DOCUMENTING THE DETAILS

DRIVING RESPONSIBLE DECISION-MAKING IN AI

A Values Canvas Case Study

TABLE OF CONTENTS

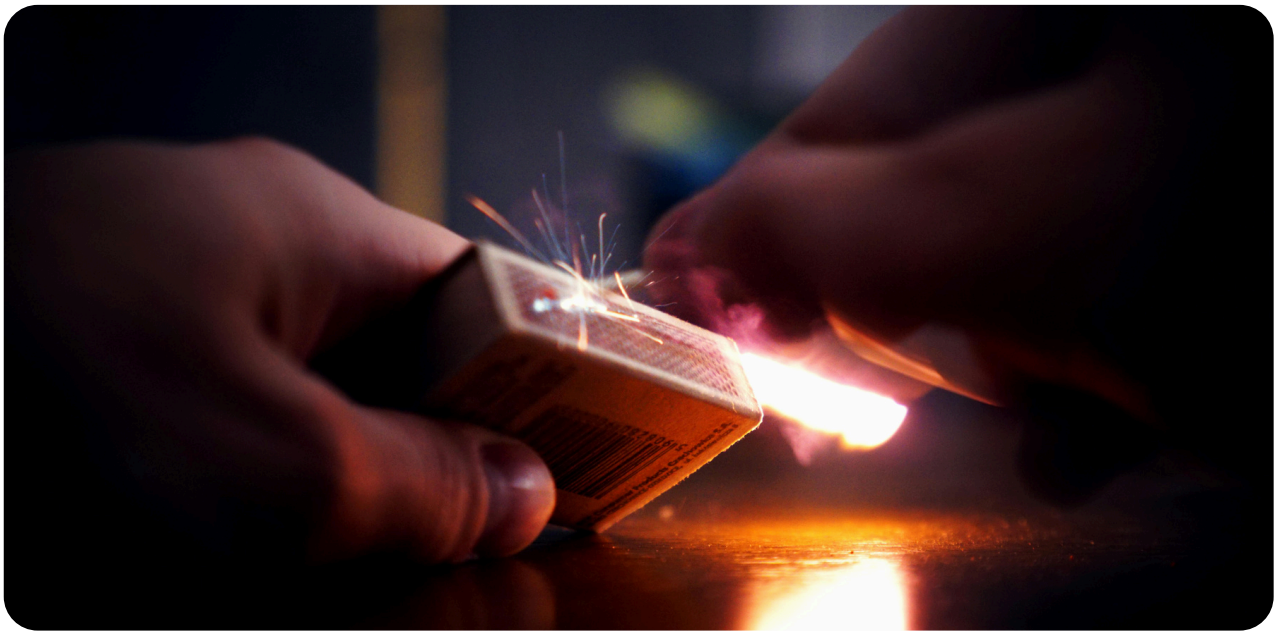
GETTING STARTED WITH RESPONSIBLE AI

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

In this case study we focus on the first of the three Technology elements: **Document**. In this element, we are looking to fill the need to create transparency of the ethical decisions being taken during the AI development lifecycle. A Document solution is the documentation of all ethics related decisions taken during the lifecycle of an AI model.

This case study is eight 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 NewsLens

NewsLens is a company that leverages cutting-edge AI technology to provide nuanced media categorization and analysis solutions. Their tools empower media outlets, educational institutions, and public sector agencies to access insights that enhance research and decision-making. Part of this involves classifying news content, enabling researchers to narrow their focus to specific areas of interest.

**NewLens is a fictional company invented for the purposes of this case study*

Setting Standards in Content Classification

The leaders at NewsLens have set high standards for excellence, including industry-leading product performance and ethical product development. For the product owners, this means crafting clear specifications that align with the needs of the end user. For the data scientists on the software development team, this means creating highly accurate classification systems, and engaging with ethical vendors for any work that needs to be outsourced.

When classifying news content, some categories in the NewsLens platform are relatively straightforward to implement, such as various types of sports, technologies, or health issues. But NewsLens started getting customer requests for more political categories, particularly a category for terrorism. They recognized that this requires careful consideration, since what is considered “terrorism” can vary depending on perspective.

The product owner researched the issue, and landed on US law as the basis for the definition: “premeditated, politically motivated violence perpetrated against noncombatant targets by subnational groups or clandestine agents.” Following standard practices in software development, the product manager included this definition in the design document that the developers would follow. The data scientists then set off on their work to build a machine learning model that would assign the appropriate articles to the “terrorism” category.

The model needed to be trained on thousands of articles, each labeled as either “about terrorism” or “not about terrorism”. A standard practice for labeling at this scale is to outsource the work to a vendor who manages large teams of annotators, who are given instructions on how to read the articles and choose the correct label. The data scientists engaged a vendor known for ensuring fair pay for annotators, and provided the news data along with labeling instructions that included the definition of terrorism provided by the product manager.

Following another standard practice, the data scientists requested that at least two or three individual annotators read and label each article. This is a method for quality assurance, because if multiple annotators independently choose the same label, it is an indication that they are each performing well on the task.

As the annotators worked through the first batch of articles, they frequently assigned conflicting “terrorism” labels, prompting data scientists to clarify the initial guidelines to better help annotators choose the correct label. Unfortunately label consistency did not improve as much as the data scientists had hoped, but with pressure to meet hard deadlines, they moved forward to train the model with the data they had collected so far.

Once the model was in production, NewsLens users began to complain about poor quality, and pointed out some specific issues. For example, when they searched news using the “terrorism” category, they were not seeing some important articles on a recognized terrorist organization’s recent bombing of train tracks on a commuter rail line. What’s worse, the category included a number of articles about Islamic culture that clearly had nothing to do with terrorism.

The team knew they needed to comb through their processes to see how this disconnect arose between user expectations and product design, and how they could fix it. Unfortunately when they reviewed the project documentation, there were no clear answers.

NewsLens’ Needs:

- Immediate: Uncover the root of the problem, and reassure customers with a plan to fix it.
- Medium term: Rerun the project with improved processes and release a new model.
- Long term: Create a new system for documenting model building policies and decisions.

THE SOLUTION

The NewsLens team had a commitment to high accuracy and ethical practices, but when this project's outcome didn't meet those standards, they struggled to find the root cause in their project documentation. In other words, NewsLens was in need of a Document solution. Document is the second element of the Technology pillar in creating responsible AI, which means that the solution for this element needs to result in consistent records of how the team meets these commitments.

A Document solution statement looks like the following:

"_____needs to document _____ decisions _____."

who *which* *where*

The product manager and data scientists knew they needed to dedicate time to a project retrospective to understand the key decisions that led to the model's poor performance. The first area they dug into was the root cause for inconsistent labeling by annotators. This was known to correspond with poor model quality, because it meant that the labels were less likely to be correct, and the model was learning from those labels.

The data scientists had followed their usual procedures for clarifying confusion about labeling guidelines, so why did the annotators continue to arrive at different labels for the same news articles? They met with the vendor's annotation team lead to hear more from the annotators' perspective, and what they heard surprised them.

The main issue for annotators during the project was not confusion around the guidelines, but rather fatigue from repeated exposure to violent and disturbing news stories. It has been well established that data workers tasked with identifying toxic content such as hate speech and abuse experience trauma as a result of their job. The annotators working on terrorism related articles were protecting themselves by not reading the articles carefully. This led to the higher rates of inconsistency, since each annotator was more likely to be making a quick guess as to the correct label.

While this explained the general quality issues, the team next wanted to understand why they were specifically not classifying the train track bombings as terrorism. They looked back at the design document and found the original definition supplied by the product manager: “premeditated, politically motivated violence perpetrated against noncombatant targets by subnational groups or clandestine agents.” This news story seemed to fit the definition — or did it? The data scientists remembered that they had relayed a question from annotators about what constituted “violence”, and the product manager had verbally responded that violence must result in injuries. The data scientists had updated the guidelines with this clarification, so annotators did not label attacks on infrastructure as terrorism, even if the attack had potential for injuries.

When the product manager looked through the annotation guidelines, he realized that there were a number of small details which did not accurately reflect his original intentions. Even though the design document ostensibly contained all the information needed to meet customer expectations, the real source of truth was in the annotation guidelines.

Finally, the team wanted to understand why articles about Islamic culture were being incorrectly assigned to the “terrorism” category. This kind of bias often arises from how the data was originally collected. The data scientists checked their project documentation and saw that they had sampled articles from the last ten years in the “world events” and “politics” categories from major news publications around the world.

On the face of it this seemed like a reasonable approach, but when the team considered it more deeply, they realized that all of the publications were in English and tended to reflect a Western viewpoint that were more likely to portray Islamic culture in a negative light. The machine learning model picked up on this pattern and amplified this bias by treating references to Islamic culture as strong signals for terrorism. This led to the inaccuracies that customers were complaining about.

These investigations led the team to realize that their commitment to ethics and to product quality go hand in hand. Furthermore, they needed to update their project documentation standards to clearly capture the ethical decisions being made. They agreed on three changes.

First, document how the well-being of the annotators is being assured. This goes beyond working with vendors who provide transparency into how annotators are adequately compensated. It should also make note of any hardships the annotators might experience, and how that is being addressed. In the case of the news content on terrorism, the team worked with the vendor to experiment with alternating short periods of work on stressful content with longer periods on more neutral content. Annotators were given full disclosure on the nature of the work, and had the opportunity to opt out without any repercussions on their employment.

Second, consider the annotation guidelines to be the final source of truth for product design when it comes to defining news categories. As such, the latest version should be easily accessible to all stakeholders, not just the data scientists and annotators. This means that the product design document should link to the current version of the guidelines, rather than just including the original high level conception of the category definition. Stakeholders are required to read and sign off on changes to the guidelines at key stages of the project, knowing that it will directly impact the functionality of the product.

Finally, data scientists are required to document any potential biases that could arise in the model training datasets, and the mitigation measures they have taken. Part of that process includes meeting with subject matter experts in the relevant domain who are aware of the ethical complexities.

The team also realized that these documents needed to be centralized so they were visible beyond the data science team. Previously, product management was the only team tasked with centralizing information in order to align teams on product specifications. They decided that an easy-to-access data governance portal would be a good place to ensure that the data scientist team's documentation on ethical AI decisions was kept updated and readily available.

If the NewsLens team were filling out the Values Canvas, their Document solution statement would look something like the following:

Data scientists need to document **(1) provisions for annotator well-being, (2) up-to-date annotation guidelines, and (3) mitigation efforts for potential training data bias** decisions **in the data governance portal.**

Product managers need to document **their approval of annotation guideline revisions** decisions **in the data governance portal.**

THE OUTCOME

Implementing the new documentation protocols and ethical guidelines had a transformative impact not only on the project, but also the organization as a whole.

Specifically for the terrorism project, the data scientists implemented several key changes to effectively manage risk. They expanded the dataset to include more diverse articles, such as positive or neutral stories related to Islamic culture. This not only addressed the issue of biased data but also created a less stressful environment for the annotators. By alternating between stressful and neutral content, the annotators could work more effectively without experiencing burnout.

These changes to the data collection process led directly to improved data quality and reduced bias. The new model achieved a much higher level of accuracy, which NewsLens' customers immediately noticed and appreciated. The ability to find reliable and relevant results in the "terrorism" category restored their confidence in NewsLens.

Beyond the terrorism project, NewsLens made the annotation guidelines the definitive source of truth. By ensuring they were linked to design documents, the entire team, including product managers and stakeholders, gained better visibility into the decision-making protocols for projects. This transparency allowed for more informed revisions and improvements across various news categories.

Building on this success, the team began exploring comprehensive data governance frameworks such as the Data Nutrition Project, whose researchers have developed a standardized format for documenting datasets to ensure transparency and mitigate bias. The data scientists at NewsLens not only adopted these frameworks but also became active contributors, helping to further develop these critical resources.

The implementation of the data governance portal played a crucial role in centralizing all relevant documentation. This portal became a repository for all updates regarding annotator well-being provisions, annotation guidelines, and bias mitigation efforts. It ensured that the information was accessible to all team members, fostering a culture of transparency and continuous improvement.

Overall, the integration of ethical guidelines and comprehensive documentation standards not only resolved the immediate issues with the terrorism classification project but also established a robust framework for future projects. The data scientists' dedication to improving annotator conditions and addressing biases led to the creation of more accurate and ethically sound models. The centralized documentation allowed for greater collaboration and oversight, enabling the team to maintain high standards of accuracy and responsibility. As a result, NewsLens has become recognized as a leader in responsible data practices, setting a new benchmark for providing high-quality, ethically developed AI solutions that meet the evolving needs of its users.

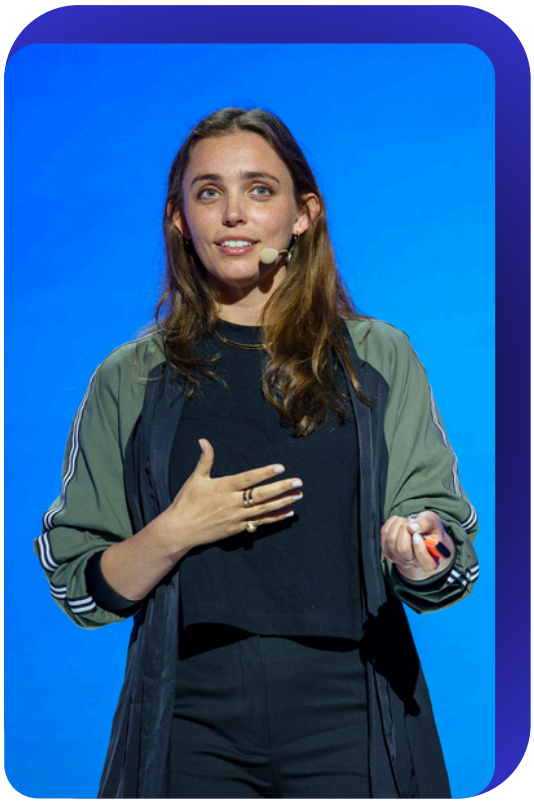
THE AUTHORS

Karin Golde

Karin Golde is the founder of West Valley AI, providing strategic advising to organizations to ensure their AI implementations are both ethical and effective. She has deep expertise in AI-driven applications, with a PhD in Linguistics and over two decades of work on natural language processing for enterprise software. Her experience includes executive leadership roles at startup companies, as well as heading the language engineering division for the AI Data team at Amazon Web Services. At the heart of her work is a passion for bridging the gap between technology and humanity. For more insights on AI data governance, check out Karin's blog, "Good Judgment".

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