A Values Canvas Case study

SEEKING DOMAIN FEEDBACK FOR AI ALIGNMENT

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 Technology pillar, the three elements are **Data**, **Document** and **Domain**.

In this case study we focus on the third of the three Technology elements: **Domain**. In this element, we are looking to fill the need to incrementally iterate and refine AI models into ethical value alignment through insights from people directly affected by the AI systems. A Domain solution is a specified time and place in the AI lifecycle for expert and customer feedback.

This case study is ninth 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

Aiforgood.asia

Aiforgood.asia is an international NGO dedicated to using AI for social good. With a mission to address global challenges such as improving health and welfare, reducing inequality, combating climate change, and promoting conservation efforts, Aiforgood.asia actively engages in research, implementation, and advisory services to ensure AI technologies are ethically deployed. Aiforgood.asia operates at the intersection of AI ethics and practical, community-based solutions. Their work spans research publications, advocacy for responsible AI, and partnerships with various organizations, governments, and communities. By collaborating with diverse stakeholders, Aiforgood.asia aims to create AI systems that serve the greater good..

Aligning LLM Integration with Organizational Values

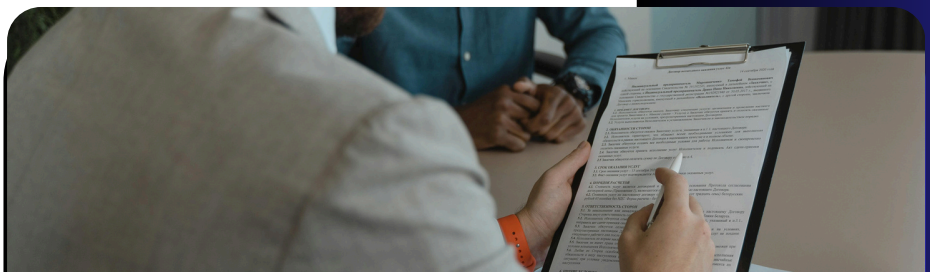
With the onset of generative AI solutions, Aiforgood.asia saw an opportunity to utilize LLM technology in order to enhance their content creation to reach wider audiences. However, the organization recognized that they needed to ensure the AI-generated content would align with their ethical standards. Aiforgood.asia aims to set an industry example in AI-generated content creation based on responsible foundations, and to drive the conversation on what standards should be expected in generated content.

Aiforgood.asia is committed to utilizing AI responsibly to support their human researchers in creating content. To do so effectively, they require a mechanism to ensure that AI-generated content reflects ethical standards while avoiding issues like bias or misinformation. In order to assess if the systems were in alignment, the organization realized they would need input from domain experts as to whether or not the content was reflective of responsible and ethical standards, as well as community members as to whether or not the content was reflective of Aiforgood.asia's mission and values.

Aiforgood.asia recognized that their AI systems must be evaluated by external experts and community members to avoid ethical blind spots and ensure the technology serves the communities for which it is intended. Without proper domain-specific feedback, Aiforgood.asia risks deploying AI solutions that fail to meet responsible and community standards. This misalignment could undermine their credibility, decrease community trust, and jeopardize funding and partnerships. The absence of a robust feedback mechanism could ultimately prevent them from achieving their mission of using AI to drive social good.

Aiforgood.asia's Needs:

- Immediate: Ensure AI-generated content aligns with community and ethical standards
- Mid-term: Drive the conversation on what standards should be expected in generated content
- Long-term: Set an industry example in AI-generated content creation based on responsible foundations



THE SOLUTION

Aiforgood.asia wanted to harness the potential of LLM technology to help drive effective content creation to enhance, rather than hinder, its human researchers. To do so, the organization needed a solution that would effectively incorporate expert and community feedback into the AI lifecycle. In other words, Aiforgood.asia was in need of a Domain solution. Domain is the third element of the Technology pillar in creating responsible AI, which means that the solution for this element needs to specify time and place in the AI lifecycle for expert and customer feedback.

A Domain solution statement looks like the following:

"_____ feedback is needed _____ during _____."

whose *how often* *when*

To start, Aiforgood.asia began with identifying the design and development stages of the AI lifecycle as where human feedback would be most critical and impactful. Throughout the design, development, testing and deployment phases, iterative feedback loops would allow for continuous refinement of the AI system.

As Aiforgood.asia looked to design the necessary feedback loops, they considered the need forthe feedback loops to scale across multiple projects, as well as the ability to maintain flexibility in deploying AI solutions in varied community contexts. Additionally, practical concerns, such as how easily human researchers could collaborate with the AI system, and how user trust could be maintained through transparency, influenced the design of the feedback mechanism.

To gather the needed domain expertise and community input, three primary feedback loops were created and deployed:

Feedback Loop 1: Internal Users

- **Why?** The internal researchers and writers at Aiforgood will be the primary users of the AI system. Their feedback is crucial to ensuring that the solution is user-friendly, effective, and aligned with the organization's objectives.
- **When?** Internal users will be involved in the design, development, and prototyping phases to ensure that the system meets both their needs and the organization's goals. Feedback is sought on a monthly basis.

Feedback Loop 2: Subject Matter Experts

- **Why?** Subject matter experts are the domain experts who will validate the content produced by the AI system. They represent the "human-in-the-loop" component that ensures the AI operates within ethical and practical boundaries.
- **When?** Subject matter experts will be consulted throughout the process, particularly during the validation and testing phases, to ensure that the content aligns with ethical standards and domain-specific expertise. Feedback is sought on a quarterly basis.

Feedback Loop 3: Readers/Consumers of Blogs & Articles

- **Why?** Readers and consumers of Aiforgood.asia's community are the end users of the AI-generated content. Their feedback on the trustworthiness and quality of the content is critical for evaluating the system's performance and ensuring that it meets societal expectations.
- **When?** Readers will be regularly consulted through tools like "Trust Surveys" to provide continuous feedback on the quality and reliability of individual articles and blog posts. This feedback will be used to iteratively improve the AI system's content generation. Feedback is sought on a quarterly basis.

These feedback loops help ensure that the AI system is developed and deployed with continuous human oversight, ensuring ethical alignment and responsiveness to both domain expertise and community expectations.

Internal User Feedback Loop In Action

Aiforgood.asia's system employs Retrieval-Augmented Generation (RAG) to ensure that content generation remains anchored in factual, verified information. By integrating RAG, the system retrieves relevant documents from a curated database uploaded from hand-picked authoritative sources to augment the LLM's generation process, constraining the model to produce content that aligns with the evidence presented in these documents. The stringent guardrails are implemented to limit the LLM's contributions outside the scope of the provided documents, ensuring that only pre-verified, fact-based information is synthesized.

An internal researcher oversees this entire process, serving as the lead author who reviews, approves, and certifies the content before dissemination. This method is designed as a continuous feedback loop to assess and gauge reader confidence in the accuracy and reliability of the information. Through this controlled and transparent process, the feedback loop works to establish a trusted knowledge generation system, presenting a robust case study in responsible AI deployment.

Tiered Feedback System Approach

The final Domain solution consisted of a multi-tiered feedback mechanism that spanned the entire AI lifecycle. Expert input was used to assess the ethical and social implications of the AI system, while community feedback ensured the AI remained practical and beneficial for users. This iterative process allowed for ongoing refinement of the system's features and outputs, ensuring that it met both ethical standards and real-world needs.

If Aiforgood.asia were filling out the Values Canvas, their Domain solution statement would look something like the following:

Internal users, external consumers, and subject matter experts feedback is needed **on a monthly and quarterly basis** during **design, development and testing phases.**

THE OUTCOME

The implementation of the Domain solution resulted in several positive outcomes for Aiforgood.asia.

First, it significantly improved the quality of AI-generated content by incorporating regular feedback from domain experts. It ensured that content not only met technical standards but also was reflective of ethical principles and the organization's values. Positive indicators included increased trust in the AI-generated content, improved user engagement, and testimonials from stakeholders praising the relevance and impact of the AI system.

Second, continuous community feedback allowed the organization to tailor their AI solutions to better meet the needs of their audience, leading to enhanced user satisfaction and trust. In the deployment phase, ongoing monitoring mechanisms such trust surveys linked at the bottom of each published blog or article allowed for direct user feedback similar to the smiley faces in an airport bathroom. In this way the organization will be able to collect direct feedback on the trust readers have in the content being published by Aiforgood and make any needed adjustments to this process to enhance trust.

By incorporating these feedback mechanisms, Aiforgood.asia was able to demonstrate a commitment to responsible AI practices, setting a positive example for other organizations looking to leverage generative AI technologies.

THE AUTHORS

Jesse Arlen Smith

Jesse is a Global Data & AI consultant, Business Leader, AI Researcher, and international speaker on the responsible development and deployment of AI. He believes in driving innovation by empowering people, companies, and governments to harness the power of AI to improve society and protect the environment. Currently, he is the Regional Business Development Lead, Data & AI, and Global Blackbelt team member at Crayon Group a worldwide leader in data and AI transformation services. He is charged with opening up emerging markets, and scaling the regional Data and AI consulting and solution delivery practice. Jesse is also the President and Founder of Aiforgood Asia, an international NGO that conducts research and implements projects that help ensure AI is being used to improve health and welfare, reduce inequality, fight climate change, and aid conservation efforts that protect people and the planet. Jesse, originally from Canada, has been in Asia since 2005 and currently works across borders, living with his wife in Hanoi, Vietnam.

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 the book:

Responsible AI: Implement an Ethical Approach in Your Organization