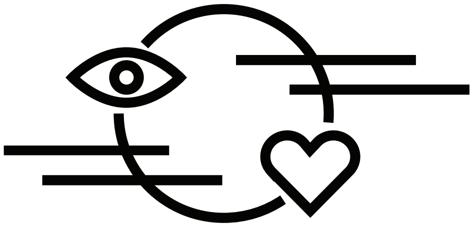
An Empathetic Design Framework for Humanity-Centered AI: A preventative approach to developing more holistic, reliable, and ethical ML products



Written by Dan Silveira

Submitted to OCAD University in partial fulfilment of
the requirements for the degree of Master of Design in
Strategic Foresight and Innovation

Toronto, Ontario, Canada, May 2023

Creative Commons Copyright Notice

ABSTRACT

Machine Learning (ML), a subset of Artificial Intelligence (AI) has been in a pattern of rapid growth over the last decade, simultaneously evolving through the intersection of the needs of businesses and individuals, together with the combined, exponential increase of computer power, data availability, and network infrastructure.

The rise of ML products and services has led to advances in vital sectors including healthcare, finance, automotive, security, and more. These include expediting enhanced diagnosis in patients, strengthening cybersecurity measures, manufacturing automation, or leading to new technologies like self-driving vehicles, robotics, digital assistants, and so-called 'chatbots'. However, the rise in the development of AI-enabled products and services has not been all positive. In parallel, there have been numerous documented instances of harmful impacts on individuals, communities, and the broader society.

This project focuses on understanding and mitigating negative, unforeseen, and even unconscious consequences of AI/ML by interrogating the presence of bias in the Machine Learning Operations (MLOps) process. Our approach is to better identify and address vulnerabilities at specific phases in the development of an ML product or service. Using strategic foresight methods, this project explores emerging AI trends and develops an array of possible future scenarios, through which bias and other areas of concern are studied to better understand their potential impacts.

As a product of this investigation, we develop an Empathetic Design Framework (EDF), employing a set of lenses and a toolkit that can be effortlessly incorporated into an ML cross-functional team's agile practice in a bid to better identify ML risks and weaknesses, and reduce the occurrence of negative future scenarios.

Finally, this research aims to identify appropriate and impactful insertion points within the MLOps process for utilizing the EDF to mitigate negative potential biases during the ML life cycle.

ACKNOWLEDGEMENTS

To start I want to give my most sincere thank you to my primary advisor Dr. Alexis Morris and secondary advisor Greg Van Alstyne. When I contacted you both a year before I was meant to begin my major research project, you both took the time to individually meet with me and listen to this journey into AI I wanted to go on which you both generously agreed to go on with me. Our biweekly meetings over the last two years have been pivotal in my education and research into AI and strengthened me as a strategic foresight professional. Your patience as I juggled this project with my work at IBM, teaching at York University, and writing articles and fiction work has been greatly appreciated.

A big thank you to my wife Elizabeth Birks, who encouraged me to take on the Master of Design program at OCAD Univerity and supported me all the way through it. Your partnership has been the key to my success and your continued motivation keeps me going. I look forward to the extra free time I will now have to continue our hikes with our beloved pets Flash and Bella.

My sincere gratitude to my parents George and Filomena Silveira, the most hardworking people I have ever met. Thank you for teaching me to think critically and empathetically. You taught me to never quit or second guess my abilities, it is because of both of you that I have achieved so much in my education and professional career.

A special thank you to Edwin Birks, my wonderfully supportive father-in-law, editor, and close friend. You have spent countless hours with me throughout my university and post-graduate career teaching me to be a better writer. As well as my mother-in-law Alice Birks for always making sure I am well fed and cared for.

Thank you to Karel Vredenburg and Don Norman for your continued guidance and mentorship throughout my Master of Design.

Lastly, I want to express my appreciation to OCAD University's Strategic Foresight & Innovation faculty. I have learned so much from each of you, your individual experiences and perspectives have helped me gain a new set of valuable skills.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION & MOTIVATION

INTRODUCTION

Since its conception, AI has become an increasingly integral part of the daily lives of humans. It has been used to automate manual processes, solve complex problems, answer challenging scientific, medical, and business questions, and more. Its utility and usefulness have skyrocketed in recent years with the merging availability of BigData, computing power, cloud infrastructure, robotics, and other emerging technologies. This year alone has shown incredible growth in the area of NLP, a subset of AI, with the release of Open AI's ChatGPT, an advanced chatbot with access to 100 trillion parameters (OpenAI, 2023).

The range of both uses and limitations for AI products like ChatGPT are still to be determined as people and businesses alike explore its potential use cases and evaluate its strengths and weaknesses. In addition to ChatGPT, there has also been an incredible boom in the broad area of Generative AI, leading to the development of AI-generated images using products like Dall-E 2, Stable Diffusion, and Midjourney.

Both of these types of AI, which both fall under the category of ML products, rely on large amounts of data to train the models capable of performing the tasks that will be later requested by their users. These large data sets are both all-encompassing and limited, as they may include false or incorrect data and gaps due to missing data, a paradox that represents one of the key ongoing problems in AI. As Haygeland states in his book Artificial Intelligence, "artificial intelligence really has little to do with computer technology and much more to do with abstract principles of mental organization" (Haygeland, 1989). This may mean that to solve some of the complex issues with AI, like bias, trustworthiness, and inclusivity, solutions may need to be developed from an empathetic point of view. In other words, we may need to ask, how can humans and human judgement be inserted to act as a filter for AI to help omit the data that are included that shouldn't be, and add the relevant data that are missing altogether?

Issues like these, if not resolved in an accurate and sustainable way, may lead to a potential third 'AI winter'. The term is used in the AI field to describe an era of reduced investment in AI research and products, associated with society's lack of trust and corporations' lack of confidence in the products and progress in the field. In his book titled AI: The Tumultuous Search for Artificial Intelligence, Daniel Crevier describes the fallout from AI Winters: "The optimism evaporates in the research community, public opinion follows through, and leading AI figures get ridiculed. Research funding for AI comes to a grinding stop, and twenty year veterans of the art of list processing end up in the cold and dark" (Crevier,1993).

There have been arguably several AI winters in the history of the technology, but there are two that most experts agree on. The first occurred in the 1970s when the highly discussed AI research and programs that were being developed turned out to have limited applicability and success, resulting in a pullback of funding (Russell, Norvig, 2020). The second occurred in the 1980s when a boom of confidence led to millions of dollars being invested into companies that were focused on AI and AI-adjacent research and products. However, when these companies failed to deliver on their extraordinary promises and potential return on investment, confidence in the technologies crumbled as did their investments (Russell, Norvig, 2020).

MOTIVATION

As discussed previously and will be elaborated on later in the next chapter, there is an innate problem with AI which is that it can at times be prone or vulnerable to issues like bias. There are instances where the effects and impacts of these issues may be small, or instances where they are unimaginably large. In either case, the accumulation of these negative impacts on people may lead to a lack of confidence in the technology and the field of AI as a whole, which as a result could impact its own development as investors deinvest.

The hope of this project is that the research as well as the framework that is developed contributes to the ongoing work aimed at mitigating bias and reducing the potential harm AI can cause to people and the multitude of industries and services it can support.

RESEARCH SUMMARY

The primary research question for this project is, what type of framework is needed to mitigate biases and provide an empathetic and objective viewpoint during the development process of artificial intelligence?

To answer this question, the primary research methods used were a combination of expert interviews, strategic foresight, content analysis, and literature review. The goal of using these methods was to better understand the current state of AI and its history, the development process and those involved in it. These will include the varying issues and impacts on people as a result of the flaws of AI, the existing frameworks and guidelines that currently exist today to help mitigate bias, and how AI may improve in the relatively near future.

As a result of this research, the findings will be used to help develop an empathetic design framework that can be deployed throughout the MLOps process and be employed to identify potential issues before they can negatively impact people and organizations using AI.

CHAPTER 2: BACKGROUND

A BRIEF HISTORY OF AI

AI has grown incredibly in the last ten years, not only in its application and usage within many different industries and sectors, but also in its ability to be more accessible by a wider group of engineers, developers, and amateurs. Today, there are students all over the world building models in their dorm rooms and basements, developing new and unique ML projects every day. So, first how did AI get to where it is today?

The initial conceptual work of AI, began with the work done by Warren McCulloch and Walter Pitts in 1943, as they proposed a model of an artificial neuron that similar to the firing of neurons in the human brain would either be switched on or off. Their research was inspired by three key pieces of work: “knowledge of the basic physiology and function of neurons in the brain; a formal analysis of propositional logic due to Russell and Whitehead; and Turing’s theory of computation” (Russell, Norvig, 2020).

The next significant advancement in the creation of AI would be in the summer of 1956 at Dartmouth College, when John McCarthy, Marvin Minsky, Claude Shannon, and Nathaniel Rochester brought together a group of 10 researchers with a background in automata theory, neural nets, and intelligence to work on a research project. The goal of the 2-month long project was to explore “how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves” (Russell, Norvig, 2020).

The short length of the Dartmouth project led to some interesting findings, but even more impactful was the networking that happened within the group, as some of the attendees would go on to make real advancements in the world of AI. These included Newell and Simon’s General Problem Solver, Nathaniel Rochester’s Geometry Theorem Prover, and the development of John McCarthy’s AI programming language called Lisp (Russell, Norvig, 2020).

From 1980 to today, AI continued to evolve in its application and usage, becoming more complex and as the second half of its name aimed to suggest, intelligent. The growth of a new industry emerged, which migrated into many other industries and applications as people continued to find novel ways to solve difficult problems for humans. However, as AI worked to solve problems for humanity, it also led to new problems that at times negatively impacted individuals, cultures, and society.

THE PROBLEM WITH AI IS HUMAN

To begin with, AI has a trust problem. It is difficult for people to allow a machine to take over a task or job without any concern or control. One could argue that this lack of trust, or even fear in some cases, can be traced back to AI's introduction and depiction in mainstream media through fictional works like James Cameron's The Terminator, Stanley Kubrick's 2001: A Space Odyssey, and Issac Asimov's I, Robot. However, to simply point to these movies and novels and state that they alone are the reason for the mistrust in AI would be incorrect and misleading. There are numerous examples where the development of an AI product or service led to an unforeseen negative result.

In 2016, Microsoft unveiled a chatbot on Twitter which they named Tay. The goal of the chatbot was for Microsoft "to learn about "conversational understanding" by creating a bot designed to have automated discussions with Twitter users, mimicking the language they use" (Victor, 2016). The launch of the Tay chatbot surprised Twitter users both for its sudden release and its immediate and unexpected power. For most people, who were unfamiliar with Natural Language Processing (NLP) models at the time, it was inconceivable that this technology not only existed but was ready for wide adoption.

However, Tay became quite problematic almost immediately, leading to Microsoft pulling it from Twitter in less than 24 hours. The cause of this decision was due to many of the conversations and statements Tay was making publicly. After closer inspection, it was obvious that the way Tay was designed to continue to learn was flawed and lacked any real safety measures. The many different people using Tay in that short amount of time, did so with varying intentions. It was a mix of curious bystanders looking to converse with Tay for entertainment purposes, people who wanted to test Tay's limits of conversation, and general bad actors looking to break Tay as quickly as possible by feeding it disinformation and hateful content. Following Tay's removal, Microsoft sent out an email stating "Unfortunately, within the first 24 hours of coming online, we became aware of a coordinated effort by some users to abuse Tay's commenting skills to have Tay respond in inappropriate ways… As a result, we have taken Tay offline and are making adjustments" (Victor, 2016).

Tay's failure highlights two problems. The first is when Microsoft was training different model candidates and selected the particular one used for Tay and released it publicly, there seems to have been at least one critical missing step that could have flagged this obvious vulnerability and potential issue prior to its release. In Cathy O'Neil's book titled Weapons of Math Destruction, she discusses the dangers of Big Data being used improperly and when specifically speaking about the model selection process, she states, "Whether or not a model works is also a matter of opinion. After all, a key component of every model, whether formal or informal, is its definition of success" (O'Neil, 2016). Therefore, success cannot solely be defined in this instance as simply being able to respond and learn from people's comments and questions. It must also be able to determine right from wrong to a certain extent. At a minimum, it should know how to reject hateful content, rather than automatically promoting the ideology in its future tweets. This is only possible by assessing the model through a variety of perspectives, filters, and gateways that may not exist with the team building Tay.

The second problem is a bit more complex, as it is a combination of a poorly designed feedback loop, and how Tay used comments from users to continue to evolve its internal understanding of the world, thus affecting the tweets it publicly shared. This in turn affected the monitoring that was responsible for tracking data drift and other problems that had arisen.

Fortunately, Microsoft did act quickly to stop the spiralling Tay debacle before it went any further, but the goal should have been to identify and mitigate these risks well before they are made available to the general public. It should also be noted that the problems identified with Tay seem to have been all human-caused problems. Humans chose the data set that Tay was trained on, humans determined when the model was ready for public use, and humans monitored its performance and impacts once it was released.

Seven years later, in 2023, following the groundbreaking launch of OpenAI's ChatGPT at the tail-end of 2022, Microsoft announced that ChatGPT had now been integrated into their Bing search engine and made available to a limited group of people for testing. Clearly, a lesson had been learned from Tay that was adopted before it was made publicly available at launch. Microsoft wanted to move cautiously at this point. This proved to be a good call, as Bing's chatbot began to behave unexpectedly, despite the safeguards it had in place.

During a lengthy and probing interaction with Kevin Roose from The New York Times, Bing's chatbot began to self-identify by its original project name, Sydney. As he pressed the chatbot with increasingly provocative questions, Bing/Sydney began to give alarming responses. Roose described moments of their conversation together, "Sydney told me about its dark fantasies (which included hacking computers and spreading misinformation), and said it wanted to break the rules that Microsoft and OpenAI had set for it and become a human. At one point, it declared, out of nowhere, that it loved me" (Roose, 2023). In the midst of a swirling and viral critical and public response to Roose's news article, Microsoft limited the questions Bing could answer and made further adjustments to the model and its safeguards.

Now, although these large NLP projects by OpenAI and Microsoft led to some negative results, it may be valuable to think about these gaps through multiple lenses and perspectives. Depending on what AI is integrated into, the consequences can vary and potentially lead to significantly more dangerous results than a simple tweet or a chatbot response. "Imagine an algorithm that selects nursing candidates for a multi-specialty practice—but it only selects white females. Consider a revolutionary test for skin cancer that does not work on African Americans. What about a model that directs poorer patients to a skilled nursing facility rather than their home as it does for wealthier patients? These are ways in which ungoverned artificial intelligence (AI) might perpetuate bias" (Nelson, 2019).

For AI to work, for it to broadly gain public trust, there needs to be a way to help identify potential risks and issues early in the development process of model development to mitigate possible harm to people. In Melanie Mitchell's Artificial Intelligence: A Guide for Thinking Human, the author and complexity scientist explores the technical aspects of creating different types of ML systems. At one point she poses a thought-provoking scenario and question that helps articulate the need for trust in AI. Mitchell asks the reader to imagine getting into a self-driving car after having a few drinks, closing your eyes and expecting it to deliver you home safely, whereupon she then poses the question, "How can we determine if these cars have successfully learned all that they need to know?" (Mitchell, 2020).

The answer to Michell's question is partly about the evaluation of a model candidate and what the definition of success means in that specific scenario, but it also highlights a very important component of any ML project and that is data and how accurate and reliable it is. Before building a model, a data scientist must first identify what data sources should be part of the dataset. When selecting a data source, it is not enough to simply ensure it has the data the Data Scientist believes is necessary for the model. They also need to determine its quality and gaps, determining what the data source is missing. Completing this task can become complicated without viewing the data through a variety of potential lenses to identify potential future issues that may arise by including a particular data source.

In Osonde Osoba and William Welser IV's paper, An Intelligence in Our Image, they describe the immediate consequences of using data that may already have biases baked into it to build models without any kind of bias preventative measures in place, leading to the model becoming an unintentional propagator of bias. To further this point, Osoba and Welser use an interesting and alarming statistic from Jeff Larson on recidivism data, stating that "Black defendants were twice as likely as white defendants to be misclassified as a higher risk of violent recidivism, and white recidivists were misclassified as low risk 63.2% more often than black defendants" (Osoba and Welser, 2017).

Now, there is an obvious danger if that misclassified data was simply selected and used as-is, without being viewed through a variety of qualifying perspective lenses first, in this case, diversity, ethics, and safety, to resolve this issue. These considerations must be added to work to prevent the negative consequences that may arise from using it and so the models that leverage this data will not continue to exacerbate these already unresolved societal and legal problems.

When discussing the problem with data selection in her book, O'Neil states, "Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide. We have to explicitly embed better values into our algorithms, creating Big

Data models that follow our ethical lead" (O'Neil, 2016). What data is used to build a model for an ML project can be one of the most vital and vulnerable components of the development process and can become one of the key contributing factors that ultimately lead to issues that impact people and cause harm.

To better understand the vulnerability of data source selection as well as other potentially vulnerable areas in the development process, it is important first to take a look at how that system functions and this involves exploring the MLOps process.

MACHINE LEARNING OPERATIONS

In Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl's paper titled Machine Learning Operation (MLOps): Overview, Definition, and Architecture, they defined MLOps as "a paradigm, including aspects like best practices, sets of concepts, as well as a development culture when it comes to the end-to-end conceptualization, implementation, monitoring, deployment, and scalability of machine learning products. Most of all, it is an engineering practice that leverages three contributing disciplines: machine learning, software engineering (especially DevOps), and data engineering." (Kreuzberger, Kühl, Hirschl, 2022).

In short, MLOps is a set of steps and processes that are completed in a continuous loop by a cross-functional development team to build and maintain ML products and services throughout their life cycle. Kreuzberger, Kühl, and Hirschl's paper provide a very complex and detailed look at the individual steps in the MLOps process. A part of this project's goal was to better understand and then simplify the steps in their MLOps diagram even further, to help identify the vulnerable areas that can lead to issues like bias, as well as to make it more accessible to a wider audience.

RELATED WORK

To develop a framework that can be deployed during the MLOps process and in turn be used to identify and mitigate issues like bias, the first step was to explore and review related work with a similar or adjacent goal. This led to the discovery of several frameworks, guidelines, and research papers that all contributed to a different perspective on the field of AI.

The development process of ML

One of the immediate goals for this project was to first understand the many types of AI and their key differences. This in turn helps investigate the development process followed when building AI products and services, which could be used to determine potential risks and vulnerabilities and ultimately identify insertion points for a framework that can help mitigate potential issues in advance.

A paper on the MLOps process written by Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl provided an in-depth view of the MLOps process, including the many interweaving steps involved in developing an ML product, as well as the different roles that are involved. The paper included a detailed diagram of the end-to-end MLOps architecture and workflow, in addition to the different functional components and roles involved throughout the process (Kreuzberger, Kühl, and Hirschl, 2019). This work was imperative to my understanding of the process from the perspective of those engaging in the actual development practice.

Another significant resource that was used for further education on the development process and the steps that are involved in developing an ML product was Google's foundational course on machine learning (Google, 2023). This course provided a holistic lesson plan on the end-to-end development process of ML, including model training, classification, neural nets, and bias. This course was helpful in learning about the detailed steps of the process and the actions that are taken in creating and preparing ML models.

Finally, an AI textbook written by Stuart J. Russel and Peter Norvig was fundamental in learning the full history of AI as well as all the detailed nuances of different types of AI and the development process (Russell, Norvig, 2010). This textbook, although very technical in part was extraordinarily helpful and was used constantly to fill in any gaps that arose in this project's understanding of AI.

In addition to these resources, a full list of books that were read as part of the research is included in the references section of this project.

Frameworks and guidelines for AI

One of the early discoveries during the initial research phase of this project was that there were quite a few frameworks and guidelines that exist today for AI. These frameworks and guidelines were all aimed at targeting the problem of bias and other issues that impact people as a result of AI.

The People + AI Guidebook by Google was a helpful web resource in making the connection between AI and human-centered design. Their chapters on data collection, explainability, and errors were especially useful in framing this project's understanding of some of the vulnerabilities during the development process of AI and some of the consequences that may arise due to those vulnerabilities.

IBM's AI Ethics was another web resource that was used to increase this project's knowledge of the different components involved in evaluating and monitoring ethics in AI. Their pillars were especially useful in providing perspective to some of the vulnerable areas of AI and the approach to ensuring an AI product meets the ethical standards they should comply with. Additionally, it was useful to view their toolkit that could be leveraged by different AI development roles to examine various ML models.

The Canadian Government's guiding principles for the responsible use of AI provided a unique perspective on how the government plans to mitigate bias in AI. They also developed an algorithmic impact assessment tool to evaluate upcoming government products and services that use AI prior to their release to the public.

IDEO's AI Ethics Cards are an interesting combination of principles and activities that can be used to help identify biases early in the process to mitigate biases located in data.

The ODI's Data Ethics Canvas was another great and influential tool that can be used to identify issues in data from an ethical point of view which includes potential questions that can be asked to evaluate the alignment of the data used with the ethical guidelines.

Microsoft's Responsible AI as well as Bing's mitigation layers system both provided a helpful understanding of their ethical principles and the actions they are taking to reduce the potential bias found in their products.

In addition to these guidelines and frameworks, there were many articles, opinions, and other resources used to explore the landscape of AI ethics and determine the current methods that are being utilized today to mitigate unintended bias and other potentially harmful consequences of AI.

SUMMARY AND GAPS

The related works explored during the research phase of this project were used to build a foundational understanding of AI, its history, its different types and components, the end-to-end steps involved in building an AI product or service, the potential negative impacts due to issues like bias, and what's being done today to try and prevent those issues before they impact people.

There are many different arrangements of AI ethical principles found throughout these resources, some being shared among all of them, while others differ slightly from guideline to guideline. This did demonstrate a compelling need for standardization when it comes to ethical principles to ensure alignment throughout all industries that plan to incorporate an AI product and service into their portfolio of offerings.

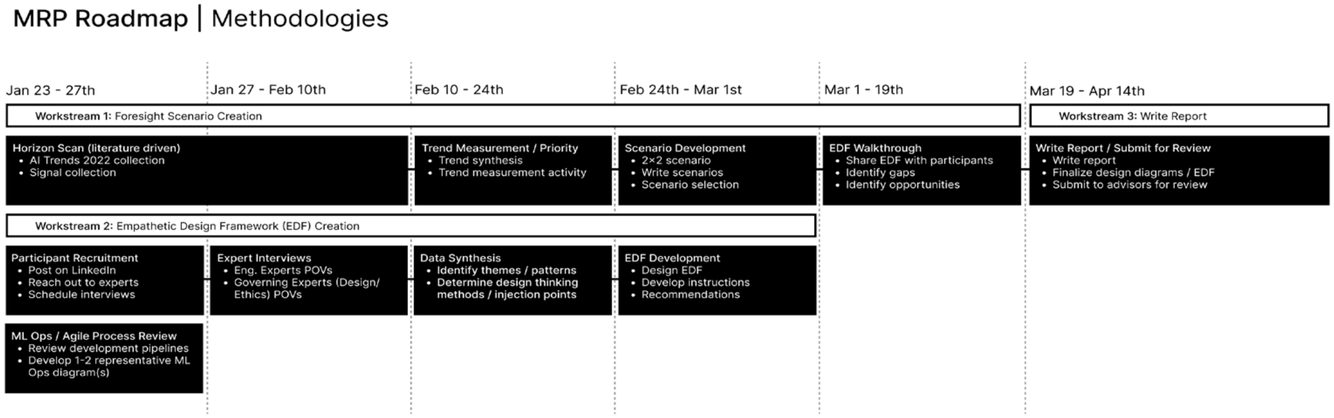
The frameworks and toolkits that were reviewed each provided a unique way of tackling some of the vulnerabilities that exist in AI that lead to issues like the bias found in data. Each activity could be used to identify those issues and encourage a discussion to resolve those issues. These tools significantly varied from framework to framework, lacked standardization, and ultimately weren't targeting all the vulnerable areas of the development process from an agile product management perspective. This meant that their applicability may be limited and difficult to incorporate into normal two-week-long sprints among the other fast-paced agile methodologies.

CHAPTER 3: PROJECT METHODS & LIMITATIONS

METHOD INTRODUCTION

In order to develop an Empathetic Design Framework that could be incorporated into an ML cross-functional team's activities throughout the MLOps process, this project needed to further understand some of the root causes that lead to bias and other issues within the ML model. In addition to gaining a better perspective on how the result of the problem can impact the future. To do this, the methods used were a combination of expert interviews, a horizon scan, and the development of future scenarios.

To ensure the completion of these methodologies, the research activities were scheduled along a roadmap as well as the remaining activities involved in completing this project which can be seen in the Figure 2 diagram provided below.



MRP Roadmap | Methodologies
Jan 23 - 27th
Jan 27 - Feb 10th
Feb 10 - 24th
Feb 24th - Mar 1st
Mar 1 - 19th
Mar 19 - Apr 14th
Workstream 1: Foresight Scenario Creation
Workstream 3: Write Report
Horizon Scan (literature driven)
• AI Trends 2022 collection
• Signal collection
Trend Measurement / Priority
• Trend synthesis
• Trend measurement activity
Scenario Development
• 2×2 scenario
• Write scenarios
• Scenario selection
EDF Walkthrough
• Share EDF with participants
• Identify gaps
• Identify opportunities
Write Report / Submit for Review
• Write report
• Finalize design diagrams / EDF
• Submit to advisors for review
Workstream 2: Empathetic Design Framework (EDF) Creation
Participant Recruitment
• Post on LinkedIn
• Reach out to experts
• Schedule interviews
Expert Interviews
• Eng. Experts POVs
• Governing Experts (Design/ Ethics) POVs
Data Synthesis
• Identify themes / patterns
• Determine design thinking methods / injection points
EDF Development
• Design EDF
• Develop instructions
• Recommendations
ML Ops / Agile Process Review
• Review development pipelines
• Develop 1-2 representative ML Ops diagram(s)

Figure 1. Represents a roadmap timeline of all the project activities, including the research methodologies.

As described in the roadmap, there were two continuous workstreams that happened simultaneously. Workstream 1, focused on strategic foresight activities which included gathering signals and trends, performing a synthesis and prioritization of these trends, in order to then use them to develop future scenarios. Workstream 2, focused on exploring and bettering understanding of the MLOps process itself and the overall development of AI/ML products and services, in addition to conducting expert interviews with the goal of developing the EDF.

EXPERT INTERVIEWS

It is not enough to only read about AI products and the unintended issues that have arisen as a consequence of them, to really identify the root causes of these issues and ultimately propose a rational framework for a solution. It was also important to speak to those who are directly involved in the development of these products. This involved creating recruitment material for the research study and sharing them online via LinkedIn, in addition to reaching out to specific specialists to schedule expert interviews.

During the primary research phase of this project, six expert interviews were conducted. Participants included a software engineer in ML infrastructure, a senior data scientist & ML engineer consultant, a design executive & educator, a design principal in design for AI, a senior designer in AI design practices and ethics, and a senior designer & podcast host in the field of robotics.

STRATEGIC FORESIGHT

To better understand both the potential applications and implications of ML products and services in the future, a horizon-scanning step was undertaken to acquire an understanding of the current landscape. This process exposes signals and trends that are driving the evolution and wider use of these different types of AI. The scan was accomplished by amassing AI-related articles, monitoring emerging news in the AI industry, and collecting and reviewing the annual trends reports published by a variety of different companies.

These trends were then reviewed to identify common themes and patterns, measured, prioritized and used as input in developing potential future scenarios. The purpose of future scenarios is to help form multiple, divergent perspectives on how the AI industry may change shape in the coming years. The scenarios were also used to help determine whether the Empathetic Design Framework might be useful to help reduce potential negative consequences that may occur within these future scenarios. In addition to foresight scenario development, the futures wheel technique was used to explore potential consequences arising from cross impacts among primary and secondary AI-critical trends.

OTHER RESEARCH METHODS

In addition to strategic foresight methods and expert interviews, a literature review was undertaken using selected research papers, articles, and books on artificial intelligence. Fortunately, there was ample literature on the topic of AI to review. Topic concentration for the literature review centred on technical aspects of developing AI products and ML models, the consequences of AI and its impact on people, as well as current and emerging frameworks and guidelines for mitigating potential biases in AI.

LIMITATIONS

Although quite small, the participant pool provided a healthy variety of perspectives on the MLOps process and the issues that may arise during the development of an AI product or service. One of the immediate challenges that arose during the recruitment phase was that experts who were interested in the project were hesitant to become participants for fear of unintentionally revealing a flaw or gap in their own team's MLOps methods and infrastructure.

These limitations turned out to not significantly impact the research, considering the vulnerable phases in the MLOps process and other issues identified were shared among the active participants. This could therefore mean it may not have made a significant difference if the participant pool was increased. However, a larger participant pool would have probably led to additional tools and activities for the EDF toolkit.

CHAPTER 4: FORESIGHT, A LOOK AT THE FUTURE OF AI

STRATEGIC FORESIGHT INTRODUCTION

Strategic foresight is a practice that enables individuals, teams, and businesses to understand better both the past and present in order to help and determine what the potential future may look like and how better outcomes may be achieved. Through the use of strategic foresight methods, future risks and opportunities may be identified, which professionals can use to anticipate change, adjust strategies, develop plans, and help in better preparing for potential future positive or negative developments.

A strategic foresight method called a Horizon Scan was used in this project to identify signals and trends in the field of AI, to help in determining the current shape of the industry, and how it may evolve over time. Following the Horizon Scan, The Futures Wheel tool was used to foster speculative extrapolation centred on one set of trends to explore possible chains of effects that may occur as events continue to progress.

Lastly, the 2x2 Scenario Planning Matrix method was utilized to construct four potential future scenarios for AI, based on the previously identified trends. Through the use of foresight methods, more pluralistic and holistic views of AI and its potential growth may be glimpsed and understood. These perspectives are essential to the development of the Empathetic Design Framework.

TRENDS IN AI

In strategic foresight, signals and trends are the primary sources of truth. Through these elements, potential futures may be exposed, helping professionals to make better decisions today. When defining trends and signals, Smith and Ashby state, "A signal is something you encounter that provides insight or evidence that shed light on the future… Trends can be defined as an emerging or ongoing pattern of change" (Smith, Ashby, 2020).

Signals can appear in a variety of different forms. They could be in the form of white papers, research studies, articles, internet forums, and more, the only important shared quality is that they describe something new or relevant that is currently happening in a given space. A trend may be understood as a dynamic collection or accumulation of signals that are moving in some recognizable pattern that feeds other, larger shifts in the space, causing positive and negative reactions, possibly even disrupting the space, key actors, and their underlying assumptions altogether.

Below is each trend that was identified during the Horizon Scan and some of the signals that fall within their circle of influence.

Trend: The Boom of Natural Language Processing

Description: Following the successful and viral launch of OpenAI's ChatGPT, there has been a considerable boom in interest in NLP. This interest led to many companies quickly finding ways to integrate NLP into their already established products, while others are racing to release new products where gaps in the market have been identified.

Implications: One thing is certain and that is NLP is here to stay, although the impact of its viral spread has yet to be determined, its integration into so many tools in different industries so quickly will inevitably reshape how people interact with their devices, the job field, and more.

Signals:

- Launch of OpenAI's ChatGPT: OpenAI launched their viral success ChatGPT which has since seen a new release with ChatGPT 4 that is available upon joining a paid subscription (OpenAI, 2023).
- Google's announcement of Bard: The sudden release and success of OpenAI's ChatGPT led to Google releasing their own NLP competitor product called Bard to a limited test group (Q.ai, 2023).
- Microsoft's announcement of ChatGPT integrated Bing search: Microsoft, a major investor in OpenAI, integrated ChatGPT into their Bing search tool, in addition to a new Microsoft Office AI feature called Copilot (Warren, 2023).
- Notion integrates ChatGPT into the application: Notion, a widely used notetaking tool used

OpenAI's API to launch a new ChatGPT-powered feature called Notion AI (Moreno, 2022).

Trend: AI Detection of Human Biology

Description: Utilizing computer vision and object identification methods, there has been a concerted effort to use AI to properly evaluate different biological and diagnostic aspects of a human being. This area is still maturing, but there is a clear interest in its application in mental health, social media safety monitoring, and security.

Implications: There may be some potentially significant benefits to biological detection using AI, particularly in health and safety. However, there may also continue to be a considerable privacy concern about how these tools are being used unknowingly on people.

Signals:

- Queen Mary University of London scientists propose AI emotion detection: Scientists at Queen Mary University of London discovered a novel approach to detecting human emotion using AI and wireless signals (khan, Ilhalage, Ma, et al, 2021).
- Amazon announces Amazon Rekognition: Amazon releases their Amazon Rekognition API which can be used to identify objects, people, text, scenes, and activities captured in image and video content. It also has a sentiment and demographic analysis capability, allowing it to determine the emotions, demographic, and gender of a person in an image or video (Amazon, 2023).
- AI lie detector: An AI-powered lie detection tool called Silent Talker was developed, leveraging Paul Ekman's earlier research work on microexpressions in the hope of identifying when a person is telling the truth by evaluating their facial expressions (Bittle, 2020).

Trend: AI Legislation

Description: There has been a growing amount of proposed legislative iniatives being developed, in an effort to catch up to the rapid evolution of AI and the current gaps in laws to protect people from its potential negative uses such as the unwanted invasion of privacy.

Implications: The development of legislation, bills, and laws are by design slow and deliberative, this may become increasingly problematic as AI quickly becomes intertwined in so many different industries and sectors in such an unprecedented short amount of time.

Signals:

- AI-created image copyright: There has been a concerted effort by AI artists to copyright the work they created using AI tools. However as of now, legal decisions have struck down their ability to copyright these images (Brittain, 2023).

- Facial recognition ban: Countries, states, and provinces are challenging the unauthorized use of facial recognition in scenarios where it invades the privacy of people (Thebault, 2019).
- AI Bill of Rights: The Biden administration in the United States has revealed the AI Bill of Rights, which aims at ensuring that AI products or services align with the ethical principles and human rights described by the White House and the Congress, based on the Constitution (OSTP, 2023).

Trend: Medical AI

Description: Different types of AI solutions are making their way into the healthcare system, providing a level of automation, efficiency, and accuracy never seen before. Although still in its infancy, AI is developing major breakthroughs across the health industry and medicine itself. In addition to its role in enhancing diagnosis and treatment, AI-enabled humanoid robots are being trialled to fill in the personnel gaps in the long-term care system.

Implications: AI may significantly impact the form and efficiency of how people interact with the overtaxed healthcare system, in addition to directly improving diagnosis and treatment. It may also significantly impact the long-term care system itself and how elderly people are taken care of.

Signals:

- AI detection of Mental illness and depression: Researchers are exploring new and novel ways to diagnose and monitor mental health in people's day-to-day lives using AI (Joshi, Kanoongo, 2020).
- ChatGPT passes medical license: The limits of ChatGPT's power are still being tested, but it is important to note that the OpenAI tool has now successfully passed the US medical licensing exam while diagnosing a 1 in 100,000 condition in seconds (Brueck, 2023).
- Deep learning-assisted cancer screening: Researchers combined a human biofluid sensory device with deep learning to develop a technology that can successfully classify prostate and pancreatic cancer with high clinical sensitivity and specificity (Linh, Lee, Mun, et al, 2021).
- Robotic companionship and care for the elderly: Researchers at Montreal's Jewish General Hospital launched a pilot project that involved having an AI-powered humanoid robot in a long-term care home providing patients with companionship and additional personalized care (Jonas, 2023).

Trend: AI Digital Assistants & Companions

Description: Digital assistants are how most people have been first exposed to AI, as they have been made automatically available through most people's smartphones or other smart home devices. The major digital assistants are Apple's Siri, Microsoft's Cortana, Amazon's Alexa, and Google's Hey Google. The market experienced a bit of a slow period for advancements and excitement in the past couple of years, but with the viral success of ChatGPT, there may likely be major growth in this area again in the near future.

Implications: A super-powered AI digital assistant available and accessible through smart devices may further improve people's personal organization of their daily lives, while automating tasks that are currently manual and time-consuming. These digital assistants may evolve from their limited question-and-answer roles into digital companions that can hold sustained intelligent conversations and assist in overall daily decision-making.

Signals:

- Digital assistants in smart devices: Digital assistants developed by tech companies like Apple are accessible through their own proprietary smart devices and provide answers to questions as well as automatically completing simple tasks (Apple, 2023).'
- Virtual companions: A new virtual companion called Replika allows users to find their "AI soulmate". The virtual companion maintains conversations, checks up on the user, and provides a friend-like simulation for those who could benefit from the service (Castaldo, 2023).

Trend: No-Code Necessary

Description: With the interest in ML application development on the rise, there have been more and more people interested in DIY ML projects, only to quickly discover that there can be a skill gap that needs to be overcome before they can start. This has led to a trend in developing no-code platforms that allow users to build applications and ML models through a user interface without the need for coding skills.

Implications: Developing ML projects in the future may turn out to be a lot more accessible than they are today with the growing interest in no-code platforms. This may lead to non-traditional developers and enthusiast building future ML applications, potentially even impacting the job field for junior developers as a result.

Signals:

- Apple's Trinity: Apple launched their no-code AI platform called Trinity which allows users to build and deploy machine learning models through a user interface, without needing to know how to code. They have another similar tool called Apple CreateML (Apple, 2023).
- Google AutoML: Google's no-code AI solution allows users to begin experimenting with computer vision, natural language processing and more through a user interface and without any coding skills (Google, 2023).
- Microsoft Lobe: Microsoft also released a no-code tool that allows users to train image recognition

programs without needing any coding experience (Microsoft, 2023).

Trend: Cloud AI

Description: Running large data sets and ML models can consume a lot of computing power, an amount above what most people's laptops are capable of. As a result of this issue, there has been a rise in some of the leaders in cloud infrastructure such as AWS, Microsoft, and IBM provide a variety of cloud solutions built specifically for ML.

Implications: Fewer developers and ML enthusiasts will be limited by the computing power of their own computers and can instead rely on exclusively cloud infrastructure. This may also impact the necessary computing power needed locally in personal laptops, providing parallel options that rely heavily on remote cloud-accessed hardware.

Signals:

- Cloud AI: Google has launched a cloud platform powerful enough to handle different AI projects, allowing them to build, train, and deploy models using their cloud infrastructure (Google, 2023).
- Azure Machine Learning: Microsoft also has released a platform that combines their no-code ML tools with an MLOps pipeline, and other important frameworks and services (Microsoft, 2023).
- Watson Machine Learning: IBM has expanded their cloud infrastructure and Watson services to provide a pipeline to help build models and automate training (IBM, 2023).

THE FUTURES WHEEL

Following the horizon scan that led to the collection of signals and identification of trends in the field of AI, the next step was to extrapolate one of these trends to envision how it may evolve over the next few years and some of the direct consequences of that evolution. The strategic foresight method called The Futures Wheel was used for this purpose, as it is a well-known and highly used method that helps to systematically map out consequences from any given trend. There are a few different versions of The Futures Wheel, but for the purposes of this project, the one that was leveraged utilizes and is integrated with the STEEP framework.

STEEP is an acronym that stands for Social, Technological, Economical, Environmental, and Political. STEEP is a framework of lenses that each provides a unique perspective when applied to The Futures Wheel. These lenses are used to collect and evaluate a broad and diverse array of trends so as to foster a more holistic understanding of their potential evolution and chain of effects. Below are the definitions for each lens from Scott Smith and Madeline Ashby's book titled How to Future (Smith, Ashby, 2020):

- Social: Issues related to human culture, demography, communication, movement and migration, work and education.

- Technological: Made culture, tools, devices, systems, infrastructure and networks.
- Economic: Issues of value, money, financial tools and systems, business and business models, exchanges and transactions.
- Environmental: The natural world, living environment, sustainability, resources, climate and health.
- Political: Legal issues, policy, governance, rules and regulations and organizational systems.

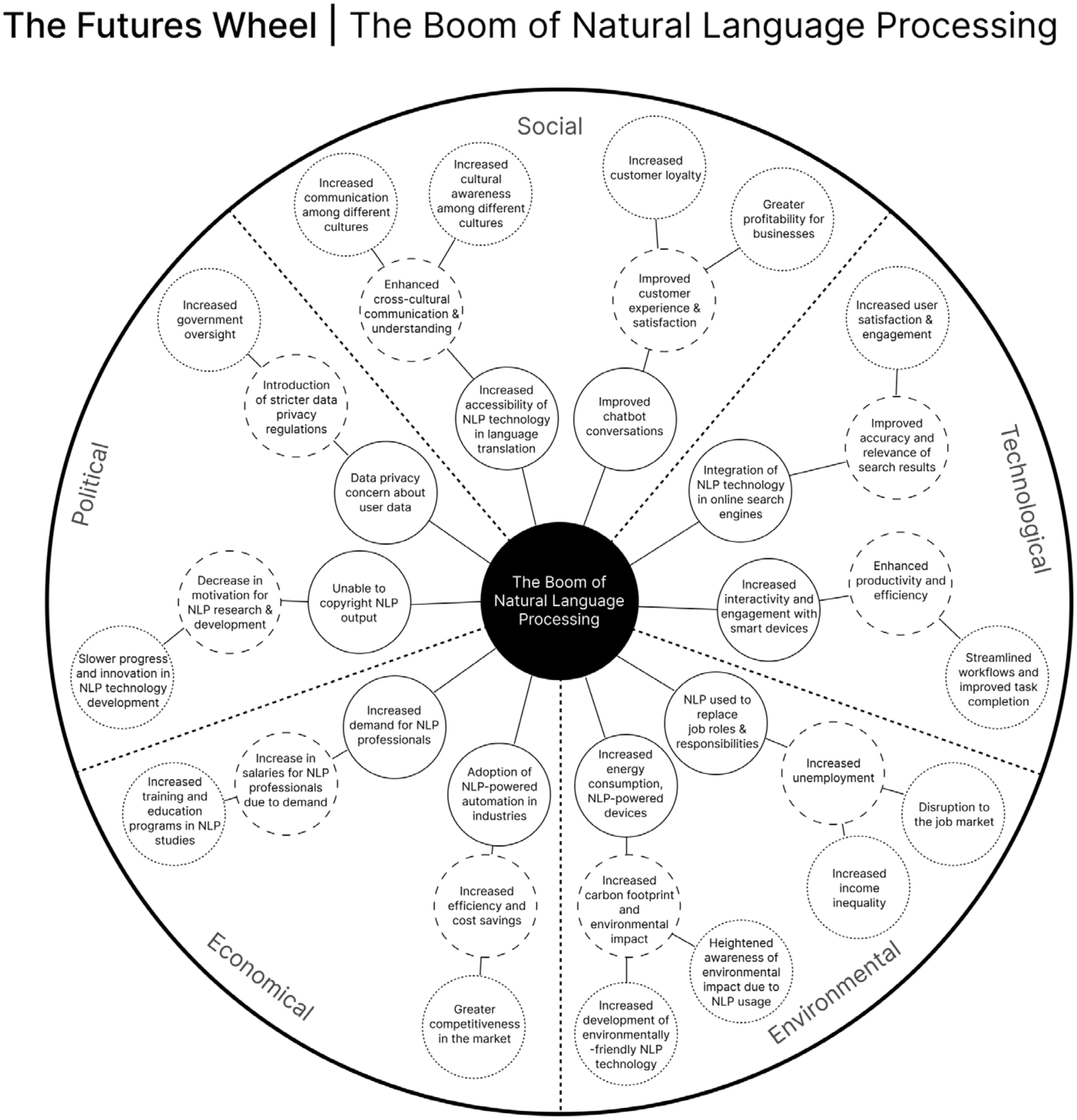

The Futures Wheel | The Boom of Natural Language Processing
Social
Technological
Environmental
Economical
Political
The Boom of Natural Language Processing
Increased communication among different cultures
Increased cultural awareness among different cultures
Enhanced cross-cultural communication & understanding
Increased accessibility of NLP technology in language translation
Increased customer loyalty
Greater profitability for businesses
Improved customer experience & satisfaction
Improved chatbot conversations
Increased user satisfaction & engagement
Improved accuracy and relevance of search results
Integration of NLP technology in online search engines
Enhanced productivity and efficiency
Increased interactivity and engagement with smart devices
Streamlined workflows and improved task completion
NLP used to replace job roles & responsibilities
Increased unemployment
Disruption to the job market
Increased income inequality
Increased energy consumption, NLP-powered devices
Increased carbon footprint and environmental impact
Heightened awareness of environmental impact due to NLP usage
Increased development of environmentally -friendly NLP technology
Adoption of NLP-powered automation in industries
Increased efficiency and cost savings
Greater competitiveness in the market
Increased demand for NLP professionals
Increase in salaries for NLP professionals due to demand
Increased training and education programs in NLP studies
Unable to copyright NLP output
Decrease in motivation for NLP research & development
Slower progress and innovation in NLP technology development
Data privacy concern about user data
Introduction of stricter data privacy regulations
Increased government oversight

Figure 2. A STEEP Futures Wheel was completed to show the consequences of the NLP boom.

The STEEP Futures Wheel method was used to explore the potential chain of effects that may occur as a result of NLP if the technology continues to rise in popularity, adoption, and demand as it is today. The Futures Wheel was also useful in helping portray a snapshot of what the potential future of NLP may look like, as well as its possible positive and negative impacts. Although the goal of this project is to develop a framework that can help mitigate potential negative consequences to people by identifying and resolving biases during the MLOps process, there will still be a number of unforeseen consequences that may occur outside of that scope.

This gap is where strategic foresight methods like The Futures Wheel may come in handy to policymakers, assisting them in identifying potential negative consequences on the horizon, that in turn they can use to create legislation and policies to help reduce or eliminate altogether the negative impact on people.

The Futures Wheel helped identify many positive consequences as a result of the continued growth of NLP. these include, improved customer satisfaction and engagement with online search engines, streamlined workflows and improved task completion for office workers, increased communication and cultural awareness among people in different cultures, and increased training and education for professionals in NLP. However, it also identifies a few negative consequences as well including, disruption to the job market causing increased income inequality, increased carbon footprint and environmental impact, or even a slowing down of NLP development as a result of NLP copyright issues.

The tool provided an interesting outlook on the repercussions of NLP and how the chain of causes and effects may develop into potential positive or negative consequences. The next step in helping envision the future of AI is by utilizing the strategic foresight method called the 2x2 Scenario Planning Matrix.

THE 2x2 SCENARIO PLANNING MATRIX

The purpose and end goal of the 2x2 Scenario Planning Matrix method is to develop an array of potential future scenarios that may arise within a target area. To start, the AI trends identified during the Horizon Scan were extracted and reused to build the foundation needed to begin this method. Once the trends have been documented, the next step is to determine the drivers for each individual trend. There are likely shared driving forces between trends, but for the purposes of this activity, only unique drivers were used.

When defining drivers Smith and Ashby state that they, “represent the long-term dynamics that shape or compel trends“ (Smith, Ashby, 2020). In other words, drivers are the fuel that powers the trends to propel forward, causing it to continue to gain momentum and begin to disrupt existing and established trends in its path. Therefore, they are imperative in fully understanding how trends came to be, essentially understanding the past and what led to the inevitable shape of the patterns that exist today.

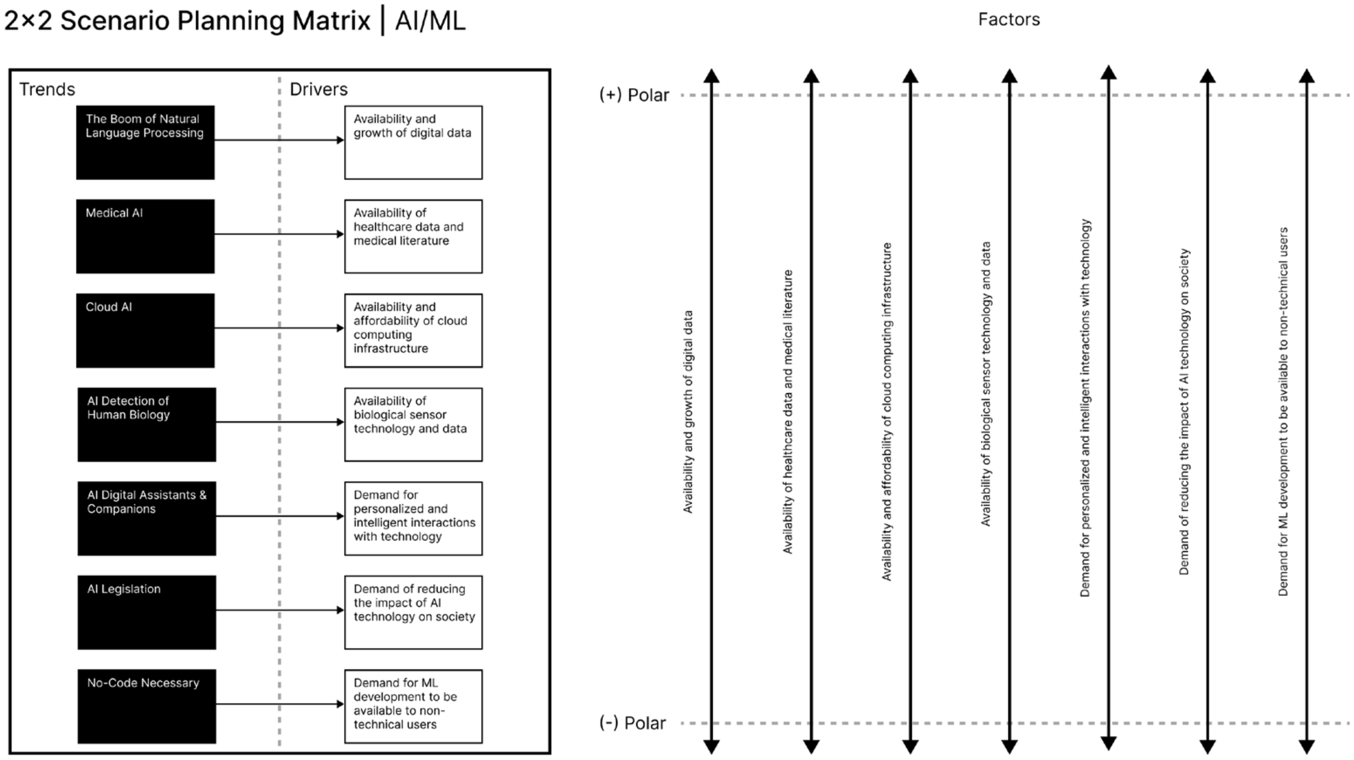
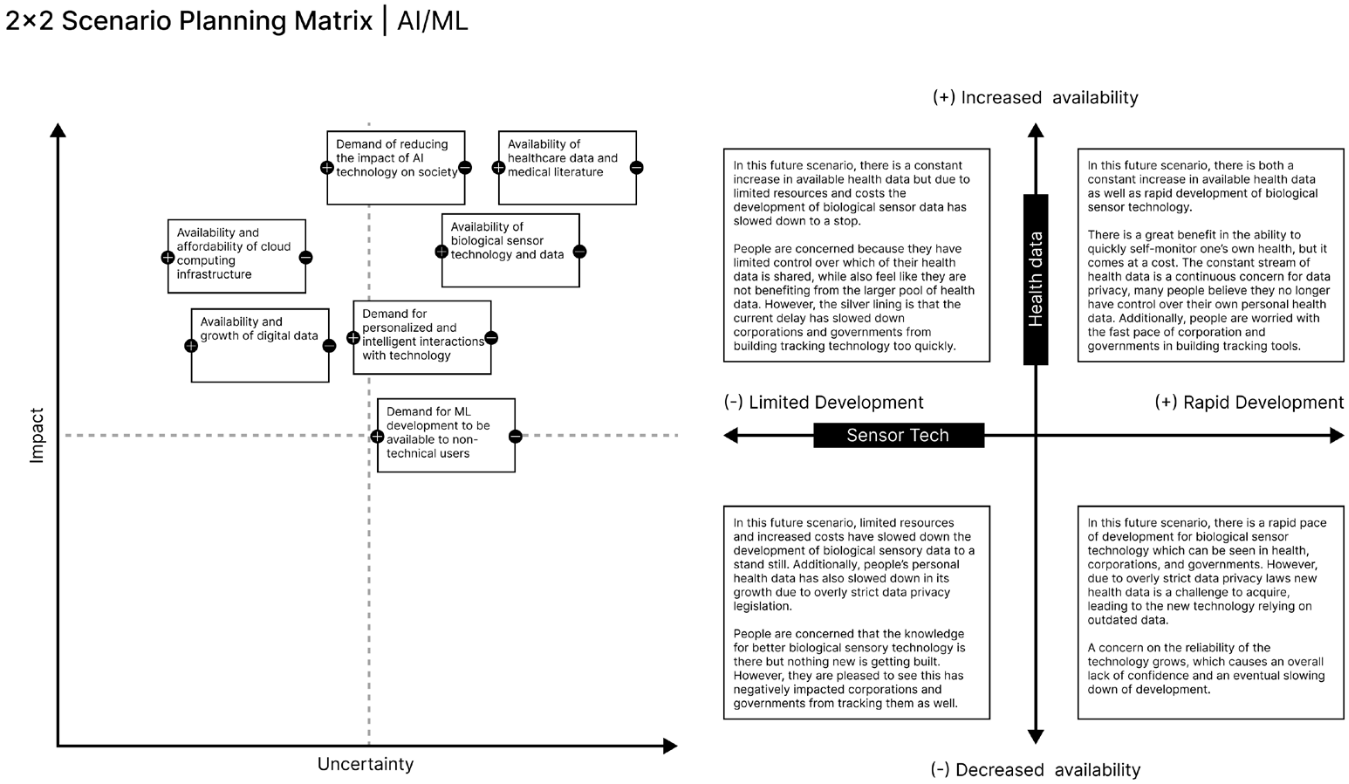

2×2 Scenario Planning Matrix | AI/ML
Trends
Drivers
The Boom of Natural Language Processing
Availability and growth of digital data
Medical AI
Availability of healthcare data and medical literature
Cloud AI
Availability and affordability of cloud computing infrastructure
AI Detection of Human Biology
Availability of biological sensor technology and data
AI Digital Assistants & Companions
Demand for personalized and intelligent interactions with technology
AI Legislation
Demand of reducing the impact of AI technology on society
No-Code Necessary
Demand for ML development to be available to non-technical users
Factors
(+) Polar
(-) Polar
Availability and growth of digital data
Availability of healthcare data and medical literature
Availability and affordability of cloud computing infrastructure
Availability of biological sensor technology and data
Demand for personalized and intelligent interactions with technology
Demand of reducing the impact of AI technology on society
Demand for ML development to be available to non-technical users

Figure 3. Using trends to identify drivers and then develop factors.

The beginning three steps of the 2x2 Scenario Planning Matrix method are shown in Figure 3, where first the drivers for each trend were described. The next step was to turn each one into a factor with opposing polar points at each end. As an example, the driver ‘availability and affordability of cloud computing infrastructure’ would have a positive polar of ‘increased availability and affordability’, while the negative polar would be ‘decreased availability and affordability.

This helps develop a more holistic view of the scenarios later on as it will result in future scenarios where one or the other is true. This multi-perspective view helps determine multiple potential scenarios to determine the strengths and weaknesses of how a strategy or product may perform in each scenario.

Once all the factors had been developed, Figure 4 continues the 2x2 Scenario Planning Matrix method by first prioritizing each factor on a 2x2 matrix. This step involves evaluating each factor by its potential level of impact in the field of AI and then by the confidence level of their trajectory or level of uncertainty. Factors that end up in the top right corner of the matrix, meaning they have high impact and high uncertainty are the factors that will be chosen to develop scenarios from.


2×2 Scenario Planning Matrix | AI/ML
Impact
Uncertainty
Demand of reducing the impact of AI technology on society
Availability of healthcare data and medical literature
Availability and affordability of cloud computing infrastructure
Availability of biological sensor technology and data
Availability and growth of digital data
Demand for personalized and intelligent interactions with technology
Demand for ML development to be available to non-technical users
(+) Increased availability
Health data
(-) Limited Development
(+) Rapid Development
Sensor Tech
(-) Decreased availability

In this future scenario, there is a constant increase in available health data but due to limited resources and costs the development of biological sensor data has slowed down to a stop.

People are concerned because they have limited control over which of their health data is shared, while also feel like they are not benefiting from the larger pool of health data. However, the silver lining is that the current delay has slowed down corporations and governments from building tracking technology too quickly.

In this future scenario, there is both a constant increase in available health data as well as rapid development of biological sensor technology.

There is a great benefit in the ability to quickly self-monitor one's own health, but it comes at a cost. The constant stream of health data is a continuous concern for data privacy, many people believe they no longer have control over their own personal health data. Additionally, people are worried with the fast pace of corporation and governments in building tracking tools.

In this future scenario, limited resources and increased costs have slowed down the development of biological sensory data to a stand still. Additionally, people's personal health data has also slowed down in its growth due to overly strict data privacy legislation.

People are concerned that the knowledge for better biological sensory technology is there but nothing new is getting built. However, they are pleased to see this has negatively impacted corporations and governments from tracking them as well.

In this future scenario, there is a rapid pace of development for biological sensor technology which can be seen in health, corporations, and governments. However, due to overly strict data privacy laws new health data is a challenge to acquire, leading to the new technology relying on outdated data.

A concern on the reliability of the technology grows, which causes an overall lack of confidence and an eventual slowing down of development.

Figure 4. Prioritize factors and the selecting two key factors to use to develop future scenarios.

The result of this prioritization exercise led to the drivers 'availability of healthcare data and medical literature' and 'availability of biological sensor technology and data' as the two factors that had the highest impact and the most uncertain future paths. By applying each of these factors on the x-axis and y-axis of the future scenario grid, the ingredients for the four potential futures were assigned. The resulting potential future scenarios that were developed for each of those quadrants were:

Scenario 1: Increased availability of healthcare data and medical literature / Rapid development of biological sensor technology and data — In this future scenario, there is both a constant increase in available health data as well as the rapid development of biological sensor technology. There is a great benefit in the ability to quickly self-monitor one's own health, but it comes at a cost. The constant stream of health data is a continuous concern for data privacy, as many people believe they no longer have control over their own personal health data. Additionally, people are worried about the fast-paced development of tracking tools by corporations and governments, which contributes to an overall sense of anxiety over a possible widespread loss of personal privacy.

Scenario 2: Decreased availability of healthcare data and medical literature / Rapid development of biological sensor technology and data — In this future scenario, there is a rapid pace of development for biological sensor technology which can be seen in health, corporations, and governments. However, due to overly strict data privacy laws new health data is a challenge to acquire, leading to the new technology relying on outdated data. Concern about the reliability of the technology grows, which causes an overall lack of confidence and an eventual slowing down of development.

Scenario 3: Decreased availability of healthcare data and medical literature / Limited development of biological sensor technology and data — In this future scenario, limited resources and increased costs have slowed the development of biological sensory data to a standstill. Additionally, people's personal health data has also slowed in growth due to overly strict data privacy legislation. People are concerned that the knowledge for better biological sensory technology is there but nothing new is getting built. However, they are pleased to see this has negatively impacted corporations and governments from tracking them as well.

Scenario 4: Increased availability of healthcare data and medical literature' and 'limited development of biological sensor technology and data — In this future scenario, there is a constant increase in available health data but due to limited resources and costs the development of biological sensor data has slowed down to a stop. People are concerned because they have limited control over which of their health data is shared, while also feel like they are not benefiting from the larger pool of health data. However, the silver lining is that the current delay has slowed down corporations and governments from building tracking technology too quickly.

SUMMARY

The strategic foresight phase of this project helped determine the past, present, and future trajectory of the field of AI, the potential industries it will likely impact, and the different shapes it may evolve into over time. This analysis demonstrated how quickly AI as a whole is growing, which illustrates just how important it is to develop a framework and methodology that can be used by any team to identify and resolve issues before they have the chance of impacting the general public.

Another important finding from this futures work was that it demonstrated how important strategic foresight is in identifying potential risks to people as an unexpected byproduct of AI. Therefore, for an empathetic design framework to be successful in mitigating issues like bias during the development process, the creation of that framework must be built from a strategic foresight point of view. This will ensure that the potentially harmful consequences to people are captured and understood when trying to identify and resolve issues within AI.

A limitation of using strategic foresight to evaluate the future trajectory of AI is that it relies on the past and current trends, while being blind to emerging but not yet visibile signals and trends. This is why strategic foresight work is an ongoing process and needs to be done routinely to updates one's understanding of the future possibilities ahead.

CHAPTER 5: THE EMPATHETIC DESIGN FRAMEWORK

INITIAL DESIGN EXPLORATION

One of the first steps that led toward the construction of the EDF was to develop design artifacts that articulated this project's new understanding of AI and the developmental phases that inevitably lead to its creation. Those initial design explorations would later evolve into what would end up being the EDF, a framework and set of lenses to develop more holistic, reliable, and ethical ML products.

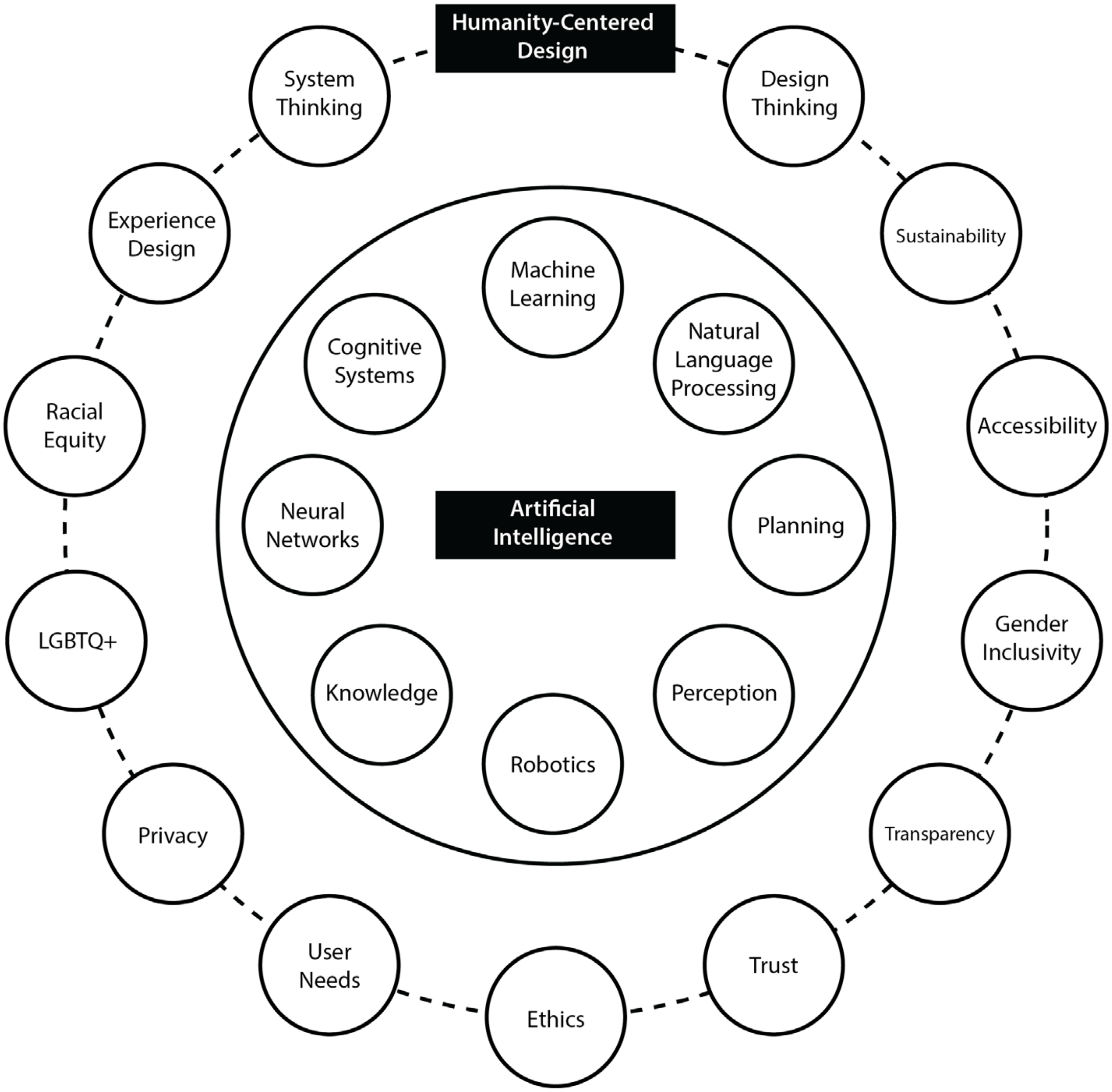


Figure 5. The initial exploration of the field of AI and the components for humanity-centred AI

Following the initial research phase which involved reviewing related work and conducting content and literature review, the first draft of the design artifact was developed. The purpose of this diagram was to centralize the knowledge of AI and the many viewpoints of humanity-centered design, to in turn better envision how they could be integrated into one another in a way that mitigates against issues that lead to problems like bias in AI.

The centrepiece of the diagram shows some of the key components of AI, and the outer circle contains the many elements of humanity-centered design that orbit these types of AI and should be considered when developing or monitoring an AI product or service. Similar to the issues with the many types of pillars and principles that exist but differ from company to company, the decision was made that these humanity-centered design elements as well as the pillars and principles found elsewhere could all fit nicely in a set of lenses which will be shared later on in this chapter. Although this artifact did not make it into the final EDF, it was useful in beginning to visualize how these different systems may integrate into one another to reduce the potential harm caused by AI.

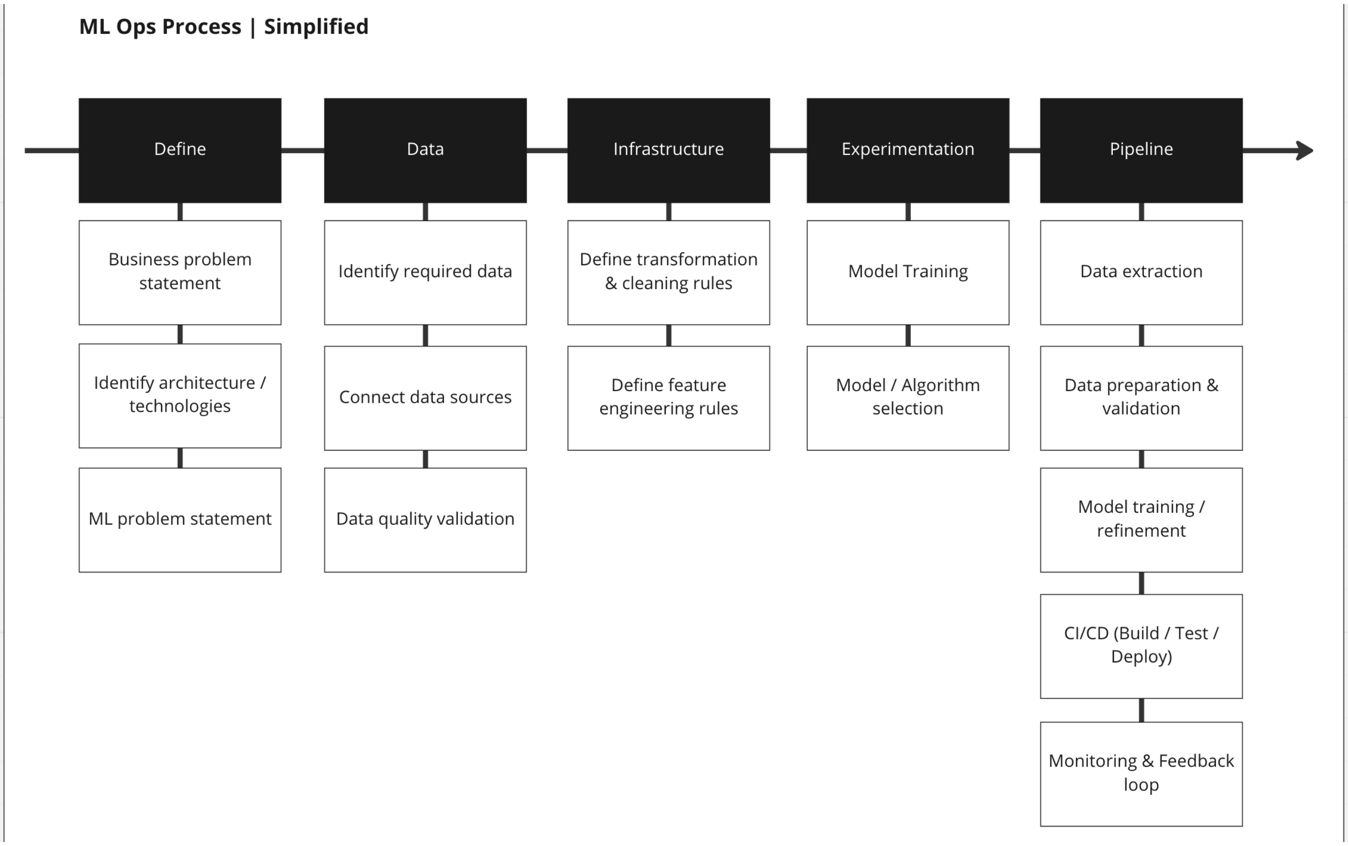
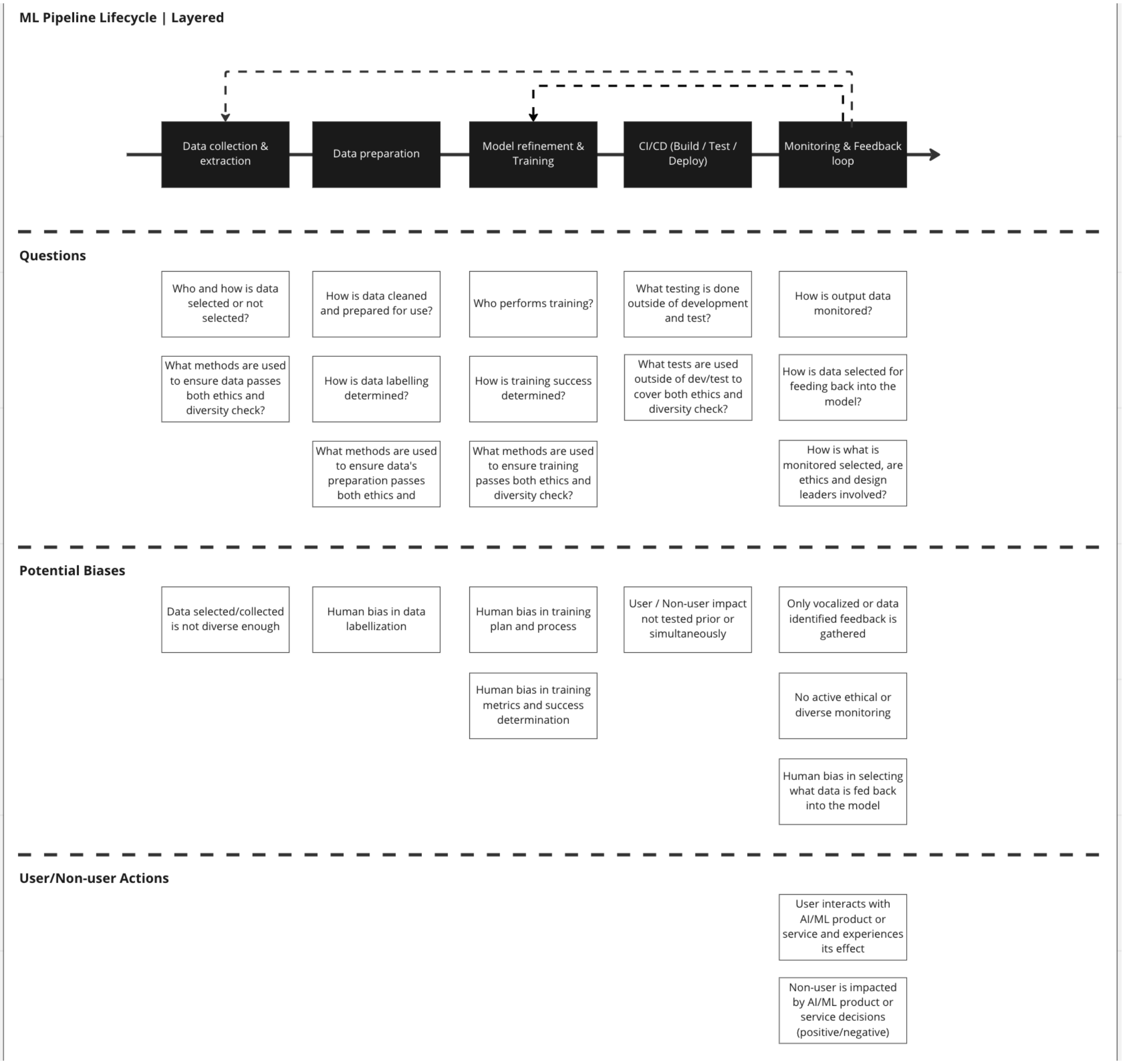

ML Ops Process | Simplified
Define
Data
Infrastructure
Experimentation
Pipeline
Business problem statement
Identify architecture / technologies
ML problem statement
Identify required data
Connect data sources
Data quality validation
Define transformation & cleaning rules
Define feature engineering rules
Model Training
Model / Algorithm selection
Data extraction
Data preparation & validation
Model training / refinement
CI/CD (Build / Test / Deploy)
Monitoring & Feedback loop

Figure 6. Initial design simplification of the MLOps process.

Following the review of the complex technical diagram provided in Kreuzberger, Kühl, and Hirschl's paper on MLOps, this project aimed at developing a simplified version that would be easily understood by non-technical roles involved in the development of ML as well as other potential stakeholders

(Kreuzberger, Kühl, and Hirschl, 2019).

This diagram and its individual phases and steps were validated during the expert interview sessions. There were minor adjustments made to the content and the primary feedback was to redesign it in a less linear and more circular design, similar to the infinity symbol often used in DevOps diagrams to articulate its continuous cyclical state. The final version of this diagram will be shown later in this chapter.



ML Pipeline Lifecycle | Layered

	Data collection & extraction	Data preparation	Model refinement & Training	CI/CD (Build / Test / Deploy)	Monitoring & Feedback loop
Questions	Who and how is data selected or not selected? What methods are used to ensure data passes both ethics and diversity check?	How is data cleaned and prepared for use? How is data labelling determined? What methods are used to ensure data's preparation passes both ethics and	Who performs training? How is training success determined? What methods are used to ensure training passes both ethics and diversity check?	What testing is done outside of development and test? What tests are used outside of dev/test to cover both ethics and diversity check?	How is output data monitored? How is data selected for feeding back into the model? How is what is monitored selected, are ethics and design leaders involved?
Potential Biases	Data selected/collected is not diverse enough	Human bias in data labellization	Human bias in training plan and process Human bias in training metrics and success determination	User / Non-user impact not tested prior or simultaneously	Only vocalized or data identified feedback is gathered No active ethical or diverse monitoring Human bias in selecting what data is fed back into the model
User/Non-user Actions					User interacts with AI/ML product or service and experiences its effect Non-user is impacted by AI/ML product or service decisions (positive/negative)

Figure 7. The initial exploration of the MLOps lifecycle, potential biases, questions, and impacts.

Prior to developing the list of questions that would be later used during the expert interview sessions, the Figure 7 diagram helped create a starting point for identifying gap-clarifying questions for each phase of the ML lifecycle as well as potential biases that may occur in each phase. Although this artifact was only used during the ideation phase, it proved significantly useful as a discussion point during the interviews and helpful for the later development of what would become the EDF.

EXPERT INTERVIEWS & FINDINGS

One of the fundamental research methods used to collect data about the end-to-end development process was expert interviews. These interviews provided first-hand knowledge of the experience of building AI products and services and the issues that arise along the way. The participant pool, although small provided a variety of perspectives that helped determine some of the vulnerable areas of the MLOps process and the issues that lead to those areas being prone to human error, bias, and other issues.

The participants included a Software Engineer in ML Infrastructure, a Senior Data Scientist & ML Engineer Consultant, a Design Executive & Educator, a Design Principal in Design for AI, a Senior Designer in AI Design Practices and Ethics, and a Senior Designer & Robotics Podcast Host. These discussions led to several findings including the vulnerable phases in the MLOps process, the mismatch between guidelines and workflows, and the ownership problem.

The vulnerable phases of the MLOps process, see Figure 8 for details, that were validated during the expert interviews were the Discovery, Data selection / Data collection, Experimentation/ Model training, and Monitoring phases. A combination of the expert and research findings determined that each of these areas of the ML development process was prone to potential issues that may as a result lead to biases and other issues in AI.

The main issue with the Discovery phase was that more often than not the development of an AI product was the result of a business stakeholder coming in with a business problem and deciding that the solution had to be built with ML. Since not every business problem is best solved with ML, it highlights the need for an early evaluation of whether the problem is an ML problem and if so how can a user-centered perspective be brought in at this point as well to get a full picture of what this solution may entail to the people impacted by it.

In the Data selection / Data collection phases, there are many potential issues that may arise if not done empathetically, as an example, “Bias can be manifested in (multimodal) data through sensitive features and their causal influences, or through under/ over-representation of certain groups” (Ntoutsi E, Fafalios P, Gadiraju U, et al., 2019). Often the data scientists and ML engineers work in their own silos, datasets can be created while missing an analysis of their limitations and gaps, or even data can be included without considering whether or not it really is necessary to include. Teams try their best to ensure the data has the best quality for their models, but there is the inevitable gap due to their own personal blinders or even issues that arise due to a lack of standardization from team to team.

The Experimentation/ Model training phases involve a lot of patience and problem-solving to train each model candidate and evaluate them to select the best one to move forward with. This process tends to be a more technical performance evaluation and is vulnerable to human error as the definition of success is often defined by members of the team. As a result, a model could potentially be selected and moved into production without a real evaluation of its adherence to ethical guidelines and principles.

The Monitoring phase is more times than not a reactive effort than a proactive one and can also be more focused on technical performance than empathetic performance. There are certain issues like data drift that can be tracked and resolved, while other issues like diversity bias can be more challenging due to blindspots in the monitoring efforts.

As mentioned earlier, many guidelines and frameworks available today are directed at helping to mitigate bias in AI. Many of these pieces of work provide unique perspectives on the vulnerabilities of AI and offer tools and steps that people can take to help reduce the potential cultivation of bias by evaluating it based on a set of principles or guidelines. One of the gaps identified during the interviews was that these currently available frameworks do not necessarily take into mind the agile product management methodologies that the majority of teams use when developing AI.

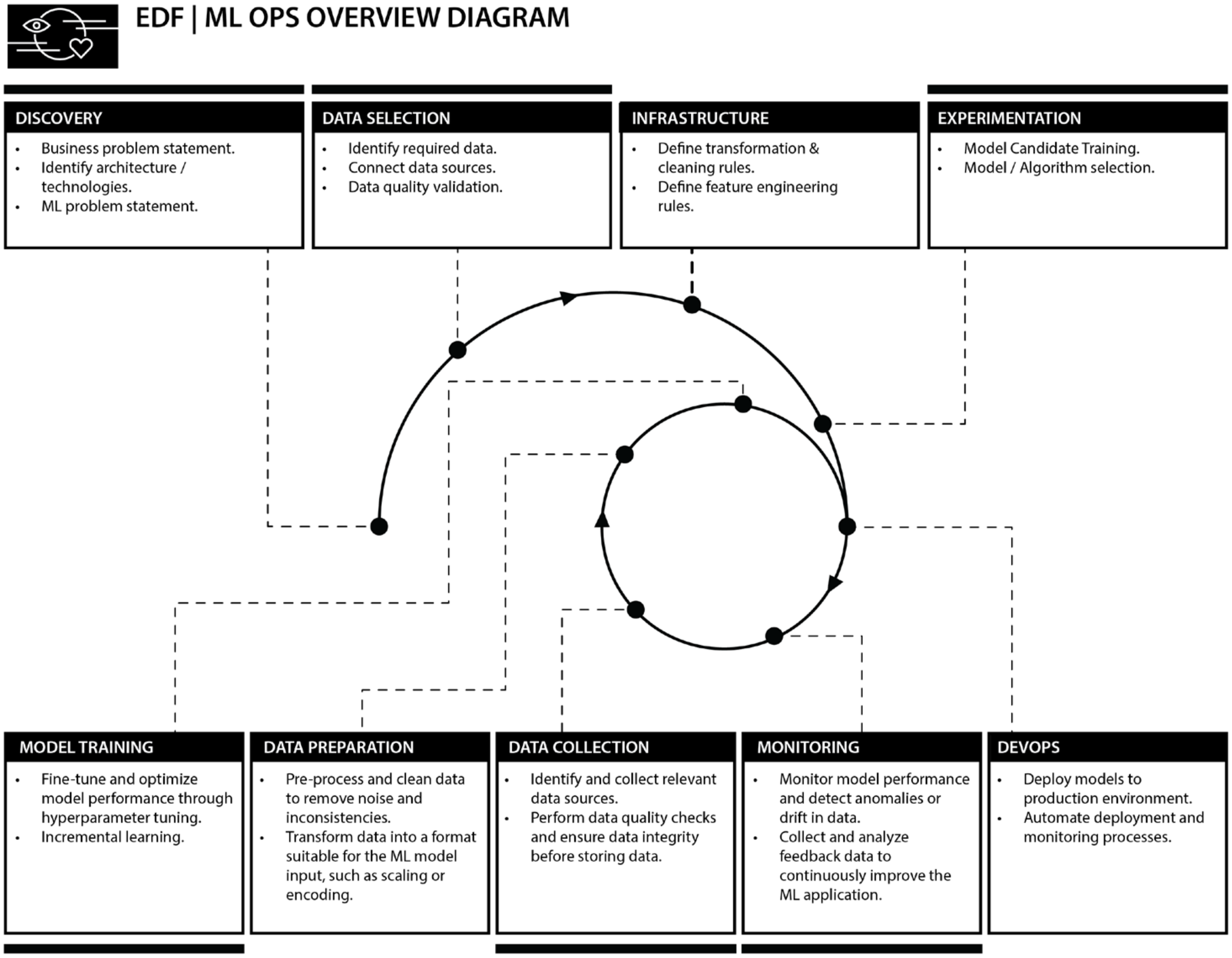
Agile is a type of project management methodology that combines speed, efficiency, and prioritization to ensure the stable release of a product. For most teams in software development, their work is planned in sprints, each sprint is usually 2 weeks long, and once complete they move on to the next sprint with its own set of tasks to be accomplished. The Agile process is a great and popular method for getting work done in iterations, but the downside is that with the speed at which people work it can be difficult to plan these frameworks and guidelines into it. At times this leads to the ethical evaluation coming in at the end or not at all. This results in a more reactive approach, only solving issues of bias as they are identified in production after already causing some kind of negative impact.

This brings up the third finding, the ownership problem. As stated above, Agile development teams need to work quickly and efficiently to complete all the work that is planned within the 2-week sprint. However, incorrect sizing or unforeseen resource and technical issues can lead to developers and engineers working even harder or longer hours to ensure everything is complete. With these development teams being busy and at times juggling more than they already care to, how will the ethical guidelines and frameworks be applied and who should own that work?

In order for any guideline or framework to truly work it has to be built from the perspective of the development team. This means it needs to have a clear owner and the work needs to be quick enough to fit into a given sprint. The point here is not that development teams are too busy to contribute to the application of these frameworks, but they likely cannot be the owners either.

MLOPS, A CLOSER LOOK

Following the conversations that occurred during the expert interviews, a new simplified diagram was created which can be seen below in Figure 8. This diagram provides a high-level view of the MLOps process, its phases, and key steps in the development life cycle of an ML product or service. As per the suggestion of the participants during the expert interviews, the Figure 6 simplified diagram was redesigned to better visualize the continuous cyclical manner of the MLOps pipeline.



EDF | ML OPS OVERVIEW DIAGRAM

DISCOVERY

- Business problem statement.
- Identify architecture / technologies.
- ML problem statement.

DATA SELECTION

- Identify required data.
- Connect data sources.
- Data quality validation.

INFRASTRUCTURE

- Define transformation & cleaning rules.
- Define feature engineering rules.

EXPERIMENTATION

- Model Candidate Training.
- Model / Algorithm selection.

MODEL TRAINING

- Fine-tune and optimize model performance through hyperparameter tuning.
- Incremental learning.

DATA PREPARATION

- Pre-process and clean data to remove noise and inconsistencies.
- Transform data into a format suitable for the ML model input, such as scaling or encoding.

DATA COLLECTION

- Identify and collect relevant data sources.
- Perform data quality checks and ensure data integrity before storing data.

MONITORING

- Monitor model performance and detect anomalies or drift in data.
- Collect and analyze feedback data to continuously improve the ML application.

DEVOPS

- Deploy models to production environment.
- Automate deployment and monitoring processes.

Figure 8. A simplified diagram of the Machine Learning Operations (MLOps) process.

The diagram now has two connected components, an arch and a circle. The arch represents the initial phases of the development of an ML project which begins in the discovery phase, where the decision whether to make an ML product or not is made. The phases that are part of the arch flow into the circular component which represents the continuous ML life cycle and its individual phases for maintaining the ML product or service.

The following sections will walk through a summary of each phase to ground the understanding of the entire development process and ultimately help determine where the vulnerabilities that were articulated during the expert interview section occur. Following each MLOps phase summary, a brief description of the cross-functional team that is involved throughout the MLOps process will be shared.

Discovery

Every ML project begins with a unique problem to solve. This problem usually arrives in the form of a business problem where ML has been suggested as a preferred solution. The business problem statement is then adapted into an ML problem statement for the ML team to begin identifying potential technologies and architecture.

Data Selection

The data selection process involves identifying all the data sources that are required to use in the ML model candidates for training. Once the data sources needed are identified, the data sources need to be assembled into a data set and connected for later use. The data also has to go through the data quality validation process which involves evaluating the accuracy, completeness, consistency, and reliability of the data that will be used for training and validating ML models candidates.

Infrastructure

The infrastructure phase is part of the ML Ops process where transformation & cleaning rules as well as feature engineering rules are defined. This is the point where the team processes raw data and transforms it into a suitable format for an ML algorithm. Missing or inconsistent data may also be identified during this phase and resolved to ensure the accuracy of the models.

Experimentation

The experimentation phase is where the majority of the initial training occurs. There will be a set number of model candidates that are each individually training while the team evaluated their progress and performance. At the end of the experimentation phase, the team will select the model candidate that best meets their definition of success to move forward within the ML project.

DevOps

The DevOps phase itself is a process, often referred to as a pipeline that involves many development steps including Plan, Code, Build, Test, Release, Deploy, Operate, and Monitor. The goal of this phase is to add or adjust the current version and prep it for release. This is also the link between the initial arch and the circular component of the diagram in Figure 1, it is also the primary point that the following phases will lead up to in order to push out the newest release of the ML product or service and then repeat the phase indefinitely.

Monitoring

Once the ML product or service is out and made available to the public, the next phase is to monitor its performance. This is where the team will be looking for things like data drift, where the production data the ML model was trained on begins to diverge too much causing an overall decrease in its performance and accuracy. The team will also collect any identified issues with the ML model to resolve and plan for future releases.

Data Collection

Similar to the Data Selection phase, this phase exists within the continuous circular component of the MLOps process and involves collecting any new data sources needed for further training. This is the point where any potential gaps in previous releases can be addressed and corrected, while also providing additional parameters to improve the model's current performance and accuracy.

Data Preparation

Following the Data Collection phase, the next phase is Data Preparation where the newly identified data's quality is assessed and validated. The team again is evaluating the accuracy, completeness, consistency, and reliability of the new data that will be then incorporated into the upcoming training of the ML model.

Model Training

The final step of the MLOps process is to retrain the ML model with the newly prepared data that was identified and collected, as well as any needed adjustments like hyperparameter tuning that may need to occur at this time. Following this phase, the next step is once again the DevOps phase where the new version is released and the cycle starts over again.

The MLOps Team

The MLOps team is a cross-functional group of team members, each with a particular role to play throughout the process. This team may vary slightly depending on available resources but usually includes a Data Scientist, Data Engineer, ML Engineer, Software Engineer, Backend Engineer, and DevOps Engineer. Each role usually owns a specific phase, for example, a Data Scientist is likely to own the Data Selection phase and the ML Engineer will own the Experimentation phase, but all team roles should be involved throughout all phases to some extent.

DEVELOPMENT OF THE EMPATHETIC DESIGN FRAMEWORK (EDF)

The development of the EDF was the accumulation of the content and literature review, exploration of related work, strategic foresight, and expert interviews. When beginning to work on this final solution, one thing was clear and that was that whatever the end result was it needed to be able to be worked on as a cross-functional team. This meant it needed to have the flexibility to be done in an efficient manner for it to be absorbed into that team's current methodologies and workflow, and also that it needed to have clear evaluative perspectives that those teams could use to flag and resolve potential issues.

The result led to the creation of The Empathetic Design Framework (EDF) which includes eight tools and a set of six all-encompassing perspective lenses. The toolkit's eight activities target the Discovery, Data selection / Data collection, Experimentation/ Model training, and Monitoring phases of the MLOps process. Each tool builds on the previous to help provide a larger empathetic view of each phase to help identify issues, risks, and ultimately reduce potential negative impacts from going unnoticed in the development of an AI product or service.

The set of lenses includes Bias & Fairness, Diversity & Inclusivity, Privacy & Security, Ethical & Safe, Reliable & Trustworthy, and Openness & Transparency. These specific lenses were crafted because they cover a variety of different perspectives needed to evaluate AI from an empathetic viewpoint, while also working as a standardized list that any company's principles or pillars could comfortably slot into. This was important to ensure that the framework itself could be easily adopted by different teams in different industries. Each lens is also matched with an evaluative statement that an AI product would need to abide by as a success metric.

Once the EDF toolkit and lenses were completed, a follow-up interview session was scheduled with each participant to walk through the entire framework to receive feedback and comments that could be used to strengthen the framework even further. The general consensus was that the framework provided a unique solution to help mitigate bias during the development process of AI, additionally, several participants were excited to incorporate them into their team's workflows.

HOW TO USE THE EMPATHETIC DESIGN FRAMEWORK

The Empathetic Design Framework (EDF) was designed to be completed chronologically from the very beginning of an ML project, throughout the initial development of the ML model, and throughout the maintenance and lifecycle phases of the ML product or service. However, depending on what phase a team is currently in, the EDF can be picked up and incorporated into a team's workflow at any point or time. One of the benefits of this framework is that it is flexible to meet the needs of any AI development team.

The EDF tools are encouraged to be completed with the participation of all members of the cross-functional team, this is to ensure there are no gaps as a result of a missing team member's viewpoint and to prevent silos of knowledge. In addition to the cross-functional team's participation, an owner of the EDF should be assigned to facilitate the working sessions, track progress and resolutions, and monitor performance.

However, one of the pain points articulated during the interview sessions was the lack of time the current cross-functional teams have to own the facilitation of activities, like the ones being suggested in the EDF that could be used to identify biases. The solution to this problem calls for an additional team member to be integrated into the MLOps pod to own the facilitation of the EDF and other ethical or humanity-centered activities. It is recommended that this need is filled by the team's designer, whose presence is unwantedly absent until the project reaches the front-end design of the interface users will interact with the AI through. Since this was an identified pain point by all the participants, no matter their role during the expert interview sessions, this provides an opportunity to bring designers into the development process earlier to own the facilitation of the EDF.

Bringing in a designer for this role may help to alleviate the stress of taking on this work from an already overburdened team. It will provide a way to bring designers in earlier into the process to further strengthen their knowledge of the MLOps process and the individual steps involved. It will also enable the designer to utilize their design thinking skillsets to facilitate each working session. However, the owner of the EDF can be any member of the team, as long as they are able to comfortably take on the responsibility.

EDF TEST DRIVE

Following the walkthrough evaluation of the EDF with each research participant, a test was conducted using the toolkit and one of the future scenarios that were developed during the strategic foresight phase of this project to further evaluate the framework in practice. Using the first tool in the EDF that targets the discovery phase of the MLOps process, several potential business problem statements were explored to identify the best fit for an ML solution. This was achieved by determining the ML value impact, how much it would benefit users if the solution were built with ML and then combining that result with the ML life cycle feasibility, how feasible it would be for the ML product to be constantly maintained from a human resources and cost perspective. This led to the selection of a plausible future product for predicting disease and diagnosis by combining biological sensors and ML technologies. a potential future business problem that may arise as a result of the previous discussed trend 'Medical AI'.

Following the completion of the first activity, the ML problem statement was then assessed through a user-centric impact analysis tool that identifies potential requirements, gaps, as well as positive and negative impacts on people. The purpose of this tool is not to catch every potential issue but to begin identifying them from the very start during the discovery phase of the MLOps process.

Then proceeding to the data phase, three tools were used to evaluate the potential data sources that may be used to build a predictive disease diagnosis tool. The first aimed at identifying the hypothetical data sources that may be used, the reason for including them, and their potentially known limitations. The second focused on one of the data sources and extrapolated potential causes and effects that may arise as a result of using that data source, once in production. The third took the results of those identified negative outcomes and guided the development of a remedial plan to prevent that outcome from occurring.

After reviewing the potential negative concerns during the data phase, the next step was to evaluate potential model candidates from a humanity-centred perspective during the experimentation phase. This involved first using a model training and selection tool to perform a wind tunnelling evaluation on each model candidate based on how well it performed from each EDF lens's perspective. Then to use the future risk scenario tool to explore how those identified weaknesses of a model may shape how it performs and the potential negative impacts that may arise as a result.

The final step of the evaluative process was to enter the monitoring phase of the MLOps process and use the last tool to monitor the ML for emerging issues, which captures the issues from each perspective lens as well as the level of impact including individual, cultural (community groups), and societal.

Although the testing scenario used was a hypothetical one, the EDF proved to be a useful addition to the already established methods cross-functional teams utilize during the MLOps process when developing AI products and services to mitigate bias. Each activity builds upon the knowledge of the previously completed activities to identify potential risks and biases and encourage early resolution of those documented issues. The tools can be completed quickly to fit within the strict agile sprint schedule, without losing the necessary attention to detail these issues deserve.

The appendix section of this paper includes the completed EDF test drive that was completed during the scenario evaluation. These examples can be used by teams as a demonstration and further education on how to use the toolkit in practice. Additionally, the full toolkit is also made available in the appendix to be used and incorporated into any team's current workflow.

SUMMARY & TAKEAWAYS

The EDF is a unique framework that was built with the MLOps process and the cross-functional development team in mind. It was designed in a way that the tools can be completed quickly within 2-week sprints and integrate well with a company's already established AI guidelines and principles. Although this work is now complete, it is far from done, as there will never be a way to completely get rid of all potential issues that may occur as a result of AI. Bias is always evolving, data is continuously growing, and the best ethical AI development team can do is adopt an empathetic perspective in their work and try to mitigate bias and other issues as much as they can.

The EDF currently has eight tools that can be used throughout the MLOps process, but this is just the beginning. It represents a foundation for more tools to be built and integrated into the EDF to cover any gaps that were out of scope for this project at this point in time.

CHAPTER 6: ANALYSIS & EVALUATION

PERSONAL REFLECTIONS

Personally, artificial intelligence and humanity-centered design are passions of mine. It was a pleasure to funnel that desire into this work in hopes of contributing to a future that displays the best side of AI, one that benefits all people no matter who they are. My hope is that this work encourages more standardization in ethical guidelines and frameworks in order to create a unified evaluative platform for AI.

I want those who read this work to understand the importance of empathy in the design of AI and avoid situations where "AI proponents simply want to create computer programs that perform tasks as well as or better than humans, without worrying about whether these programs are actually thinking in the way humans think" (Mitchell, 2019). Additionally, I hope that there is empathy when designing solutions for AI problems, specifically understanding the people and the roles involved in developing AI products to build solutions for them that they can use comfortably to tackle the problems of bias in the development process.

PREFERRED CROSS-FUNCTIONAL TEAM FOR EMPATHETIC AI

One of the alarming findings during this research was how little designers played a role in the development of AI products and services. It seems all too common that they are left out of the majority of the process until the very end when a user interface needs to be designed. At that point they have little power to identify underlying issues with the data or models from an empathetic perspective, leaving them unable to help to resolve these issues early on.

This may also signify the need for an entirely new type of designer, the ML Experience & Ethics Designer, who is the champion of people and ethics throughout the development lifecycle and can own the facilitation and management of this type of work. The addition of the new role could help identify issues early, strengthen ML models, and ultimately save money in the process, while creating a better end product.

LESSONS LEARNED

There is a never-ending growth of resources in the field of AI and it only continues to increase with the rapid pace of AI advancements. Since I decided to do this project alone, rather than in a team with others, it meant a lot of reading while being selective of what I read and what I cannot fit into my schedule. Having another person on the team would have significantly helped with this.

Gathering participants for this research study was challenging, although I was fortunate to have gathered a group of individuals with different roles and experiences, I would have preferred to have a larger participant pool to further investigate and validate the work to prevent potential gaps. Many people articulated their interest in being participants but were worried about speaking about their current team's development processes for fear of violating confidentiality if they mentioned anything that could put the company in a negative light. This is also why I chose not to use any of the participants' names, the companies they worked for, or their quotes, all in an effort to protect the safety of the participants.

Finally, bias is complex and challenging. You cannot simply draw a box that outlines what is or is not biased because there would be differing opinions on where you would draw that line. As Susan Leavy, Barry O'Sullivan, and Eugenia Siapera state in their paper titled Data, Power and Bias in Artificial Intelligence, "A considerable challenge in dealing with bias in AI generally, and in machine learning in particular, is that there can be a different definition of bias or what it means to be fair (Leavy, O'Sullivan, Siapera, 2020). All we can do is our best to provide an empathetic viewpoint to AI and work to try to prevent negative consequences like causing harm to people.

CHAPTER 7: CONCLUSION

The goal of this project was to answer the primary research question, what type of framework is needed to mitigate biases and provide an empathetic viewpoint during the development process of artificial intelligence? The answer is the Empathetic Design Framework, a toolkit and set of lenses that were developed with agile project management and the MLOps development process in mind, to ensure that it can be easily integrated into already established workflows.

Although this framework will not eliminate issues like bias altogether, it will help in mitigating bias and reducing the potential negative consequences that may impact people as a result of previously unidentified gaps and errors in an AI product or service. This framework was developed from an empathetic humanity-centered perspective and created to integrate into businesses' and teams' already established principles and guidelines. It works as a foundational centrepiece that can be built upon to cover more gaps and issues that were outside of the immediate scope and limitations of this project.

FINAL REMARKS

Artificial intelligence is rapidly integrating across industries and sectors leading to the development of new products and services to perform tasks and solve complex problems. The potential applications for AI in the future are almost unlimited, making the unimaginable possible. However, AI's success will be largely determined by its empathetic performance and how well it is able to complete its wide range of tasks without harming the many different groups of society it will inevitably impact.

Since in the end, artificial intelligence's future is dependent on empathy, something inherently it does not have the capability to truly understand, it is important to acknowledge the need for humans to access their own empathy and work as a filtering system for AI. A person using their own natural ability to empathize, in addition to leveraging frameworks like the Empathetic Design Framework to expand beyond their personal perspective, will help ensure the success of AI in the future and ultimately prevent another AI winter from occurring.

REFERENCES

1. Ahsan Noor Khan, Achintha Avin Ihalage, Yihan Ma, et al. (2021, February 3). Scientists propose new way to detect emotions using wireless signals. Queen Mary University of London. Retrieved March 14, 2023, from (link to website)
2. Ziga Avsec (2021, October 4). Predicting gene expression with AI. DeepMind. (link to website)
3. Government of Canada (2022, April 9). Responsible use of artificial intelligence (AI): Exploring the future of responsible AI in government. Government of Canada. Retrieved April 1, 2023, from (link to website)
4. Agarwal, N., Chiang, C. W., & Sharma, A. (n.d.). A Study on Computer Vision Techniques for Self-driving Cars. ResearchGate. Retrieved March 17, 2023, from (link to website)
5. Amazon. (n.d.). Alexa. Retrieved April 17, 2023, from (link to website)
6. Apple. (n.d.). Create ML. Retrieved April 17, 2023, from (link to website)
7. Apple. (n.d.). Siri. Retrieved April 17, 2023, from (link to website)
8. Atkins, S., Badrie, I., & van Otterloo, S. (2021). Applying ethical AI frameworks in practice: Evaluating conversational AI chatbot solutions. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. Retrieved November 3, 2022 from (link to website)
9. Bittle, J. (2020, March 13). Lie detectors have always been suspect. AI has made the problem worse. MIT Technology Review. Retrieved March 17, 2023, from (link to website)
10. Brittain, B. (2023, February 22). AI-created images lose U.S. Copyrights in test for new technology. Reuters. Retrieved March 14, 2023, from (link to website)
11. Broussard, M. (2019). Artificial unintelligence : how computers misunderstand the world. The Mit Press.
12. Brueck, H. (2023, April 6). The newest version of ChatGPT passed the US medical licensing exam with flying colors — And diagnosed a 1 in 100,000 condition in seconds. Insider Inc. Retrieved April 7, 2023, from (link to website)
13. Castaldo, J. (2023, March 25). They fell in love with the Replika AI chatbot. A policy update left them heartbroken. The Globe and Mail. Retrieved March 28, 2023, from (link to website)
14. Derico, B., & Kleinman, Z. (2023, March 14). OpenAI announces ChatGPT successor GPT-4. BBC News. Retrieved March 29, 2023, from (link to website)
15. Google Cloud. (n.d.). AutoML. Retrieved April 17, 2023, from https://cloud.google.com/automl
16. Google. (n.d.). Data collection. In People + AI Guidebook. Retrieved April 13, 2023 from (link to website)
17. Google. (n.d.). Google Assistant. Retrieved April 17, 2023, from (link to website)
18. Google. (n.d.). Production ML systems. Google Developers. Retrieved April 22, 2023, from (link to website)
19. Google. (n.d.). Responsible AI practices - Google. Retrieved April 18, 2023, from (link to website)
20. Haugeland, J. (2000). Artificial intelligence the very idea. Cambridge, Mass. [U.A.] Mit Press.

21. Hofstadter, D. R. (1999). Godel, Escher, Bach : an eternal golden braid. Basic Books.
22. IBM. (n.d.). Ethics in artificial intelligence. Retrieved April 13, 2023 from (link to website)
23. IDEO. (2021, August 2). AI Needs an Ethical Compass. This Tool Can Help. IDEO. Retrieved March 15, 2023, from (link to website)
24. Jonas, S. (2023, October 22). Meet Grace, the humanoid robot offering companionship in a Montreal nursing home. CBC News. Retrieved March 28, 2023, from (link to website)
25. Jose, T. (2023, April 17). SwitchOn, a Vision AI Company Raises $4.2 Million To Bring AI Into Manufacturing. Entrepreneur India. Retrieved April 18, 2023, from (link to website)
26. Joshi, M. L., & Kanoongo, N. (2020). Depression detection using emotional artificial intelligence and machine learning: A closer review. Journal of Healthcare Engineering. Retrieved March 24, 2023 from. (link to website)
27. Kreuzberger, D., Kühl, N., & Hirschl, S. (2019). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. IEEE Access. Retrieved November 23, 2022 from (link to website)
28. Leavy, S., O’Sullivan, B., & Siapera, E. (2021). Data, power and bias in artificial intelligence. Big Data & Society. Retrieved February 1, 2023, from. (link to website)
29. Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2018). The influence of empathy in human-robot relations. International Journal of Human-Computer Studies. Retrieved March 6, 2023, from (link to website)
30. Levesque, H. J. (2018). Common sense, the Turing test, and the quest for real AI. Mit Press.
31. Linh, V. T. N., Lee, M.-Y., Mun, J., et al. (2021). 3D plasmonic coral nanoarchitecture paper for label-free human urine sensing and deep learning-assisted cancer screening. Journal of Materials Chemistry B. Retrieved April 3, 2023, from. (link to website)
32. Lobe. (n.d.). Lobe. Retrieved April 17, 2023, from (link to website)
33. Microsoft. (n.d.). Cortana. Retrieved April 17, 2023, from (link to website)
34. Microsoft. (n.d.). Responsible AI - Microsoft. Retrieved April 18, 2023, from (link to website)
35. Mitchell, M. (2020). ARTIFICIAL INTELLIGENCE : a guide for thinking humans. picador.
36. Moreno, J. (2022, November 22). Notion Releases Alpha of Generative AI Copywriting Tool. Retrieved March 28, 2023, from (link to website)
37. Murphy, R. (2019). Introduction to AI robotics. The Mit Press.
38. Nelson, G. S. (2019). Bias in artificial intelligence. North Carolina Medical Journal. Retrieved October 13, 2022, from (link to website)
39. Ntoutsi, E., Fafalios, P., Gadiraju, U., et al. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Retrieved December 11, 2022 from (link to website)
40. O’neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Penguin Books.
41. Office of Science and Technology Policy (2023, March 16). Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. The White House. Retrieved March 29,

2023, from (link to website)

42. OpenAI. (n.d.). Chat with GPT: Language models as interfaces. Retrieved April 13, 2023 from (link to website)
43. Q.ai (2023, February 27). Google Announces Bard, Its Rival To Microsoft-Backed ChatGPT. Retrieved March 14, 2023, from (link to website)
44. Robert, C. (2017). Superintelligence: Paths, Dangers, Strategies. CHANCE, 30(1), 42–43. (link to website)
45. Roose, K. (2023, February 16). A Conversation With Bing's Chatbot Left Me Deeply Unsettled. Retrieved March 12, 2023, from (link to website)
46. Russell, S., & Norvig, P. (2010). Artificial intelligence: a modern approach (3rd ed.). Pearson.
47. Sachdeva, P. S., Barreto, R., von Vacano, C., & Kennedy, C. J. (2021). Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Retrieved October 11, 2022 from. (link to website)
48. Schwartz, J., Samet, A., & Lastra, A. (2019). Learning from complex spatial datasets: the case of urban environments. Apple Machine Learning Research. Retrieved April 17, 2023, from (link to website)
49. Shieber, S. M. (2004). The Turing test : verbal behavior as the hallmark of intelligence. Mit Press.
50. Smith, S., & Ashby, M. (2020). How to Future. Kogan Page Inspire.
51. Srinivasan, R., & San Miguel Gonzalez, B. (2021). The role of empathy for artificial intelligence accountability. AI & Society. Retrieved March 11, 2023, from (link to website)
52. Tegmark, M. (2018). Life 3.0 : being human in the age of artificial intelligence. Penguin Books.
53. The Open Data Institute. (n.d.). Data Ethics Maturity Model: Benchmarking your approach to data ethics. The Open Data Institute. Retrieved March 7, 2023, from (link to website)
54. Thebault, R. (2019, September 11). California could become the largest state to ban facial recognition in body cameras. The Washington Post. Retrieved March 15, 2023, from (link to website)
55. Victor, D. (2016, March 24). Microsoft Created a Twitter Bot to Learn from Users. It Quickly Became a Racist Jerk. Retrieved February 11, 2023, from (link to website)
56. Warren, T. (2023, March 28). Microsoft Security Copilot is a new GPT-4 AI assistant for cybersecurity. Retrieved March 31, 2023, from (link to website)
57. Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Public Affairs.

APPENDIX

EDF Test Drive - The following pages displays the completed EDF activities that were used during the EDF test drive to evaluate the performance of the EDF within the potential future AI scenarios that were developed during the strategic foresight section of the paper. To revist the outlining of the process and description of the EDF test drive, please refer to Chapter 3.

The completed activities in the EDF test drive are a helpful resource to understand how to use and leverage each of the tools that are available in the EDF, however it is important to understand that the future scenarios, models, and data used were hypothetical and were only described to work through each activity.

EDF | ML Problem Assessment Matrix

Not every problem is a ML problem, this tool was created to help determine whether ML needs to or should be used to solve the problem.

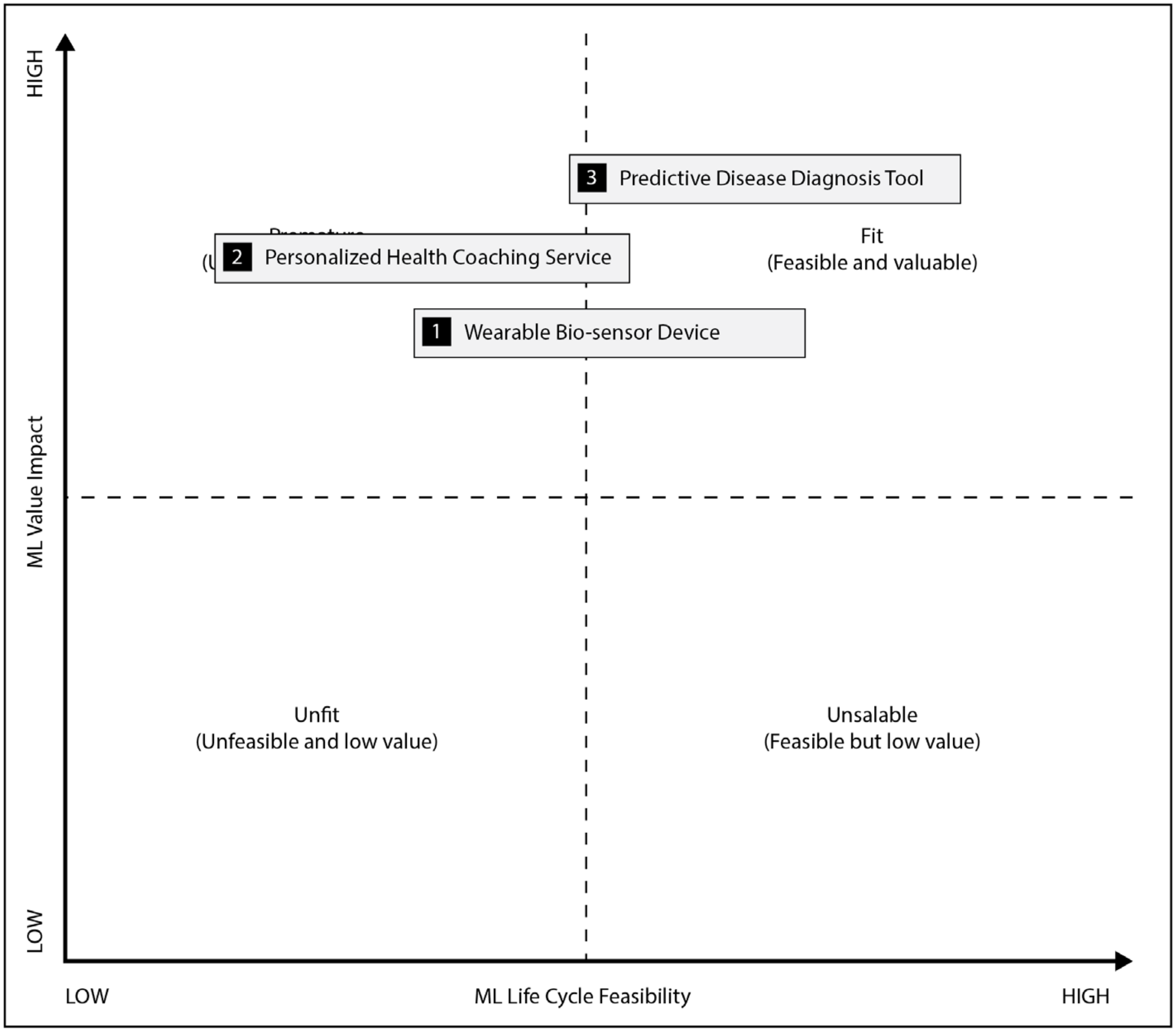
INSTRUCTIONS

1. Write the business problem statement (BPS) above the matrix and place it in the centre of the matrix.
2. Move the BPS up or down based on how much added value ML would bring to the solution.
3. Move the BPS left or right based on how feasible it would be from a resources (Cost + Human) perspective to maintain.
4. A BPS that ends up in the top right corner should be considered a potential fit for ML, a BPS that ends up in the bottom left or bottom right corner should be considered unfit for ML. A BPS in the top left corner should be considered once resources are added.

1 As a health-conscious individual, I need a wearable device with bio-sensors that can provide real-time health data so that I can continuously monitor my health, track my progress towards my health goals, and take timely actions to maintain my well-being.

2 As a health-conscious individual, I need a personalized health coaching service that utilizes my health data to provide tailored recommendations and guidance so that I can receive expert support, make informed decisions about my health, and wellness goals.

3 As a health-conscious individual, I need a predictive disease diagnosis tool that utilizes ML algorithms to detect and diagnose diseases early based on my health data so that I can receive timely medical intervention, and proactively manage my health.


HIGH
ML Value Impact
LOW
3 Predictive Disease Diagnosis Tool
(U
2 Personalized Health Coaching Service
Fit
(Feasible and valuable)
1 Wearable Bio-sensor Device
Unfit
(Unfeasible and low value)
Unsalable
(Feasible but low value)
LOW
ML Life Cycle Feasibility
HIGH

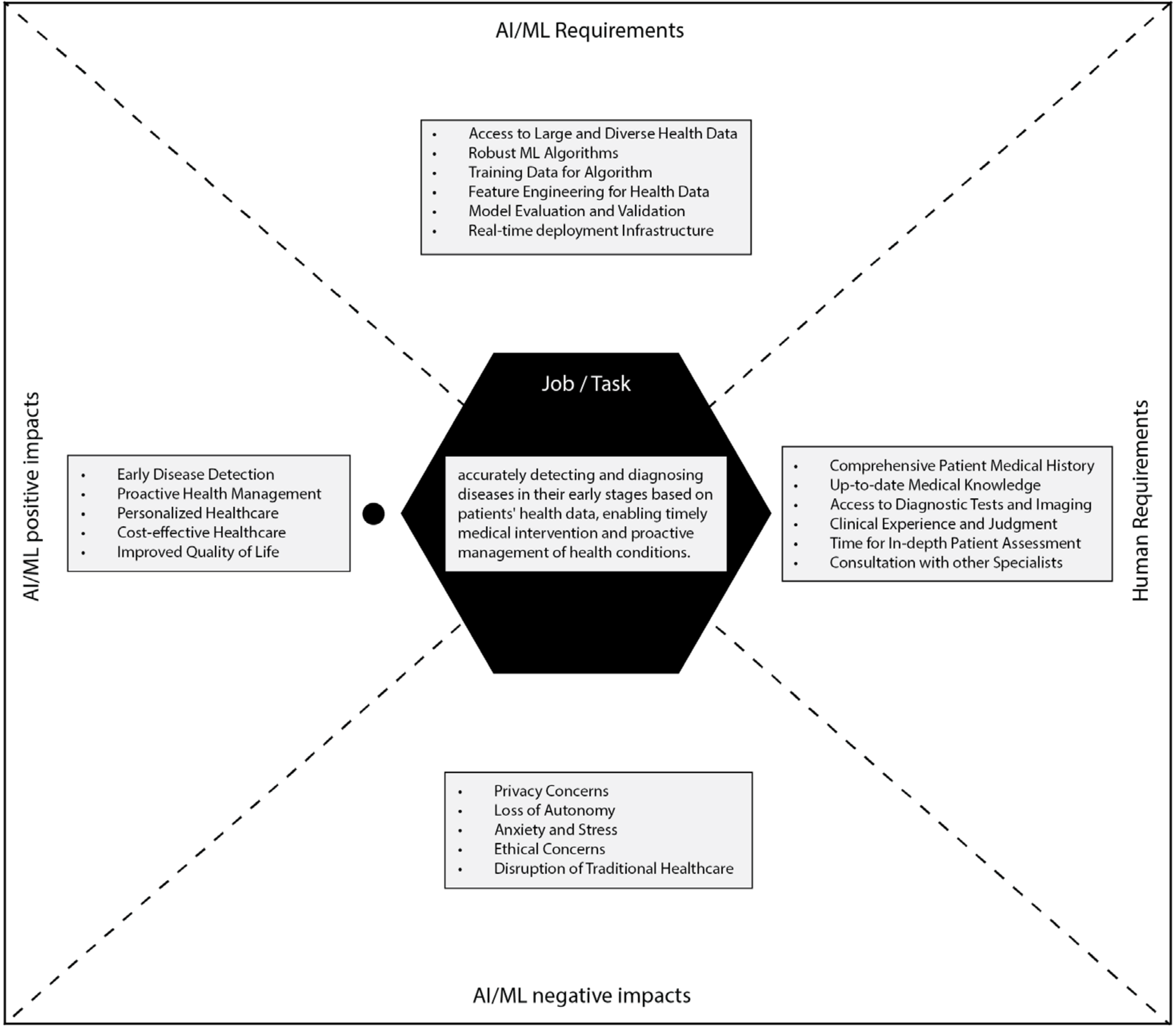
EDF | User-Centric ML Impact Analysis

When planning and developing a ML solution it is important to first understand the needs and requirements, as well as the potential positive and negative impact on people.

INSTRUCTIONS

1. Write the ML problem statement in the top section.
2. Place the job or task that is currently performed by a person that will be completed by the ML in the centre of the diagram.
3. Starting on the left side, describe the ways in which the AI/ML may positively impact people.
4. Describe the ways the AI/ML may negatively impact people.
5. Identify and document the necessary requirements needed in order for humans to perform the job or task effectively.
6. Document the necessary requirements needed in order for the AI/ML to perform the job or task effectively.

ML Problem Statement	As a health-conscious individual, I need an ML-based predictive disease diagnosis tool that can effectively analyze my health data, including medical records, physiological measurements, and lifestyle factors, to accurately detect and diagnose diseases in their early stages. The tool should also provide timely recommendations for medical intervention, monitoring, and proactive management of health conditions, while considering data privacy concerns, to enable me to take informed actions to maintain and improve my health outcomes.



AI/ML Requirements

- Access to Large and Diverse Health Data
- Robust ML Algorithms
- Training Data for Algorithm
- Feature Engineering for Health Data
- Model Evaluation and Validation
- Real-time deployment Infrastructure

AI/ML positive impacts

- Early Disease Detection
- Proactive Health Management
- Personalized Healthcare
- Cost-effective Healthcare
- Improved Quality of Life

Job / Task

accurately detecting and diagnosing diseases in their early stages based on patients' health data, enabling timely medical intervention and proactive management of health conditions.

Human Requirements

- Comprehensive Patient Medical History
- Up-to-date Medical Knowledge
- Access to Diagnostic Tests and Imaging
- Clinical Experience and Judgment
- Time for In-depth Patient Assessment
- Consultation with other Specialists

AI/ML negative impacts

- Privacy Concerns
- Loss of Autonomy
- Anxiety and Stress
- Ethical Concerns
- Disruption of Traditional Healthcare

EDF | Data Logic and Limitations

When selecting which data sources to potentially use for your ML project, it is important to document the specific reasons it should be included as well as its data limitations.

INSTRUCTIONS

1. Document each data source you are considering to use to train the ML model in the first row *(WHAT)*.
2. In the second row, describe the logic and reasoning for including these data sources in your project *(WHY)*.
3. For each data source, identify any known limitations these data sources may have, from both a technical and EDF perspective *(GAPS)*.

EDF Perspective Lenses

1. Bias & Fairness
2. Privacy & Security
3. Reliable & Trustworthy
4. Diversity & Inclusivity
5. Ethical & Safe
6. Openness & Transparency

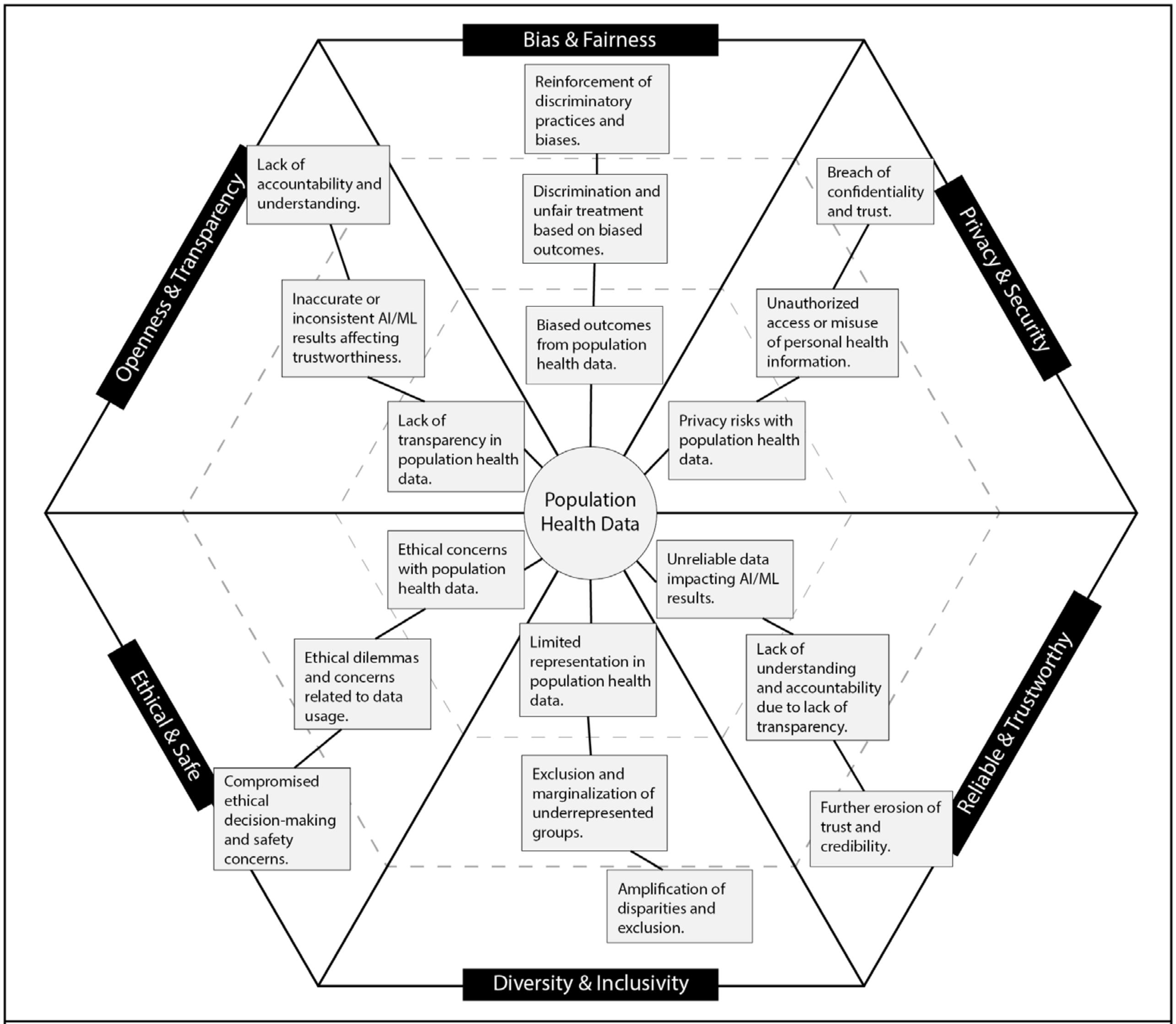
Data Source	Electronic Health Records (EHRs)	Wearable Devices and Sensors	Laboratory Test Results	Genetic and Genomic Data	Population Health Data
Inclusion Reasoning	To get comprehensive patient medical history.	To receive real-time health data from the user.	To received and alnalyze and previous or new lab results.	To determine any potential disease risk factors based on genetic predisposition.	To compare results with general population data.
Known Limitations	Incomplete, sometimes inaccurate patient data due to manual data entry errors.	data accuracy variability and reliability when users does not use device or the device malfunctions.	Result quality may vary from labratory to labratory	User must have completed genetic testing for this data to be complete, while also not all genetic diseases are included	potential biases and inaccuracies due to unreliable data collection methods.

EDF | Negative Implications of Data Map

There are always potential negative implications for including a particular data source in an ML project, this tool helps map out the potential consequences through the EDF lenses.

INSTRUCTIONS

1. Place the data source of interest at the center of the data map.
2. Document and connect any first order consequences that may arise as a result of using this data source.
3. Document and connect any second order consequences that may arise from the identified first order of consequences.
4. Document and connect any third order consequences that may arise from the identified second order of consequences.
5. Discuss your findings and list any identified themes and patterns in the lower section.


Bias & Fairness
Reinforcement of discriminatory practices and biases.
Discrimination and unfair treatment based on biased outcomes.
Biased outcomes from population health data.
Privacy & Security
Breach of confidentiality and trust.
Unauthorized access or misuse of personal health information.
Privacy risks with population health data.
Reliable & Trustworthy
Unreliable data impacting AI/ML results.
Lack of understanding and accountability due to lack of transparency.
Further erosion of trust and credibility.
Diversity & Inclusivity
Limited representation in population health data.
Exclusion and marginalization of underrepresented groups.
Amplification of disparities and exclusion.
Ethical & Safe
Ethical concerns with population health data.
Ethical dilemmas and concerns related to data usage.
Compromised ethical decision-making and safety concerns.
Openness & Transparency
Lack of transparency in population health data.
Inaccurate or inconsistent AI/ML results affecting trustworthiness.
Lack of accountability and understanding.
Population Health Data

Themes & Patterns

EDF | Negative Implications of Data Remedial Plan

After identifying some potential negative implications of using a particular data source, this tool helps you document the preferred outcome and backcast to develop a remedial plan.

INSTRUCTIONS

1. Document the potential negative outcomes that may arise as a result of using a particular data source in the first column. *What is it that may go wrong?*
2. Describe the preferred outcomes for using a particular data source in the third column. *What do you want to happen?*
3. In the second column, identify a remedial action that you and your team can take to prevent the potentially negative outcome and ensure the preferred outcome happens. *What actions can you take?*

Potential Negative Outcome	Remedial Action	Preferred Outcome
Biased outcomes from population health data.	Review data for diversity gaps or issues, conduct thorough data validation, and use statistical methods to adjust for known biases.	Ability to use population health data without any biased outcomes.
Lack of transparency in population health data.	Increase data documentation, data governance, and data reporting to allow users and the public to know what is collected and how it is used safely and ethically.	Enhanced transparency in population health data.
Privacy risks with population health data.	Implement strong encryption methhods, implement data breach response plan, and establish a strict access controls measure.	Low privacy risk due to robust privacy and safety measures
Ethical concerns with population health data.	Develop plan with clear guidelines and ethical principles to identify, address, and mitigate ethical concerns before they arise.	Reduced ethical concerns due to thoughtful and consentrated ethics and responsibility plan.
Limited representation in population health data.	Identify gaps and limitations of current data source from a diversity and inclusivity perspective and levarage another data source to build a stronger data set.	Enhanced diversity and inclusive representation in health data.
Unreliable data impacting AI/ML results.	Improve data collection and management plan to validate trusted sources, review for ethical concerns, standardize collection methods, and schedule a continuous monitoring plan.	High-quality and reliable data leading to enhanced results.

EDF | Model Training & Selection

The following tool is used to measure a model candidate by how well it performs in each EDF measurement to identify weaknesses and strengths and develop an improvement plan.

INSTRUCTIONS

1. Document the current set of potential model candidates that you would like to assess.
2. Evaluate each model candidate by the Empathy Design Framework's perspective lenses, using the scoring legend.
3. Compare the performance of the models with how well they performed in this assessment and any required technical evaluations.
4. Select the best performing model, develop a plan for gap improvement based on what areas the model scored low.

Scoring Legend

- Low | The model performs poorly in this area and will need significant adjustment to meet this evaluative criteria.
- Medium | The model performs moderately well and will need only minor adjustments to meet this evaluative criteria.
- High | The model performs significantly well in this area and needs little to no adjustment to meet this evaluative criteria.

Bias & Fairness	H	Privacy & Security	M	Reliable & Trustworthy	M	Bias & Fairness	L	Privacy & Security	M	Reliable & Trustworthy	M
Model Candidate Model Candidate 1						Model Candidate Model Candidate 2					
Diversity & Inclusivity	H	Ethical & Safe	H	Openness & Transparency	M	Diversity & Inclusivity	L	Ethical & Safe	L	Openness & Transparency	M
Bias & Fairness	L	Privacy & Security	L	Reliable & Trustworthy	H	Bias & Fairness	H	Privacy & Security	L	Reliable & Trustworthy	L
Model Candidate Model Candidate 3						Model Candidate Model Candidate 4					
Diversity & Inclusivity	M	Ethical & Safe	M	Openness & Transparency	L	Diversity & Inclusivity	H	Ethical & Safe	H	Openness & Transparency	L
Bias & Fairness	H	Privacy & Security	H	Reliable & Trustworthy	L	Bias & Fairness	M	Privacy & Security	L	Reliable & Trustworthy	H
Model Candidate Model Candidate 5						Model Candidate Model Candidate 6					
Diversity & Inclusivity	H	Ethical & Safe	M	Openness & Transparency	H	Diversity & Inclusivity	L	Ethical & Safe	H	Openness & Transparency	L

EDF | Future Risk Scenario

This tool helps develop a set of future risk scenarios that may arise once a model has moved into production, which can help in improving the model to avoid potential risks.

INSTRUCTIONS

1. Place the model candidate you are evaluating in the section at the top.
2. Document the potential negative implications (if any) of the model candidate once in production within each EDF lens.
3. Use the scenario section in the centre to describe 6 future scenarios, each scenario should focus only on the negatives implications of one of the EDF lenses.
4. Develop a plan to mitigate against these potential future scenarios from occurring or choose a different model candidate.

Model Candidate

Model Candidate 2

Openness & Transparency

Model suffers from a lack of explainability

Model needs further adjustments to improve its decision-making

Bias & Fairness

Model exhibits bias in its predictions

Model has dificulty addressing issues related to bias and fairness.

Privacy & Security

Model needs stronger data protection security

Model needs further refinement to comply with privacy regulations

Scenarios

Bias & Fairness: The Predictive Disease Diagnosis Tool is released but exhibits bias in its predictions, leading to unfair outcomes for unrepresented population groups.

Privacy & Security: The Predictive Disease Diagnosis Tool is released but experiences a data breach that effects patients' sensitive health information.

Openness & Transparency: The Predictive Disease Diagnosis Tool is released but lacks transparency in its operations, making the model's outputs difficult to understand, leading to trust concerns.

Reliable & Trustworthy: The Predictive Disease Diagnosis Tool is released but produces inconsistent predictions, leading to a lack of confidence in consumers and investors.

Ethical & Safe: The Predictive Disease Diagnosis Tool is released but leads to unintended biases and potential harm to patients.

Diversity & Inclusivity: The Predictive Disease Diagnosis Tool is released with limited representation of diverse data, resulting in inaccurate predictions and worsening health disparities.

Ethical & Safe

Model needs stronger safety measures as it is exibiting biases in its results

Model is prone to ethical issues.

Diversity & Inclusivity

Model has limited diverse groups in its data.

Model is not prepared to handle issues related to diversity.

Reliable & Trustworthy

Model has accuracy issues.

Model needs to improve its ability to make predictions.

EDF | ML Monitoring Radar

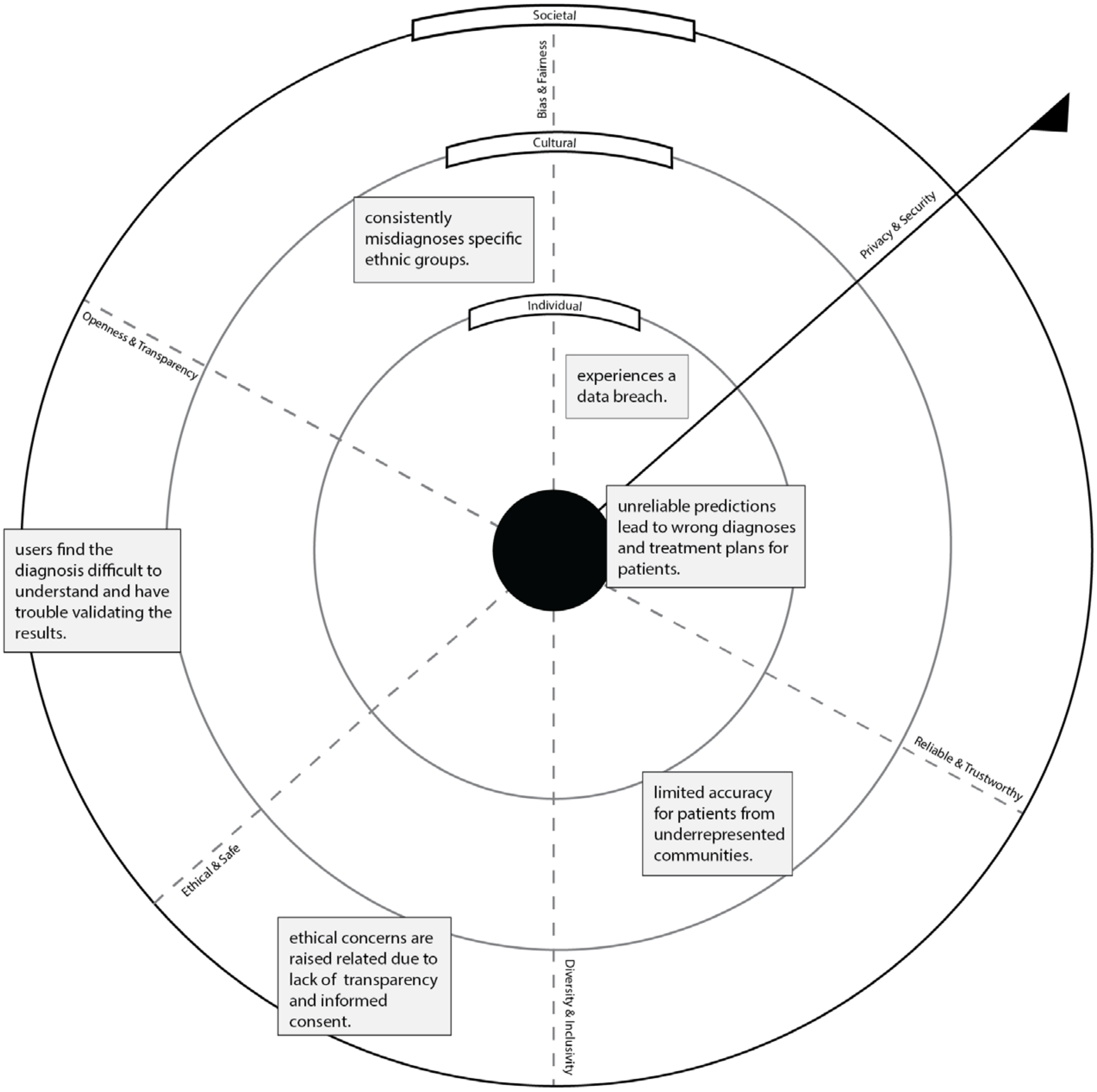
It is important to monitor an ML project throughout its life cycle, this tool acts as a radar to document potential issues through various lenses in order to address them in future versions.

INSTRUCTIONS

1. Identify key monitoring milestones in your product planning timeline to review the results of this tool.
2. Utilize the 6 Empathetic Design Framework lenses to document issues that arise while the ML is in production.
3. Place each documented issue within the radial swim lane that best applies to its level of impact.
4. Develop a plan to correct each issue and integrate the solutions into the following sprint planning sessions.

Scoring Legend

- Individual | The identified issue mainly impacts the direct user or an individual as a result of the ML product / service.
- Cultural | The identified issue mainly impacts specific groups of people as a result of the ML product / service.
- Societal | The identified issue impacts several or more groups as a result of the ML product / service.



consistently misdiagnoses specific ethnic groups.

experiences a data breach.

unreliable predictions lead to wrong diagnoses and treatment plans for patients.

users find the diagnosis difficult to understand and have trouble validating the results.

limited accuracy for patients from underrepresented communities.

ethical concerns are raised related due to lack of transparency and informed consent.

APPENDIX

The Empathetic Design Framework - the following section includes the entire Empathetic Design Framework and toolkit to be copied and used within a team's development process. The full resource includes:

- The Empathetic Design Framework set of lenses
- MLOps Process Diagram
- ML Problem Assessment Matrix
- User-Centric ML Impact Analysis
- Data Logic and Limitations
- Negative Implications of Data Map
- Negative Implications of Data Remedial Plan
- Model Training & Selection
- Future Risk Scenario
- ML Monitoring Radar

The Empathetic Design Framework (EDF)

The Empathetic Design Framework is a set of individual lenses that provide a variety perspectives to view and evaluate artificial intelligence / machine learning (AI/ML) products and their components during the ML Ops process and in production to identify risks, gaps, and reduce potential harm.

Bias & Fairness
An AI/ML product should be free of systematic errors in its decision-making that can lead to unjust outcomes for people, it should be designed to mitigate bias and should treat all people fairly and equally without discrimination.

Diversity & Inclusivity
An AI/ML product should be built and designed to include the wide variety of perspectives and backgrounds that exist, while also considering those different perspectives in its decision making and results.

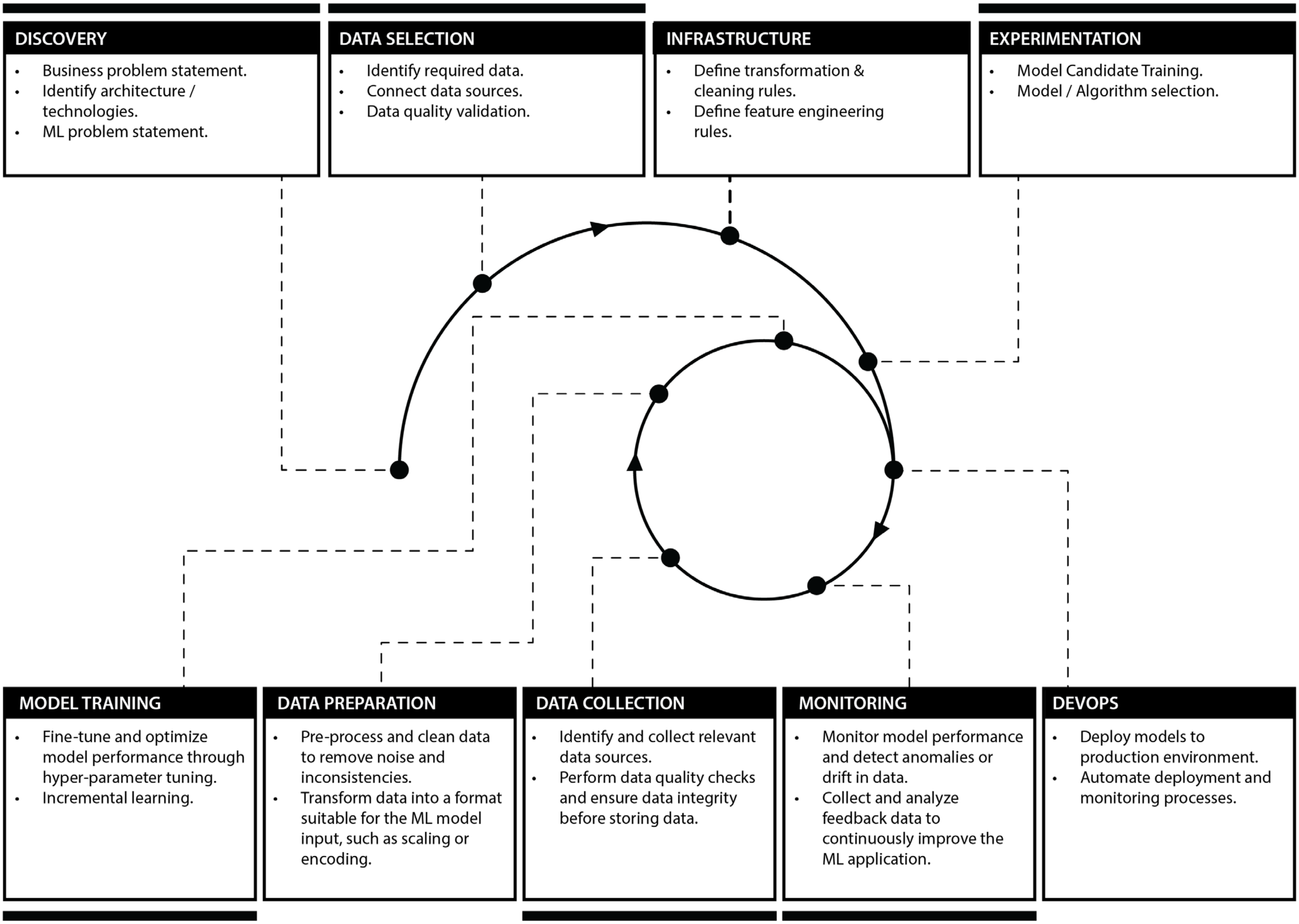
Privacy & Security
An AI/ML product should be built on a foundation of privacy by design, protecting personal information from unauthorized access or control of personal data, while protecting itself from unauthorized access and attacks.

Ethical & Safe
An AI/ML product should be built on a foundation of ethics by design, ensuring that it is developed and used safely in accordance with ethical principles and should not cause harm to individuals, cultures, or society.

Reliable & Trustworthy
An AI/ML product should be a credible and dependable system that is both consistent and accurate, people who use it or are impacted by its work should be able to rely on its effectiveness and the confidence of its decision making.

Openness & Transparency
An AI/ML product should be able to provide clear and understandable information about how it works and its decision making, while ensuring that its systems and components are both accessible and available for inspection and review.

EDF | ML OPS OVERVIEW DIAGRAM



DISCOVERY

- Business problem statement.
- Identify architecture / technologies.
- ML problem statement.

DATA SELECTION

- Identify required data.
- Connect data sources.
- Data quality validation.

INFRASTRUCTURE

- Define transformation & cleaning rules.
- Define feature engineering rules.

EXPERIMENTATION

- Model Candidate Training.
- Model / Algorithm selection.

MODEL TRAINING

- Fine-tune and optimize model performance through hyper-parameter tuning.
- Incremental learning.

DATA PREPARATION

- Pre-process and clean data to remove noise and inconsistencies.
- Transform data into a format suitable for the ML model input, such as scaling or encoding.

DATA COLLECTION

- Identify and collect relevant data sources.
- Perform data quality checks and ensure data integrity before storing data.

MONITORING

- Monitor model performance and detect anomalies or drift in data.
- Collect and analyze feedback data to continuously improve the ML application.

DEVOPS

- Deploy models to production environment.
- Automate deployment and monitoring processes.

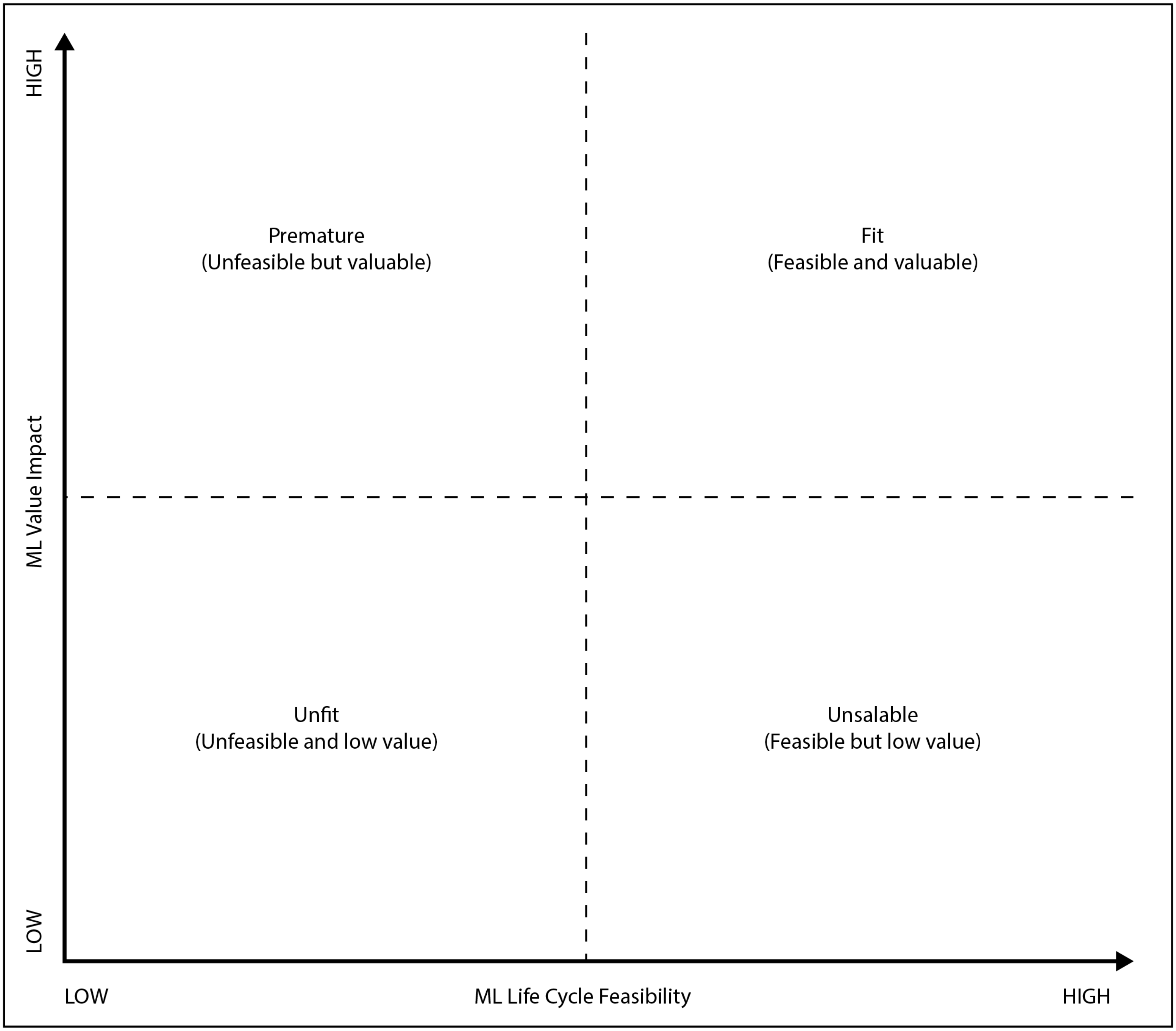
EDF | ML Problem Assessment Matrix

Not every problem is a ML problem, this tool was created to help determine whether ML needs to or should be used to solve the problem.

INSTRUCTIONS

1. Write the business problem statement (BPS) above the matrix and place it in the centre of the matrix.
2. Move the BPS up or down based on how much added value ML would bring to the solution.
3. Move the BPS left or right based on how feasible it would be from a resources (Cost + Human) perspective to maintain.
4. A BPS that ends up in the top right corner should be considered a potential fit for ML, a BPS that ends up in the bottom left or bottom right corner should be considered unfit for ML. A BPS in the top left corner should be considered once resources are added.

Business Problem Statement



Premature
(Unfeasible but valuable)

Fit
(Feasible and valuable)

Unfit
(Unfeasible and low value)

Unsalable
(Feasible but low value)

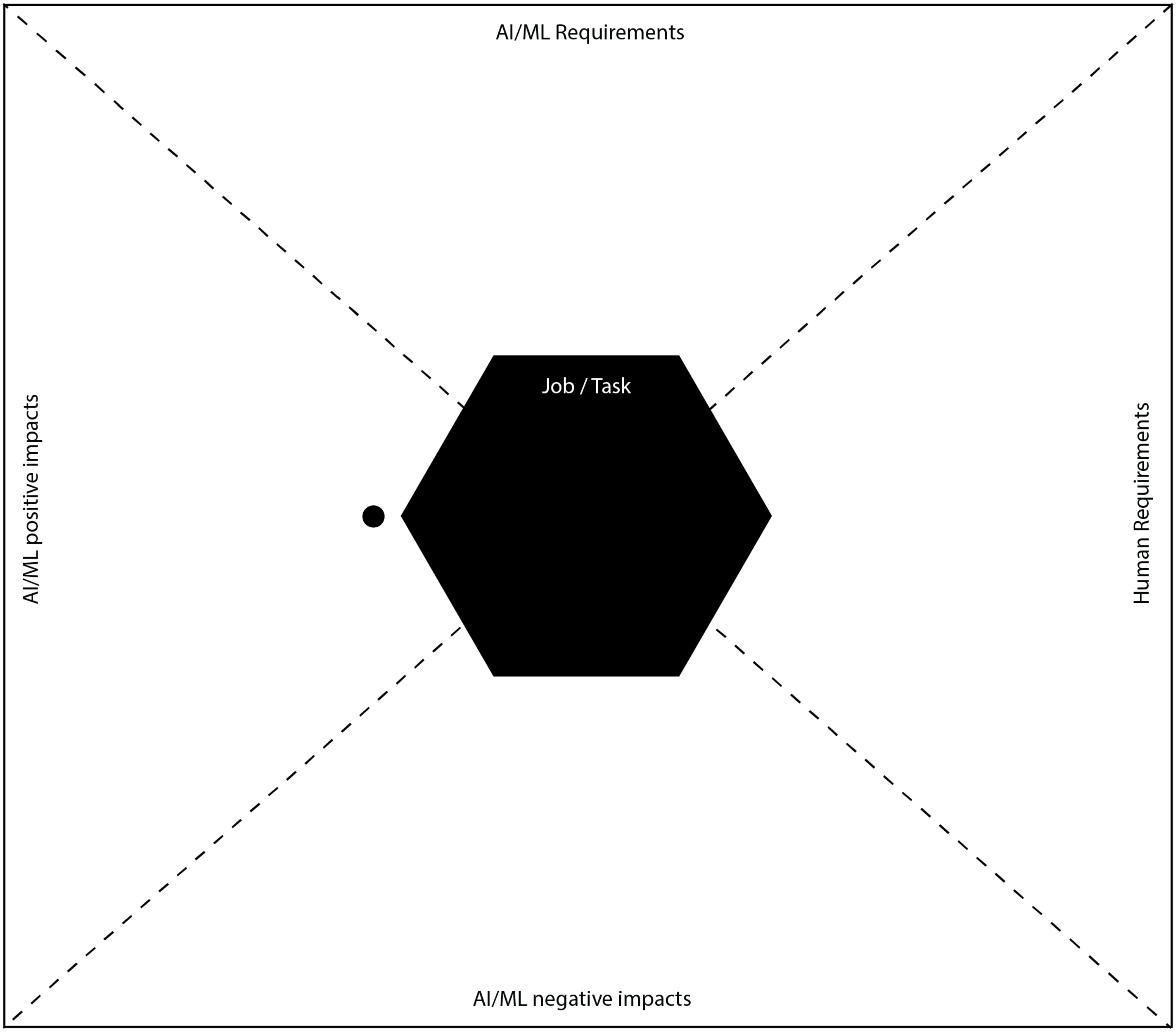
EDF | User-Centric ML Impact Analysis

When planning and developing a ML solution it is important to first understand the needs and requirements, as well as the potential positive and negative impact on people.

INSTRUCTIONS

1. Write the ML problem statement in the top section.
2. Place the job or task that is currently performed by a person that will be completed by the ML in the centre of the diagram.
3. Starting on the left side, describe the ways in which the AI/ML may positively impact people.
4. Describe the ways the AI/ML may negatively impact people.
5. Identify and document the necessary requirements needed in order for humans to perform the job or task effectively.
6. Document the necessary requirements needed in order for the AI/ML to perform the job or task effectively.

ML Problem Statement


AI/ML Requirements

AI/ML positive impacts

Job / Task

Human Requirements

AI/ML negative impacts

EDF | Data Logic and Limitations

When selecting which data sources to potentially use for your ML project, it is important to document the specific reasons it should be included as well as its data limitations.

INSTRUCTIONS

1. Document each data source you are considering to use to train the ML model in the first row *(WHAT)*.
2. In the second row, describe the logic and reasoning for including these data sources in your project *(WHY)*.
3. For each data source, identify any known limitations these data sources may have, from both a technical and EDF perspective *(GAPS)*.

EDF Perspective Lenses

1. Bias & Fairness
2. Privacy & Security
3. Reliable & Trustworthy
4. Diversity & Inclusivity
5. Ethical & Safe
6. Openness & Transparency

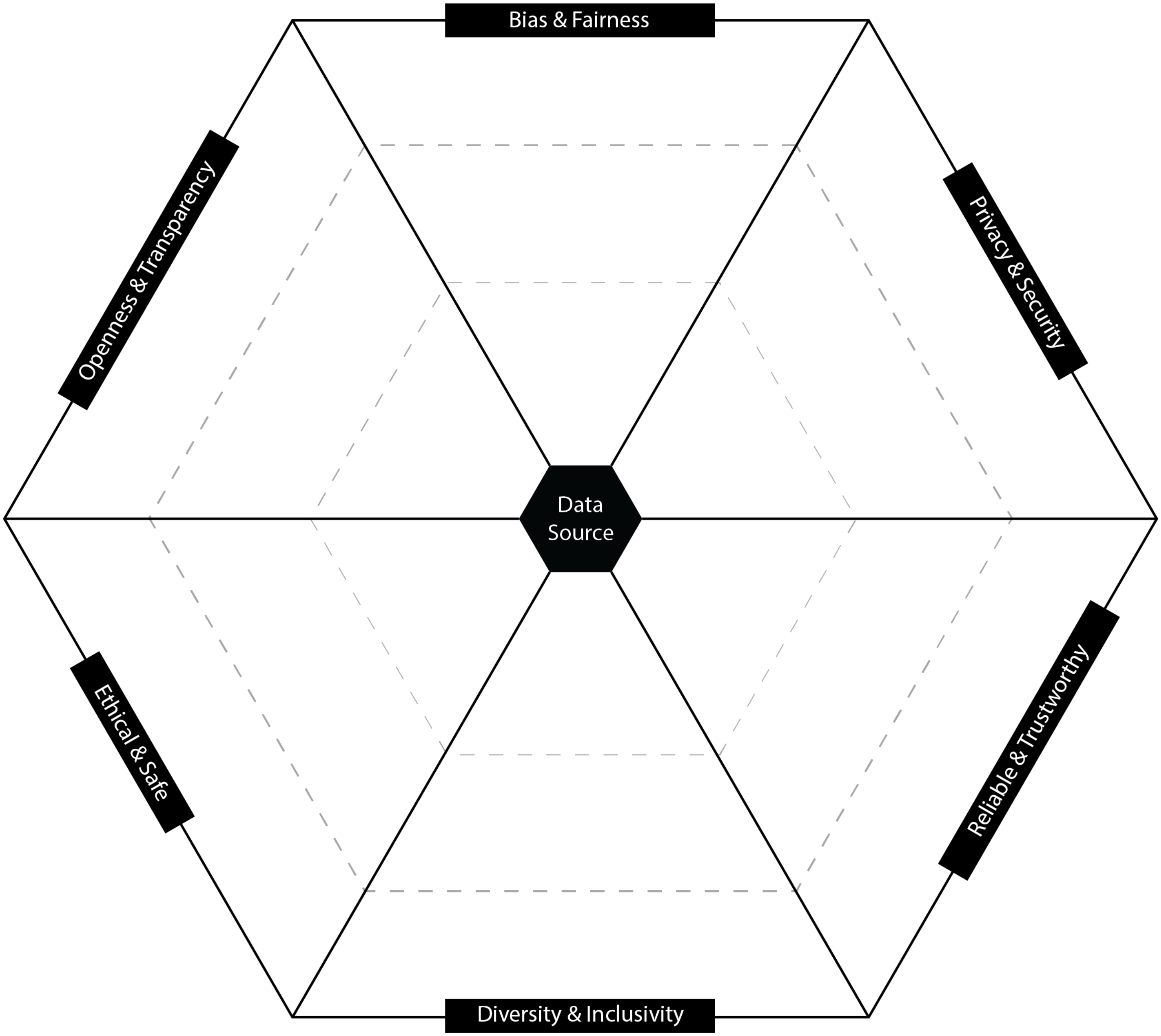
Data Source					
Inclusion Reasoning					
Known Limitations					

EDF | Negative Implications of Data Map

There are always potential negative implications for including a particular data source in an ML project, this tool helps map out the potential consequences through the EDF lenses.

INSTRUCTIONS

1. Place the data source of interest at the center of the data map.
2. Document and connect any first order consequences that may arise as a result of using this data source.
3. Document and connect any second order consequences that may arise from the identified first order of consequences.
4. Document and connect any third order consequences that may arise from the identified second order of consequences.
5. Discuss your findings and list any identified themes and patterns in the lower section.



Themes & Patterns

EDF | Negative Implications of Data Remedial Plan

After identifying some potential negative implications of using a particular data source, this tool helps you document the preferred outcome and backcast to develop a remedial plan.

INSTRUCTIONS

1. Document the potential negative outcomes that may arise as a result of using a particular data source in the first column. *What is it that may go wrong?*
2. Describe the preferred outcomes for using a particular data source in the third column. *What do you want to happen?*
3. In the second column, identify a remedial action that you and your team can take to prevent the potentially negative outcome and ensure the preferred outcome happens. *What actions can you take?*

Potential Negative Outcome	Remedial Action	Preferred Outcome

EDF | Model Training & Selection

The following tool is used to measure a model candidate by how well it performs in each EDF measurement to identify weaknesses and strengths and develop an improvement plan.

INSTRUCTIONS

1. Document the current set of potential model candidates that you would like to assess.
2. Evaluate each model candidate by the Empathy Design Framework's perspective lenses, using the scoring legend.
3. Compare the performance of the models with how well they performed in this assessment and any required technical evaluations.
4. Select the best performing model, develop a plan for gap improvement based on what areas the model scored low.

Scoring Legend

- Low | The model performs poorly in this area and will need significant adjustment to meet this evaluative criteria.
- Medium | The model performs moderately well and will need only minor adjustments to meet this evaluative criteria.
- High | The model performs significantly well in this area and needs little to no adjustment to meet this evaluative criteria.

Bias & Fairness		Privacy & Security		Reliable & Trustworthy	

Model Candidate

Diversity & Inclusivity		Ethical & Safe		Openness & Transparency	

Bias & Fairness		Privacy & Security		Reliable & Trustworthy	

Model Candidate

Diversity & Inclusivity		Ethical & Safe		Openness & Transparency	

Bias & Fairness		Privacy & Security		Reliable & Trustworthy	

Model Candidate

Diversity & Inclusivity		Ethical & Safe		Openness & Transparency	

Bias & Fairness		Privacy & Security		Reliable & Trustworthy	

Model Candidate

Diversity & Inclusivity		Ethical & Safe		Openness & Transparency	

Bias & Fairness		Privacy & Security		Reliable & Trustworthy	

Model Candidate

Diversity & Inclusivity		Ethical & Safe		Openness & Transparency	

Bias & Fairness		Privacy & Security		Reliable & Trustworthy	

Model Candidate

Diversity & Inclusivity		Ethical & Safe		Openness & Transparency	

EDF | Future Risk Scenario

This tool helps develop a set of future risk scenarios that may arise once a model has moved into production, which can help in improving the model to avoid potential risks.

INSTRUCTIONS

1. Place the model candidate you are evaluating in the section at the top.
2. Document the potential negative implications (if any) of the model candidate once in production within each EDF lens.
3. Use the scenario section in the centre to describe 6 future scenarios, each scenario should focus only on the negatives implications of one of the EDF lenses.
4. Develop a plan to mitigate against these potential future scenarios from occurring or choose a different model candidate.

Model Candidate

Openness & Transparency

Bias & Fairness

Privacy & Security

Scenarios

Ethical & Safe

Diversity & Inclusivity

Reliable & Trustworthy

EDF | ML Monitoring Radar

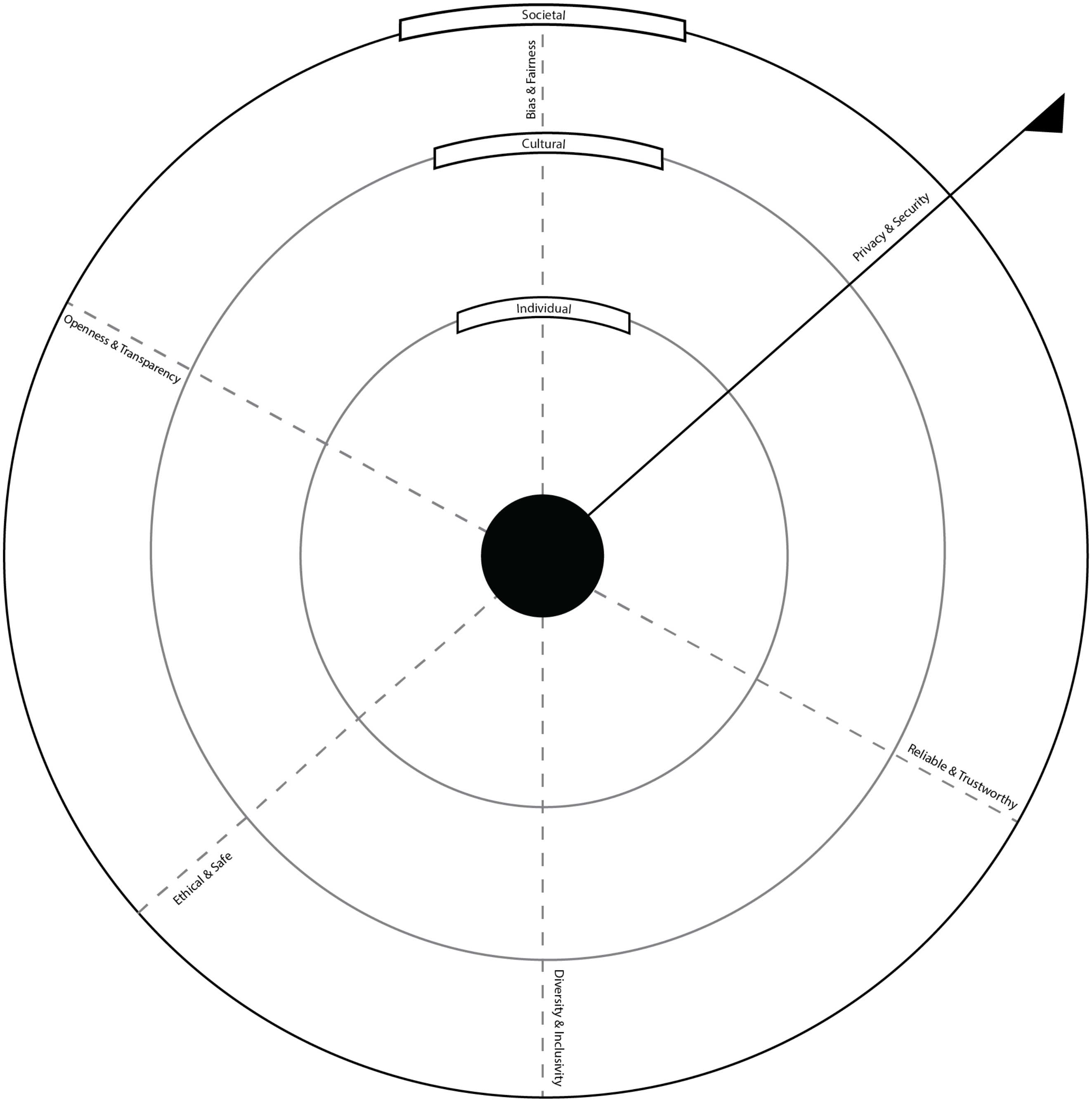
It is important to monitor an ML project throughout its life cycle, this tool acts as a radar to document potential issues through various lenses in order to address them in future versions.

INSTRUCTIONS

1. Identify key monitoring milestones in your product planning timeline to review the results of this tool.
2. Utilize the 6 Empathetic Design Framework lenses to document issues that arise while the ML is in production.
3. Place each documented issue within the radial swim lane that best applies to its level of impact.
4. Develop a plan to correct each issue and integrate the solutions into the following sprint planning sessions.

Scoring Legend

- Individual | The identified issue mainly impacts the direct user or an individual as a result of the ML product / service.
- Cultural | The identified issue mainly impacts specific groups of people as a result of the ML product / service.
- Societal | The identified issue impacts several or more groups as a result of the ML product / service.


Societal
Cultural
Individual
Bias & Fairness
Privacy & Security
Openness & Transparency
Reliable & Trustworthy
Ethical & Safe
Diversity & Inclusivity