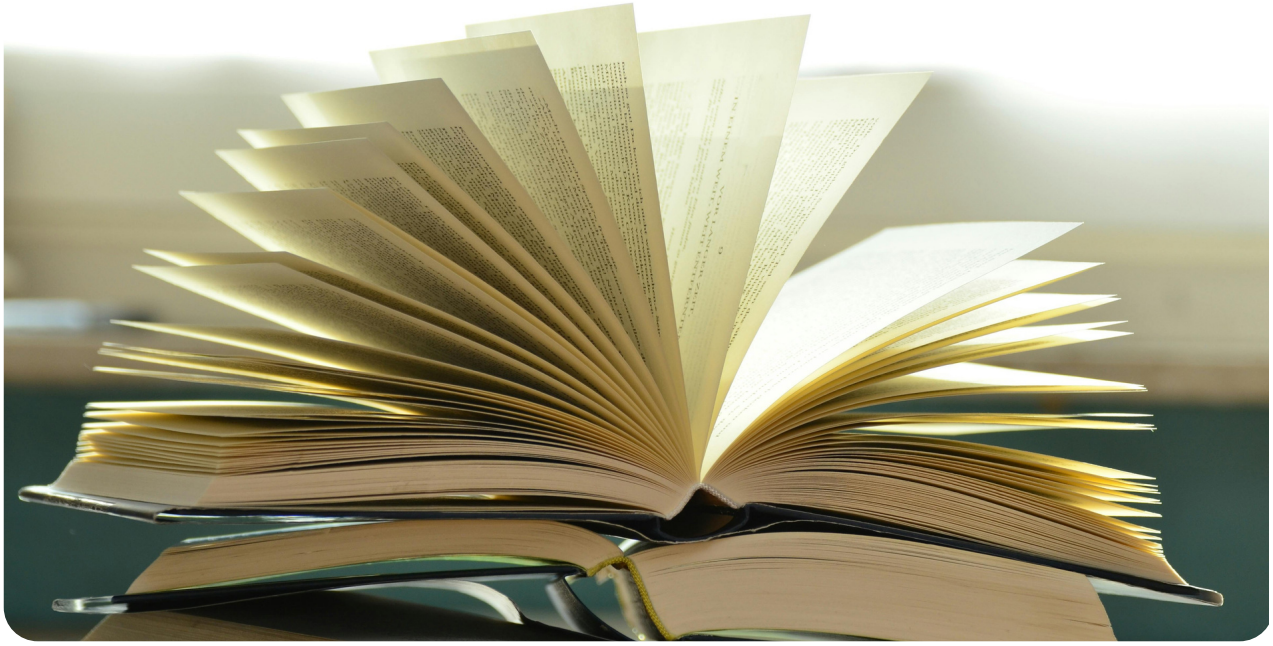
EDUCATING YOUR PEOPLE ON RESPONSIBLE AI

A Values Canvas Case study

TABLE OF CONTENTS

GETTING STARTED WITH RESPONSIBLE AI



Embracing AI is no longer an option, it is an expectation. However, AI is known to be risky business, as it comes with significant investment requirements, up to 93% failure rates, and a concerning lack of confidence in today's context of countless AI mishaps. There are many ways that AI can go wrong, but in a world demanding the adoption of this cutting-edge tool, how can companies ensure it goes right?

This is where Responsible AI & Ethics comes in. The only way to consistently grow customer trust, mitigate unnecessary harmful risks, and get the most out of an investment in this technology, Responsible AI practices are quickly becoming the standard of operations for success in AI.

So, where do you start?

Originating from the book *Responsible AI* by Olivia Gambelin, **the Values Canvas** is a holistic management template for developing Responsible AI strategies and documenting existing ethics efforts. Designed to drive success in developing and using AI responsibly, it brings clarity on where to start and if something is missing in a company's journey to becoming Responsible AI-enabled.

THE VALUES CANVAS

The Values Canvas is made up of three pillars: **People**, **Process**, and **Technology**.

People looks at who is building or using AI, Process is focused on how AI is being built or used, and Technology is about what AI is being built or used. Each pillar is broken down into three elements, with each element capturing a specific need that your Responsible AI initiatives must fill. Another way to think about this is that the elements highlight the impact points in which you can translate your ethical values into reality for your company and technology through strategic solutions. You can hone in and work on a single element solution, or zoom out to understand how all the element solutions work together to create an efficient and effective Responsible AI strategy. In the case of the People pillar, the three elements are **Educate**, **Motivate** and **Communicate**.

In this case study we focus on the first of the three People elements: **Educate**. In this element, we are looking to meet the need of ensuring your people have the skillsets and knowledge bases necessary to critically engage with using ethics in the context of AI. An Educate solution is then any effort that is designed to up-skill individuals and teams in the practical application of ethics in AI.

This case study is one of a nine-part series on the Values Canvas. To explore the Values Canvas, access the full case study series, and discover further resources, visit ***www.thevaluescanvas.com****.*

THE VALUES CANVAS

The Values Canvas is made up of three pillars: **People**, **Process**, and **Technology**.

People looks at who is building or using AI, Process is focused on how AI is being built or used, and Technology is about what AI is being built or used. Each pillar is broken down into three elements, with each element capturing a specific need that your Responsible AI initiatives must fill. Another way to think about this is that the elements highlight the impact points in which you can translate your ethical values into reality for your company and technology through strategic solutions. You can hone in and work on a single element solution, or zoom out to understand how all the element solutions work together to create an efficient and effective Responsible AI strategy. In the case of the People pillar, the three elements are **Educate**, **Motivate** and **Communicate**.

In this case study we focus on the first of the three People elements: **Educate**. In this element, we are looking to meet the need of ensuring your people have the skillsets and knowledge bases necessary to critically engage with using ethics in the context of AI. An Educate solution is then any effort that is designed to up-skill individuals and teams in the practical application of ethics in AI.

This case study is one of a nine-part series on the Values Canvas. To explore the Values Canvas, access the full case study series, and discover further resources, visit ***www.thevaluescanvas.com****.*

THE NEED

Introducing Koa Health

Koa Health is a digital mental health company that provides a range of mental health services from supporting mental well-being through to supporting the treatment of anxiety, depression, and PTSD. Koa operates across 60 countries, with offices in Spain, the UK, and the US.

Koa Health's mission is to provide mental health for everyone, using technology to address the huge gap between the demand for mental health support and the limited supply of mental health professionals. AI is an important part of Koa's offering, as it helps to improve the care recommendations that the service provides, leading to better outcomes and more personalized support.

Building Trust in AI for Mental Health

From the very outset, Koa recognized that using AI in mental health required a lot of trust. Users would need to trust Koa to share data about their mental health, to trust in the recommendations from Koa, and to trust that following the recommendations would improve their mental health. Essentially, Koa was dealing in the business of user trust, and that is not something you can take shortcuts to achieve.

For this reason, Koa decided to establish an ethics strategy very early on in its operation in order to work on building the long term capital of user trust. The strategy encompassed five principles: improving your health and happiness, putting you in control, being understandable and transparent, securing your data, and being accountable.

Each of Koa's principles was made concrete by a number of commitments. For instance: one of the commitments under 'improving your health and happiness' is to not make recommendations based on discriminatory bias; and under the 'being accountable' principle is a commitment to regularly hold external audits of progress against the ethics strategy.

Staying true to the commitment of external audits, Koa shortly after embarked on its first external audit. To complete the audit, Koa had to take part in a series of interviews, including the team members on the Research and Development team responsible for creating Koa's AI. It was soon discovered that some members of the team tended to view algorithmic bias as being something that only related to the data, and so there was a limit to what could be done to address it. When it was explained that there was a lot more to algorithmic bias than the quality of the data set, and that there were lots of mitigation strategies, the team was eager to learn more.

Unwanted bias in mental health services is a quick path to failure, and possibly even lawsuits. Without a robust understanding of the different types of bias and mitigation techniques, the Koa team was opening itself up to unnecessary risks, and so jeopardizing their users' trust. Clearly, the team at Koa had uncovered an important and urgent need for education on bias mitigation.

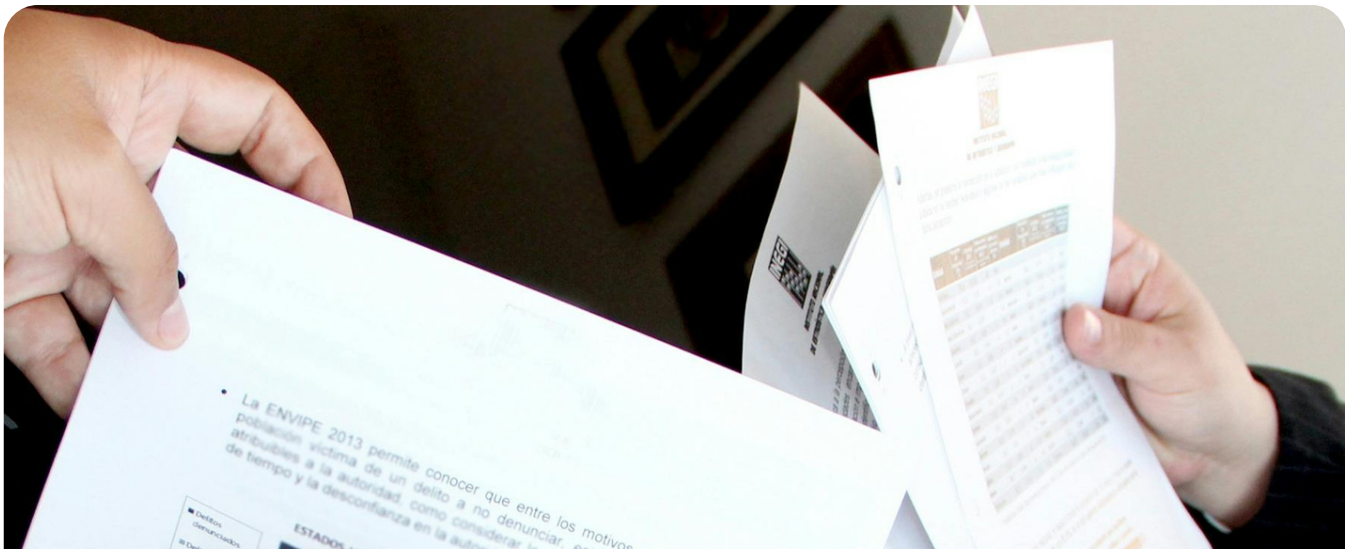
THE SOLUTION

The Koa team was in need of training on bias mitigation, or, in other words, the Koa team was in need of an Educate solution. Educate is the first element of the People pillar in creating responsible AI, which means that the solution for this element needed to upskill the Koa team.

To address this educational need, Koa engaged Eticas, the AI auditing company who undertake Koa's external audits, and Professor Carlos Castillo at Universitat Pompeu Fabra in Barcelona to help develop a solution. This led to the creation of a substantial guidance document, accompanied by training sessions. This document was a compilation of the evidence on algorithmic bias: how to define it, how to measure it, and the various methods to mitigate and manage it. It was an incredibly thorough resource for the Koa team, although it is fair to say that it was very academic.

If Koa were filling out the Values Canvas, their Educate solution statement would have looked like the following:

Koa R&D team needs **training in bias mitigation** that will be delivered **through a robust guidance document.**



Things seemed to be running smoothly after the guidance document was finished. However, when progress was checked in the next external audit, it was found that the R&D team were struggling to translate the information in the guidance document into practice. From discussions with the team it turned out that, even though they are academically minded people, the guidance that had been produced as a first attempt at an Educate solution was just too dense. It was great for an academic seminar, not so useful when trying to incorporate it into a busy work day.

Although Koa understood **who** needed training on **what** information, they had miscalculated **how** it would be delivered.

Diving back into the thick of it, Koa reassessed what kind of format was needed to effectively deliver the necessary bias mitigation training to the R&D team. Working again with Eticas and Professor Castillo, Koa this time gave clarity to their brief, asking a step-by-step 'how to' guide for bias mitigation that would be developed iteratively alongside members of the Koa R&D team to ensure that the guide met their needs.

After multiple discussions and iterations, the team successfully created a ‘how to’ guide based on a seven-step process:

1

Contextual analysis
To develop an initial set of hypotheses for algorithmic bias

2

Mapping the user population
The aim here is to flesh out which groups of users might be at risk of bias

3

Prioritizing protected group
Based on the level of potential harm (or lack of benefit) that may befall a protected group as a result of algorithmic bias

4

Selecting Fairness Metric
Understanding that some trade-offs may need to be made to prioritize the protected groups

5

Calculating bias
Applying the fairness metric to data collected

6

Analyzing results
To understand if initial hypotheses developed in Step 1 have been validated, and testing new hypotheses

7

Mitigating bias
Making the changes required to mitigate any bias identified and being explicit about trade-offs made with other needs

To ensure that it was as user-friendly as possible, real-world examples using Koa Health's products and previous audits were included in the guide. Not to get too metaphysical, it also included a section on how to use the 'how to' guide. This responded to one of the most challenging aspects of the previous guidance, which was that R&D members found it difficult to easily understand which aspects were most relevant in different circumstances. Here's a peek inside from the guide:

a) When planning an algorithm

The first stage when this document is used is as part of algorithmic planning. The "problem" to be addressed by the system is framed and examined, which requires **figuring out possible sources of algorithmic bias** to be considered within the training data and model design, particularly if they negatively affect women, minorities, or other groups that experience structural discrimination.

Related steps: 1 & 2.

b) During algorithm development

During algorithmic development, **exploratory algorithmic bias analysis** can be conducted in parallel to the iterative analysis process addressing how efficient the code is in solving the problem. First, biases in training sets should be explored, and if found, characterized. Second, algorithmic bias metrics should be computed alongside standard efficacy metrics for a modeling problem (e.g., accuracy, precision, recall, area under ROC curve).

Related steps: 2 to 7.

c) Before deploying an algorithm into production*

The next critical moment when this guide can be used is before the algorithm is put into production. As part of validation processes, this document should be used to characterize algorithmic biases and **ensure algorithmic fairness before deployment**. To do so, you can select relevant metrics and conduct tests to produce information about biases; this may lead to mitigation strategies, such as changes in the model.

Related steps: 5 to7.

d) After making major changes to an algorithm

Lastly, the above analysis should be applied after significant changes are made to a model. These include extending a training set substantially, introducing new types of data, or changing the learning scheme or target for optimization.

Related steps: all steps.

The new how-to guide, along with a further training session on algorithmic bias from Professor Castillo, and his materials for future use, created the perfect combination for translating the necessary knowledge into action for the R&D team.

If the Koa team were to edit their Educate solution statement to reflect the new delivery method, it would look something like the following:

Koa R&D team needs **training in bias mitigation** that will be delivered **through a step-by-step how-to guide and training sessions.**

THE OUTCOME

The R&D team found the new how-to guide and training much more accessible, with both receiving positive feedback. More importantly, the guide was actually used by the team, impacting how they both created new algorithms and continued to evaluate and evolve existing ones. In fact, Koa has even published a paper on how it used the 7-step process to assess the fairness of its mental wellness app, Foundations. The assessment found no disparate impact nor undesired bias.

The team has even used the guide beyond its original use case of creating and evaluating algorithms. For instance, it significantly aided the team when writing the fairness analysis that accompanied a paper that Koa published in Nature Medicine on using machine learning to predict mental health crises.

In terms of wider impact on Koa, in 2022 it was awarded a Welcoa Well-being Trailblazer award. The judges explicitly referenced Koa's efforts on ethics and trust, within which the 7-step guide is an important component.

Speaking to Dr Aleksander Matic, R&D Director at Koa Health, he was clear about the importance of the guide.

"The 'how-to' guide stands out not only for its comprehensive coverage of Koa's ethical pillars but also for respecting UX principles. This was key for its adoption and impact in our team. Firstly, it was easy to use — it provides clear instructions on what needs to be done and why it's important. Secondly, it adopts a user-centric approach, considering both us - the researchers who are applying the guide - and the end users who are the ultimate beneficiaries of the ethical principles. Thirdly, it addresses context, recognizing our challenges in developing novel algorithms, especially within high-stakes domains like mental health care. This guide is not just a manual, it ended up being a testament to our commitment to ethical innovation in algorithmic design."

An important takeaway to emphasize here is that the best solutions take time to refine, and sometimes you need to challenge assumptions in order to find the right fit for your team. In Koa's case, there had been the assumption that the R&D team would prefer to be presented with all of the evidence for algorithmic fairness within an academic-style document. The team members did want to see the evidence, however they needed to deploy the knowledge in a busy work environment that was not necessary conducive to heavy academic papers. Thus the information needed to be robust, but also presented in a rapidly accessible manner that aligned to the different tasks that they faced within their work

So although Koa had correctly identified the need for an Educate solution from the start, the original solution did not fit the format the teams needed. It was not the case that the education on bias mitigation was useless, it was instead the case that the delivery format of the education needed to be adjusted to better fit with the team's specific needs.

THE AUTHORS

Oliver Smith

Oliver is a member of the Ethical Intelligence Network and the Founder and Director of Daedalus Futures, a Responsible AI consultancy that supports organizations to create strategic advantage by using AI responsibly. He was one of the founding executive team at Koa Health, which span out of Telefonica Alpha in 2020 with a €50m Series A, where he was responsible for strategy and ethics. Prior to Koa Health he was Director of Strategy and Innovation at Guy's and St Thomas' Charity, responsible for investing £100m over five years in innovations across acute, primary, and integrated care, biomedical research and digital health start-ups. He was a Senior Civil Servant in the UK Department of Health; responsible for UK Tobacco Control Policy, and UK Obesity Policy before that. Oliver was also a Policy Adviser in the Prime Minister's Strategy Unit under Tony Blair. He has an MA in Politics, Philosophy, and Economics from Oxford University.

Olivia Gambelin

One of the first movers in Responsible AI, Olivia is a world-renowned expert in AI Ethics whose experience in utilizing ethics-by-design has empowered hundreds of business leaders to achieve their desired impact on the cutting edge of technological innovation. As the founder of Ethical Intelligence, the world's largest network of Responsible AI practitioners, Olivia offers unparalleled insight into how leaders can embrace the strength of human values to drive holistic business success. She is also the author of the book *Responsible AI: Implement an Ethical Approach in Your Organization* with Kogan Page Publishing, and the creator of The Values Canvas, which can be found at www.thevaluescanvas.com.

To access the Values Canvas download and further case studies, visit:

www.thevaluescanvas.com

–

To learn more about why, how and when to use the Values Canvas, read:

Responsible AI: Implement an Ethical Approach in Your Organization