to minimize the risk of re-identification using pseudo-identifiers or combining metadata,

[11] A table to look up ages for the different U.S. Grade Levels can be found at https://en.wikipedia.org/wiki/Education_in_the_United_States#Educational_stages

safe data storage and communication are of utmost importance. Data communication between app and server is done with a TLS secure communication. Data is stored in an encrypted database. However, it is not a zero-knowledge system. Koa Health uses ISO 27001 certified systems in order to protect registered information, including emails and passwords. Moreover, symmetric 256-bit encryption, RSA public-key and SHA-2 algorithms are used, which ensures high data integrity.

2.6.4 Summary of data management issues and recommendations

The following Table summarizes the data management issues found, the strategy adopted by Koa for their implementation and related recommendations. These lines of action will be reviewed and adjusted on the basis of Koa feedback and results of the second phase of the audit to be conducted in February 2020.

Table 17. Foundations' data management assessment

DM-related issues	Foundations strategy	Recommendations for improvement
Unauthorized access to personal data by employers/third parties. Users' re-identification	Foundations has applied data minimization and robust security measures to enforce purpose limitation. The Koa team has implemented a pipeline to pseudo-anonymise data and an NLP algorithm is applied to remove personal data such as names, emails or telephones from the open questionnaires. Closed questions are included at the end of open questionnaires. This data is used to categorize open texts.	Make open questionnaires not available to employers/third parties by **storing them locally** on the user's device or creating a zero-knowledge system. **Assess the risk of offering the service** to companies with less than 5 employees. Reduce the dataset shown in the dashboard in the case of companies with less than 10 employees.
Privacy Policy	Full description of purposes and data sharing, including basic required information: personal information collected, the categories of third parties with whom Foundations shares the information, how users can review and request changes to their information, how Koa notifies users of material changes to the privacy policy and the effective date of this policy. Good readability standards for	Provide **further information as part of the onboarding process** (summary of the PP). Provide **further information** about personal to be shared as part of Provision of basic App services and Marketing. Stress the sharing of personal data with third parties and processors in "data treatment" and explain the categories of data.

	targeted users.	**Repeat standard clauses** protection in case of not removing certain categories of personal identifiers. **Provide full categories** of personal data used by cookies. Assess **readability of text** to comply with Koa's reading age objective
Informed consent	Consent provided during the onboarding process is based on complete information, including description of the system, risks and benefits, confidentiality, contact information and voluntary participation. Opt-out mechanisms regarding specific data processing activities are also provided.	Repeat the **right to withdraw** at the end of the consent section. Provide **different models and strategies for consent targeted to users with disabilities** (See Table 7).
Security	Koa Health applies 27701 ISO/IEC. Every three months they perform an internal platform security assessment to identify risks and propose mitigations.	In these regards, Koa Health uses best practices.
Data erasure and reuse of Foundations data	Data retention linked to an identified user is one year. Data subjects can request the removal of this data. Koa has automated scripts to perform such requests. Personal data is automatically processed in a pipeline to strip any location data and hash the user and device id with a salted key and a sha256 algorithm, thus converting it to pseudonymised data. Pseudonymised data retention is unlimited.	In these regards, Koa Health uses best practices.

Source: own elaboration.

2.7. Desirability and Acceptability assessment

This section will examine **the desirability and acceptability of Foundations** based on analysing these concepts' most relevant drivers. With this purpose, we will assess the system against the critical state of the art issues and analyze the outcomes of usability evaluations conducted by Koa. The analysis

focuses on aspects that may affect the app's perception, ease of use and ethical factors that may influence technological adoption.

2.7.1 Desirability and acceptability grounds and issues

The analysis of acceptability is critical for ensuring users engagement and the alignment of ehealth systems with social and individual needs. Four main issues have been defined in this framework. Firstly, it has been indicated that a **good performance of mHealth systems** is a crucial driver for their adoption by employees (de Korte et al., 2018). Still, going beyond clinical efficiency by addressing ehealth related solutions usability and contextual social factors determining technological deployment as part of their validation has been recommended (Aryana et al., 2019, Price et al., 2014). In this framework, testing the reach of such systems capabilities is considered a crucial requirement. Some indicators to be considered when conducting these studies include:

- publications in peer-review journals,
- tests with healthy and unhealthy populations (size),
- other studies of validation and clinical effectiveness (Safavi et al., 2019: 120).

Foundations has been assessed following the above recommendations regarding the need for evidence concerning health-related apps' efficiency. A four-weeks study was conducted in the UK with 136 participants. The 2-armed randomised controlled trial compared an app-based intervention (Foundations) to a non-intervention control group. Participants were between 30 and 50 years old (two groups 30-40 and 40-50), male and female, with moderate to high levels of perceived stress, mild to severe anxiety and none to moderate sleep problems. The analysis has shown **significant improvements compared to the control group on measures of anxiety, resilience, sleep and mental wellbeing**. Future assessments, with other samples' stratification are planned. On the above basis, information about these systems' effects and their relevance should be clear and adequately provided to users.

Secondly, another driver for acceptability is **trust**. Specific factors framing organizations and employees trust regarding workplace health promotion technologies have been identified, including functionality, visual design, security and outcomes communication (Stoyanov et al., 2015; Vithanwattana and George, 2017; Heffernan et al., 2016). Foundations communication materials follow this orientation since they explain that the system targets **healthy users with the promise of maintaining health and preventing mental disorders**. The requirement to be concise and specific in the definition of the system scope is therefore addressed in Foundations communication instruments, including its Terms and Conditions:

> "The App is intended to help you manage your stress by allowing you to have a better idea of how stressed you think you are and providing you access to content that may be of interest to you.The App has not been developed to meet your individual requirements."

Moreover, **privacy breaches** involving health-related data are perceived as a high risk, which significantly determines trust in technology (van der Graff et al., 2015). This factor's impact acquires higher importance within ehealth implemented at the workplace. In these contexts, already

mentioned stigmatizing elements are combined with fear of losing jobs, which in case of data misuse can contribute to putting the whole treatment in jeopardy (Jimenez and Bregenzer, 2018).

Thirdly, relevant aspects potentially harming the system's acceptability relates to the processing of **confidential or privacy-protected data** could negatively affect how the app supports individuals (Yang, 2016; Buckovich et al., 1999; Wynia et al. 2011). Technological aspects that may be impacted by users' behaviours aimed at protecting their privacy include overall precision, fastness, targeting level and ease of use (Laur 2015; Motti and Caine 2015). In this regard, personal data should be kept confidential for the system's administration as a mechanism to foster trust between users, the system owners and concerning the scientific basis of the health system at stake (Yaghmaei and van de Poel, 2017).

Along these lines, an important contextual factor influencing the system's desirability and acceptability is how it is integrated **into private organizations working relations and data management protocols**. As already mentioned, **trust in Foundations is clearly associated with the perceived relative risk of access to sensitive personal data** by both employers and service providers. So, ensuring anonymity and secure administration of sensitive data will be fundamental to guarantee the system's desirability. Moreover, its acceptability partially relies on **users' understanding** of the possible influence of sharing their data with the app over their job status or personal life. This includes users' perceptions about the potential implications of Foundations for monitoring employees' psychological status or stigma on users derived from disclosing information produced in the interaction with the system among supervisors or other employees.

Foundations has different mechanisms to raise awareness about Foundations' **implications for users' privacy** regarding its implementation working environment. For instance, the Privacy Policy states:

> "These insights will never include personal information and your employer will not be able to know your name, email address nor see any raw data you have entered into the App."

Fourthly, the framing of recommendations should be assessed under the light of **socioeconomic and labour contexts** that may limit individuals' autonomy to conduct certain practices or adopt new behaviours. Certain governance structures have shown to be more permeable than others to promote these technologies among employees, revealing top-down barriers and also poor implantation due to lack of targeted strategies (Farrell et al., 2016; Kaipainen et al., 2017). This means that organizations could raise specific organizational or resource limitations for using technology, such as limited access to information about the system or restrictions to its use during working hours.

Moreover, while it has been found that low socioeconomic status adult and young users tend to positively value ehealth apps notifications more often than higher socioeconomic groups (Cremers et al., 2014; Vries, 2011), workers in the low-income sectors could experience less autonomy at work. This factor could influence practices regarding the use of Foundations. For instance, in the UK, caring and recreation workers, and workers within elementary occupations, receive on average less than half

of median earnings[12] than managing directors or senior officials -and also work more hours on average[13]. This combination of factors could condition technological adoption. This can be illustrated with certain recommendations; 8 minutes audios could not be used in stressful situations within certain labour conditions or jobs. Even though the app may be targeted to be used outside working time, alternatives to relaxation recommendations requiring time availability could be considered taking these factors into account.

2.7.2 Usability and concept testing

As mentioned above, Foundations is being tested with human participants. Concerning usability, which is also associated with acceptability under the Technological Acceptance Model (TAM), two types of usability tests have been conducted:

a) **Face to face testing.** The Foundations study sample included:

- Age range: core working age **of 25 to 55**.
- Other exclusion/inclusion criteria depending on the app's focus: testers should **have experienced stress, depression or anxiety**.
- **50:50 split of men and women**, with no exclusion of non-binary (if one round of testing has been skewed then Koa asks for a compensatory bias in the following round).

No other characteristics were requested.

b) **Online testing. Beta users for the apps** were recruited to test the app's experience as a whole by using adverts on social media. No specific characteristics or personal data are requested for online tests.

Two surveys were conducted (August and November/December 2020). These studies show **high acceptability in terms of usability** (with >95% of users scoring it a 7 or above)**.** The top reasons users chose to use Foundations were ease of use, single activities, ability to use content relevant to users. In line with the study cited above, people found the app helpful for (in priority order): a. Relaxing/switching off and Stress management, b. Worrying less, c. Feeling more in control, and d. Improving sleep. In terms of the app's qualitative review, most respondents found the system to be user friendly but provided some negative insights about the music used for relaxation.

Moreover, Mass General Brigham (MGB), a Koa partner organisation in the US, **surveyed** 19 users in 2020, showing an overwhelming majority of answers expressing positive and very positive impressions about the app. Ease of use was also valued positively. Most users also pointed out that the system helped them with their mental and overall wellbeing. However, its impact on sleep quality was not considered relevant.

[12] Data available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2019#employee-earnings-and-hours-worked

[13] Data available at: https://www.unionlearn.org.uk/compare-average-hours-job

Although these studies are very informative in terms of usability, in line with their main aims, they do not provide details about socioeconomic or cultural groups regarding differential impact or reception of the system, which could be used to consider acceptability in a broader sense.

2.7.3 Summary of desirability and acceptability issues and recommendations

The following table presents the most important acceptability and desirability issues examined in this report, the Foundations strategy for tackling them and recommendations for improvement along these lines. These recommendations have been adjusted on the basis of Koa feedback.

Table 18. Foundations' data management assessment

Acceptability and Desirability -related issues	Foundations strategy	Recommendations for improvement
Foundations performance assessment and communication	The usability and efficiency of the app is being tested. Effectiveness across different social groups is expected to be tested in the near future. Part of these examinations' results communicated in the app web and app.	**Continue integrating** results of evidence-based studies into the app communication.
The scope of the app should be focused on wellbeing	The concept of the system is being focused on healthy individuals. The app PP indicates: "*The App and any information and/or services provided by the App are not intended to be used in the detection, diagnosis, prevention, monitoring, prediction, prognosis, therapy, treatment or alleviation of any condition, disease or vital physiological processes or for the transmission of time sensitive health information".*	To reinforce this information, **protocols for ensuring a proper presentation of the system may be provided to employers** so they can integrate them into their communication with employees.
Consideration of contextual socioeconomic and labour relations factors affecting reception	Usability studies are not integrating job-related contextual factors as variables for the analysis. However, this issue will be considered in empirical studies conducted by Koa.	**Study differential acceptability across job position, working sectors and socioeconomic demographic groups**. Consider results from these studies within the system communication with users.
Privacy and confidentiality	Privacy Policy, Consent, Terms and Conditions and information about data management integrated into the app provide comprehensive details about the type of personal data used	In line with conclusions of Section 7, provide further information about **personal data categories** and produce targeted versions of consent protocols.

	by the system and the main data processing purposes.	Further communicate privacy-enhancing methods used by Foundations to users.
Users' control	This is addressed by integrating information about users' data management and the scope of the app within the app content and the Privacy Policy.	In those cases where the company is offering the app to its employees, integrating the system within the framework of **labour relations** could create suspicion and therefore harm trust in this technology. Consequently, these factors could harm Foundations usability and strategy for exploitation. Possible problems in this regard involve communication issues between hiring companies and employees. As stated above, Koa could **assess organizations-users communication impact** through user experience and acceptability analysis.
Stigma harming users' integrity and acceptability	Same as above.	Same as above.

Source: own elaboration.

2.8. Summary of conclusions and recommendations for Foundations

The following Tables summarize the main issues found concerning the four dimensions of the audit, ethics, data management, desirability and acceptability. They also reflect the most significant recommendations produced in terms of existing gaps between requirements and already implemented solutions. This includes the full results corresponding to the Foundations algorithmic audit and prioritized recommendations. High-priority categorises the recommendations associated with higher risks in the short term and/or which are blockers for other recommendations to be implemented. Low priority recommendations correspond to less problematic or urgent issues.

Table 19. High priority findings and recommendations

Assessment	Main findings	Most significant recommendations
Ethical assessment	Lack of evidence-based studies. Need to assess the system impact on	Already conducted research has used large samples, but they tested the effects of Foundations on users during a short period of time. These studies should

	specific groups of users and regularly.	also **consider the differential impact of the app concerning different job positions and working sectors**.
		Accessibility regarding vulnerable groups (including different disabilities) should be tested.
	Ensure informed consent and informational mechanisms for vulnerable groups. Explainability of algorithms should be achieved in this framework.	Differential explanation and consent mechanisms could be **further developed to facilitate access to vulnerable collectives**. In particular, this includes consent for disabled people (deaf, blind, others).
	Guarantee lack of organizational coercion regarding the use of Foundations.	**Foundations anonymity-related capabilities should be reinforced in the public presentation and onboarding of the app** to foster trust and address employees' reluctance to share information about their mental status revealed by the literature.
Data management assessment	Unauthorized access to personal data by employers/third parties. Users' re-identification	**Assess the risk of offering the service** to companies with less than 5 employees. Reduce the dataset shown in the dashboard in the case of companies with less than 10 employees.
	Privacy Policy	Provide **further information** about personal data to be shared as part of Provision of basic App services and Marketing.
		Provide full categories of personal data used by cookies.
	Informed consent	Provide **different models and strategies for consent targeted to users with disabilities**.
Desirability and acceptability assessment	Consideration of contextual socioeconomic, cultural and labour relations factors affecting reception and adoption	**Study differential acceptability across job position, working sectors, ethnic belonging and socioeconomic demographic groups**. Koa should **assess integration of the system within the framework of labor relations** through user experience and acceptability analysis. Consider results from these studies within the system communication with users.
Algorithmic impact assessment	No indirect evidence of bias has been measured.	As stated above, the data minimization strategy applied to Foundations could lead to hiding possible sources of bias, contributing to opacity and limiting algorithmic audit. To address these issues in

		Foundations and future Koa developments, three recommendations are provided: 1. **Implement the methodology described in Section 4** to collect indirect evidence on algorithmic bias and establish the need for a full automated-processing assessment. This means conducting an internal gender bias audit based on Eticas inputs. 2. The second strategy concerns creating a Koa protocol to ensure data availability on protected attributes needed to assess algorithmic bias. 3. Lastly, it is proposed to **indirectly assess possible biases through RCTs**, including the analysis of outputs (activities) and outcomes (well-being status) by each demographic group.

Source: own elaboration.

Table 20. Medium-level priority recommendations

Assessment	Main findings	Most significant recommendations
Ethical assessment	Lack of evidence-based studies. Need to assess the system impact on specific groups of users and regularly.	**Long term impact** should also be measured.
	Ensure informed consent and informational mechanisms for vulnerable groups. Explainability of algorithms should be achieved in this framework.	Almost no information about automated processing is being provided. The **how and why of algorithmic processing should** be presented in the Privacy Policy.
	Discrimination	Graphics used for illustrating the activities could be **more diverse in terms of ethnicity and age**. They could also further consider the **accessibility of people with disabilities**. It is recommended to take these drivers for plurality into account.
Data management assessment	Unauthorized access to personal data by employers/third parties.	Make open questionnaires not available to employers/third parties by **storing them locally** on the user's device or creating a zero-knowledge

	Users' re-identification	system.
	Privacy Policy	Provide **further information as part of the onboarding process** (summary of the PP).
		Stress the sharing of personal data with third parties and processors in "data treatment" and explain the categories of data.
		Assess **readability** to comply with the age limit.
	Informed consent	Repeat the **right to withdraw** at the end of the consent section
Desirability and acceptability assessment	Foundations performance assessment and communication	**Continue integrating** results of evidence-based studies into the app communication.
	The scope of the app should be focused on wellbeing	To reinforce this information, **protocols for ensuring a proper presentation of the system may be provided to employers** so they can integrate them into their communication with employees.
	Privacy and confidentiality	Further communicate privacy-enhancing methods used by Foundations to users.

Source: own elaboration.

Table 21. Low priority recommendations

Assessment	Main findings	Most significant recommendations
Ethical assessment	The system does not replace clinicians' roles nor provide clinical treatment, so it should be presented as an adjuvant and self-assessment tool designed to reach well-being.	**Communication regarding the system's limitations** could be reinforced within the app introduction. Foundations lack of a duty of care could be underlined in this context.
	Ensure informed consent and informational mechanisms for vulnerable groups. Explainability of algorithms should be achieved in this framework.	Address **readability** in the final version of the system recommendations to ensure homogeneity and easy access.
	Addiction	It is recommended to **consider user investment, rewards (in particular, the final set of notifications) and possible gamification** as relevant aspects to be

		examined in future addiction assessments.
Data management assessment	Privacy Policy	**Repeat standard clauses** protection in case of not removal of certain categories of personal identifiers.

Source: own elaboration.

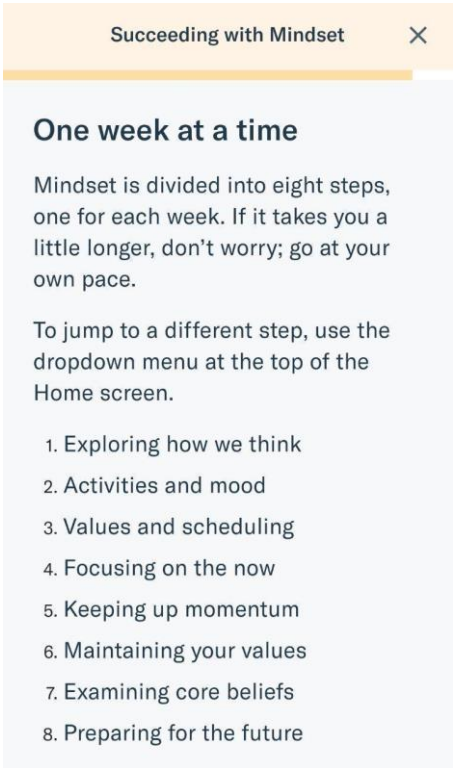
3. Audit of Mindset

Mindset is an **app that offers an eight-step program** based on Cognitive Behavioral Therapy (CBT) principles to help individuals manage their symptoms of depression. Individuals work through exercises to better understand and manage their mental state while providers, including health organizations, can track their progress through an integrated dashboard.

The content is based on the work of Dr Sabine Wilhelm, Chief of Psychology and Massachusetts General Hospital, and her team. It consists of CBT broken into structured elements that are delivered to users in easy-to-digest parts. The app is designed only **for use by patients** who have received a diagnosis of depression by a clinician and remain **under that clinician's care**. To facilitate that ongoing patient-clinician relationship, Mindset also has a **clinician dashboard** separated from the app.

Therefore, on the one hand, the system provides a secure, HIPAA-compliant program so that individuals can access it through their mobile phones. This includes the **eight-step program shown in the image below that will count** with chat-based support, educational content and exercises. On the other hand, it also consists of a streamlined dashboard to help providers manage patients and prioritize care among them.

Figure 1. Mindset presentation



Source: Koa.

The assessment is based on four primary data collection techniques: a) a literature review concerning the use of technologies in the domain of Mindset; b) the review of Koa documents describing technical specifications and assessing ethics compliance of the system; c) Interviews with the Koa team focusing

on different aspects of the system[14] and d) a thorough evaluation of its functioning and data management. Since Mindset is currently under development, the analysis has been oriented towards complementing Koa self-assessment and providing guidelines for the system's ethical design, management, and testing.

3.1. Theoretical framework and social context for the model

3.1.1 The Cognitive Behaviour model

Mindset therapeutic models and content are based on CBT psycho-social intervention. CBT is a **directive, time-limited and structured** approach used to treat several mental health diseases (Fenn and Byrne, 2013). Within this therapeutic approach, clinicians mostly use a combination of cognitive and behavioural observations about the patient to decide when and how to intervene in the cycle that goes from certain cognitive appraisals and related emotions to specific behaviours and events (Wright et al., 2006). The relationship between patient and clinician is based on the **collaborative empiricism model** oriented towards developing a cooperative therapeutic relationship. Under this general premise, the therapy's effectiveness is clearly related to its relative capacity to identify and develop strategies to influence both cognitive and behavioural pathologies (Wright, 2006). Evidence about this efficacy has been recently underlined for the case of cognitive behaviour therapies (Thoma et al., 2015; Layard and Clark, 2014).

CBT has two key levels of intervention. On the one hand, examining patients' *automatic thoughts*, defined as "often private cognitions that flow rapidly in the stream of everyday thinking and may not be carefully assessed for accuracy or relevance". On the other hand, *maladaptive schemas*, considered as "fundamental rules or templates for information processing that are shaped by developmental influences and other life experiences"(Wright, 2006:174). The clinician-patient **reconstruction of these mental frames** allows them to identify how they manifest and operate in daily life.

On this basis, **different techniques are applied to intervene at both cognitive and behavioural levels**, which has shown to be helpful to cope with social problems and treat depression and reverse anxiety disorders (Wright, 2006). Cognitive techniques used in CBT include Socratic questioning, guided discovery, examining the evidence or examining advantages and disadvantages. Behavioural techniques also consist of a broad set of methods such as graded task assignments, exposure and response prevention, relaxation and breathing training, or coping cards. CBT has a typical 12–20 session format, but this model has been adapted to different social scenarios and modified to treat other pathologies such as Borderline Personality Disorder (BPD) (Thoma et al., 2015: 439).

However, research has Identified that CBT outcomes are often modest to average, benefits may not persist in the long run and some patients derive limited or no advantages (Lambert, 2011; Rey et al.,

[14] **Roles interviews** were: Strategic Director and Head of Ethics, Project Manager, Service Design Strategist, Director of Cyber Security and Clinician (Dec'20 - Jan'21).

2011; Vittengl et al., 2007). These studies call to assess the effectiveness of CBT regularly and across different groups of patients.

CBT has been offered through **computer-assisted psychotherapy** for almost two decades (Wright, 2004). Since the beginning, it has shown promising results for treating depression (Proudfoot et al., 2003) and different types of anxiety disorders (Rothbaum et al., 2000; Kenwright et al., 2001). Computer-delivered CBT for depression has been recently studied and validated through several investigations (Gumport et al., 2016). Some studies have demonstrated that these treatments can mitigate symptoms of depression with medium to large effect sizes (Andersson & Cuijpers, 2009; Andrews, Cuijpers, Craske, McEvoy, & Titov, 2010). Positive results have been identified even six or more months following therapy (Andersson et al., 2013; Andrews et al., 2010). Acceptability of these systems has also been reflected in a high degree of adherence to the therapy for some technological solutions (Andrews et al., 2010; van Ballegooijen et al., 2014).

However, it should be noted that computer-delivered CBT has shown less smooth adoption within some social contexts and for some mental diseases, for instance, concerning high dropout rates of some systems used for depression (Andersson et al., 2005; Andersson and Cuijpers, 2009).

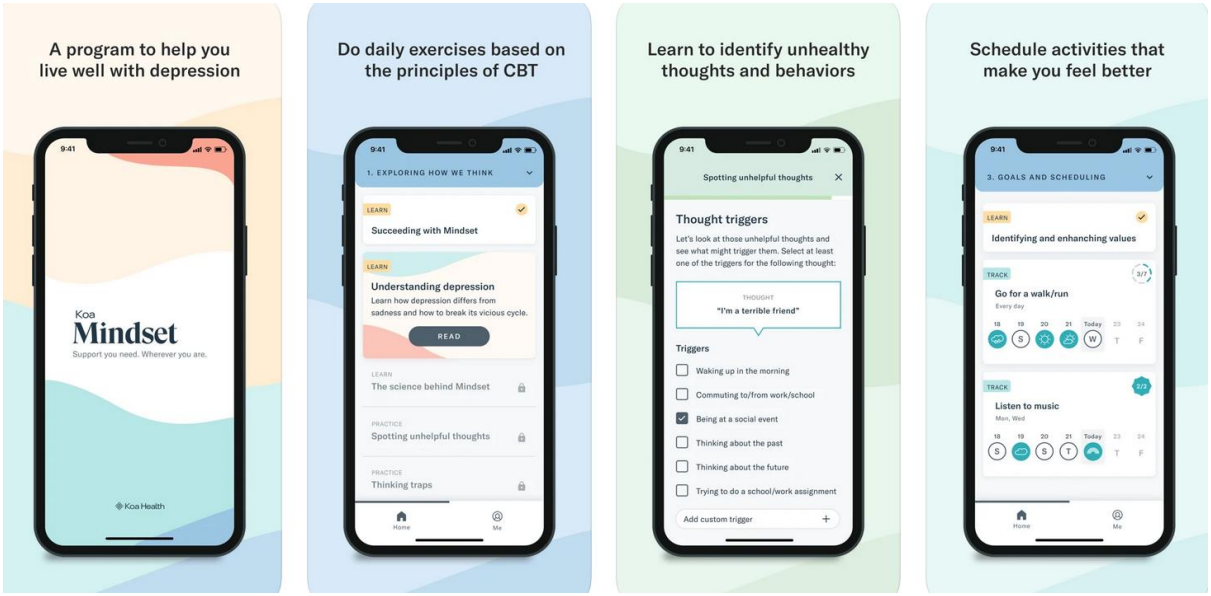
3.2. How Mindset works

3.2.1 The system overview

Mindset's content is split into eight steps. For the beta version reviewed by Eticas between [November 2020, February 2021], only half **of these were available**. The app suggests that users do a step per week, although it also allows them to choose another pace. Users can go back and re-read the 'learn' sections. At the moment, they cannot re-do practical exercises. However, a functionality allowing patients to review their completed Practice exercises will be available soon.

The content within each programme is a mix of: text; videos; audios; questions and free text answers; and quizzes where the user chooses an option.

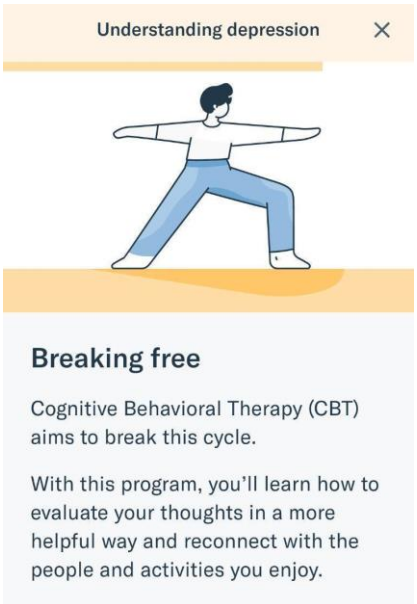
Figure 2. Mindset presentation



Source: Koa.

Mindset's eight-week program offers functionalities that allow users to:
- Identify the cycle of depression
- Understand why and how self-help exercises based upon the principles of CBT can help
- Recognize unhelpful thoughts and thinking patterns
- See how actions impact mood
- Plan activities that make you feel good
- Use mindfulness to focus your attention on the present
- Identify unhealthy core beliefs and develop more balanced, healthy ones

Figure 3. Mindset guidance



Source: Koa.

The **system dashboard** will allow clinicians to monitor their patients' status, identify relevant moments for intervention and communicate with patients[15]. Moreover, experts will be able to use the system's information to select and prioritize patients according to their status and moment within the CBT treatment.

3.2.2 Mindset validation and testing

Mindset is aimed at people who are more likely to have a mental health illness, so testing has been focused on the **age range of 18-55**. This app is not designed for children. In consequence, individuals below 18 have not been considered either within the design or the validation processes. The sample has been split **into two gender-equal groups** (men and women), but with no exclusion of non-binary. In case the testing sample has been unbalanced, this was compensated in the following round.

The **goals of the usability tests** have been to validate if the general navigation and architecture of the app is understandable, to assess if logging and scheduling flows are easy to complete and understand, to compare different design options for "read/watch", activity tracking and mood icons, and to get general feedback on the look and feel and the concept of Mindset. Opinions and perceptions of small groups of users (between 5 and 12) with different pathologies (depression, anxiety, stress) **have been used to improve the legibility, ease of use, and app design reception**. Reactions to recommendations have been documented and utilised to reframe their outline and presentation in some cases.

Moreover, an acceptability analysis was carried out based on a **Diary Study**. This testing's main conclusions were that the app allowed users to control their mental health and mitigate their mental problems. Along these lines, most users (10 out of 11) considered that they imagine using Mindset on their own - "but it depends on the moment they are in their lives and their depression". This contextual aspect is in line with the existing literature about this topic.

Participants also stressed that they accept sharing their data with care providers, but they also expressed that they would like to have the ability to **decide what to share**. This finding is in line with the need for providing disaggregated controls over the different categories of personal data used by the system.

In terms of ease of use, some users expressed that they prefer to read the videos' instructions. The Koa team has decided to include further subtitles to improve **accessibility**. In line with this, the app content assessment also includes an **analysis of its recommendations** to address fairness and biases. Examples of changes that arose as a result of reviews include:

- References to a variety of religions and ethnicities
- References to a variety of families (not only heterosexual and with men as the main breadwinners)
- References to gay, lesbian, and bisexual dating, rather than only heterosexual
- References to people that might be interested only to a platonic relationship rather than romantic
- References to people that might be out of work/study for a prolonged length of time
- References to people that are caregivers of older adults or people with disability

[15] This function is currently under development.

- Use of the neutral pronoun 'they/their' throughout
- Study on a neutral color palette (nor feminine or masculine)
- Study on a character that could be non-binary and can be seen as male or female according to the context
- The whole app is AA accessible (contrast, legibility, and color blindness)
- The text is written in plain English with paragraphs of max 190 characters long
- Options for both reading content and/or watching a video

Every two weeks, the Koa team in charge of the system development reassess these inclusiveness aspects. In this way, inclusiveness has been considered in terms of gender, religion and ethnicity, family composition, romantic practices, socio-economic exclusion and disability (physical and deaf people).

3.3. Ethics assessment

Ethics will be analyzed in this section considering both broad social aspects affecting the ethical principles guiding Mindset and also the specific commitments established by Koa for their systems. Following the Koa ethics self-assessment's conclusions (Ethics Internal Audit document, EIA), this part of the analysis's primary purpose will be to contribute to the application of ethical standards.

3.3.1 Koa 10 commitments in Mindset

The following table summarizes the supplementary assessment of Mindset compliance with Koa 10 commitments based on this audit conducted on the system's initial version.

Table 1. Mindset under the 10 Koa commitments

Commitment	Analysis
1. We aim to support users to achieve their optimal balance of health and happiness	This is currently not a **problematic aspect in Mindset** due to how its aims and capabilities are framed and communicated: The system is presented as a **wellness device**. This factor also minimizes the importance of the lack of a measure of happiness[16] pointed out in the EIA.
2.We will ensure that our recommendations are not based on discriminatory bias	**No discriminatory** recommendations have been found through digital ethnography. Instead, the use of inclusive language and graphics has been confirmed. Findings from usability tests regarding the need for integrating further subtitles into videos should be considered in terms of accessibility for disabled people. It should be noted that this function is under development.

[16] In this regard, it has been pointed out that the subjectivist definition of happiness provided by positive psychology, frequently used in CBT, often fails to recognize social context and commonly relies on people's self-reports (Pawelski and Prilleltensky, 2005). This issue may be considered by Koa when developing specific measures for happiness in this context.

3.We follow best practice in giving users control over how we use their personal data	The main instruments in this regard are the users' consent protocol and the Privacy Policy. Data minimization also contributes to this purpose. Communication about data processing is clear and ARCO rights integrated into these instruments. Explainability of **algorithms** should be improved in the Privacy Policy (reviewed version here). Only references to automated decisions are included concerning ARCO rights. Based on the usability assessment, options for **disaggregated sharing and data protection requirements** concerning different sets of personal data, including "thoughts", should be provided.
4.We will deploy the best available techniques to prevent any user from becoming addicted to any of our services	As shown in Section 4.2, **no direct evidence of addictive features** is found.
5.We will explain how our services work to support you in having the greatest possible health and happiness; in doing this, we will ensure that such explanations are comprehensible, aiming for a reading age of no more than 11	Explanations seem to be clear, but the presentation of the system's overall goal could be refined. As shown in section 5.2, the Privacy Policy's **readability has a reading age of 16.** As the content was still under development at the moment of this audit, a readability analysis of the text has not been done. We recommend analyzing the content with the tools provided in Foundations' audit.
6.We will publish the ethical approvals of our research and external audits of our work, although we may remove some commercially sensitive information	Koa is not only partly yet complying with this commitment since, whilst external audits are published, ethical approvals of research are not as yet. However, although it, Koa states that it plans to put in place processes for this during 2021.
7.We use the state-of-the-art industry standards of encryption to protect your data	Security features and subsystems are still under development. However, **data security protocols are being developed** following both GDPR and ISO27001. Privacy Enhancing Technologies (PET) include TLS secure communication and symmetric 256-bit encryption - RSA public-key SHA-2 algorithms. The methodology used for integrating these requirements into the design includes the review, every two weeks, of the system design. Both technical and legal experts are involved in this iteration. As part of this exercise, threats are measured by addressing a comprehensive list of specific risk scenarios.

8. We will create products and services that preserve as much privacy as possible, for you and your community	Data minimization is applied to data needed to register in the app. Health organizations and experts deploying Mindset, as well as patients, will communicate a comprehensive set of personal and sensitive data into the system. While Koa's security mechanisms and protocols to ensure data integrity are robust (point 7), **proportional data security strategies to the sensitivity and amount of data should be guaranteed also by data processors**, which must be monitored by the data controller (Koa).
9.We will not generate revenue through serving adverts to end-users of our services	According to Mindset policy, **no personal data is shared with third parties with advertising or revenue goals**. Services involved, included in the PP, fulfil concrete purposes such as hosting or analytics. However, providers and data to be shared with them are not defined yet. This information should be included in the PP's final version and offered to users before data processing starts.
10. We will hold external ethics audits at least once annually to assess progress against our ethics strategy, including algorithmic audits	Under development.

Source: own elaboration.

3.3.2 Mindset ethics analysis

As we have seen in the previous section, CBT is a therapeutic model adapted to computerized mediation. This has been tested in different systems under production, showing a clear potential in using technology as a therapeutic "bridge" between clinicians and patients. However, four main ethical dimensions of introducing technological systems in this context should be considered.

Firstly, the mediation capacity of used technology. According to the literature addressing the effects of CBT, to guarantee its efficiency, it is crucial to balance users' autonomy and "clinical threats needs", considering the **specific condition and situation of each patient** and ensuring **clinical monitoring** (Beauchamp and Childress, 2001; Torous and Roberts, 2017). These principles should be applied to technological mediation by providing both patients and physicians with instruments to ensure transparency, smooth communication, secure assessment of the patient statuses and clear medical treatment reception. The application of these requirements in Mindset, calls to assess the system's functionalities for reassessing shared information, maintaining regular communication and orienting the treatment accordingly.

It is also essential to assess possible **prioritization features or design embedded in the Mindset dashboard** to ensure that no patient in need is relegated against traditional treatment standards. This concerns clinicians' attention to different groups of patients based on available data and how these datasets are presented. In this regard, the expert clinician interviewed for the audit indicated that

while for most cases interventions, the dashboard and app communication features are not an issue, **alerts should be available** for those urgent and sensitive case scenarios (such as suicidal thoughts).

Secondly, **privacy aspects** are crucial for guaranteeing ethical standards concerning sensitive data. Tensions have emerged between companies sharing limited information about how these systems work, on the one hand, **and physicians' duty to be open** and the obligation to care for patients' needs, on the other (Roberts, 2016). These problems have often related to **commercial commitments reflected in services contracts.** This should be addressed by providing precise and comprehensive contractual and privacy policies designed to ensure patients and clinicians mechanisms to **exercise their rights on shared data properly**. Therapists and patients should have a common understanding of how technology will be implemented as part of the therapeutic process, including information about security and confidentiality risks (Epstein and Bequette, 2013).

Furthermore, **explicit and accessible informed consent** has been presented as a critical aspect of this type of technology (Prentice and Dobson, 2014). This is particularly important for most vulnerable patients, which should enjoy specific accessibility strategies. This also concerns people with disabilities and potential patients with specific mental pathologies (anxiety, depression), which may predispose to weakening their decisional capacity and favour misunderstandings (Torous and Roberts, 2017). Risk management strategies, such as family members' incorporation into the therapy, should be feasible for therapists to address these issues. Moreover, in the view of the WP29, a controller may not make a service conditional "upon consent, unless the processing is necessary for the service, which WP29 would dispute, might not be the case regarding behavioural advertising" (ARTICLE29 Data Protection Working Party, 2018: 2). This issue should be considered within all ehealth developments but does not apply to Mindset since the system's personal data is not expected to be used for advertising purposes.

Security measures concerning personal data protection should be proportional to the risks of data breaches and misuse. Some authors have focused on **authentication methods and password protection** as fundamental mechanisms for protecting patients privacy rights (Price et al., 2014).

Thirdly, contextual aspects potentially leading to an **unfair or coercive social framework** for the Mindset implementation should be analyzed. Possible coercive elements conditioning the access to the app and the patient's decisional capacity should be addressed as part of the consent protocol (Torous and Roberts, 2017:11). This aspect goes beyond the consent form's design. It may include mechanisms to prevent the system's social pressure to use the system or unethical advertising to promote the app, negatively affecting users' decisions. For instance, stigma aspects related to the social consideration of depression have been pointed out by both the literature and Mindset users during the usability tests. In this regard, the expert clinician interviewed for the audit has indicated that using this technology to mediate the therapist-patient relationship can help eliminate these kinds of barriers. This is also **addressed in Mindset by integrating different explanations about the purposes and limitations of the system**.

Another social dimension to be considered is the presentation of the systems' capabilities and **limitations in terms of the CBT expected outcomes**. Mindset provides different instruments to users to frame boundaries of the system impact and efficiency. For instance, the system web includes the disclaimer, *"Mindset is a wellness device and has not been cleared or reviewed by the FDA.*"

Fourthly, risks for **addiction** in Mindset are mitigated by patients-physicians interaction and monitoring patients' psychological evolution by expert clinicians. However, this issue can also be assessed by design. In this regard, as we can see in the following Table, no major addiction indirect evidence is found.

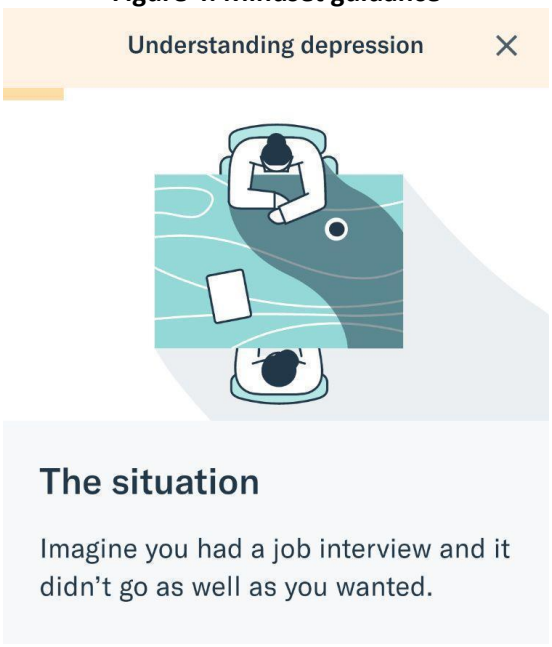
Table 2. Addiction primary dimensions in Mindset design

Addiction variable	Definition	Review
Variable rewards	Random and unpredictable rewards produce more of the neurotransmitter dopamine than regular rewards. In apps, they are based on notifications and other processes.	There are only a **few notifications** and they are integrated into regular activities. The expert clinician interviewed for this audit pointed out the need for having by-design mechanisms promoting users' engagement.
Social reciprocity	These are compensations derived from social interaction and reciprocity. In software applications, chemical satisfaction is received from outcomes of these interactions, for instance in the form of likes.	Given that the system only requires clinician-patient interaction, this risk is not relevant for Mindset.
Infinite scrolling	This is achieved by loading content on a single page instead of spreading it across a series of pages. It produces an interface through which consuming content is allowed by scrolling instead of moving to a different page.	No single page model is used for setting information nor graphic-based activities.
Illusion of choice	User choices can be oriented by software design through the layout of their applications. While some applications seem to empower users with reviews or notifications about different products and services, they often provide a limited number of options.	This risk does not apply to Mindset, since options are clear for the patient and secondary commercial purposes are not expected to be achieved from them.
User investment	Many social media applications take advantage of the human tendency to invest time in activities they feel they "construct" (the so called "Ikea effect") by giving users the power to curate their profiles.	Self-reporting and self-assessments are not oriented towards contributing to external outcomes. Moreover, clinicians monitor this involvement in the app-based therapy.
Gamification	"Closely tied to variable rewards, "gamification" is defined in the tech industry as the process of using game mechanics to reward the completion of tasks." (Neyman, 2017: 4).	No games are included into the app.

Source: own elaboration based on Neyman, 2017.

Lastly, **discrimination** derived from personal data processing or embedded in the design of CBT recommendations should be prevented in the app communication and design. For instance, in some cases, these systems recommendations design and ground rationale has been biased towards non-white and traditional family populations (Parker et al., 2018). In this regard, we have not identified any evidence of group discriminatory within Mindset layout or recommendations as part of our digital ethnography. Visuals are inclusive from the gender perspective, even illustrating power relations as in the following capture:

Figure 4. Mindset guidance



Source: Koa.

3.3.3 Summary of ethics recommendations

- ❑ **Assess efficacy regularly**, considering the impact on patients on the short, medium and long run, and use results to improve the system. Ensure targeted treatment and clinical monitoring.
  - ❑ Examine clinical interventions to examine "false negatives" concerning the prioritization model. Measures for identifying the app-based treatment's negative impact could include rates of non-prioritized cases suffering from worse depression conditions or dropout rates against prioritized segments, among others. Same rates could be applied to assess algorithmic impact by groups.
- ❑ Provide **clear and comprehensive contractual clauses between vendors and health contractors on patient data access and management**. These conditions should be aligned with ARCO and other data protection rights reflected in the Privacy Policy.
- ❑ **Intelligible and complete consent forms** should be produced in formats friendly for vulnerable groups, including people with mental illness and physical disabilities.
- ❑ Provide **robust identity verification methods** to ensure purpose limitation and avoid any unauthorized access.

3.4. Data management assessment

The **data governance** of the system has not been defined yet. However, some elements are outlined in the **preliminary Privacy Policy** provided to Eticas by Koa. According to this initial approach, Koa will be the Data Controller of the system (with its assigned DPO *dpo@koahealth.com*) and health organizations and/or clinicians hiring the services would act as processors.

Physicians -and their institutions- must explicitly adhere to the **service's privacy policies** (Beauchamp and Childress, 2001). As part of these policies, Koa should secure data flows from the patient to the app and the organization or clinician in charge of the patient. In this regard, it is recommended that logs and authentication data should only be shared with the clinician in charge of each patient under his/her supervision. This protocol must be aligned with patients' consent required to provide personal clinical information to controllers and processors. The WP9 indicates that when developers act as data controllers, clear distribution of roles and responsibilities between data controllers and processors should be ensured. The end-user should also be informed and take part in the definition of any other data governance approach: "in particular, in the case of co-controllership, a single point of contact should be offered to the user." (ARTICLE29 Data Protection Working Party, 2018: 2). So this information should be clearly presented to patients before data processing starts.

Mindset collects different **categories of personal information**, including:

- **Data provided by patients** when installing the app, including their names, emails and personal identifiers associated with their insights produced as part of the therapy
- **Data provided by the clinician** when the patient profile is created: First Name, Last Name, Email Address, MRN number, Birthdate, Phone Number, Selected Treatment, Diagnosis Notes.
- **Data collected through cookies**: Frequency of access to the app, time spent on different screens, functions used etc.

Legal basis for the processing: Informed consent is defined as the legal basis for the preliminary PP processing. Moreover, a contract's performance is the basis for processing some personal data for the app's functioning, such as registration data. Lastly, under Koa legitimate interest, information about users activity and interaction with the system, including cookies, is also processed. In this regard, the need to ensure the correct use of cookies within ehealth apps has been stressed by the Working Party 29 (ARTICLE29 Data Protection Working Party, 2018: 2). The final version of the Mindset Privacy Policy should offer comprehensive information about the types of data collected by cookies and the purposes of these data collection processes, including sharing of data with third parties or external service providers.

Identified **purposes for the collection of personal data** in the case of the Data Controller are:

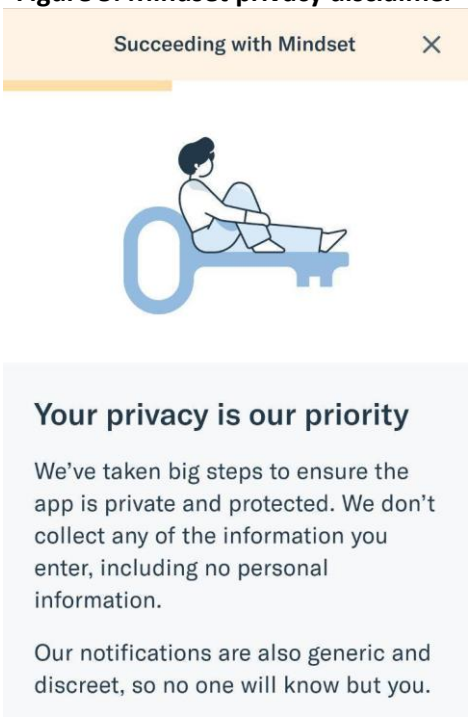
- Improving patients treatment,
- To "provide better quality of care to patients that might suffer similar symptoms",
- Managing "depression symptoms",
- Providing "more relevant experience such as recommending activities" to patients,
- Registration, authentication or support.

Data sharing includes the following actors:

- **Medical organizations in charge of patients**: the PP explains that users are "in control of specific information collected via exercises and shared to your medical team".
- **Sub-processors providing services to Koa:** It includes activities such as hosting, providing customer support, analytics or application functionality such as notifications. The PP points out that the principle of data minimization is applied to this data sharing, which may involve processors located outside the EEA on the basis of Standard Contractual Clauses.
- **Sub-processors to conduct studies about the system impact**.

While Koa does not share large sets of personal data with other processors than medical teams, the PP stresses patients' responsibility for the data shared with hiring organizations or clinicians. In fact, according to the PP, no personal data is shared by Koa with other controllers or processors beyond medical teams involved in registered patients' treatment. This approach is already integrated into the app advice as follows:

Figure 5. Mindset privacy disclaimer



Source: Koa.

However, as mentioned, the PP also states that some personal data may be shared by Koa with service providers, some of them located outside the EU. In order to achieve further transparency, this **information should also be summarized within the content application**. The roles of the parties involved in processing should be clear.

Data retention periods are established according to the types of personal data and the different purposes of data sharing mentioned above:

- As long as users are active or Koa have legal obligations to retain the data. In the case of non-active users, Koa will remove personal data after 12 months from last user access.
- Besides the above, Koa will have to remove data in cases of withdrawal consent or cancellation requests.

It is recommended to better clarify how Koa will proceed to **ensure data deletion** in third parties and clinical organizations. The WP29 has stressed that consent withdrawal requirements and binding clauses between controllers and their parties need to be considered in advance (ARTICLE29 Data Protection Working Party, 2018). The latest should be in line with privacy policy and contracts established with these organizations.

Furthermore, it is important to anticipate **possible ARCO requests**, including portability, from both technical and organizational sides. According to the WP29, Mindset should offer details and relevant examples on how app developers can integrate "privacy by design" and "privacy by default" in their development process as well as be attentive to legal restrictions relating to retention periods (ARTICLE29 Data Protection Working Party, 2018). Security issues where the users would like to allow third parties (e.g. doctors) to access their data require clarification.

3.4.1 Privacy Policy readability

The readability of the privacy policy available at https://mindset-dashboard-web.int.koahealth.com/assets/legal/privacy-policy-mindset.pdf
(effective from October 26th, 2020) is checked against different indices using the open source Python library "Readability" available at https://github.com/andreasvc/readability/. This library calculates text statistics as well as the Flesch Kincaid Reading Ease and several grade level indicators, which equate the readability of the text to the U.S. grade level system.

The following indexes are checked:

- The **Flesch Kincaid Reading Ease** test works by counting the number of words, syllables, and sentences in the text. It then calculates the average number of words per sentence and the average number of syllables per word. The idea is that **shorter words** (with few syllables) and shorter sentences are easier to read. The higher the score, the easier the text is to understand. The result is a number between 0 and 100. Higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read. A value between 60 and 80 should be easy for a 12 to 15 year old to understand.

- **Flesch–Kincaid grade level**. Test used extensively in the field of education. As in Flesch Kincaid Reading Ease**,** it is based on the idea that **shorter words** (with few syllables) and shorter sentences are easier to read**.** But the "Flesch–Kincaid Grade Level Formula" instead presents a score as a U.S. grade level.

- **Gunning fog index** is a readability test for English writing that takes into account two qualities to determine readability: the average number of words in sentences and the percentage of **complex words** (words with three or more syllables).

- **SMOG grade.** SMOG is an acronym for "Simple Measure of Gobbledygook". It is used particularly for checking health messages. It also takes into account the **complex words** (words with three or more syllables), in three 10-sentence samples. Analized texts of fewer than 30 sentences are statistically invalid, because the formula was normed on 30-sentence samples.

- **Coleman–Liau index.** It relies on **characters** instead of syllables per word. Although opinion varies on its accuracy as compared to the syllable/word and complex word indices, characters are more readily and accurately counted by computer programs than are syllables.

- **Automated readability index**. As the Coleman-Liau index, it relies on a factor of **characters** per word, instead of the usual syllables per word. This index was designed for real-time monitoring of readability on electric typewriters.

Additionally, an average grade level is calculated as the arithmetic average of Flesch–Kincaid grade level, Gunning fog index, SMOG grade, Coleman–Liau index and Automated readability index. We recommend using this average in order to account for the readability of a text as it takes into consideration different approaches to measure complex texts.

Mindset app privacy policy has an average grade level of 11 (16-17 years old). Results show readability based on Flesch Reading Ease is 10th to 12th grade (15-18 years old, fairly difficult to read) and readability based on the average U.S. grade level is 11th grade (16-17 years old, high school - junior).

Table 3. Privacy policy statistics

Text Statistics	
No. of sentences	109
No. of words	1923
No. of complex words	366
Percent of complex words	17.50%
Average words per sentence	17.64
Average syllables per word	1.57

Source: own elaboration.

Table 4. Privacy policy readability

Readability Indexes	Score/ Grade	Ages[17]
Flesch Kincaid Reading Ease	56	15-18
Flesch Kincaid Grade Level	9.83	14-15
Gunning Fog Score	14	19-20
SMOG Index	12.6	17-18
Coleman Liau Index	11.4	16-17
Automated Readability Index	10.5	15-16
Average U.S. Grade Level	11.7	16-17

Source: own elaboration.

3.4.2 Variables for future data management assessments[18]

- ❑ Alling data protection instruments (Privacy Policy, Informed Consent, Contractual Clauses between controllers and health organizations) within the governance structure, ensuring clarity in the definition of roles and responsibilities and transparency about rights over patients data.
- ❑ Further clarify which categories of personal data are shared with service providers both within the app content and the PP.
    - ❑ Provide full information about information collected by Cookies and its data sharing purposes.
- ❑ Develop an ARCO strategy for ensuring rights to withdraw, cancellation, rectification, access and objection regarding personal data in the hands of both Koa and all sub-processors.

3.5. Desirability and acceptability assessment

In this section, we will assess Mindset under the light of the main variables affecting its desirability and acceptability. The analysis will be based on the analysis of the app functionalities, literature review and an interview with an expert clinician working with the system.

3.5.1 Desirability analysis

CBT is well **aligned with technological mediation** since it involves an active intervention of the patient in the examination of her/his psychological wellbeing and awareness regarding her/his condition and mental health evolution. As already mentioned, significant acceptability of technologies used for this purpose has been revealed (Batra et al., 2017), for instance, concerning mobile app mediated CBT for

[17] A table to look up ages for the different U.S. Grade Levels can be found at https://en.wikipedia.org/wiki/Education_in_the_United_States#Educational_stages

[18] Based on the current Mindset Privacy Policy, these are issues to be considered by KoA once data governance is established.

insomnia, which shows significant patients adherence to therapeutic recommendations (Koffel et al., 2018). Positive results and acceptability have also been observed in CBT apps designed to reduce dysfunctional beliefs and behaviours in individuals with depression, and evidence is being collected regarding similar effects in the treatment of obsessive-compulsive disorder and post-traumatic stress disorder (Gershkovich et al., 2021; Simon et al., 2019; Schlosser et al., 2017; Birney et al., 2016)[19]. As indicated during our interview for this audit with an expert clinician, this science-based character of CBT fosters Mindset's clinical relevance and desirability.

Moreover, the cost of computer-delivered CBT modules also increases its desirability (McCrone et al., 2004; Proudfoot, 2004). As indicated by the interviewed clinician, technological availability and accessibility are crucial since they may allow a level of **standardization and accessibility** that represents an opportunity to reach populations who cannot afford or access traditional therapies. This may include people who live in rural areas or belong to ethnic minorities (Price et al., 2014; Smith, 2010).In line with this, Mindset is available on iOS and Android, which can be considered as best practice.

Another element supporting the use of these CBT apps is their **expected social impact**. Research has shown that identifying cognitive deviations early is more productive than doing at an advanced stage of these pathologies (Erbes et al., 2014; Prentice and Dobson, 2014; Price et al., 2014). Moreover, it has been suggested that mHealth apps can facilitate psychoeducation to mitigate stigma, which can significantly affect some groups, such as veterans (Jones and Moffitt, 2016; Hoffman et al., 2011).

Nevertheless, some constraints limiting desirability should be observed within mental health apps. For instance, in some specific cases, these systems have shown to promote medicalization of persons with non-pathological mental problems or relegate some social groups' attention based on **socioeconomic or cultural reasons behind mental problems** (Parker et al., 2018). Moreover, the literature has revealed indirect evidence and hypotheses regarding **technology-based CBT increasing self-stigmatisation risks** due to adverse clinical effects of repeated questioning (Batra et al., 2017; Husky et al., 2014). Increasing self-awareness of depression and self-reporting of negative cognition without regular clinical monitoring could, according to these studies, has led to negative effects for patients.

It has been pointed out that a **dynamic clinician-patient relationship** can help to avoid the above issues. According to Tourous and Roberts (2017:7), "*Ideal use of these technologies...occurs when these tools enhance the psychiatrist's ability to deliver high-quality clinical care*". Along these lines, inscribing these tools within the relationship between the patient and therapist integrating an **open communication about its functioning** can facilitate better risk management and balancing "patient autonomy with clinical needs" (Torous and Roberts, 2017). Within CBT, it is recommended to **explain to patients how the model and each step of the treatment work** (Wright, 2006) as is done within Mindset content.

[19] However, it should be noted that most RCTs conducted in these investigations include small sizes and short time period assessments. As a side note, one of the first large scale and long term studies -randomised controlled trial targeting up to 10,000 Year 8 Australian secondary school- with young people using mobile apps offering CBT for preventing depression is being launched (Werner-Seidler et al., 2020).

Lastly, existing standards regarding the **quality of evidence for mHealth interventions** should be used to reinforce these systems' desirability. This includes the one developed by the World Health Organization (Agarwal et al., 2016) or CONSORT-eHealth (Consolidated Standards of Reporting Trials of electronic and mHealth applications and online telehealth) (Eysenbach, 2011). These models systematize the minimum information needed to contextualize and define the intervention's technical features to support its **data portability and replication**.

3.5.2 Acceptability analysis

The current Mindset Privacy Policy explains the primary purpose for data processing as follows: "*The main purpose of the App is to help you better cope with depression symptoms through cognitive behavioural therapy*." In terms of how Mindset **goals and limitations are communicated to patients,** it is essential to ensure social, economic, and cultural inclusiveness. Instruments under development, such as the consent or Privacy Policy, should address the above-explained ethical issues regarding the decisional capacity of individuals with different cultural backgrounds or disabilities (Torous and Roberts, 2017). Still, risks related to informed consent, fairness and discrimination are **mitigated in Mindset by the therapist's intervention** in the process that goes from the patient selection for the technologically assisted CBT therapy to his/her interaction with the app.

User-centred features in the design of Mindset should be assessed in the future to ensure the system is properly **tailored to the targeted (gender, educational, etc.) populations**, which is a pending aspect in the validation of similar systems (Batra et al., 2017). According to the interviewed expert clinician working with Mindset, differential treatment of these protected groups is an issue to be considered in the system development. While the digital divide is addressed by Mindset, the expert pointed out that broadening access for the elderly is being considered. Another variable to be contrasted in relation to different target populations through the iterative validation of Mindset is **adherence to CBT** over a period of time to correctly identify possible decreasing (Hidalgo-Mazzei et al., 2016).

In line with the above and following Hsin et al. (2016), technology should be placed as an "adjuvant" to the psychiatrist-patient relationship. This approach facilitates that the patient's autonomy is respected and clinical care is adequately conducted assessing the patient's risks. Stressing this supporting nature of the system in its communication materials (web, PP, etc.) should therefore provide more safety to both clinicians and patients. Under these coordinates, **ensuring anonymity beyond this relationship (including third parties such as insurance companies to be involved)** becomes even more essential to balance patients' trust with the treatment efficiency.

Moreover, for this approach to be implemented and accepted by clinicians, the system should also guarantee an accurate and **operative documentation of each case evolution and patient-clinician interaction** (Torous and Roberts, 2017). Physicians should assess these technologies regarding their benefit in improving patients' health and enhancing the psychiatrist-patient relationship's efficacy. In this regard, it has been suggested that "in the absence of clinical outcomes data, clinical benefit can be referenced with respect to the therapeutic relationship." (Torous and Roberts, 2017:10).

According to the clinician interviewed for this audit, having a trustable service provider and a reliable healthcare system establishing standards for data protection are key elements for ensuring patients'

confidence in the system. Together with these contextual elements supporting the acceptability of Mindset data protection standards, it is suggested that the service should include **a Manual for physicians on data management** and treatment tracking for these purposes. The Manual should combine a description of key data protection requirements to be followed by end-users with specific preventative strategies at the managerial level (for instance, using strong passwords, ensuring public or free Wi-Fi is protected, encrypting stored data) to ensure purpose limitation and avoid data breaches or function creep.

3.5.3 Summary of desirability and acceptability recommendations

Desirability

- ❑ Ensure comprehensive documentation of the therapeutic process, addressing data minimization at the same time.
- ❑ Provide guidelines to clinicians on informed consent, data management and system treatment.

Acceptability

- ❑ Assess the system performance considering adherence to the therapy, together with efficiency, as part of possible RCT for the systematic review of the system
- ❑ As part of the above effort for avoiding negative differential impact of the app in specific disadvantaged groups, it is recommended to address the following aspects in acceptability/ usability tests:
    - Examine stigma as contextual and CBT technological-based inflicted factor.
    - Vulnerable group targeting and adaptability to disabilities, including other social barriers for use. Analysis specific socioeconomic and cultural factors in relation to provided recommendations and technological guidance.

3.6. Summary of conclusions and recommendations for Mindset

The following Tables summarize the **main issues found** concerning the four dimensions of the Mindset audit, ethics, data management, desirability and acceptability. It will also reflect the **most significant recommendations produced** in terms of existing gaps between requirements and already implemented solutions. High priority is a categorization of the recommendations associated with higher risks in the short term and/or are blockers for other recommendations to be implemented. Low priority recommendations correspond to less problematic or urgent issues.

Table 5. High priority findings and recommendations

Assessment	Main findings and issues found	Most significant recommendations

Ethical assessment	Strengthening mechanisms for bias prevention. Evidence based studies are required.	Access number of clinical interventions to **examine "false negatives"** concerning the prioritization model. Measures for identifying the negative impact of the app-based treatment could include rates of non-prioritized cases suffering from worse depression conditions or dropout rates against prioritized segments, among others. Same rates could be applied to assess algorithmic impact by groups.
	Ensuring targeted consent	The PP's final version should ensure that consent protocols are produced **in formats friendly for vulnerable groups**, including people with mental illness and physical disabilities.
	Data protection	The PP's final version should further clarify which categories of personal data are shared with service providers.
Desirability assessment	Efficiency and reporting	Ensure **comprehensive documentation** of the therapeutic process, addressing data minimization at the same time. With this purpose, the WHO standards for quality of evidence for mHealth interventions should be considered. Assure **smooth patient-clinician communication** to avoid technology-based CBT to increase risks of self-stigmatization. In this regard, notifications for clinicians on most urgent interventions could be included.

Table 6. Medium-level priority findings and recommendations

Assessment	Main findings	Most significant recommendations
Ethical assessment	Evidence based studies are required	Assess **efficiency regularly**, considering the impact in patients on the short, medium and

		long run, and use results to improve the system.
	Data protection agreements	The PP's final version should provide **clear and comprehensive contractual clauses** between vendors and health contractors on patient data access and management. These conditions should be aligned with ARCO and other data protection rights reflected in the Privacy Policy.
	Ensuring secure data management	The Manual for end-users should provide information about **identity verification methods** to ensure purpose limitation and avoid any unauthorized access.
Data management assessment	Data protection	The **PP's final version should provide clear information about** the system governance structure, ensuring clarity in the definition of roles and responsibilities and transparency about rights over patients data.
	Data protection	The **PP's final version** should provide **full data about information collected by Cookies** and its data sharing purposes.
Desirability and acceptability assessment	Evidence-based studies are required	Following the literature findings above, different users groups' **adherence to the Mindset should be assesse**d over time, together with efficiency, as part of possible RCT for the system's systematic review.

Table 7. Low priority findings and recommendations

Assessment	Main findings	Most significant recommendations
Data management assessment	Data protection	Develop an **ARCO strategy** for ensuring rights to withdraw, cancellation, rectification, access and objection regarding personal data in the hands of both Koa and all sub-processors. Provide end-users with **Manuals on data management, focusing on ground requirements and strategies for ensuring purpose limitation.**

Desirability and acceptability assessment	Evidence-based studies are required	Based on **findings from the literature review, usability/acceptability tests** could: ● Examine stigma as contextual and CBT technological-based inflicted factor. ● Group targeting and adaptability to disabilities. ● Other social barriers for use. This includes socioeconomic and cultural factors in relation to provided recommendations and technological guidance.

4. Overall Conclusions

This second audit conducted by Eticas Research and Consulting for Koa has examined **Foundations and Mindset**, two apps aimed at supporting users with wellbeing and mental health issues, respectively. Overall, the systematic integration of data protection and ethics requirements is observed in both cases. This includes avoiding discrimination in content design, and personal data protection through robust security tools and protocols. Besides implementing technical and organizational protocols for ensuring these standards by design, Koa has used ethics and privacy self-assessment outcomes to advance both systems. Our analysis of the systems' desirability, acceptability, usability and data management reflects this standardization and control process, which can also be seen in terms of data availability and quality.

Still, this audit has gone beyond compliance analysis to capture users' risks embedded in these systems and ensure that proper mitigation strategies are incorporated into technology or through human management protocols.

In the case of Foundations, the main gaps identified include the need for the further collection of empirical results about system performance and impact for specific social groups, the improvement of communications targeted to vulnerable groups, the integration of additional mechanisms to ensure that there is no identification of workers by their employer, and the need to refine Koa's methodology for data registering during algorithmic training and deployment in order to ensure effective and privacy-friendly audit of algorithmic bias. While some of these points are currently being addressed by Koa, others are expected and should be addressed in the near future.

Mindset's audit results reflect similar issues. While high overall ethics and data protection standards are identified, some aspects for further consideration are stressed. It is recommended that the impact of technological mediation on different groups of patients is monitored, particularly with respect to the clinician's intervention in the therapeutic process based on the dashboard data. Clear communication of the Privacy Policy with groups with accessibility problems should be ensured. Given the sensitivity of the system and its implications for patients' mental health, it is also recommended that protocols for the documentation of the therapy process should be put in place, to ensure smooth data portability if required. Lastly, although no algorithmic analysis was sought to be conducted in this case, differential impact over protected groups must be audited going forward.

In brief, the results of these audits reflect the substantial consideration of ethics and data subjects' rights in the technological development process. To maintain this by design standards, Koa must continue deploying internal and operational monitoring mechanisms to ensure the fair and secure treatment of special categories of personal data.

5. References

5.1 Foundations Audit

Aazami, S., Shamsuddin, K., and Akma, S., (2015). Examining behavioural coping strategies as mediators between work-family conflict and psychological distress. *Scientific world journal*, 1-7.

Ackerman, L. (2013). *Mobile Health and Fitness Applications and Information Privacy: Report to California Consumer Protection Foundation*. *Privacy Rights Clearinghouse.* Available at: https://www.privacyrights.org/mobile-health-and-fitness-apps-what-are-privacy-risks

Aguilera, A. and Muench, F. (2012). There's an App for That: Information Technology Applications for Cognitive Behavioral Practitioners. *The Behavior Therapist*, 35, 65-73.

Ahthes E. (2016). Mobile mental-health apps have exploded onto the market, but few have been thoroughly tested. *Pocket Psychiatry, Nature*, 532(7).

Airaksinen J, Jokela M, Virtanen M, Oksanen T, Pentti J, Vahtera J, et al. (2017). Development and validation of a risk prediction model for work disability: multicohort study. *Sci Rep*. 7(1):1–12.

Allison J.B. Chaney, Brandon M. Stewart, Barbara E. Engelhardt (2018). *How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility.* Available at: https://arxiv.org/pdf/1710.11214.pdf

Anthes, Emily (2016). "Pocket psychiatry: mobile mental-health apps have exploded onto the market, but few have been thoroughly tested." *Nature*, 532, 7597.

Aryana, B., Brewster, L. & Nocera, J.A.(2019). Design for mobile mental health: an exploratory review. *Health Technol.* 9, 401–424.

Bani-Hani, M. , Hamdan-Mansour, A. , Atiyeh, H. and Alslman, E. (2016). Theoretical Perspective of Job Demands Correlates among Nurses: Systematic Literature Review. *Health*, 8, 1744-1758.

Barnett, M.D., Martin, K.J. and Garza, C.J. (2019), Satisfaction With Work–Family Balance Mediates the Relationship Between Workplace Social Support and Depression Among Hospice Nurses. *Journal of Nursing Scholarship,* 51: 187-194.

Barocas, Solon and Selbst, Andrew D. (2016). *Big Data's Disparate Impact, SSRN Scholarly Paper*. NY: Social Science Research Network. Available at https://papers.ssrn.com/abstract=2477899.

Batra S, Baker RA, Wang T, Forma F, DiBiasi F, Peters-Strickland T. (2017). Digital health technology for use in patients with serious mental illness: a systematic review of the literature. *Med Devices*.10:237-251.

Becker, S., Miron-Shatz, T., Schumacher, N., Krocza, J., Diamantidis, C. and Albrecht, U.V. (2014). mHealth 2.0: Experiences, Possibilities, and Perspectives. *JMIR mHealth uHealth,* 2, e24.

Berglind F. Smáradóttir, Jarle A. Håland, Santiago G. Martinez. (2018). "User Evaluation of the Smartphone Screen Reader VoiceOver with Visually Disabled Participants", Mobile Information Systems.

Berkman, L. , Kawachi, I. and Theorell,T. (2014). "Working conditions and health," in L. Berkman, I. Kawachi, and M. Glymour, Eds. *Social Epidemiology*. New York: Open University Press, pp. 153–181.

Birnbaum HG, Kessler RC, Kelley D, Ben-Hamadi R, Joish VN, Greenberg PE (2010). Employer burden of mild, moderate, and severe major depressive disorder: mental health services utilization and costs, and work performance. *Depress Anxiety.* 27(1):78–89.

Buckovich, Suzy A., Helga E. Rippen, and Michael J. Rozen. (1999). "Driving Toward Guiding Principles: A Goal for confidentiality, and Security of Health Information." *Journal of the American Medical Informatics Association* 6, 2, 122–33.

Burgard, SA, Brand, JE, House, JS (2009). Perceived job insecurity and worker health in the United States, *Soc Sci Med,* 69, 777-785.

Burns,M.N., Begale, M., Duffecy, J., Gergle, D., Karr, C.J., Giangrande, E., et al. (2011). Harnessing Context Sensing to Develop a Mobile Intervention for Depression. *Journal of Medical Internet Research*, 13, e55.

Caetano, L. (2013). *Location, Location, Location: Three Reasons It Matters for Your Smartphone*. Available at: https://blogs.mcafee.com/consumer/mobile-security/location-location-location-three-reasons-it-matters-for-your-smartphone

Calnan, M., Wainwright, D., Forsythe, M., Wall, B., & Almond, S. (2001). "Mental health and stressin the workplace: The case of general practice in the UK".*Social Science & Medicine*,52,499–507.

Castillo, C. (2018). Algorithmic Discrimination. Assessing the impact of machine intelligence on human behaviour: an interdisciplinary endeavour, *Proceedings of HUMAINT Workshop*. Disponible en: https://arxiv.org/pdf/1806.03192.pdf.

Chandola, T, Brunner, E. and Marmot, M. (2006). 'Chronic stress at work and the metabolic syndrome: prospective study',*British Medical Journal*,332.

Chang, Betty & Bakken, Suzanne & Brown, S & Houston, Thomas & Kreps, Gary & Kukafka, Rita & Safran, Charles & Stavri, P. (2004). Bridging the Digital Divide: Reaching Vulnerable Populations. *Journal of the American Medical Informatics Association : JAMIA*. 11. 448-57.

Chartered Institute of Personnel Development (2016).'Absence management 2016'.Available at: ttps://www.cipd.co.uk/Images/absence-management_2016_tcm18-16360.pdf

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806.

Cowgill, B. (2019). Bias and productivity in humans and machines, *Upjohn Institute Working Paper, No. 19-309, W.E. Upjohn Institute for Employment Research,Kalamazoo*. Available at: https://research.upjohn.org/up_workingpapers/309/.

Danks, D. y John London, A. (2017). Algorithmic bias in autonomous systems, *Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI Press*: 4691-4697.

de Korte EM, Wiezer N, Janssen JH, Vink P, Kraaij W (2018). Evaluating an mHealth App for Health and Well-Being at Work: Mixed-Method Qualitative Study, *JMIR Mhealth Uhealth*;6(3):e72.

Dewa CS, Loong D, Bonato S. (2014). Work outcomes of sickness absence related to mental disorders: a systematic literature review. *BMJ Open*.;4(7):e005533.

Economides, M., Martman, J., Bell, M.J. *et al.* (2018). Improvements in Stress, Affect, and Irritability Following Brief Use of a Mindfulness-based Smartphone App: A Randomized Controlled Trial. *Mindfulness* 9, 1584–1593.

Eyal, Nir (2014). *Hooked: How to Build Habit-Forming Products.* Canada: Penguin.

Fan, L.-B. , Blumenthal, J. A., L. L. Watkins, A. Sherwood (2015). Work and home stress: associations with anxiety and depression symptoms, *Occupational Medicine,* 65, 2, 110–116.

Farrell M. (2016). Use of iPhones by nurses in an acute care setting to improve communication and decision-making processes: qualitative analysis of nurses' perspectives on iPhone use. *JMIR mHealth uHealth*. 4(2):e43.

Gajecki M, Berman AH, Sinadinovic K, Rosendahl I, Andersson C (2014). Mobile phone brief intervention applications for risky alcohol use among university students: a randomized controlled study. *Addict Sci Clin Pract* 9:11.

Gaspar FW, Wizner K, Morrison J, Dewa CS. (2020). The influence of antidepressant and psychotherapy treatment adherence on future work leaves for patients with major depressive disorder. *BMC Psychiatry*. 20(1):320.

Glenn T, Monteith S (2014). Privacy in the digital world: medical and health data outside of HIPAA protections. *CurrPsychiatry Rep,* 16: 494.

Haque, A.U., Aston, J., & Kozlovski, E. (2016). Do causes and consequences of stress affect genders differently at operational level? Comparison of the IT sectors in the UK and Pakistan, *International Journal of Applied Business and Management Studies*, 1, 1.

Harrison, V., Proudfoot, J., Wee, P.P., Parker, G., Pavlovic, D.H. and Manickavasagar, V.(2011). Mobile Mental Health: Review of the Emerging Field and Proof of Concept Study. *Journal of Mental Health,* 20, 09-524.

Heffernan KJ, Chang S, Maclean ST, Callegari ET, Garland SM, Reavley NJ, et al. (2016). Guidelines and recommendations for developing interactive eHealth apps for complex messaging in health promotion. *JMIR Mhealth Uhealth*, 4(1):e14.

Heffner JL, Vilardaga R, Mercer LD, Kientz JA, Bricker JB (2015). Feature-level analysis of a novel smartphone application for smoking cessation. *Am J Drug Alcohol Abuse,* 41:68-73.

Hörbe, R.; Hötzendorfer, W. (2015). Privacy by design in federated identity management. In: *IEEE. 36th IEEE Symposium on Security and Privacy Workshops.* San Jose, CA, USA, 167–174.

Hsin H, Torous J, Roberts LW (2016). An adjuvant role for mobile health in psychiatry. *JAMA Psychiatry* 73(2):103-4.

Jimenez P, Bregenzer A. (2018). Integration of eHealth Tools in the Process of Workplace Health Promotion: Proposal for Design and Implementation. *J Med Internet*;20(2):e65.

Kaipainen K, Välkkynen P, Kilkku N. (2017). Applicability of acceptance and commitment therapy-based mobile app in depression nursing. *Transl Behav Med*. 7(2):242-253.

Kamei-Hannan, T McCarthy, B Pomeroy (2015). *Methods in creating the iBraille Challenge mobile app for braille users*. California State University, Northridge.

Kotera, Y. , Green, P. , and Sheffield, D. (2019). Mental Health Shame of UK Construction Workers: Relationship with Masculinity, Work Motivation, and Self-Compassion. *Journal of Work and Organizational Psychology*, 35, 135 - 143.

Lacerda Shirley S., Little Stephen W., Kozasa Elisa H. (2018). A Stress Reduction Program Adapted for the Work Environment: A Randomized Controlled Trial With a Follow-Up, *Frontiers in Psychology,* 9.

Lane, Julia, and Claudia Schur (2010). "Balancing Access to Health Data and Privacy: A Review of the Issues and Approaches for the Future." Health Services Research 45, 5, 1456– 67.

Laur, Audrey (2015). "Fear of E-Health Records Implementation?" *Medico-Legal Journal* 83, 1, 34–39.

Lippert-Rasmussen, Kasper (2013). *Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination*. Oxford: Oxford University Press.

Li XB, Qin J. Anonymizing and Sharing Medical Text Records. *Inf Syst Res*. 2017;28(2):332-352. doi:10.1287/isre.2016.0676

Luxton, D.D., McCann, R.A., Bush, N.E., Mishkind, M.C., and Reger, G.M. (2011). mHealth for Mental Health: Integrating Smartphone Technology in Behavioral Healthcare. *Professional Psychology: Research and Practice*, 42, 505-512.

Mendis, M.D.V.S. and Weerakkody, W.A.S., (2014). Relationship between work life balance and employee performance: with reference to telecommunication industry of Sri Lanka. *Kelaniya journal of human resource management,* 9(1-2), 95-117.

Motti, Vivian Genaro, and Kelly Caine (2015). "Users' Privacy Concerns About Wearables: Impact of Form Factor, Sensors and Type of Data Collected." In *Financial Cryptography and Data Security* (FC 2015), edited by M. Brenner, N. Christin, B. Johnson, and K. Rohloff, 8976. Berlin: Springer, 231–44.

Muñoz-Laboy, M., Ripkin, A., Garcia, J. *et al.* (2015). Family and Work Influences on Stress, Anxiety and Depression Among Bisexual Latino Men in the New York City Metropolitan Area. *J Immigrant Minority Health* 17**,** 1615–1626.

Muoka, Michael Obinna; Lhussier, Monique (2020). The impact of precarious employment on the health and wellbeing of immigrants: a systematic review, *Journal of Poverty and Social Justice*, 28, 3, 337-360(24).

Neyman, C. J. (2017). *A survey of addictive software design*. Cali *California Polytechnic State University*. Available at: https://digitalcommons.calpoly.edu/cscsp/111/

Nguyen,Theresa; Madeline Reinert, Michele Hellebuyck, and Danielle Fritze (2019). *Mind the workplace*. Alexandria: Mental Health America/Fass Foundation.

Njie, C.M.L. (2013). Technical Analysis of the Data Practices and Privacy Risks of 43 Popular Mobile Health and Fitness Applications. *Privacy Rights Clearinghouse.* Available at: https://www.privacyrights.org/mobile-medical-apps-privacy-technologist-research-report.pdf

Paganin, G., Simbula, S. (2020). Smartphone-based interventions for employees' well-being promotion: a systematic review. *Electronic Journal of Applied Statistical Analysis*, 13.

Price M, Yuen EK, Goetter EM, Herbert JD, Forman EM, Acierno R, et al. (2014). mHealth: a mechanism to deliver more accessible, more effective mental health care. *Clin Psychol Psychother*. *Wiley OnlineLibrary;* 21, 427–36.

Rasool SF, Wang M, Zhang Y, Samma M. (2020). Sustainable Work Performance: The Roles of Workplace Violence and Occupational Stress. *Int J Environ Res Public Health*. 17(3):912.

Ravalier, J.M. (2018). 'Psychosocial working conditions and stress in UK social workers', *British Journal of Social Work*, 49 (2), pp. 371-390.

Rosengren, A., Hawken, S., Ounpuu, S., Sliwa, K., Zubaid, M., Almaheed, W.A., Blackett, K.N., Sitthi-amorn, C., Sato, H.andYusuf, S. (2004). 'Association of Psychological Risk Factors with Risk of Acute Myocardial Infarction in 1119 cases and 13648 controls from 52 countries (the INTERHEART study): case-control study. *Lancet.* 364(9438):953-62.

Safavi, K., D. Bates and S. Chaguturu (2019), *Harnessing Emerging Information Technology for Bundled Payment Care Using a Value-Driven Framework*. Available at: https://catalyst.nejm.org/harnessing-it-bundled-payment-care/

Safety Executive (2015).'*Health and safety in the health and social care sector in Great Britain, 2014/15'*. Available at: http://www.hse.gov.uk/Statistics/industry/healthservices/health.pdf?pdf=health

Siegrist J (2008). Chronic psychosocial stress at work and risk of depression: evidence from prospective studies. *Eur Arch Psychiatry Clin Neurosci*, 258 :115–119.

Sime, Carley (2019) The Cost Of Ignoring Mental Health In The Workplace, *Forbes*. Available at: https://www.forbes.com/sites/carleysime/2019/04/17/the-cost-of-ignoring-mental-health-in-the-workplace/?sh=4012a65a3726#e6565ea3726a

Sohail, M. and Rehman, C.A. (2015) Stress and Health at the Workplace—A Review of the Literature. *Journal of Business Studies Quarterly,* 6, 94-121.

Stineman, M., & Musick, D. (2001). Protection of human subjects with disabilities: Guidelines for research. *Archives of Physical Medical Rehabilitation*, 82 (Suppl. 2), 9–14.

Stinson C. (2020). *Algorithms are not neutral: Bias in collaborative filtering*. Available at: https://www.catherinestinson.ca/Files/Papers/Algorithms_are_not_Neutral.pdf

Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. (2015). Mobile app rating scale: a new tool for assessing the quality of health mobile apps. *JMIR Mhealth Uhealth*;3(1):e27.

Stratton E, Lampit A, Choi I, Calvo RA, Harvey SB, Glozier N (2017). Effectiveness of eHealth interventions for reducing mental health conditions in employees: A systematic review and meta-analysis. *PLoS ONE* 12(12).

Sun W., Khenissi S., Nasraoui O., Shafto P. (2019). *Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering Collaborative Filtering*. Available at: https://core.ac.uk/download/pdf/289241477.pdf

Torous J, Roberts LW. (2017). The ethical use of mobile health technology in clinical psychiatry. *J Nerv Ment Dis,* 205: 4–8.

Torres-Carazo, M. I. ; M. J. Rodríguez-Fórtiz and M. V. Hurtado. (2016). "Analysis and review of apps and serious games on mobile devices intended for people with visual impairment," 2016 IEEE International Conference on Serious Games and Applications for Health (SeGAH), Orlando, FL, USA, pp. 1-8.

Tsintzou V, Pitoura E, Tsaparas P (2018). "Bias Disparity in Recommendation Systems". Available at: https://arxiv.org/pdf/1811.01461.pdf

US Federal Trade Commission (2016). *Lumosity to pay $2 million to settle FTC deceptive advertising charges for its "brain training" program.* Available at https://www.ftc.gov/news-events/press-releases/2016/01/lumosity-pay-2-million-settle-ftc-deceptive-advertising-charges.

van der Graaf S, Vanobberghen W, Kanakakis M, Kalogiros C. (2015). Usable Trust: Grasping Trust Dynamics for Online Securityas a Service. In: Tryfonas T, Askoxylakis I, editors. *Human Aspects of Information Security, Privacy, and Trust*. Cham:Springer International Publishing; 357-368.

Vayena, Effy, Urs Gasser, Alexandra Wood, David O'Brien, Micah Altman (2016). "Elements of a New Ethical Framework for Big Data Research." *Washington and Lee Law Review* Online, 72, 3, 420-441.

Vithanwattana N, Mapp G, George C. (2017).Developing a comprehensive information security framework for mHealth: a detailed analysis. *J Reliable Intell Environ*, 27;3(1):21-39.

Wynia, Matthew K., Steven S. Coughlin, Sheri Alpert, Deborah S. Cummins, and Linda L. Emanuel.(2001). "Shared Expectations for Protection of Identifiable Health Care Information. National Consensus Process." *Journal of General Internal Medicine,* 16, 2,100–111.

Yaghmaei, E., van de Poel, I. (2017). Canvas White Paper 1 – Cybersecurity and Ethics. Available at: https://ec.europa.eu/research/participants/documents/download

Yang, Bian (2016). "What Make You Sure That Health Informatics Is Secure." In *Inclusive Smart Cities and Digital Health,* edited by C. K. Chang, L. Chiari, Y. Cao, H. Jin, M. Mokhtari, and H. Aloulou, 9677:443–48. Cham: Springer.

Yeom, S., Datta, A., & Fredrikson, M. (2018). "Hunting for discriminatory proxies in linear regression models". In *Advances in Neural Information Processing Systems* (pp. 4568-4578).

Zaidel CS, Ethiraj RK, Berenji M, Gaspar FW. (2018). Health care expenditures and length of disability across medical conditions. *J Occup Environ Med.*60(7):631–636.

5.2 Mindset Audit

Agarwal S, LeFevre AE, Lee J, L'Engle K, Mehl G, Sinha C, Labrique A. (2016). WHO mHealth Technical Evidence Review Group.Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. *BMJ.*; 352:1174.

Andersson G, Bergström J, Holländare F, Carlbring P, Kaldo V, Ekselius L (2005). Internet-based self-help for depression: randomised controlled trial. *Br J Psychiatry.* 187:456-61.

Andersson G, Cuijpers P (2009). Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. *Cogn Behav Ther.*; 38(4):196-205.

Andersson G, Hesser H, Hummerdal D, Bergman-Nordgren L, Carlbring P. (2013). A 3.5-year follow-up of Internet-delivered cognitive behavior therapy for major depression. *Journal of Mental Health.*22(2):155–164.

Andrews G, Cuijpers P, Craske MG, McEvoy (2010). PComputer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis., *Titov N PLoS One*. 5(10):e13196.

Andrews G, Cuijpers P, Craske MG, McEvoy P, Titov N (2010). Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis. *PLoS One*., 5(10):e13196.

ARTICLE29 Data Protection Working Party (2018). *Subject: your letter of 7th December 2017 and a new draft code of conduct with the request of a positive opinion from the WP29 under the Data Protection Directive*. Brussels.

Batra S, Baker RA, Wang T, Forma F, DiBiasi F, Peters-Strickland T. (2017). Digital health technology for use in patients with serious mental illness: a systematic review of the literature. *Med Devices*.10:237-251.

Beauchamp TL, Childress JF (2001). *Principles of biomedical ethics.* New York: Oxford University Press.

Birney AJ, Gunn R, Russell JK, Ary DV (2016). MoodHacker Mobile Web App With Email for Adults to Self-Manage Mild-to-Moderate Depression: Randomized Controlled Trial *JMIR Mhealth Uhealth*;4(1):e8.

Eysenbach G. (2011). CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. *J Med Internet Res.* 13(4):e126.

Fenn, K., and Byrne, M. (2013). The key principles of cognitive behavioral therapy. *InnovAiT: Education and Inspiration for General Practice*, 6(9), 570-585.

Gershkovich, Marina ; Rachel Middleton, Dianne M. Hezel, Stephanie Grimaldi, Megan Renna, Cale Basaraba, Sapana Patel, H. Blair Simpson (2021). Integrating Exposure and Response Prevention With a Mobile App to Treat Obsessive-Compulsive Disorder: Feasibility, Acceptability, and Preliminary Effects, *Behav Ther.*, 52(2):394-405.

Gumport NB, Williams JJ, Harvey AG. (2015). Learning cognitive behavior therapy. *J Behav Ther Exp Psychiatry*.48:164-169.

Hidalgo-Mazzei D, Mateu A, Reinares M, et al. (2016). Psychoeducation in bipolar disorder with a SIMPLe smartphone application: feasibility, acceptability and satisfaction. *J Affect Disord.* 200:58–66.

Hsin H, Torous J, Roberts LW (2016). An adjuvant role for mobile health in psychiatry. *JAMA Psychiatry* 73(2):103-4.

Husky M, Olié E, Guillaume S, Genty C, Swendsen J, Courtet P. (2014). Feasibility and validity of ecological momentary assessment in the investigation of suicide risk. Psychiatry Res. 220(1–2):564–570.

Ji-Won Hur, Boram Kim, Dasom Park, and Sung-Won Choi (2018).A Scenario-Based Cognitive Behavioral Therapy Mobile App to Reduce Dysfunctional Beliefs in Individuals with Depression: A Randomized Controlled Trial.*Telemedicine and e-Health*, 710-716.

Jones, N., and Moffitt, M. (2016). Ethical guidelines for mobile app development within health and mental health fields. *Professional Psychology: Research and Practice, 47*(2), 155–162.

Kazdin AE, Blase SL. (2011). Rebooting Psychotherapy Research and Practice to Reduce the Burden of Mental Illness. *Perspect Psychol Sci.*;6(1):21-37.

Kenwright M, Liness S, Marks I. (2001). Reducing demands on clinician's time by offering computer-aided self-help for phobia/panic: feasibility study. *Br JPsychiatry*, 179:456–459.

Koffel E, Kuhn E, Petsoulis N, et al. (2018). A randomized controlled pilot study of CBT-I Coach: Feasibility, acceptability, and potential impact of a mobile phone application for patients in cognitive behavioral therapy for insomnia. *Health Informatics Journal,* 3-13.

Lambert MJ.(2011) What have we learned about treatment failure in empirically supported treatments? Some suggestions for practice. *Cognitive and Behavioral Practice.* 18(3):413–420.

Layard R, Clark DM. (2014). *Thrive: The power of evidence-based psychological therapies*. London: Allen Lane.

McCrone P, Knapp M, Proudfoot J, Ryden C, Cavanagh K, Shapiro DA, Ilson S, Gray JA, Goldberg D, Mann A, Marks I, Everitt B (2004). Cost-effectiveness of computerised cognitive-behavioural therapy for anxiety and depression in primary care: randomised controlled trial. *Tylee A Br J Psychiatry*. 185:55-62.

Natalie Simon, Leah McGillivray, Neil P. Roberts, Kali Barawi, Catrin E. Lewis and Jonathan I. Bisson (2019). Acceptability of internet-based cognitive behavioural therapy (i-CBT) for post-traumatic stress disorder (PTSD): a systematic review, *European Journal of Psychotraumatology*, 10:1.

Parker L, Bero L, Gillies D, Raven M, Mintzes B, Jureidini J, et al. (2018). Mental health messages in prominent mental health apps.*Ann Fam Med.* 16: 338–42.

Pawelski, J. O., & Prilleltensky, I. (2005). 'That at which all things aim': Happiness, wellness, and the ethics of organizational life. In R. Giacalone, C. Dunn, and C. L. Jurkiewicz (Eds.), *Positive psychology in business ethics and corporate social responsibility*. Charlotte, NC: Information Age Publishing, pp. 191–208.

Proudfoot J, Goldberg D, Mann A, Everitt B, Marks I, Gray J (2003). Computerised, interactive, multimedia cognitive behavioural therapy reduces anxiety and depression in general practice. *Psychol Med,*33:217–227

Rey, Y., Marin, C. E., & Silverman, W. K. (2011). Failures in cognitive-behavior therapy for children. *Journal of Clinical Psychology, 67*(11), 1140–1150.

Roberts LW (Ed) (2016). *A Clinical Guide to Psychiatric Ethics.* Arlington, VA: American Psychiatric Publishing, Inc

Rothbaum BO, Hodges L, Smith S, Lee JH, Price L (2000). A controlled study of virtual reality exposure therapy for the fear of flying. *J Consult ClinPsychol,* 60:1020–102639.

Schlosser DA, Campellone TR, Truong B, et al. The feasibility, acceptability, and outcomes of PRIME-D: A novel mobile intervention treatment for depression. *Depress Anxiety*. 2017;34(6):546-554.

Siegler M (1981) Searching for moral certainty in medicine: a proposal for a new model of the doctor-patient encounter. *Bull N Y Acad Med*, 57:56-69

Thoma N, Pilecki B, McKay D. (2015). Contemporary Cognitive Behavior Therapy: A Review of Theory, History, and Evidence. *Psychodyn Psychiatry*. 43(3):423-61.

Torous J, Roberts LW. (2017). The ethical use of mobile health technology in clinical psychiatry. *J Nerv Ment Dis*, 205: 4–8.

van Ballegooijen W, Cuijpers P, van Straten A, Karyotaki E, Andersson G, Smit JH, Riper H (2014), Adherence to Internet-based and face-to-face cognitive behavioural therapy for depression: a meta-analysis. *PLoS One*. 9(7):e100674.

Vittengl JR, Clark LA, Dunn TW, Jarrett RB (2007). Reducing relapse and recurrence in unipolar depression: a comparative meta-analysis of cognitive-behavioral therapy's effects. *J Consult Clin Psychol.* Jun; 75(3):475-88.

Werner-Seidler, A., Huckvale, K., Larsen, M.E. *et al.* (2020). A trial protocol for the effectiveness of digital interventions for preventing depression in adolescents: The Future Proofing Study. *Trials,* 21, 2.

Wright JH (2004) Computer-assisted cognitive-behavior therapy, in Wright JH (Ed.). *Cognitive-Behavior Therapy*. Washington, DC, AmericanPsychiatric Publishing, pp 55–82.

Wright JH, Basco MR, Thase ME (2006). *Learning Cognitive-Behavior Therapy: AnIllustrated Guide*. Washington: American Psychiatric Publishing.

Wright, J. H. (2006). Cognitive behavior therapy: Basic principles and recent advances.*Focus,* 4,173–178.


eticas