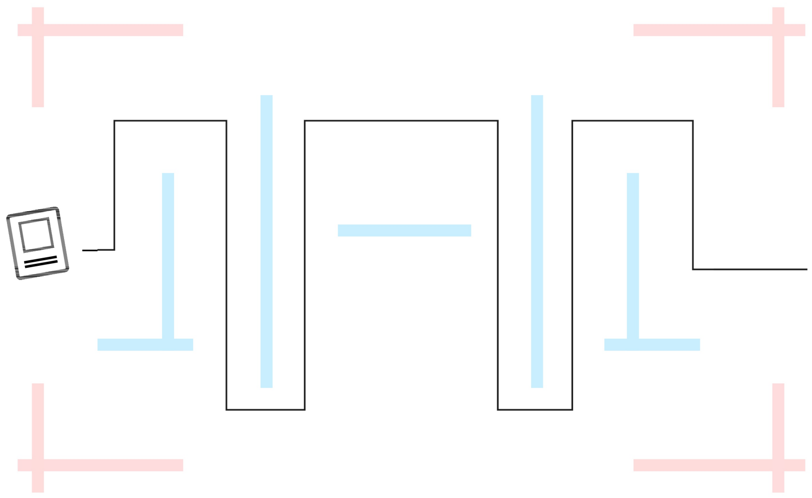
InnovAIting Responsibly

Discovering the Opportunities & Risks of Foundation Models



Master Thesis
Strategic Product Design
Delft University of Technology

Prathamesh Patalay

INNOVAITING RESPONSIBLY
Discovering the Opportunities and Risks of Foundation Models

MASTER THESIS
MSc. Strategic Product design
Faculty of Industrial Design Engineering
Delft University of Technology

Prathamesh Patalay
Delft, August 2023

GRADUATION COMMITTEE

CHAIR
Prof. Dr. Gerd Kortuem
Department of Sustainable Design Engineering

MENTOR
MSc. Jeroen Coelen
Department of Design, Organisation and Strategy

EXTERNAL MENTOR
Jon Ezeiza
CTO, Scitodate B.V.

TU Delft

ACKNOWLEDGEMENTS

The village that helped raise this thesis

I’m extremely grateful to Jon & Petek at Scitodate for their support. Thank you so much for taking out so much time to help me. Jon, this thesis was impossible without your technical expertise. Petek, thank you for all your help to make this thesis more designerly. I would also like to thank Mehdi, Amauri, Joanna, Arthur & Pooya for making me feel at home over the past 5 months.

Gerd & Jeroen, thank you for your wisdom and honesty. Thank you for noticing my blind spots and helping me see past them. Gerd, thank you for having faith in me especially when I needed it the most. Jeroen, thank you for your resourcefulness, and memes.

Super thanks to Ayla, Niya, Kars & Dasha for the incredible conversations about Responsible AI in practice. Your contributions to this thesis are irreplaceable. They were the missing link I needed to ground the responsible innovation aspect of this thesis in practice.

Mahan, Charlotte, Pepijn and Ioannis, thank you for your enthusiastic inputs, and invaluable insights. They helped me find a way forward through uncharted territory.

So many people I can call friends, were there in times of need, in big and small ways. I’ll try to include as many as I can in reverse alphabetical order : Yin, Yallaling, Vaibhav, Shital, Shamir, Sarah, Parshwanath, Meghana, Kalyani, Jayneel, Isha (my sister!), Eren, Carolina, Arjun and Abhijith.

Most importantly, I want to thank my mother. Aaii, thank you for everything.

The giants whose shoulders it stands on

This thesis has benefitted immensely from practitioners like Marty Cagan, Timnit Gebru and Steve Blank, among others, who continue to share their learnings from practice with the broader community. Along with the researchers cited, notable academic mentions include researchers at Stanford HAI for their pioneering work on the concept of Foundation Models, and the HCI Institute at Carnegie Melon University for their research into Human-AI Interaction Design. Everybody at TU Delft who’ve made my past two years transformative also belong here.

The open-source community has contributed incredibly to the latest developments in language models. This thesis also benefits from their selfless efforts to create and share knowledge. The report and all deliverables have been designed in Scribus, an open-source alternative to Adobe InDesign. IBM Plex is the only typeface used for all content. Wikipedia, obviously, has been instrumental in helping me learn about multiple topics.

On a personal level, I’ve had the privilege to learn from Naval Ravikant and Nassim Nicolas Taleb among many others, and their philosophy has helped shape many aspects of this thesis. They helped me tackle the unique challenges of working in a rapidly evolving field of technology.

This thesis tries to be an academic project inspired from the work of practitioners, intended to support practitioners. It is inspired by, and intends to support the contrarian thinkers & bold doers, who continue to change the world against all odds. This is for you.

READING GUIDE

Browsing through this report in 10 mins

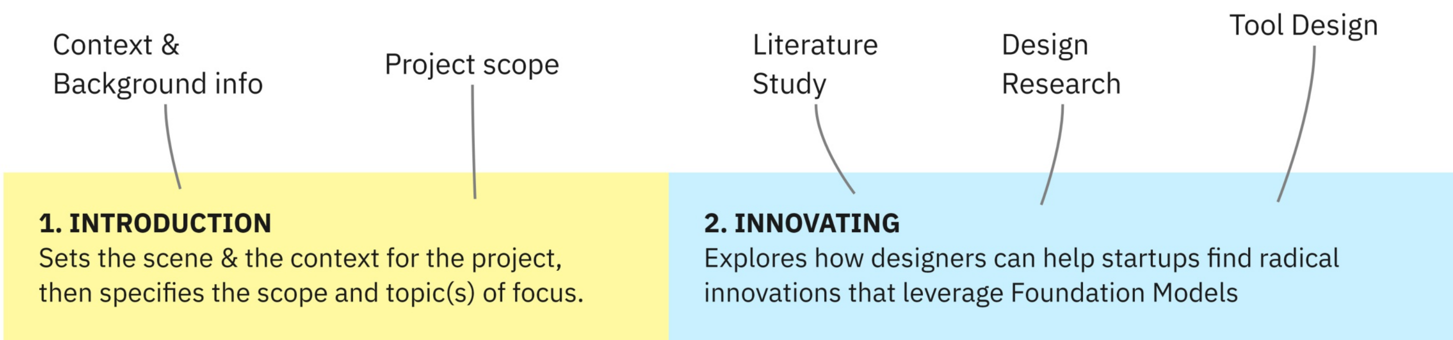
1) Read the abstract (pg. 4)
2) Skim through the visuals of the outcomes of this thesis in Toolkit (pg. 104)
3) If you want an overview of the process I followed, check the summaries at the beginning of every chapter.
4) If you find any chapter particularly interesting, glance over the section titles, subtitles & visuals.
5) At the end of 10 minutes, decide if you want to come back later.

Reading this report effectively

1) Read the abstract (pg. 4) to get a short overview of focus of this thesis.
2) Check the initial Research Questions (pg. 22) and understand how I approached the work (pg. 23).
3) For every chapter, I suggest reading the initial summary. Then glance over the section titles and subtitles. See if you find the visuals or quotes interesting. Then read the rest of the text.
4) Go through Introduction (pg.8) if you are new to the domain of AI startups or Foundation Models.
5) If you just want to find out whether the outcomes of this project are useful for you, jump to Toolkit (pg. 104)
6) To understand the reasoning behind their design, take a look at Innovating (pg. 26) and Responsibly (pg. 66).
7) For reflections and discussion on the work done in the thesis, jump to Discussions (pg. 124)

Report structure

The report is divided into 5 chapters, the topics that these chapters discuss and their key elements are highlighted in this visual



Variation in text formatting

Titles for sections

"Quotes for interesting things people said."

Main body text to convey the bulk of the content and the arguments in this thesis.

> Text inside a colourful box wherever it is important or more insightful than the rest of the text.

`LLM text where I use a Language Model to generate “Low Novelty Text”. Although the text is machine generated, I take responsibility for it’s accuracy and have cited relevant sources. I do not use this style when using LLMs to paraphrase or polish content I wrote, as recommended in the ACL 2023 Policy.(Chairs, 2023).`

Colours

Colours signify a distinction between different parts of this report. The report takes a monochromatic approach to its visual elements and every chapter is assigned its own colour. That colour reflects in the visuals, infographics and graphic elements.

Two of the same colours are also used in the design outcomes. The colours related to Chapter 2 & 3 are also used in other parts where related visuals are used.

The colours are chosen to be colour-blind friendly

Literature Study

Practitioner Interviews

Designing process & tool

3. RESPONSIBLY
Explores what responsibility could mean in this context and what designers can do to help AI startups be responsible.

Deliverables

How to use them

5. TOOLKIT
A short guide on how the proposed design tools work & how to use them

Reflections

Recommendations

4. DISCUSSION
Reflections on the work done during this project & Recommendations for practitioners & researchers

ABSTRACT

Problem

Foundation Models are emerging as a new paradigm in AI research & commercialisation. While this opens up possibilities for radically innovative solutions and significant value creation, startups are challenged with finding unique & differentiated ways to leverage the technology, while simultaneously mitigating potentially negative consequences.

Despite the rising prevalence of machine learning (ML) in products, designers face challenges in creating solutions that make the best use of the new possibilities. Designers building AI products are rarely involved in problem setting & value finding, primarily solving human-AI interaction problems. Additionally, common human-centered, customer feedback based innovation approaches hinder radical innovation.

Lastly, AI Ethics and responsible innovation continues to be an afterthought. Product designers are seldom involved in mitigating the potentially negative consequences of such products.

Approach

This thesis combines literature study with a "research through design" approach to explore ways to address these problems. Working in collaboration with a startup trying to leverage Large Language Models (LLMs) in their products,I use empirical research to engage in the process of finding radically innovative opportunities for using LLMs to create value for customers.

I study risk assessment practices to explore how designers can anticipate risks in the discovery phase of product design. I then design & test tools that can support other designers in the future.

Outcomes

This thesis makes multiple contributions to further the research on technological & responsible innovation. It documents my process for finding & analysing user insights to support the discovery of potentially radical innovations. Additionally, I explore the process of finding value propositions that leverage Foundation Models, and their potential risks, early in the design process. The thesis also records how I design a process for discovering and anticipating potential risks of harm, and how I developed a pair of canvases and card decks to support future designers.

The thesis supports product designers in repeating these processes through the pair of canvases and card decks. These help them to collaborate with engineers, and contribute to the innovation & risk mitigation processes more effectively. The discussions focus on observations & recommendations that can further aid them.

To help designers innovate responsibly, this thesis brings the discovery of opportunities and risks of using Foundation Models into the same conversation. The designed deliverables and processes showcase how both aspects of technology innovation can be tackled in similar ways.

Through all of the above, this thesis showcases the relevance of designerly ways of thinking and doing to the fields of radical innovation, risk management, and Foundation Model based product development.

TABLE OF CONTENTS

Introduction

Let's get started !

Contents

Summary

This chapter introduces Foundation Models as a technology and how it has evolved from developments in AI over multiple decades. It then explores the context of this project, with rapid developments in LLMs and how that has attracted attention from academia as well as industry.

After reflecting on the negative consequences of past technological developments, I introduce Scitodate, the industry client for this graduation project. The chapter concludes with a discussion on the scope of the project and the approach taken while performing the research and design activities.

1.1 FOUNDATION MODELS

Foundation Models are machine learning algorithms that are developed with a focus on versatility, so that they can be used for a variety of downstream applications. That makes them easy to adapt and use in different domains and for different use cases. They become versatile from finding and establishing patterns between parameters within a large amount of varied data, enabling them to predict patterns in a similarly broad space. It is relatively easy to build products that leverage the abilities of Foundation Models because they don't need to be significantly modified to work well for a specific application.

The term "Foundation Models" (Bommasani, 2021) was coined by researchers at Stanford University's Institute for Human-Centered AI (HAI). They went on to setup the Center on Foundation Models (CRFM), that focuses specifically on research & development on this technological paradigm.

Why are they a big deal ?

"A foundation model is any model that is trained on broad data that can be adapted to a wide range of downstream tasks"

What makes Foundation Models interesting is the ease with which they can be adapted to different usecases and applications. That makes extremely capable AI systems easily accessible to teams who do not have the expertise or resources required to develop such technology on their own. Theyse teams can now directly take these pre-trained models and use them with a drastically low amount of adaptation in their own products.

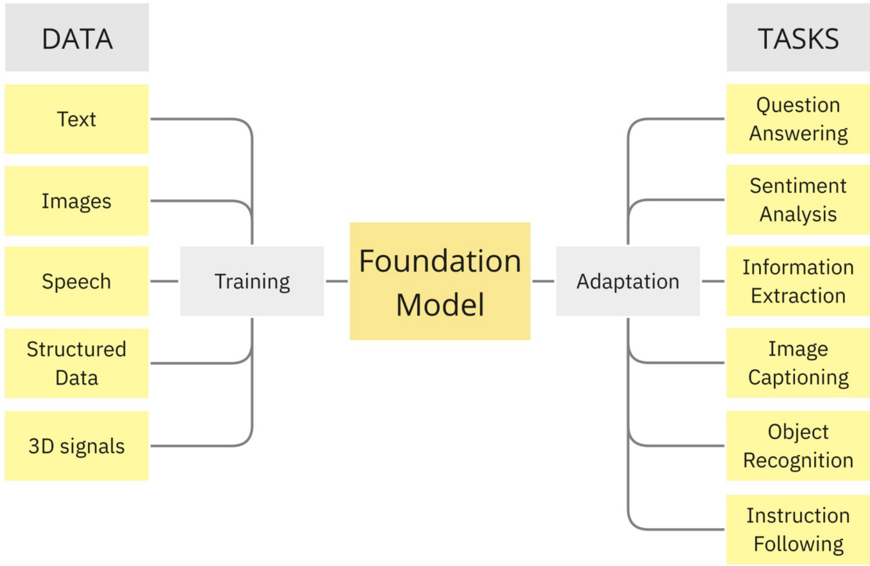


Fig. 1 : Foundation Models can be used for multiple tasks. Adapted from Bommasani (2021)

Evolution of Machine Learning

Machine learning is a branch of artificial intelligence that focuses on developing algorithms and statistical models to enable computer systems to learn and make predictions or decisions without being explicitly programmed. It involves the use of algorithms to analyze and learn from large amounts of data, enabling computers to identify patterns, make predictions, and improve performance over time.

The scientific origins of machine learning can be traced back to the development of neural networks, which were inspired by the structure and function of the human brain. Neural networks are interconnected layers of artificial neurons that can process and learn from data. They were initially developed in the 1950s and 1960s but faced limitations due to the lack of computational power and large datasets.

However, a significant breakthrough occurred in 2012 with the introduction of the ImageNet dataset (Deng et al. 2009) and the AlexNet (Krizhevsky et al., 2012) architecture. The ImageNet dataset consisted of millions of labeled images, which served as a benchmark for training and evaluating computer vision algorithms. AlexNet, a deep convolutional neural network, outperformed all previous methods by a large margin, demonstrating the potential of deep learning in computer vision.

Following the success of AlexNet, there were rapid developments in machine learning based computer vision. Researchers began exploring deeper and more complex neural network architectures, which further improved the accuracy of image recognition tasks. These advancements led to significant progress in various computer vision applications, including object detection, image segmentation, and image generation. That was partly due to capable "pre-trained" models that were already trained at general image classification, and could be further trained or tuned for other applications.

The success of pre-trained models in computer vision prompted similar developments in the field of Natural Language Processing (NLP). Language models such as GPT (Generative Pre-trained Transformer) emerged, leveraging large-scale pre-training on vast amounts of text data to achieve impressive results in tasks such as language generation, sentiment analysis, etc. (Radford & Narsimhan, 2018)

Based on the trends observed in computer vision and NLP, scholars hypothesised the rise of versatile, pre-trained machine learning models that can handle different types of input/output media and be applicable across various domains. These models would leverage the power of deep learning and large-scale pre-training to generalize and adapt to different downstream tasks and data types, thereby reducing the need for extensive task-specific training. We've already witnessed this in the fields of Computer Vision and Natural Language Processing.

The team at Stanford HAI coined these versatile, reusable ML models as Foundation Models.

How ML works in simple english

Here's ChatGPT explaining Machine Learning with "an" example :

```
Imagine you have a clever friend who loves to read books. They've
read many sentences and noticed patterns in how words come after
other words. Now, you give them a sentence that starts with "an"
and ask them to guess the next word.

To help them guess, they remember all the times they've seen the
word "an" at the beginning of sentences. They notice that quite
often, words like "apple," "animal," and "amazing" come after
"an." They also see that words like "the," "and," and "is" don't
appear much after "an". So, based on all these observations, your
friend makes an educated guess. They might say something like "an
apple" or "an animal," because those are common choices. They're
using their learning from past examples to predict the most
likely next word.

Machine learning works in a similar way. Instead of a friend, we
use a computer program. Instead of reading books, the program
looks at a lot of data. It searches for patterns and
relationships between words. When you ask the program to predict
the next word after "an," it looks at all the words that came
after "an" before and makes a guess based on which words appeared
more often.

Just like your friend gets better at guessing with more sentences
they read, the machine learning program gets better with more
data it analyzes. It can learn to make smarter predictions over
time. This is how machine learning helps computers do tasks like
predicting words in sentences, recognizing images, or making
decisions based on patterns it finds in data.
```

It's important to note that this algorithm does not know *why* we use an "an" before words that start with vowels. Depending on the training data, it might not know that "an" can also come before hour, honest, LLM, etc. The algorithm functions based on correlations between data points, not cause-effect relationships.

How Foundation Models work

When machine learning algorithms are trained on a very large amount of data, with a complex algorithm that can find patterns between multiple characters in a large text, they become capable of finishing incomplete sentences. With the correct adjustments to the algorithms, they become capable of generating answers to questions as inputs. If the data contained two languages with examples of translation between them, the algorithm could become capable of translating input in one language into output in the other language.

When such a model is allowed to find patterns in data from a variety of sources to form a large dataset, they become extremely useful for a variety of tasks. For example, GPT-3, the first Large Language Model (LLM) that powered ChatGPT, was trained on 45 TB of text data, with a model size of 175 billion parameters. Once an LLM like this becomes good at common language tasks, it can be used for a variety of downstream applications that need understanding and generation of text. These can vary from writing emails to summarizing blogs, etc. A similar behaviour has already been observed with pre-trained computer vision models, and similar ML models are expected to emerge in other media and specific domains. (Bommasani, 2021)

It is important to note that these algorithms are still statistical prediction machines that rely on pattern matching and correlation. At the time of this writing, we still do not know if these LLM algorithms can reason like humans to get to an answer. (Huang, 2022) They can only predict the most likely output to an input, but with a very high level of accuracy. Their high accuracy only serves as a proxy for understanding. Despite this limitation, these systems are still quite useful in many applications.

1.2 PROJECT CONTEXT

"There are decades where nothing happens; and there are weeks where decades happen"

- Vladimir Lenin

This graduation thesis was undertaken through early-mid 2023. The project brief was discussed and finalised in December 2022. This was just a few weeks after the launch of ChatGPT, OpenAI's conversational A.I. product. While the underlying technology that powered it was present for the past few years, ChatGPT exhibited how useful it could be and how capable is already was.

That led to large and small companies, academia, and open-source communities going all out on experimenting with this technology. Many people also raised concerns about the challenges and potential negative consequences of releasing this technology into society without sufficient evaluation and safety guardrails.

1.2.1 Large Language Models come of age

Large Language Models (LLMs) have been around for quite some time, but they've mostly not been very reliable or directly useful. However, with the release of ChatGPT, that changed. ChatGPT enabled people to directly extract value from the technology, and the product showcased that the technology is potentially ready for mainstream adoption. It marked a turning point where LLMs became more accessible and demonstrated their practical applications.

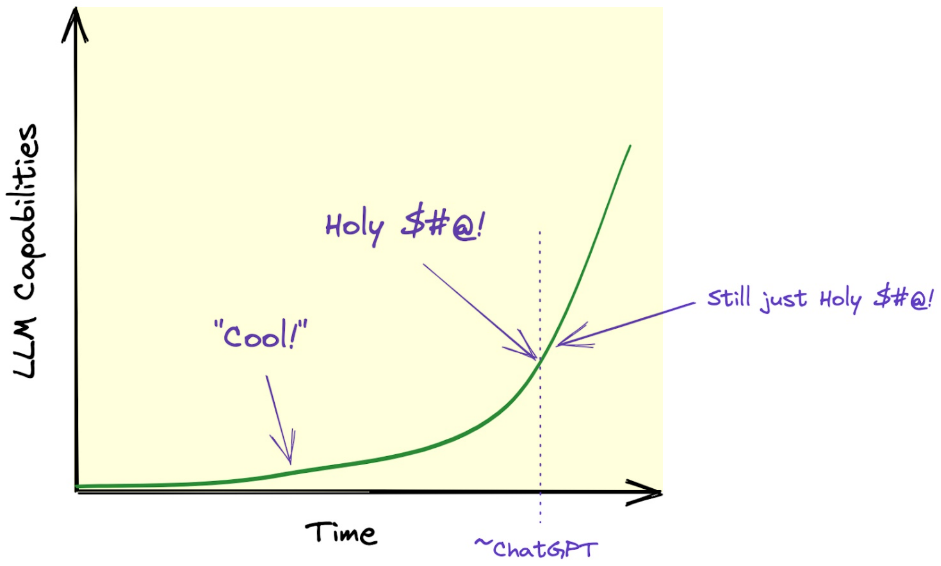


Fig. 2 :How the industry experienced the improvements in LLMs (Hershey & Oppenheimer, 2023)

The state of LLMs before ChatGPT

Language models are based on the idea of using neural networks to understand and generate human-like text. This concept became feasible in the 2010s when powerful GPUs (Graphics Processing Units) enabled efficient parallel processing. Previously, in the 1950s, the lack of such technology prevented these developments in language learning. After the success of large neural networks at image recognition with AlexNet in 2012 (Krizhevsky et al., 2012), researchers started applying them to other domains. Natural Language Processing (NLP) was one of them.

In 2017, researchers at Google Research introduced the Transformer architecture (Vaswani, 2017), which could process input sequences in parallel, allowing the training and use of much larger models. In terms of language processing ability, it could track where some word or phrase appeared in a sentence.

The Transformer architecture laid the foundation for language models like BERT (Bidirectional Encoder Representations from Transformers)(Devlin, 2018) from Meta & GPT (Generative Pretrained Transformer)(Radford & Narsimhan, 2018) from OpenAI. Over time, GPT models evolved and improved to provide advanced conversational abilities. Going from GPT-1 in 2018, to GPT-2 in 2019(Radford et al. 2019) and then GPT-3 in 2020 (Brown, 2020), OpenAI's LLMs kept getting larger and more capable.

With an improved model called InstructGPT, (Ouyang et al., 2022) GPT-3 was tuned to match human preferences, making the model better at having conversations and reducing toxic output. That led to the creation of ChatGPT. Other tech companies also took notice of GPT-3 and it capabilities. This led to an acceleration in the development of LLMs, with many teams building their own.

While the term Large Language Models has itself emerged around 2018, it gained visibility in 2019 and 2020, with the release of DistilBERT (Sanh, 2019) and Stochastic Parrots (Bender et al., 2019) papers respectively. Both focused on the "Large-scale pretrained models", citing the BERT family as an example of LLMs.

From 2018 to 2020, the usual approach to using LLMs like these involved fine-tuning the model with task-specific training for a particular task. For example, using an LLM for understanding a specific language would need the pre-trained LLM to be further trained on data in that specific language. However, it was later found that LLMs like GPT-3 can tackle different tasks without needing specific training for each one. Instead, they can be guided by providing a prompt that includes a few examples of similar problems and their corresponding solutions. That opened up the possibility for smaller teams & individuals to use LLMs for different downstream applications with very little effort.

LLM developments after ChatGPT

After the release of ChatGPT, not only did people in the Artificial Intelligence, computer science and broader scientific domain take note, but everyone else also saw the capabilities & relevance of the breakthrough LLM product. Seeing the value that ChatGPT could create for a large variety of people, industry, academia and the open-source community took notice. Access to the research behind the underlying technology, open-source models, and OpenAI's GPT-3 & GPT-4 API (Application Programming Interface) led to an explosive growth in research and development.

BIG TECH COMPANIES

Google, Facebook & Meta had already been working on Language models over the past few years, with many significant developments coming out from them. OpenAI, with Microsoft's support continued to improve the LLMs underlying ChatGPT and released GPT-4, their updated SOTA (State of the art) model with significantly better capabilities than the initial ChatGPT release.

Bubeck et. al. (2023) showcased the remarkable capabilities of GPT-4 that extended beyond its proficiency in language. They demonstrated its ability to solve complex and unprecedented challenges across various domains such as mathematics, coding, vision, medicine, law, psychology, and more, without requiring any specific instructions. They observed that GPT-4's performance in these tasks is remarkably close to that of humans and frequently exceeded the capabilities of previous models like ChatGPT.

In February 2023, Meta released a family of LLMs called LLaMA (Large Language Model Meta AI) in a range of sizes (Touvron et al., 2023). Since their release to the academic community (and subsequent torrent leak that made it accessible to everyone), LLaMA models gained a lot of attention from both researchers and the open-source community. Many researchers worked on extending these models by making further modifications or giving them specific instructions in the inputs. (Zhao et al., 2023)

ACADEMIA

Universities also continued to make significant contributions to the development of LLMs. Stanford's Center for Research on Foundation Models (CRFM) released Stanford Alpaca 7B, an "instruction tuned"(Wang, 2022) version of LLaMA 7B. Similar to how OpenAI tuned GPT-3 to work better in a conversational context to create, ChatGPT, Alpaca 7B improves upon LLaMA's performance to make it better suited for human interaction.

Building upon the progress from Alpaca, The Sky Computing Lab at UCBerkeley released Vicuna (LMSYS Org.,2023) improving performance over the base LLaMA model and Alpaca. Their performance is comparable to that of Bard & ChatGPT, despite being significantly smaller and cheaper to train & run. Later in May, they released Gorilla (Patil et al.,2023), "a finetuned LLaMA-based model that surpasses the performance of GPT-4 on writing API calls."

OPEN-SOURCE COMMUNITY

The open-source community around the world also joined the party, and contributed significant breakthroughs to the discussion. A week after LLaMA was released to academic researchers, it got leaked to the rest of the Internet. A week after the leak, Georgi Gerganov created llama.cpp, a downsized C/C++ version that could run an a Macbook M1 computer. (Ggerganov, 2023) The next day, Artem Andreenko got LLaMA-7B to run slowly on a Raspberry Pi 4 single board computer. (Artem Andreenko on X) Four days later, Gerganov, who first wrote llama.cpp got the same model to run locally on a Google Pixel 5 smartphone. (Georgi Gerganov on X, 2001)

STARTUPS

All this progress led many other organisations building and releasing their own language models. An example is Databricks, an enterprise software company, releasing Dolly (Databricks), an open-source LLM, available for commercial use. Such open-source LLMs, along with the OpenAI's APIs to its language models, have made it significantly more feasible for startups to build their own tools and products. The next section expands on this.

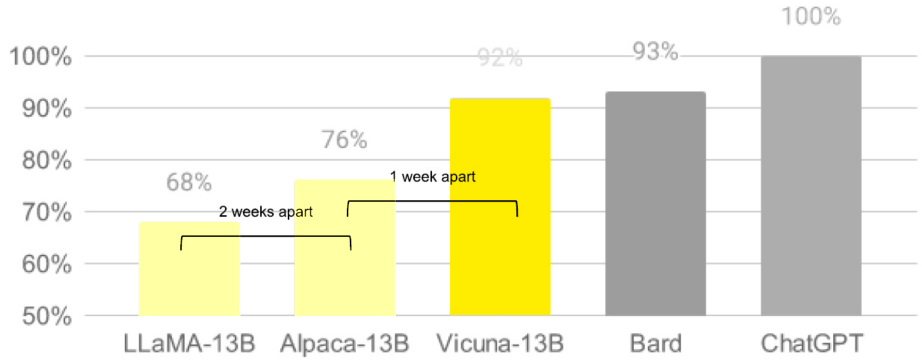


Fig. 3 : Open-source LLMs as compared to those from private firms (LMSYS Org.,2023)

Fig. 4 : Meme about startups using OpenAI's GPT-3 API (Sofiia Shvetson X, 2023)

1.2.2 The Gold Rush

ChatGPT was the fastest growing consumer software application to date, reaching 1 million users in 5 days (Greg Brockman on X, 2022), and 100 million users in 2 months (Hu, 2023) after launch. This viral consumer adoption of ChatGPT and the tangible added value that it created for users made everyone else accelerate their AI development efforts.

Over the first half of 2023, the big tech corporations have been scrabbling to integrate language models into their solutions and be the first to introduce their products to the market and capture the created value.(Bing, 2023) Along with that, startups are also building products that leverage this technology. But while many new projects and companies will emerge in the near future, few of them will survive in the longer run.

Investment Frenzy

Investments in AI startups, especially in the generative AI space had been picking up steam in late 2022, due to the progress from image generation startups like Stability AI, and OpenAI's GPT-3 LLM. After ChatGPT took off, things got even more interesting. Generative AI including text, image, sound and video, started getting hyped by investors as well as tech companies :

"We are at the beginning of a platform shift in technology. We have already made a number of investments in this landscape and are galvanized by the ambitious founders building in this space."

"Absolutely, the hype is high. I think it's absolutely justified given the results that we're seeing."

- Sonya Huang, Sequoia Capital (Huang and Grady, 2022)

"There are a few times in technology where you really see a generational leap forward with revolutionary technology, these companies are the next trillion-dollar opportunities in software."

- John Somorjai, leads Salesforce's VC (Griffith & Metz, 2023)

"Whatever the case, one thing we're certain about is that generative AI changes the game. We're all learning the rules in real time, there is a tremendous amount of value that will be unlocked, and the tech landscape is going to look much, much different as a result. And we're here for it!"

- Andreessen Horowitz (Bornstein et al., 2023)

"There's a company called Essential AI. It was founded by two former Google AI researchers... Before that company even had a name or a business plan or a way to generate revenue, essentially, venture capitalists were hounding these two founders trying to lodge an early investment in their company."

- Berber Jin, Startups Reporter (Thomas, 2023)

"We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run"

- Roy Amara, scientist & futurist (2006)

Startups need to find success

"I think one thing that's important to note is that there hasn't actually been a clear path to success that any of these startups have proven. So I think there are a lot of question marks around who is going to capture a lot of the value around this technology, and that's the big risk that these VCs are taking by paying these high prices to get into these very young and unproven companies."

- Berber Jin, Wall Street Journal (Thomas, 2023)

Many existing startups with products in the market started exploring how they could integrate LLMs into their offerings. Many others started building LLM powered products from scratch. A majority of them though, were only integrating OpenAI's GPT-3 API into their products. Through an API (application programming interface), a software program can interact with and leverage the functionality of another software prorgam.

That meant that the technology around LLMs wasn't unique for them and anyone could copy it. Apart from that, Incremental product improvements might not be sufficient to ensure a competitive edge, as other startups & incumbents are also capable of developing them, with better resources and distribution/market presence.

It isn't just startups that are facing this challenge of differentiating their product offerings and building unique long term advantage. A leaked internal Google document (Patel & Ahmad, 2023) claimed that developments from the open source community will eventually outcompete Google and OpenAI. The writer claimed that Google shouldn't be worried about OpenAI as their competitor.

"Open-source models are faster, more customizable, more private, and pound-for-pound more capable."

That claim serves as an opportunity for startups, in the sense that the most capable LLM technology is openly available for anyone who wants to use it. But, it also serves as a challenge, making it harder for them than tech giants like Google or OpenAI to find a powerful long-term competitive advantage.

Beyond Exponential growth

Due to this continued increase in interest from academia, industry and the broader open-source community, the rate of progress in this field keeps increasing. That attracts more interest from the rest. This thesis itself is an example to this phenomenon.

Progress in AI is propelled by three main factors: innovative algorithms, various types of data (including supervised data or interactive environments), and the computational resources dedicated to training.(Open AI, 2018) Improvements in all 3 will affect the cumulative improvement in the broader domain.

AI computing hardware performance has been observed to double every ~2.5 yrs, slightly slow than Moore's Law, but still following an exponential curve (Hobbhahn & Besiroglu, 2022) There has been an exponential rise in published academic research and capital invested into the AI domain. (Giattino et al., 2022) Increased investment in AI continues to increase the hardware resources available to train AI systems beyond Moore's Law (Unit performance X quantity of processing units) (Sevilla et al., 2022). Increasing academic research and R&D talent in industry contribute to improved software algorithms.

In “The AI Dilemma”,(2023) Tristan Harris & Aza Raskin expand on how progress in AI is accelerating beyond the rate of an exponential curve. As multiple factors that contribute to AI advancement, like data, hardware and algorithm efficiency continue progressing at an exponential rate, that leads the technology itself to develop over a double exponential curve.

They go on to propose 3 Rules of Humane Technology :

1) When we invent a new technology, we uncover a new class of responsibility.
2) If that new technology confers power, it will start a race.
3) If we don’t coordinate, the race will end in tragedy.

The rules reflect that this rate of progress in technology & its adoption makes accidents & unintended consequences very likely.

1.2.3 Collateral Damage

Murphy’s Law does not care about good intentions.

Damage from past technologies

Previous technological developments & breakthroughs have led to unintended negative consequences. Coad et. al.(2020) provide a broad discussion about the negative consequences of technology, with a list of indicators that show a steady deterioration of human life & progress as a result of innovation:

Global issues stemming from this pose significant challenges. Air pollution continues to burden communities worldwide, affecting the health and well-being of countless individuals. Insect populations are collapsing at an alarming rate, disrupting ecosystems and biodiversity. The pervasiveness of fluorinated chemicals in our environment and the accumulation of plastic waste in our oceans are further sources of concern.

Human-induced disasters, such as the Chernobyl and Fukushima nuclear accidents, have left significant portions of the Earth's surface desolate and uninhabitable. Additionally, polluted battlefields like Verdun in France, the Union Carbide disaster site in Bhopal, and areas with contaminated groundwater pose ongoing threats.

SOCIAL MEDIA

Mirroring the concerns of The Center for Humane Technology (www.humanetech.com), Coad et al.(2020) also mention the risks of harm from social media. While social media has a significant impact on various aspects of our lives today, an important area it affects is our self-perceptions and social interactions. People's perceptions of themselves can be influenced by the images and narratives they encounter on social media platforms.

Moreover, social media has the potential to threaten democracy, especially during elections. The spread of micro-targeted fake news can manipulate public opinion and disrupt the democratic process. It also provides a platform for nations to interfere in the politics of other nations, creating further challenges to democratic governance.

The business model employed by many social media platforms is driven by "click-bait," which favors content that polarizes and provokes anger. This approach undermines the necessary conditions for democratic deliberation, which require respectful and balanced consideration of different perspectives.

Amnesty International published a report (Amnesty International, 2022) showcasing the role of social media platform Facebook in the violent atrocities perpetrated against Rohingya Muslims in Myanmar.

The actions of the Myanmar security forces and radical Nationalist groups, including using Facebook to spread disinformation and hatred, led to mass violence resulting in the unlawful death of thousands and displacement into Bangladesh of over 700,000 Rohingya Muslims. These were linked to recommendation systems on the platform, which are often based on machine learning algorithms (Portugal et al., 2018) :

> *“We have evidence from a variety of sources that hate speech, divisive political speech, and misinformation on Facebook and the family of apps are affecting societies around the world. We also have compelling evidence that our core product mechanics, such as virality, recommendations, and optimizing for engagement, are a significant part of why these types of speech flourish on the platform.”*
>
> - A Facebook Paper, titled “What is Collateral Damage?” (Amnesty International, 2022)

Current state of LLM riskS

Coming to LLMs, Timnit Gebru with Emily Bender discussed the potential risks and challenges of making Language Models “Large” in a paper they wrote in 2020 while Gebru was co-leading Google’s ethical AI team. (Bender et al., 2019) Internal conflicts within the Google AI team arising from this paper led to Gebru resigning from her position at Google. In March 2023, Microsoft fired it’s Responsible AI team, while it continued to invest into OpenAI and its own product development efforts.

Microsoft’s 2016 Twitter chatbot Tay malfunctioned after interacting with people on Twitter, resulting in extremist, racist & sexist comments and tweets. (Vincent, 2016) Microsoft’s new Bing search with chat has struggled with similar hallucinating outcomes. (Roose, 2023)

Snap (previously Snapchat), a social media platform targeted specifically towards youngsters, released its own AI chatbot in April and people were able to elicit harmful output from it while pretending to be a13 yr old girl. (Tristan Harris on X, 2023)

Call to action

With Foundation Models, its important to ensure that as the technology becomes more accessible to a large group of enthusiastic innovators, they also become capable of addressing the potential harms of using this technology, and that they do take responsibility for the consequences of their work.

“It feels like a gold rush. In fact, it is a gold rush. And a lot of the people who are making money are not the people actually in the midst of it. But it’s humans who decide whether all this should be done or not. We should remember that we have the agency to do that.”

- Timnit Gebru, May 2023 (Harris, 2023)

“Hope here is not enough . . AI is moving incredibly fast, with lots of potential — but also lots of risks. We have unprecedented opportunities here, but we are also facing a perfect storm, of corporate irresponsibility, widespread deployment, lack of adequate regulation, and inherent unreliability . . The choices we make now will have lasting effects, for decades, even centuries.”

- Prof. Gary Marcus, U.S. Senate hearing, May 2023 (G. Marcus, 2023)

1.2.4 Scitodate

Scitodate develops software tools to help manufacturers of scientific instruments find the right data & insights about their target markets and customers. The tools give businesses access to vast scientific data points and benchmarks, helping them create personalized sales campaigns and find potential customers and business opportunities.

Scitodate primarily offers 2 products to their customers :

1) Market Landscape is tailored for science industry marketers and product teams, It helps them define precise market segments by product attributes. The platform enhances campaign efficacy by offering trend insights. Using an large database of scientific data, it identifies potential clients and partners, enabling personalized engagement and research updates. This tool empowers scientific marketing, supporting campaign success through market understanding.

2) Intelliscope is designed for science sales. It leverages AI to help customers forge connections with scientists by uncovering shared projects, offering funding insights, and enabling personalized outreach. It helps streamline sales, providing granular data for effective engagement and efficient lead qualification.

Interest in LLMS

Many of the problems that Scitodate's customers struggle with could probably be addressed using generative LLM solutions like ChatGPT. After ChatGPT was released and they witnessed the performance of language models, the team started seriously exploring integrating the technology into their existing products.

Multiple aspects of how their existing products worked at a technological level were also significantly relevant to the technology behind language models (embeddings, etc.)

Motivation for this thesis

While hundreds of "GPT-powered" tools were getting built across different industries, Scitodate wanted to explore the potential for finding long term value in the technology and using it to tackle specific customer problems while leveraging their existing unique value propositions and strengths.

Considering their customers' expectation for products that function as intended and not causing them harm in anyway, both aspects of the graduation brief were relevant to the team.

During the span of this thesis, Scitodate introduced a new product that tried to leverage the rapid developments in LLMs. Many of the outcomes of this thesis were able to contribute to this new product during its initial development and evolution.

The product is called MirrorThink. It is powered by GPT-4 and integrated with reliable sources of information like Wolfram and Pubmed, helping accelerate scientific research by being able to address intricate queries and offering tools for academic exploration, mathematical computations, and market trend identification.It aims to serve as an efficient research ally, providing scientists with timely insights and dependable data.

Role of Scitodate

This thesis aims to generate generalisable insights, recommendations and design outcomes while rooting them to an empirical context and supporting it with academic literature.

Scitodate serves as a case study for performing the required design activities and using that process to create knowledge that can be valuable to the broader design community working in the field of AI. The findings from the process will be relevant and hopefully valuable to the company too.

1.3 SCOPE

Startups

This thesis focusses primarily on the challenges of new entrepreneurial firms, colloquially called startups. For the context of this thesis, I would like to use this definition of a startup from Steve Blank (2010)

"A startup is an organization formed to search for a repeatable and scalable business model."

A business model consists of how a company can create, deliver & capture value. A startup is then an experiment to find out what that business model can be, starting with a hypothesis and either validating it or invalidating it to form and test a new hypothesis. The process of "starting up" then becomes a process of finding the hypothesis that ends up being correct. This often takes the form of a product or service that creates value for a market of customers, that the startup can deliver to them and capture a part of the created value in the process. For products or services that leverage some technology, the process of technological innovation, then becomes a process of finding out what that product or service could be. In this way, innovation forms an integral part of entrepreneurship.

WHY STARTUPS

Startups have been the source of a large fraction of past innovations, especially those that leverage technology in a unique and novel way. Thus, it made a lot of sense to focus on the startup context of technology innovation for thethesis. I intended that the outcomes help other startups in the future. Another underlying assumption here was that design outcomes that work for a startup will very likely work for corporate innovation teams too, whereas the opposite might be less likely. These assumptions later nudged me to propose tools that are generalisable & versatile.

The unique opportunities & the accompanying risks that Foundation Models unlock are most relevant for startups : they opens up opportunities for startups that do no have the resources to train ML models from scratch. At the same time, these startups also often do not have the resources to manage the risks of such ML systems, unlike larger tech companies. Foundation Models enable startups to build powerful ML applications and also challenge them to take responsibility for mitigating potential harm

Large Language Models

As Scitodate only finds Language Models (and not other Foundation Models) currently relevant for its value propositions and customer problems, the majority of work done at Scitodate focuses purely on Language Models.

While that is the case, the synthesis from the findings and the final design outcomes are designed with the intent of being relevant to other types of Foundation Models as well. During the course of the thesis, I've tried to get feedback on the proposal from people working beyond LLMs, to try my best to ensure broader relevance.

For AI product designers

The thesis focusses primary on the role of designers in this context of technological innovation at startups. And the role designers can play in the current context of adoption of Foundation Models. While designers can contribute to the product design-development-deployment lifecycle in multiple ways, this thesis explores how they can help startups find ways to leverage Foundation Models most effectively and do that responsibly.

Fuzzy front end of innovation

This thesis focuses primarily on parts of the fuzzy front end of the innovation process, (Reid & De Brentani, 2004) the initial steps towards finding ways to leverage emerging technology into solutions. The reason to focus on this early stage of the innovation process was the challenges that startups currently face with using Foundation Models : Finding a long-term unique value proposition and ensuring that the technology does not lead to unintended negative consequences.

My hypothesis was that the process of finding the right solution to design & later develop might have the most significant impact towards addressing both of those challenges. Considering that the later stages of the product design process for designing a solution often starts with a set of design requirements, I assumed that the process of defining those design requirements holds great promise.

Marty Cagan (2017b) refers to this as Product Discovery, the process of finding the right solution to build. He breaks it down into problem discovery & solution discovery. He refers to them as finding the right problem to solve and finding the right solution to build. He then contrasts product discovery with product delivery, that focuses primarily on design & development of the proposed solution. For a majority of this thesis, I refer to Product Discovery as my scope.

Product Discovery helps teams find the most promising solution to build. Product Delivery is about building the solution in the most effective and efficient way. The difference between Discovery & Delivery is that of processes that help teams “Building the right thing” (finding out what is the right solution) as compared to “Building the thing right” (developing the solution correctly).

Comparing this framework with the Double Diamond of discover-define-develop-deliver in the user-centered design process (Design Council), problem discovery seems to align the most with the discovery stage and solution discovery aligns with the develop stage.

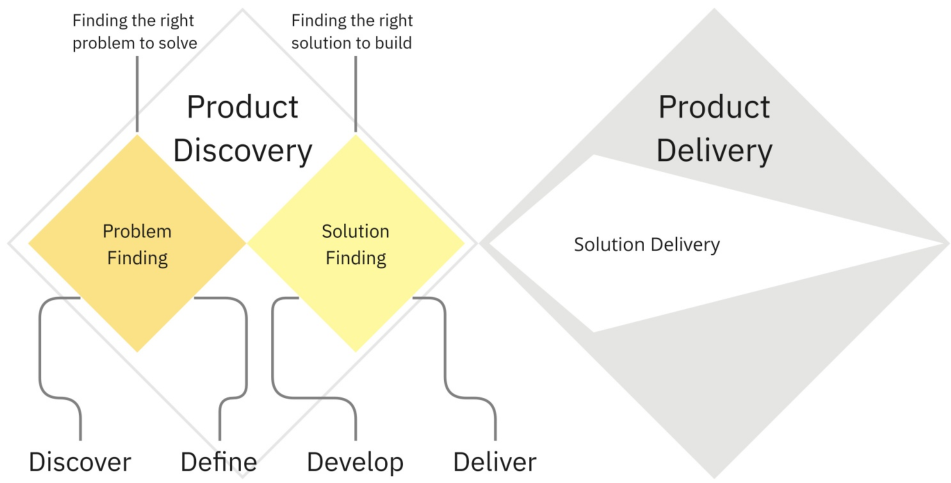


Fig. 5 : Showcasing the Double diamond design process along with Product Discovery

This thesis explores how designers can contribute to Problem Discovery. Referring to the Double diamond, I focus on the discover & define stages of the design process.

WHY DISCOVERY

The Product discovery process is where innovative product ideas will differ in how novel, incremental and valuable they are. That’s because the delivery phase purely consists of building the product proposal which is the outcome of the discovery phase. Thus for technological innovation, especially when trying to create value using emerging technologies, Product discovery becomes extremely valuable, and designers can make a significant contribution to this discovery process using design research practices.

Within Product Discovery, the Discover & Define stages of the double diamond will be the focus in the thesis.

1.4 APPROACH

Research Questions

The context and the scope of the thesis led me to frame 2 research questions that guided my initial enquiry into the topic. As stated previously, the goal is to find how designers can contribute to responsible technological innovation, in the context of Foundations Models. I focus primarily on their role at startups. I break down this broader question into finding opportunities to innovate, and finding ways to mitigate harm.

How can designers help startups :

RQ1: Find innovative solutions that leverage Foundation Models

RQ2: Mitigate the negative consequences of using Foundation Models

Target Outcomes

Apart from contributing to Scitodate's journey of integrating LLMs into their product offerings, the intent of this thesis is to create outcomes that benefit the broader design community, beyond the designers involved in this specific project. One potential way of achieving that goal is to create a proverbial map by documenting and reflecting on this journey. Designers who want to take the same journey in the future can use it as a reference.

This map can take the form of a design artefact and it can take the form of knowledge, as also discussed by Zimmerman et al. (2007). While exploring and trying to find the answers to the primary research questions, I kept an eye out to find out the following :

1) What can I design as part of the project to support designers in this process ?
2) What can the design community learn from the outcomes and findings of this thesis ?

General Methodology

To answer both research questions and acheive the target outcomes, I followed a "Research through Design" (RtD) approach of design research (Frayling, 1994) to structure the design research & design activities throughout the course of the thesis, and find useful insights through the design practise. This approach uses design activities, along with designed artefacts, as the chief elements in the process of generating and communicating knowledge.

The primary intent of such a process is generate knowledge, not work to support the development of a commercial product. (Zimmerman et al., 2007) I've supported my observations and design decisions with academic literature and insights from practitioners working on those topics. I later reflect on the process I followed and my observations to help future design practitioners benefit from it.

The process of working with Scitodate serves as a means towards the end of generating knowledge about that same process. I use the guidelines from Zimmerman et. al.(2007) to try to achieve sufficient rigor to the RtD process:

1) Process : Have a rigorous rationale for the methods selected in the process, along with documenting the process in sufficient detail, so that the process employed can be repeated by other designers.

2) Invention : Demonstrate how the findings advance the current SoTA of design research in the research community and how the produced outcome is a novel integration of various subject matters. Have an extensive literature review and sufficiently detailed articulation of the invention.

3) Relevance : Articulate the preferred state the design process aims to acheive. Support why the design community should consider this outcome preferable.

4) Extensibility : Ensure that the design community can leverage and build on top of the findings from this thesis by describing & documenting the research appropriately.

WHY RESEARCH THROUGH DESIGN ?

The Research through Design approach enabled me to pursue 2 different goals through the same design project :

1) Explore how Scitodate could innovate responsibly
2) Creating knowledge & artefacts about the process

Research Approach

For RQ1, the focus is more on the empirical research at Scitodate. Literature was primarily used to drive the initial exploration of the research question and serve as a starting point for the research activities at Scitodate and designing the artefacts.

For RQ2, I started with exploring the literature to understand how organisations have tried to develop AI solutions responsibly. I complimented that with interviews of practitioners working in the field of Responsible AI. I used this research to propose an effective way for designers to answer RQ2 in the context of this thesis. I then supported those insights with researching the current methods of risk management to propose a new risk discovery process. I then design tools to help designers in this process.

I followed a triple diamond approach to answer both research questions. That meant focussing on Problem Finding, then Solution Finding, and finally Solution Development with multiple iterations along the way. There was plenty of overlap & back-and-forth between the three stages during those iterations.

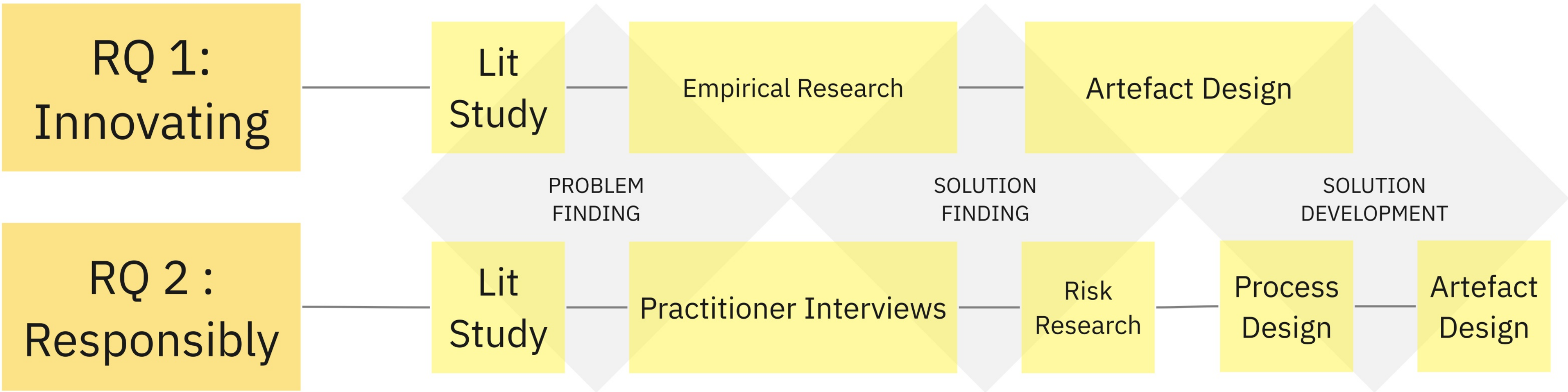


Fig. 6 : Triple Diamond process for answering both research questions

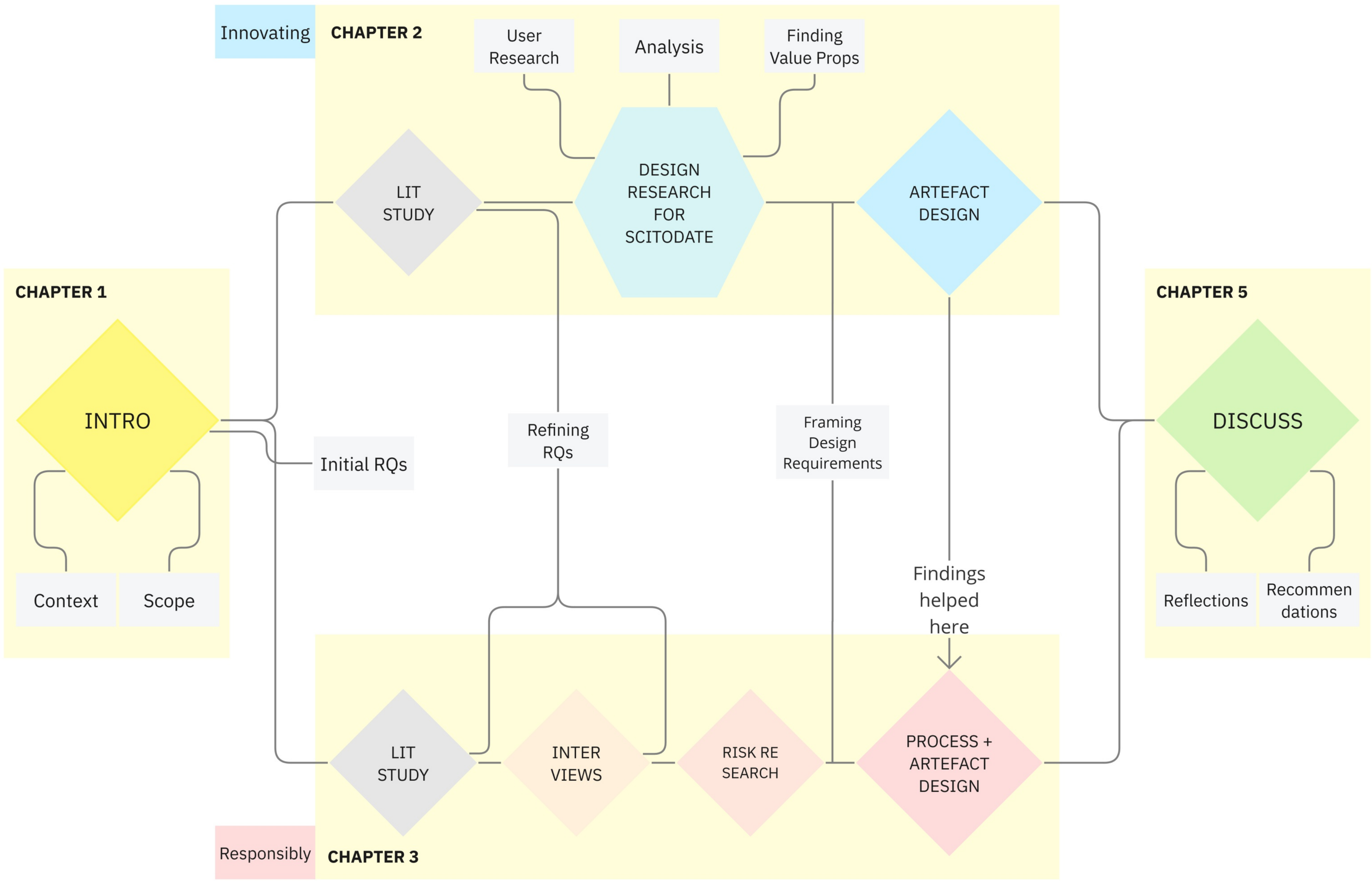


Fig. 7 : Overview of the approach for the entire thesis

Innovating

How can designers help startups find innovative solutions that leverage Foundation Models ?

Contents

Summary

In this chapter, I used the literature review process to build an understanding of technological innovation and how designers can contribute to radical innovation and ML product design. That served as a starting point for the design research at Scitodate. Through design research, I build customer understanding that can help in value finding.

The user insights are combined with an understanding of the capabilities of LLMs to find how to leverage LLMs and propose value propositions that address customer challenges. I use the insights from this process to design a canvas accompanied by a card deck to support designers who want to design products using Foundation Models.

2.1 LITERATURE STUDY

2.1.1 Technological Innovation

Innovations frequently arise as a result of technological advancements such as scientific progress, research, and development, rather than being deliberately created to meet a specific market need. (Nemet, 2009) Examples of these innovations include digital photography, the internet, magnetic resonance imaging, and consumer global positioning systems. They are built upon the growth of scientific knowledge, with market application and development considered later on.

Foundation Models like Large Language Models (LLMs), Computer Vision (CV) Models, etc. are similar technological advancements that have led and continue to lead to innovative solutions. (Rishi Bommasani et. al. 2022) While LLMs and CV models were initially developed as part of scientific research efforts (Vaswani, 2017)(Krizhevsky et al.,2012), they have been adopted in commercial products after the research efforts got the technology to the point where it could be implemented in products.

Scientific research in the field of Natural Language Processing led to the development of the Transformers model architecture, leading to Large Language Models that performed sufficiently well to be valuable in commercial applications like ChatGPT. A similar trend was previsouly followed with research in Computer Vision, where scientific research led to the AlexNet Convolutional Neural Network, leading to computer vision ML models that had a sufficiently high accuracy to be useful in consumer applications like smile recognition and autonomous driving.

This kind of technological innovation and technological change have been studied and discussed for decades, from railroad industries & computer manufacturers(Myers, 1969) to software development in corporations (Brem & Voigt, 2009) and startups (Guo et al., 2020).

Technology push & Market pull

Over that time, scholars in the field of investigating the economic of such technical change examined two contrasting perspectives that shaped their research on the sources of innovation.

One perspective, known as the technology-push approach, emphasized the crucial role of science and technology in the development of technological innovations and their adaptation to evolving industry structures. In contrast, some scholars adopted a demand-pull or market-pull approach, and identified a wider range of market factors, including the characteristics of end markets (especially the users) and the overall economy, that influenced the success and performance of innovation. (Guo et al., 2020)

From their research, science and technology seemed to be the primary source for the majority of technological innovations, while demand acted as the most effective catalyst for driving innovation towards the appropriate economic and institutional paths. Even for a technology push approach to technological innovation, finding eventual market pull is necessary to ensure market adoption of a new technology product.

TECH PUSH	Originates from new technological breakthroughs	Push to apply new technical capabilities	Emphasis on new technology over current market needs
MARKET PULL	Originates from unmet market needs	Pull from market for new solutions	Emphasis on existing market demands

Fig. 8 : Overview of Tech Push vs Market Pull

Brem & Voigt (2009) further explain the two approaches and expand on the differences :

1) Technology push : New products and processes are driven by the stimulus of research, whether conducted internally in an organisation or externally, with the objective of commercially utilizing newly acquired knowledge. The impulse for innovation arises from the push to apply technical capabilities, regardless of the existence of a specific demand.

2) Market pull/demand pull/need pull : The source of innovations lies in the existing insufficient fulfillment of customer needs, leading to the emergence of new demands for problem-solving. This prompts individuals or groups to articulate their subjective demands and seek to create or order a product that addresses their specific needs.

They also make it clear that although various approaches exist, the line between technology-induced and market-induced distinctions is not always clearly defined. Strong interdependencies exist between technology push and market pull models, making it impossible to make straightforward determinations that enable or disable a particular approach in a simple black-and-white manner

Drawing a distinction between the two approaches helps to look at the different causal drivers of innovation. Technology Push is an attempt at commercialising new possibilities unlocked by new technology. Market pull attempts to commercialise solutions that address an existing or underlying market demand. The two approaches have different goals and challenges and thus the innovation techniques and methods that work in a Market pull context will not be the best solution in a Technology push situation.

The current wave of innovations enabled by technological advancements of Foundations Models appears to primarily have a Technology Push approach that is trying to find Market Pull. Big & small organisations are designing solutions that are leveraging LLM technology and trying to develop products that their customers find valuable.

Radical Innovation

Multiple authors (Guo et al., 2020) categorize new technologies into two classes: 1) disruptive, radical, emergent, or step-function technologies, or alternatively, 2) evolutionary, sustaining, incremental, or "nuts and bolts" technologies. Within that context, Gerpott (2005) draws a distinction between innovations of high and low 'newness', categorizing them as radical innovations (stemming from 'technology push') and incremental innovations (stemming from 'market pull')

Norman & Verganti (2014) frame the difference between innovations in the way they change what we do :

1) Incremental innovation involves making improvements within a given frame or problem space, essentially "doing better what we already do."
2) Radical innovation, on the other hand, entails changing the problem space and "doing what we did not do before."

In this thesis, I refer to Radical Innovations as those with a high degree of newness, in a way that the underlying technology is a step change from the previous solutions leading to significant, discontinuous changes in what’s possible and what people can do with a product.

“We're seeing new use cases every day that demonstrate how AI will change the way we work, create and play.”

- Konstantine Buhler, Sequoia Capital (2023)

The context of Foundation Models appears to fit into this class of step-function technologies because of the way their capabililites are very different from existing solutions. Similar to past technological breakthroughs, radical innovations resulting from them are likely going to be more valuable to organisations and society than incremental innovations. But radical innovation is difficult and faces multiple barriers.

Challenges with Radical Innovation

Sandberg & Aarikka-Stenroos (2014) categorize these barriers, grouping them into external & internal factors. Organisations often have direct control over influencing the internal barriers and therefore overcoming them. The lack of competencies and especially, the lack of discovery and incubation competencies seem especially relevant to designers working in technology innovation.

O'Conner & DeMartino (2006) further expand on them :

1) Discovery competencies are activities that generate, identify, develop, and express research and innovation opportunities. The necessary skills involve exploring and conceptualizing, both in terms of technical and scientific exploration, as well as actively seeking external opportunities. These skills contribute to recognizing opportunities.

2) Incubation competencies are activities carried out to develop radical opportunities into business proposals. A business proposal represents a practical hypothesis regarding the potential market impact of a technology platform, the future state of the market, and the corresponding business model. The incubation process remains incomplete until the proposal, or more commonly multiple proposals stemming from the initial discovery, have been tested in the market using a functional prototype.

For the scope and focus of this thesis, discovery competencies within organisations seem to be the most relevant barrier to radical innovation. This align with the Divergent stages of Product Discovery, as discussed in the previous chapter.

INTERNAL BARRIER CATEGORY	CATEGORY DESCRIPTION	ILLUSTRATIVE EXAMPLE
1. Restrictive mindset	Fear, resistance, and conservative culture	Employee resistance to radical innovations
2. Lack of competences	Inadequate abilities for radical innovations	
a. Lack of discovery competences	Inability to create and recognize opportunities	Overemphasis on current customer needs
b. Lack of incubation competences	Inability to convert opportunities to proposals	Difficulties in building effective business model
c. Lack of commercialisation competences	Inability to ramp up business and disseminate	Difficulties in identifying new partners
3. Insufficient resources	Lack of internal resources	High visibility of innovating teams and cutbacks
4. Unsupportive organizational structure	Hierarchical arrangement challenges	Separation causing coordination difficulties

Fig. 9 : Overview of Internal challenges to Radical Innovation (Sandberg & Aarikka-Stenroos, 2014)

2.1.2 How startups innovate

New firms have been found to be better at commercialising new technologies and innovating radically than incumbents. (Bower & Christensen, 1995) (Almus & Nerlinger, 1999). Therefore it is probable that they will be the source of the next wave of radical innovations that leverage Foundation Models. But this promise comes with its perils. Startups fail often (Kotashev, 2022), for multiple reasons. (CB Insights, 2022)

To help startups maximise their chances of survival and success, entrepreneurs, scholars, venture capitalists and innovation experts have proposed multiple frameworks, techniques and business concepts. Among them, the Lean Startup methodology has been highly influential and widely adopted for developing and validating business ideas and products (Blank & Eckhardt, 2023). For that reason, I chose to focus on understanding it better.

The Lean Startup Methodology

The Lean Startup methodology emphasizes the importance of experimentation rather than detailed planning, prioritizes customer feedback over intuition, and advocates for iterative design instead of traditional extensive upfront development. Various types of new ventures strive to enhance their likelihood of success by adhering to its principles of quickly identifying and learning from failures. Over time, it has gained significant importance and popularity in professional circles. (Ghezzi et al., 2018)

Blank (2013) explains that the Lean Startup Method comprises three principles:

1) Instead of dedicating months to planning and research, entrepreneurs embrace the fact that their initial state is a collection of untested hypotheses. Founders condense these hypotheses into a framework known as a business model canvas, which visually represents how the company generates value for itself and its customers.

2) Entrepreneurs prioritize listening to their customers. Through customer development, start-ups actively search for a functional business model. If customer feedback indicates that their initial hypotheses are incorrect, they either modify them or pivot towards new hypotheses. Once a model is validated, the start-up commences execution and constructs a formal organization. Each stage of customer development follows an iterative process, with start-ups likely encountering failures before discovering the right approach.

3) Lean start-ups actively employ a methodology called agile development, which complements customer development. Agile development minimizes wastage of time and resources by progressively and incrementally developing the product. It serves as the process through which start-ups create minimum viable products that are subsequently tested.

REVELANCE OF THE LEAN STARTUP TO PRODUCT DISCOVERY

Within the context of product discovery and this thesis, the process of creating untested hypotheses and validating them as early and iteratively as possible, and getting rapid customer feedback are extremely relevant. These topics play a significant role in how the Lean Startup methodology affects the initial stages of the technological innovation process, the focus of this chapter.

Marty Cagan (2017c) further discusses how the Lean Startup principles support successful Product Discovery. He states that the Lean Startup principles makes it easier to identify & tackle potential risks of failure early, focus on solving customer problems instead of building product features, and building solutions collaboratively through customer development + agile development.

The problem with Lean Startup

The Lean Startup method proposes an iterative approach to finding business models with a focus on customer feedback to test and validate customer demand and business viability before investing resources into product development or marketing. Product development then also follows an iterative approach with minimum viable products that complement customer development. In that way, this recommendation focusses on finding “market pull” and makes it less wasteful, thereby helping reduce the probability of startup failure. That might not support radical innovations.

Mollick (2019) further expands on that, pointing out the problem. The method urges founders to proactively engage with customers outside of the office and initiate conversations at the earliest opportunity. However, the emphasis on rapidly obtaining feedback from customers for Minimal Viable Products exposes start-ups to the tendency of pursuing incremental enhancements, concentrating on meeting present customer demands rather than envisioning future possibilities. Furthermore, research conducted by scholars like Clay Christensen (Bower & Christensen, 1995) on disruptive innovation highlights that customers often have an initial aversion to novelty. Consequently, seeking validation from early customers can be even more challenging when introducing a groundbreaking idea compared to a readily understandable, incremental product.

Customer focussed radical innovation

Felina et al.(2020) further expand on the challenge with focussing too much on the customer feedback. The significance of prioritizing customer focus is widely recognized, and it is challenging to dispute its importance. However, the precise timing and relevance of engaging with customers, particularly for certain product types, remains a question. Additionally, the effectiveness of observing or surveying customers in assisting start-ups to learn and create radically innovative products is not clearly understood.

To quote the authors :

“The eagerness to get customer feedback assumes that customers know what they might want in the future. Again, this might be true in some situations, particularly in cases of incremental innovation. But as quipped by Steve Jobs, “it isn't the consumers' job to know what they want.”...
The problem is that customer imagination is delimited by what is presently there or what is presented to them. As captured by Henry Ford, “if I'd asked customers what they wanted, they would have told me, ‘a faster horse’!” Startup founders need to, in some sense, look beyond the present and into some unknown future—beyond existing products and realities. Thus there is a gap in whatever informational signals and validation that might be available from interacting with and surveying customers—and the future.”

The question then becomes :
(how) Can we understand the present to look into the future ?

2.1.3 Designing radical innovations

Donald Norman has played a pivotal role in pioneering the category of design investigation that is widely recognized today as user-centered or human-centered design (HCD). User Centered Design focusses on product usability and ease of user adoption. Human Centered Design helps ensure that the design aligns with the requirements and abilities of the individuals it is intended for. (Norman, 2002)

Similar to Felina et. al. (2020) Norman(2010) also recognized that consistently consulting with the intended users would ultimately result in incremental improvements being made to the product. Therefore, the HCD approach might be only suited for incremental innovation.

Norman & Verganti (2014) try to differentiate incremental & radical innovation in the context of technology and meaning, comparing 4 ways in which they could overlap. Along with Technology push & Market pull innovation, they propose Technological epiphanies & meaning driven innovation.

They compare them as follows :

1) Market-pull innovation begins with an analysis of user needs, followed by the development of products to fulfill those needs. This category encompasses both Human-Centered Design (HCD) and traditional market-pull methods, as both approaches start by considering users to identify avenues for innovation.

2) Meaning-driven innovation, on the other hand, begins with an understanding of subtle and unspoken dynamics within socio-cultural models. It leads to the creation of entirely new meanings and languages, often resulting in a shift in socio-cultural norms. The invention of the mini-skirt in the 1960s serves as an illustration: It represented more than just a different style of skirt; it symbolized women's liberation and signaled a radical societal change. No new technology was involved in this case.

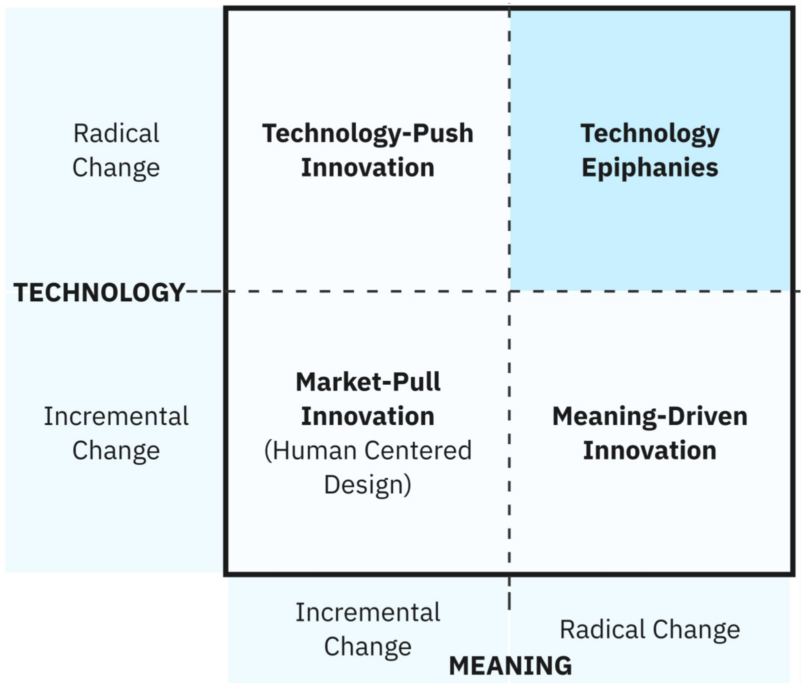


Fig. 10 : 4 types of Innovation from Norman & Verganti (2014)

3) Technology-push innovation arises from radical changes in technology, without altering the meaning of products. An example of this is the invention of color television sets alongside the existing black and white sets. Technology-push innovation is clearly not driven by users.

4) Technology epiphanies occur when new technologies emerge or existing technologies are applied in entirely new contexts, resulting in a profound change in meaning. The term "epiphany" refers to a meaning that surpasses others and provides insight into the essential nature or significance of something. This innovative application of technology is often not immediately apparent because it doesn't fulfill existing needs. It is not user-driven, but rather a dormant meaning that becomes apparent only when a design challenges the prevailing interpretation of a product and creates new, unexpected offerings that people are not actively seeking.

Design research for innovation

They then go on to introduce how Design Research could contribute to incremental & radical innovations :

1) Basic design research : involves the exploration of new meanings, without specific consideration for their application in products. This research activity is purely focused on fundamental research and is not intended for the mass market.

2) Design-driven research : aims to envision new meanings that can be applied in products. It involves seeking a deep understanding of why people purchase products and how existing solutions can be transformed into items that people buy for their emotional, playful, and symbolic aspects as much as for their functional use.

3) Human-centered research : involves exploring the current meanings that people assign to products and aims to identify existing meanings and needs in order to design products that align with those meanings and needs. Applied ethnography and user-centered observation are key research methods utilized in this approach.

4) Tinkering : refers to the act of experimenting or playing with a product or technology without a specific goal in mind, neither for enhancing its meaning nor for practical purposes. Tinkering can lead to unexpected insights and the development of new products, although such outcomes are entirely accidental.

They propose that two primary factors drive radical innovation: the advancement of a new enabling technology and the alteration of the meaning associated with the object. While the technological pathway towards radical innovation is relatively well comprehended, it is important to note that many of these innovations initially face failure upon introduction. On the other hand, the exploration of meaning as an approach to innovation has not received significant attention and remains in its early stages of research development.

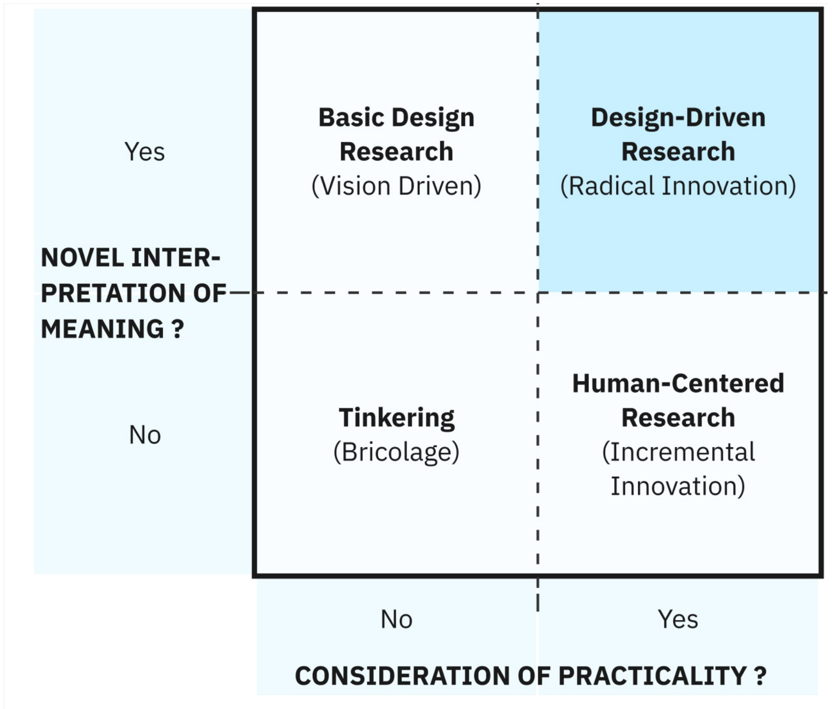


Fig. 11 : 4 types of Design Research from Norman & Verganti (2014)

Design research for radical innovation

They contend that design research has the potential to drive radical product innovation, although it is improbable to achieve such innovation through the methods of Human-Centered Design (HCD). To truly establish a new paradigm or breakthrough outcome in the realm of solutions, a profound reinterpretation of the product's meaning is necessary. Therefore, the objective of design research should be to foster a vision that enables this deep reinterpretation and paves the way for radical innovation.

To achieve this, they recommend that the research should focus on exploring fresh interpretations of what holds meaning for individuals. Traditional ideation processes and other creative methods often overlook the significance of interpretation processes, although these procedures can be appropriately adjusted. Engaging in research centered around interpretation processes has the potential to result in recognizable and reproducible radical changes.

Verganti(2006) & Pannozzo(2010) further conceptualise design research as a way to identify emerging behaviors and behaviour patterns. Verganti considers it a process in which designers actively contribute to generating and overseeing information, creating perceptions rather than solely focusing on producing products and services. Pannozzo suggests identifying opportunities to innovate, that are interpretations of a need that customers have. Design research can help discover the customer insights that can be further interpreted into opportunities to use technology to address their needs.

The challenge for designers is to be able to achieve this outcome in a repeatable, effective way.

Need Finding to find opportunities

Patnaik & Becker (2010) propose using design research to study people and identify their unmet needs. That helps companies find important new problems to work on. While these needs are often not customer feedback or feature requests, they are still opportunities to be exploited, not guesses at what customers might want in the future. They claim the following advantages of this approach :

1) Needs last longer (stay relevant for a longer timescale) than specific solutions
2) Needs can be disconnected from the current solutions
3) Need Finding helps look beyond immediately solvable problems
4) Needs help identify opportunities to use the enabling technology and propose solutions

Pannozzo (2010) recommends using the outcomes of need finding to discover Design Innovations, “products that didn’t invent new technologies but used enabling technologies to meet unfulfilled opportunities in the market”, in the following way :

1) Identify needs of customers & emergent market behaviour from research
2) Interpret the needs to identify opportunities to use enabling technology that address the needs
3) Design products to serve that opportunity such that they enable & support the emergent behaviour

Jobs to be done

The Jobs To Be Done theory tries to uncover user needs by understanding their motives behind buying a product and using those insights to improve product offerings to better serve them.

Christenson et al.(2016) emphasize that when we buy something, we're essentially using it to help us with a specific task. If the product does the job well, we're likely to use it again in the future. On the other hand, if it doesn't perform well, we'll stop using it and look for a better alternative.

They propose that what companies should really focus on is understanding the progress customers aim to achieve in a particular situation – their goals and aspirations. This is often referred to as the "job to be done.”

They argue that this way of looking at customer choice focusses on the causal driver behind a purchase decision. This helps uncover latent, unspoken needs that product developers can then try to tackle. In the general context of startups, following this approach helps founders uncover user problems they can solve and will probably get paid for. The JTBD framework helps them in avoiding building products nobody would want to buy.

In the context of this thesis, taking this approach can serve as a good foundation of user understanding, over which I could try to find value propositions for LLM solutions. Understanding of latent needs and the causal reasoning behind user behavior and choices opens up the possibility of finding radical solutions that can be designed from first principles, starting with the user needs and challenges.

Practitioners' point of view

In interviews with Gordon Murray & Kenneth Grange, Cross (2011) observes three strategies that helped them find radically innovative solutions :

1) First Principles approach
2) Creating a unique framing to the problem
3) Taking a systems approach to looking at the problem

Murray recommends “Considering the problem situation from first principles”. Grange tries to find-create a “fundamental reassessment of the purpose, function & use of the product”. That resonates with Verganti’s proposal to explore fresh interpretations of what holds meaning for individuals.

“Some people say give the customers what they want, but that’s not my approach. Our job is to figure out what they’re going to want before they do. I think Henry Ford once said, ‘If I’d ask customers what they wanted, they would’ve told me a faster horse.’ People don’t know what they want until you show it to them. That’s why I never rely on market research. Our task is to read things that are not yet on the page.”

- Steve Jobs

Reflecting on Steve Jobs’ quote on customer understanding, Katie Dill, a design leader previously at frog, AirBnB, Lyft and now Stripe (Smith, 2019), argues that it’s often misinterpreted as user research is not useful. Her perspective is that simply asking customers for their preferences may not always yield clear answers as they might struggle to express their needs well.

However, she strongly advocates for understanding customers, relying on the potency of user research and gathering qualitative insights. We shouldn't guess what customers want; instead, we must learn and understand their lives to know their present and future needs. She believes that specific kinds of user research is very effective and wouldn't disregard it. In the case of the team at Apple, she argues that they themselves are users of their products.

In Zero to One (Thiel & Masters, 2014), entrepreneur and investor Peter Theil argues that successful companies are built on secrets and that some of them can be secrets about people. By people, he’s referring to customers and specific markets. Secrets about people are unique insights about customers that are not common knowledge.

He says that discovering these secrets can help find solutions to people’s problems, leading to business success. In his opinion these secrets are often things people don’t know themselves, or hide, or are not allowed to speak. These secrets could very well be unique insights discovered through user research that would otherwise stay hidden.

2.1.4 Designing AI solutions

Designers have struggled with a unique set of challenges when it comes to designing AI products & solutions. These make it difficult for them to contribute effectively to the design process. These challenges vary across different stages of the user-centered design process (Yang et al. 2020).

For the scope of this thesis, I focused only on the discovery phase, because that aligns most with the unique challenge of finding opportunities for radical innovation, as discussed previously.

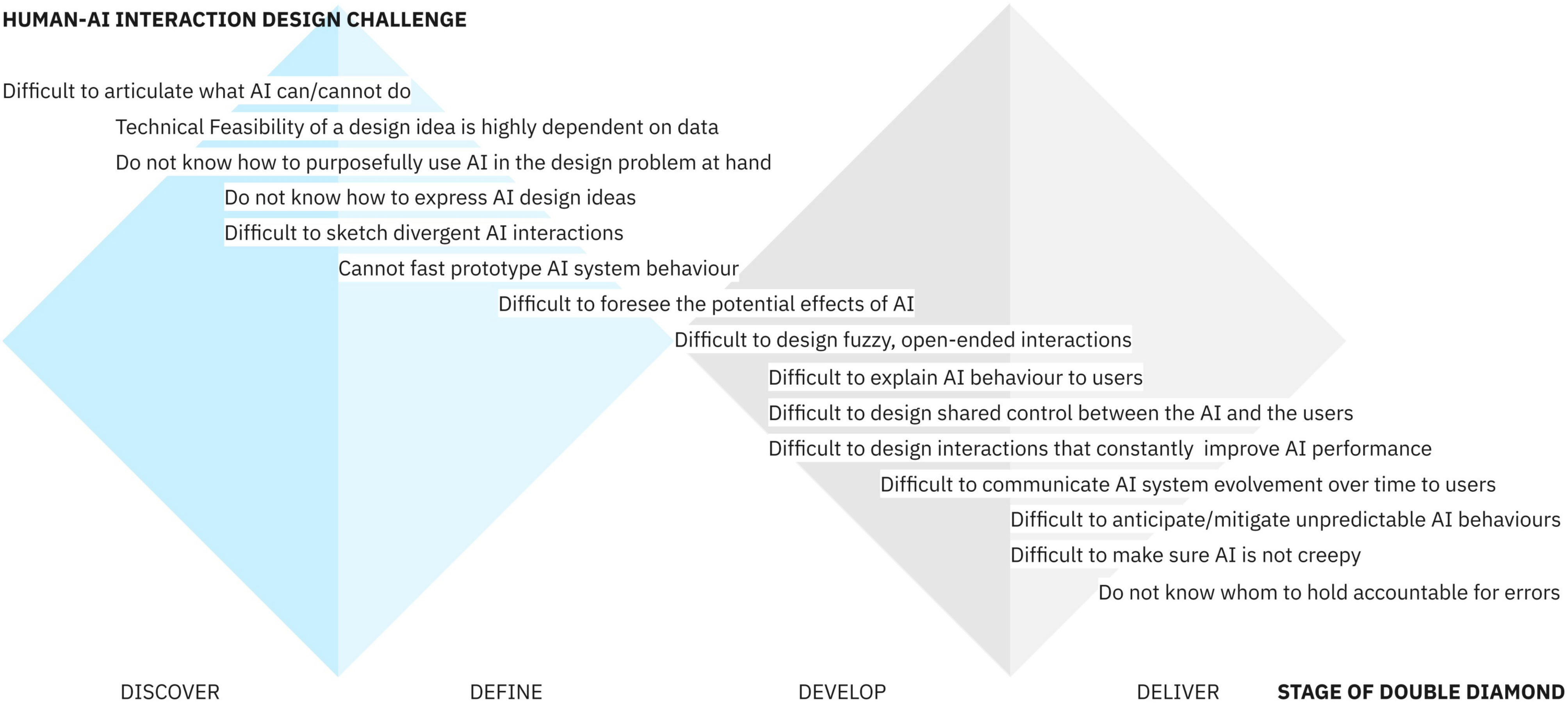


Fig. 12 : Mapping human-AI Interaction Design challenges onto the Double Diamond(Yang et al. 2020)

Uncertainty around AI capabilities

Yang et. al.(2020) further expand on the challenge that designers face with understanding different AI capabilities and dealing with the uncertainties around it : In the early design ideation stage, designers face a peak of uncertainty regarding AI's capabilities as they strive to comprehend the design possibilities AI can offer in general. This task is challenging due to the absence of a catalog listing the available AI capabilities.

An AI design concept that may initially appear unrealistic can suddenly become feasible thanks to the emergence of a new dataset. The performance of a deployed AI system can continuously fluctuate and diverge as it acquires new data to enhance its learning. This substantial uncertainty in AI's capabilities poses difficulties for designers in assessing the feasibility of their emerging ideas, thereby impeding their creative processes.

Nur Yildirim et. al.(2022) found that designers who had developed an extensive collection of "designerly abstractions" pertaining to AI's capabilities demonstrated higher levels of success and comfort when working with AI. These designers engaged in reflective conversations with AI and utilized the technical expertise of data scientists as a means of obtaining feedback on the possibilities. Participants possessed an inherent comprehension of AI's capabilities and consistently identified opportunities where AI could provide value. Alongside a broad understanding of AI capabilities, designers actively sought to acquire an in-situ understanding of the AI system, including its knowledge and actions within a specific context and with particular data.

In their study, designers conceptualized AI capabilities as active verbs representing human functions (such as reading, seeing, listening), rather than focusing on the technical mechanisms behind them (such as neural networks or collaborative filtering).

Collaborating with technologists helps

Previous studies indicate a disconnect between design and data science practices, where designers conceive AI advancements that are impractical to develop, and data scientists suggest AI innovations that users do not desire. (Yang et al., 2019).

Nur Yildirim et. al. (2022) found from their workshops that when initiating a new project, designers dedicated a substantial amount of time to comprehend the AI outlined in the design brief. They collaborated closely with data scientists, software engineers, and AI engineers to gain an understanding of the functioning of the proposed AI system and the necessary data for its operation.

They further expand on the value of this collaboration adding that, the design and data science teams managed to bridge this gap through informal collaboration and working together in the same location. Data scientists assisted designers in evaluating the technical feasibility of their concepts, while designers supported data scientists in engaging with end users to extract relevant knowledge.

Boundary objects support collaboration

Nur Yildirim et. al. (2022) reflect from their findings that boundary objects play a crucial role in enabling effective collaboration between designers and data scientists. Various artefacts such as flow diagrams, system maps, and service data blueprints were instrumental in assisting participants in envisioning and establishing a shared understanding, as well as in prototyping to specify data dependencies.

Although their study primarily concentrated on designers and data scientists, participants frequently acknowledged the involvement of other roles, such as business managers and software developers. They propose the possibility to create taxonomies and resources that explicitly document AI capabilities, accompanied by exemplars, to aid designers in putting AI concepts into practice. Their study demonstrated that these resources also benefit non-experts who actively engage in collective AI ideation.

While visual canvases are a fairly common boundary object used by a variety of design, innovation practitioners and are considered during this project as one of the candidates for boundary objects for collaboration, Avdiji et al. (2018) point out that many such visual canvases are not developed following a scientific process, but rather a trial and error approach. Thoring et al. (2019) propose that conducting a "reflective development" using an action research approach can help establish a good foundation for future research on this topic.

2.1.5 Findings

Innovation processes can be divided into a technology push and market pull approach, although they have strong interdependencies, and successful innovation needs an understanding of both, technology & markets. Innovations are often grouped into incremental & radical innovations, differing in whether they're an improvement over existing solutions or if they enable completely new possibilities.

Radical Innovations arise from a step change in the underlying technology or the product's meaning to its users, but they face multiple challenges, with designers possibly struggling from a lack of discovery & incubation competencies.

The popular Lean Startup methodology, while being significantly helpful to startups in overcoming multiple challenges, potentially inhibits their ability to look beyond incremental innovations.

Human Centered Design practices also promote incremental improvements in products, but there is potential for designers to contribute to radical innovations through design research and interpreting the findings in ways that enable radical innovations.

While designing and developing AI products, designers benefit from building an understanding of the capabilities of AI technologies, even if they don't have built a functional understanding of how the technology works. They also benefit from collaborating with technologists in the process.

Boundary objects can help designers address the challenge of collaborating with engineers that develop AI solutions and potentially also helping them understand the capabilities of AI technologies. But despite their popularity, there are gaps in the research on boundary objects like visual canvases.

Reframing sub RQs

The literature helped me understand the nuances of the initial research question RQ1 and go deeper into framing more specific challenges and try to find a way to answer them in the next stage of the thesis. I framed these questions to reflect the outcomes of the literature review and guide the Research through Design process.

1) How can we perform design research to understand users & their context in a way that can contribute to finding radical innovations ?

2) How can we support designers in understanding the capabilities of AI systems and collaborate with engineers and other stakeholders ?

2.2 METHODS

For this part of the thesis, I started out with semi-structured interviews with the team at Scitodate and with Scitodate's customers. I analysed the findings using the Grounded Theory Method to frame user needs and Jobs to be Done. I used these insights to find opportunities to leverage the capabilities of Large Language Models to create new value propositions. I then use this process to design a canvas and a card deck to support designers in following a similar process in the future.

Qualitative research

The findings from the literature review, as well as the scope of the thesis focussing on the discovery stage, led to the decision to perform qualitative research. Considering the limitations of time & access to Scitodate's customers, I chose semi-structured interviews for conducting qualitative research as they are a flexible yet reliable way of uncovering user motivations and latent needs.

The interviews with the team at Scitodate were conducted in-person with a formal interview guide (Appendix A). The guide had few opening questions and the conversations led to follow-up questions. These interviews lasted 30 to 60 minutes. I took notes on my observations and findings from the interviews and later followed them up with informal chats during office hours and lunch breaks. I interviewed 5 different people in the organisation, spread across product development, sales and customer success.

That was later followed up with interviews with Scitodate's customers. These were all conducted over video call, again with a formal interview guide (Appendix B). These interviews lasted 60 minutes each. They were all recorded and later transcribed. I interviewed 3 different customers twice, leading to 6 interviews. All three of them worked in different roles and used Scitodate products for different applications.

Research analysis

To analyse the data collected during the interviews, I used the Grounded Theory Method (Strauss & Corbin, 1997). The coding followed 3 stages : open coding, where I converted findings from the interviews into short lines of text, called codes. That was followed by arranging them into categories, that I marked with different colours. In the end, I found relations and connected different codes and categories, to help me understand dependencies and causal factors between them.

I used the outcomes of the GTM process to propose Jobs from the Jobs To Be Done theory. For every Job, I identified user needs and challenges that customers faced. I then structured all these findings across the "BowTie Growth funnel" to further analyse the user needs and Jobs to find common themes and patterns.

Finding value propositions

To find out how LLMs could help address user needs and challenges, I built an understanding if what the current LLMs are capable of. I used that understanding to identify how they could to be used to design value propositions that Scitodate could offer. Multiple discussions and feedback from the engineers at Scitodate helped me get a better understanding of what the technology can do and how that could be translated into product features.

Tool prototyping

To design artefacts that can help designers follow a similar process, I tried to map out the process I had followed into a canvas and tested that proposal with the team at Scitodate. My primary goal here was to find out what form the tool could take. I did multiple iterations to the canvas prototypes depending on the feedback from the team and introduced a card deck to support the canvas. I facilitated every test of using these tools to guide the team and identify where the canvas and cards were falling short. I also got feedback from practicing innovation consultants working in technology focussed innovation.

Design optimisation

After the canvas & cards had reached the point where they were evidently able to support the process, I worked on optimising the details of the design. I first wanted to get the tool prototype to be effective and finalize its structure and content before investing time into optimising their design. Considering my limited expertise at visual design, I deemed that more efficient and faster, from past experiences of similar design projects. Due to time limitations, I could not refer to literature around canvas design to support this process of optimisation. Many ideas and motivations for these improvements also came from the previous process of testing the prototypes.

2.3 DESIGN RESEARCH

2.3.1 Qualitative Research

What is important for Scitodate's customers and how does Scitodate currently creates value for them ?

To find answers to that question, I began with interviewing the team at Scitodate to get a preliminary understanding and later spoke with Scitodate's existing customers.

Interviews with Scitodate

To get up-to-speed with Scitodate's products and value proposition, I first had discussions with multiple people in the team, including the engineering lead, product designer, product manager, growth team lead and customer success manager. That helped me build an initial understanding of the value proposition of Scitodate's current products and the challenges that customers face in their work. I used these insights to frame questions for the customer interviews.

FINDINGS

These internal interviews primarily helped me form an overview of the customers' workflow. The bowtie funnel (Jacco J. Van Der Kooij, 2023) as a general marketing & sales funnel was extremely helpful to get a brief overview of the customers' activities as well as how the Growth team at Scitodate engaged with their customers. The fact that Scitodate's customers had similar job roles and duties to their/our own Growth team made these conversations extremely helpful for building user empathy and understanding.

Scitodate's customers mainly operated in the first half of the bowtie funnel, focussing on marketing and sales. That covered 3 stages of the bowtie funnel: Awareness, Education & Selection.

The sales and marketing bowtie funnel is a framework used to visualize the customer journey and understand the various stages they go through before and after making a purchase decision. Here's a brief explanation of the awareness, education and selection stage:

1) Awareness: This is the top of the funnel where potential customers become aware of your brand, product, or service. It involves creating brand visibility through various marketing channels such as advertising, content marketing, social media, and public relations. The goal is to generate initial interest and attract a wide audience.

2) Education: Once prospects are aware of your brand, the next stage is to educate them about the value and benefits of your offering. This involves providing relevant and informative content, such as blog posts, videos, whitepapers, or webinars. The goal is to build credibility, establish thought leadership, and address the pain points or needs of the target audience.

3) Selection: In this stage, prospects have gathered information and are considering different options. The focus shifts to demonstrating why your product or service is the best choice for their needs. This can be done through case studies, product demos, testimonials, comparisons, or reviews. The goal is to differentiate your offering from competitors and convince prospects to choose your solution.

Interviews with customers

SETTING UP INTERVIEWS WITH CUSTOMERS

The team at Scitodate was kind enough to reach out to customers in the first weeks of the project to setup online interviews with me where I could talk to them. I was able to collect insights from 3 customers with different roles and responsibilities. I had the opportunity to interview all three of them twice, allowing me to base my followup interviews based on our initial discussions.

That was extremely helpful to cover the broad range of activities that Scitodate's customers used their products for. The interview candidates were part of teams of different sizes therefore had different kind of activities under their scope. That also helped me understand how they worked in a solo role and as part of a team of marketers & sales representatives.

THE INTERVIEW PROCESS

The interviews were done in a semi-structured manner, driven by an interview guide. The Interview guide had a set of open-ended questions about specific topics which gave me the freedom to steer the conversation depending on the responses from the interviewees. This gave me the opportunity to dig deeper into the conversation and ask follow-up questions, wherever I found some interesting insight or point-of-view.

For the interviews, the goal was to explore the different activities they did and find out what caused them distress. Along with that, I also tried to explore what they desired their workflow to be like. Some questions also focussed on their use & expectations of AI tools, especially ChatGPT. Being able to do follow-up interviews helped me to first analyse the findings from the first conversations and then ask follow-up questions and focus on specific topics that were relevant for the thesis. Being able to use Scitodate's own LLM tools as prototypes during the follow-up interviews was useful in getting their first impressions and opinions about the new technology.

FINDINGS

The interviews helped me get a deeper understanding of the specific challenges and nuances of the marketing and sales process for Scitodate's customers.

Many insights were very context specific to the specific customers I talked to and the organisations they worked at. These differences among the interview candidates helped me gather a broad view of the problems from different point of views. Some of the most interesting quotes were :

"We keep users occupied with material, so when they are going to make a purchase, they have everything they need"

"Account based marketing takes up 50% of our time"

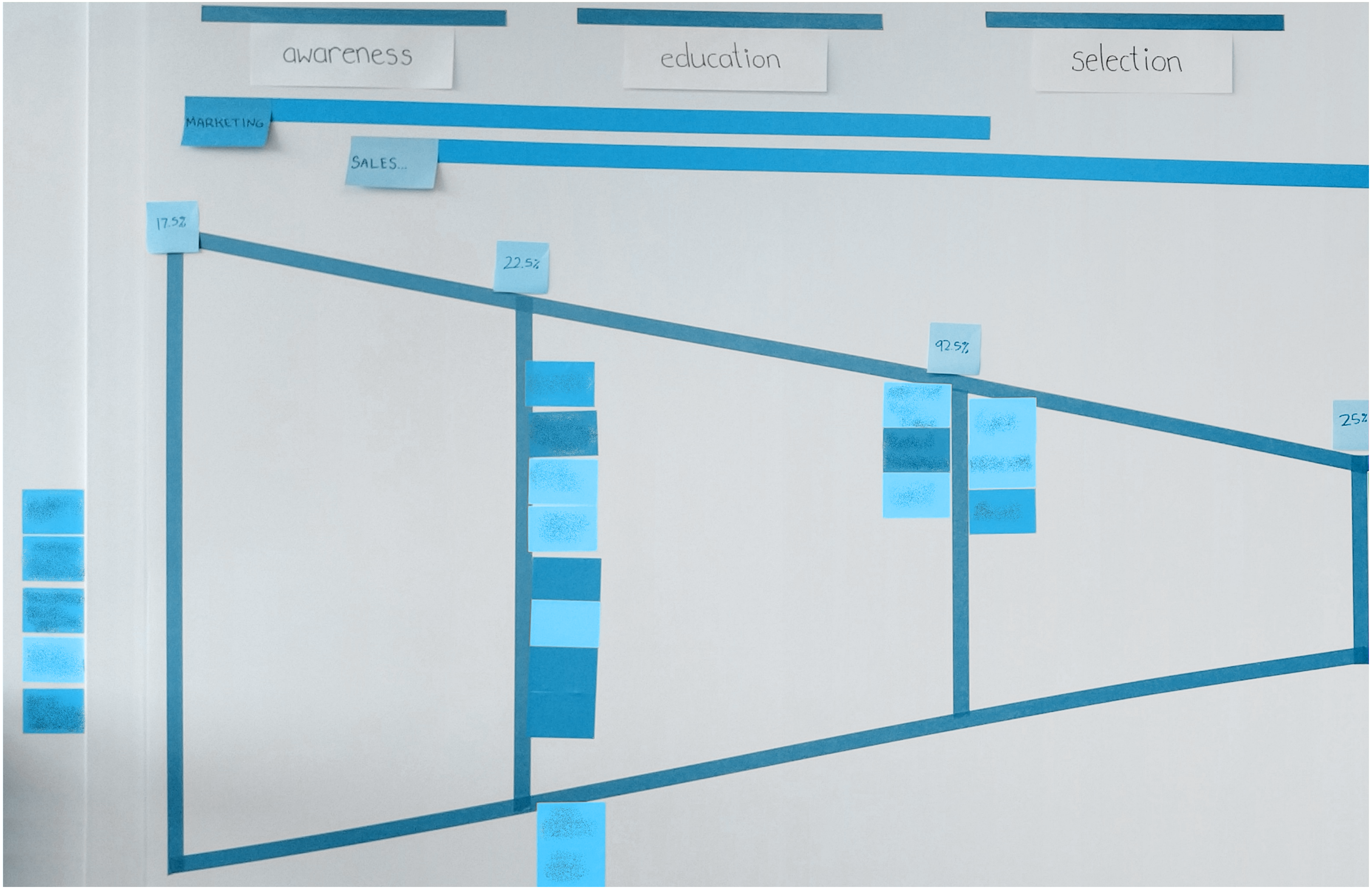
"We don't know if targeted emails would work, because we never tried it"

"The more you personalise an email, the longer it takes"

"Our job : Get users interested, find their pains, offer a solution"

"Hierarchy within a corporation is just a blackbox."

"There were multiple instances where the AI output wasn't accurate or reliable."


awareness
education
selection
MARKETING
SALES...
17.5%
22.5%
92.5%
25%

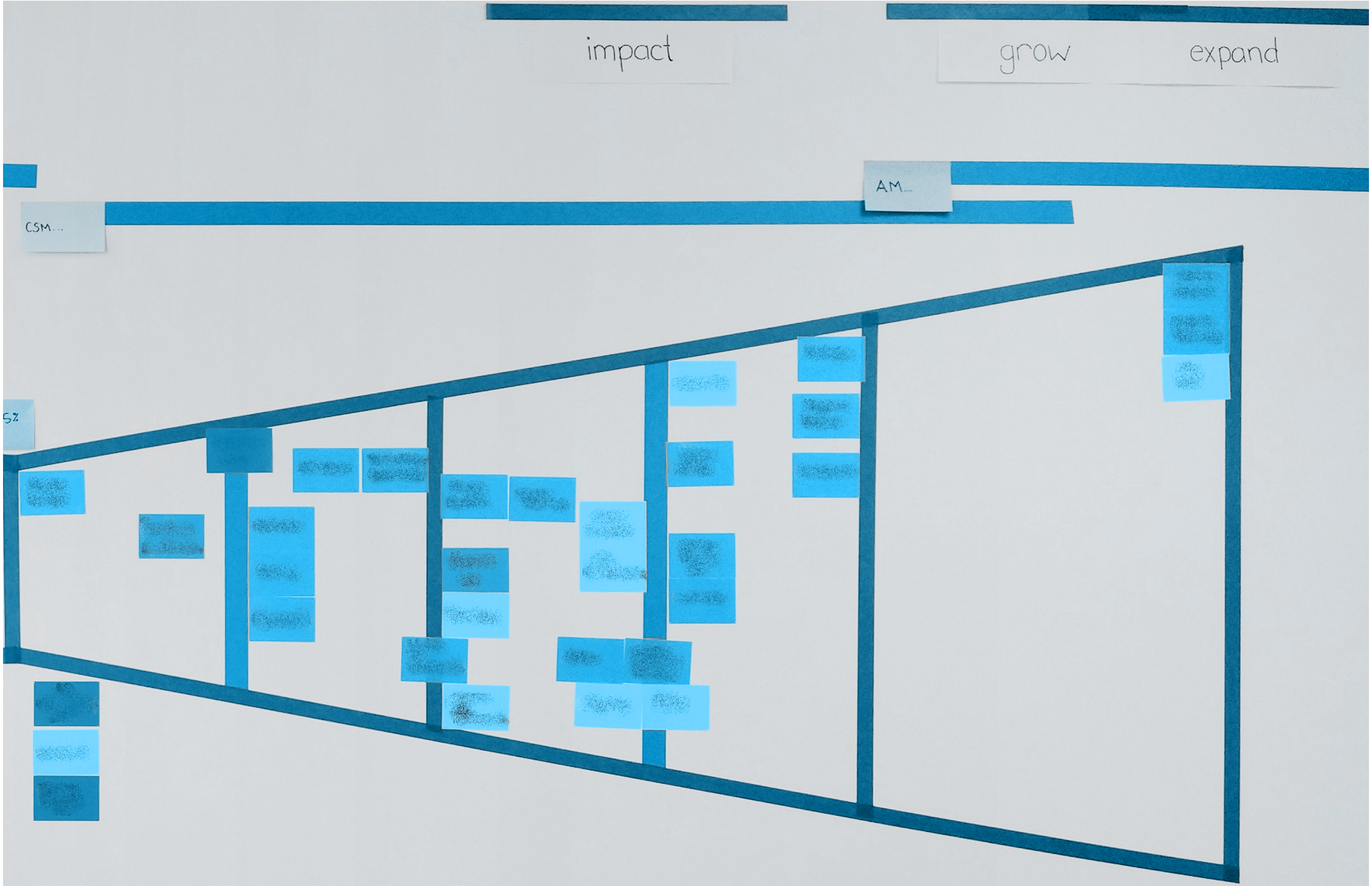


Fig. 13 : Bowtie funnel used by the team at Scitodate to track their interactions with clients. Client names blurred for confidentiality

2.3.2 Research Analysis

In order to turn the raw information from the interviews into structured insights that can be effectively used to find radically innovative solutions, I followed a multi-step process of analysing and interpreting the interview data. I employed the Grounded Theory Method (GTM) to first structure the raw data and later analysed them from a Jobs-To-Be-Done point of view.

PROCESS OF ANALYZING THE INTERVIWS

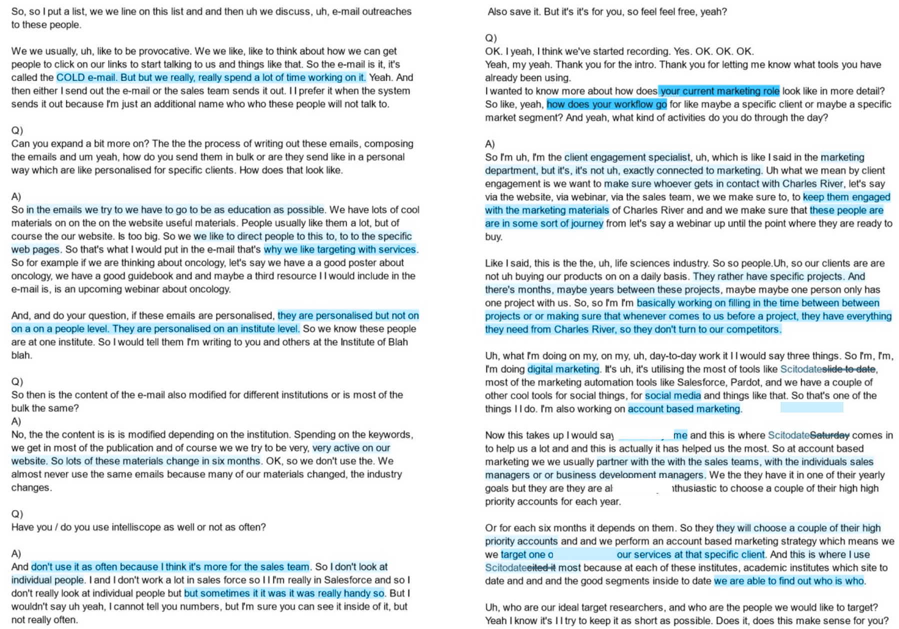
1) Coding the interviews from its transcripts
2) Structuring the codes to find clusters & patterns
3) Find causal relations, connections and dependencies between them
4) Frame them as Jobs-To-Be-Done & challenges connected to every job
5) Cluster these Jobs along the BowTie funnel
6) Find overarching patterns across different Jobs & challenges
7) Hypothesize overarching causal drivers & latent needs

Grounded Theory Method

I used the Grounded Theory Method (GTM) to analyse the data collected during user research.GTM is a qualitative research approach that aims to develop theories grounded in empirical data.The central idea behind Grounded Theory is to allow theories to emerge from the data itself, rather than imposing preconceived notions or existing theories on the research.

The process involves collecting qualitative data, constantly comparing and analyzing the data to identify patterns and themes, using theoretical sampling to gather data that refines the emerging theory, coding the data to create categories and concepts, reaching theoretical saturation when no new insights emerge, integrating and refining the concepts to develop a coherent theory, and finally, documenting the theory in a research paper or thesis. GTM allows theories to emerge directly from the data, making it a powerful method for studying complex social phenomena and human behavior in natural settings.

Due to time constraints & other limitations of this thesis, I was not able to collect sufficient data to reach theoretical saturation. Theoretical sampling was not possible due to access to a limited number of available customers to interview. Being able to do followup interviews with customers helped in approaching the GTM approach iteratively. But despite those limitations, the clusters and theories that emerged from this process helped me frame user needs and their Jobs to be done.



So, so I put a list, we we line on this list and and then uh we discuss, uh, e-mail outreaches to these people.

We we usually, uh, like to be provocative. We we like, like to think about how we can get people to click on our links to start talking to us and things like that. So the e-mail is it, it's called the COLD e-mail. But but we really, really spend a lot of time working on it. Yeah. And then either I send out the e-mail or the sales team sends it out. I I prefer it when the system sends it out because I'm just an additional name who who these people will not talk to.

Q)
Can you expand a bit more on? The the the process of writing out these emails, composing the emails and um yeah, how do you send them in bulk or are they send like in a personal way which are like personalised for specific clients. How does that look like.

A)
So in the emails we try to we have to go to be as education as possible. We have lots of cool materials on on the on the website useful materials. People usually like them a lot, but of course the our website. Is too big. So we we like to direct people to this to, to to the specific web pages. So that's what I would put in the e-mail that's why we like targeting with services. So for example if we are thinking about oncology, let's say we have a a good poster about oncology, we have a good guidebook and and maybe a third resource I I would include in the e-mail is, is an upcoming webinar about oncology.

And, and do your question, if these emails are personalised, they are personalised but not on on a on a people level. They are personalised on an institute level. So we know these people are at one institute. So I would tell them I'm writing to you and others at the Institute of Blah blah.

Q)
So then is the content of the e-mail also modified for different institutions or is most of the bulk the same?
A)
No, the the content is is is modified depending on the institution. Spending on the keywords, we get in most of the publication and of course we we try to be very, very active on our website. So lots of these materials change in six months. OK, so we don't use the. We almost never use the same emails because many of our materials changed, the industry changes.

Q)
Have you / do you use intelliscope as well or not as often?

A)
And don't use it as often because I think it's more for the sales team. So I don't look at individual people. I and I don't work a lot in sales force so I I I'm really in Salesforce and so I don't really look at individual people but but sometimes it it was it was really handy so. But I wouldn't say uh yeah, I cannot tell you numbers, but I'm sure you can see it inside of it, but not really often.

Also save it. But it's it's for you, so feel feel free, yeah?

Q)
OK. I yeah, I think we've started recording. Yes. OK. OK. OK.
Yeah, my yeah. Thank you for the intro. Thank you for letting me know what tools you have already been using.
I wanted to know more about how does your current marketing role look like in more detail? So like, yeah, how does your workflow go for like maybe a specific client or maybe a specific market segment? And yeah, what kind of activities do you do through the day?

A)
So I'm uh, I'm the client engagement specialist, uh, which is like I said in the marketing department, but it's, it's not uh, exactly connected to marketing. Uh what we mean by client engagement is we want to make sure whoever gets in contact with Charles River, let's say via the website, via webinar, via the sales team, we we make sure to, to keep them engaged with the marketing materials of Charles River and and we make sure that these people are are in some sort of journey from let's say a webinar up until the point where they are ready to buy.

Like I said, this is the the, uh, life sciences industry. So so people.Uh, so our clients are are not uh buying our products on on a daily basis. They rather have specific projects. And there's months, maybe years between these projects, maybe maybe one person only has one project with us. So, so I'm I'm basically working on filling in the time between between projects or or making sure that whenever comes to us before a project, they have everything they need from Charles River, so they don't turn to our competitors.

Uh, what I'm doing on my, on my, uh, day-to-day work it I I would say three things. So I'm, I'm, I'm doing digital marketing. It's uh, it's utilising the most of tools like Scitodate~~slide to date~~, most of the marketing automation tools like Salesforce, Pardot, and we have a couple of other cool tools for social things, for social media and things like that. So that's one of the things I I do. I'm also working on account based marketing.

Now this takes up I would say me and this is where Scitodate~~Saturday~~ comes in to help us a lot and and this is actually it has helped us the most. So at account based marketing we we usually partner with the with the sales teams, with the individuals sales managers or or business development managers. We the they have it in one of their yearly goals but they are they are al thusiastic to choose a couple of their high high priority accounts for each year.

Or for each six months it depends on them. So they they will choose a couple of their high priority accounts and and we perform an account based marketing strategy which means we we target one o our services at that specific client. And this is where I use Scitodate~~cited it~~ most because at each of these institutes, academic institutes which site to date and and and the good segments inside to date we are able to find out who is who.

Uh, who are our ideal target researchers, and who are the people we would like to target? Yeah I know it's I I try to keep it as short as possible. Does it, does this make sense for you?

Fig. 14 Coding the interview transcripts to find out interesting parts to analyse further

Fig. 15 Analysing codes into clusters, patterns and causal relations

Finding Jobs, causal factors & challenges

I framed the “theories” that were outcomes from the Grounded Theory Method as the customers’ Jobs To Be Done (JTBD). The clusters pointed me to reasons why customers did what they did. The clusters helped me identify what outcomes customers were trying to achieve during some activity. I used the findings from the analysis to also identify the different challenges they faced and their needs that aligned with every Job they had to do.

After I had analysed the interview data and clustered them to frame Jobs and challenges, I arranged the findings across the BowTie Growth funnel. Using the funnel as a framework helped me visualise how different Jobs were relevant to different activities and how they related to each other. Another advantage I realised was that this framing helped me identify recurring themes and common challenges across different Jobs and activities. I discovered how few causal factors led to multiple challenges.

This process also led to the customer insights getting abstracted away from the specific products they used, specific actions they performed and specific goals they wanted to achieve. That helped me identify insights that applied to multiple parts of the funnel, and thus relevant for multiple user Jobs.

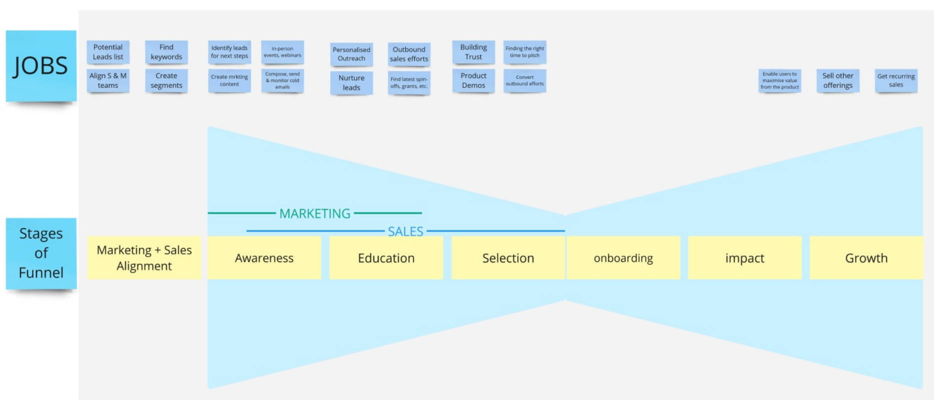


Fig. 16 Framing JTBDs and compiling them across the BowTie Funnel

STAGE IN BOWTIE FUNNEL	JOB TO BE DONE	USER CHALLENGES
AWARENESS	JTBD 1	Challenge 1
		Challenge 2
	JTBD 2	Challenge 3
		Challenge 4
EDUCATION	JTBD 3	Challenge 5
	JTBD 4	Challenge 6
		Challenge 7

Fig. 17 : Table showing JTBDs and User challenges across different stages of BowTie funnel

2.3.3 Finding Value Propositions

Understanding what's possible

After completing the user research came time to find ways to leverage Language Models to find better solutions. But before that it was important to understand what Language Models are capable of doing. While being busy with user research, I was also parallelly trying to understand what Large Language Models can do, trying my best to follow new developments every week. I

At the beginning of the thesis and before the official project kick-off, the product team at Scitodate discussed some potential ideas with me for using ChatGPT and the ChatGPT API (Application Programming Interface) to create solutions for their customers. They mainly focussed on automated text generation, to create personalised emails and LinkedIn messages for cold outreach. The idea was to use information from a person's LinkedIn page and other profiles on the internet to create messages that are tailored to them and focus on the services that user wants to provide them.

As the broader LLM landscape kept evolving with new research findings, open-source projects and products from private organisations (GPT-4 API from OpenAI), we kept finding out new opportunities for using LLMs. Along with the ability to generate customised text content, we also explored the potential of information retrieval from specific sources (Guides | Langchain), using chatbot instances that could act like automated Agents, and access APIs that connected LLMs to Scitodate's products, etc.

In the context of Scitodate, that meant giving a language model access to the company's database of scientific publications, etc. and letting it use that information to generate more factually correct and relevant output. That could open up potential possibilities like generating content that's personalised for specific scientists, research domains or instruments. Some other possibilities included automating a sequence of repetitive tasks in our customers' workflow and making the current software easier to use by introducing a personal assistant.

Finding opportunities to use LLMS

Once both sides of the equation (Customer Jobs & Technology) were fairly well understood, I could focus on finding opportunities to address those customer challenges with the possibilities unlocked by LLMs. The process seemed like a question of finding an alignment and fit between what the customers would find valuable and what new possibilities Language Models could unlock. This was similar to the common challenge of aligning technology-push with market-pull (Brem & Voigt, 2009))

After mapping the "Jobs to be Done", the related problems, and the rest of the insights at different levels of abstraction, it became easy to identify which of them could be addressed using language model technology. I found it helpful to be aware of both sides of the equation to find a good fit: Being aware of the State of the Art in the rapidly evolving field of LLMs, and emerging developments for their commercial use, as well as customer problems.

I used the Value Proposition Canvas from Strategyzer (Osterwalder et al., 2014) to visualise how customer problems and Jobs could be addressed using LLMs to develop new features and products. I listed down all possibilities to use LLMs at different stages of the BowTie funnel. This contributed as a list of ideas for Scitodate, which can serve as a starting point for designing new features/products that leverage LLMs.

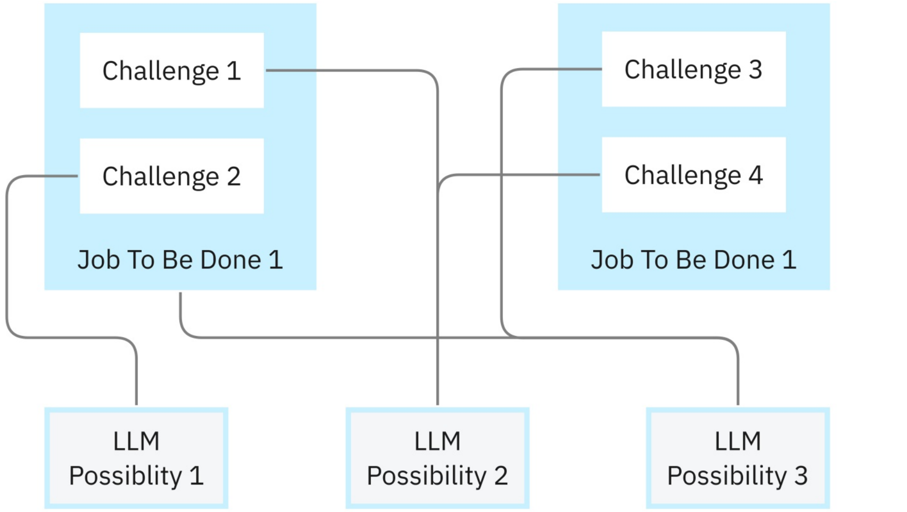


Fig. 18 Schematic showing how User research findings connected to new possibilities of LLMs.

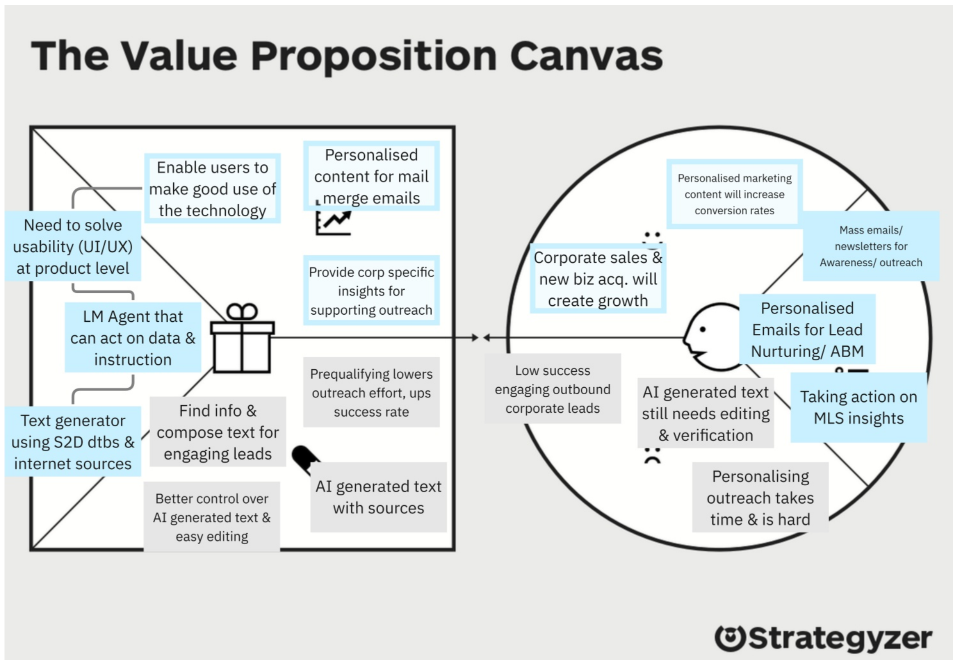


Fig. 19 Using the Value Proposition Canvas to map out how customer Jobs can be supported with LLMs

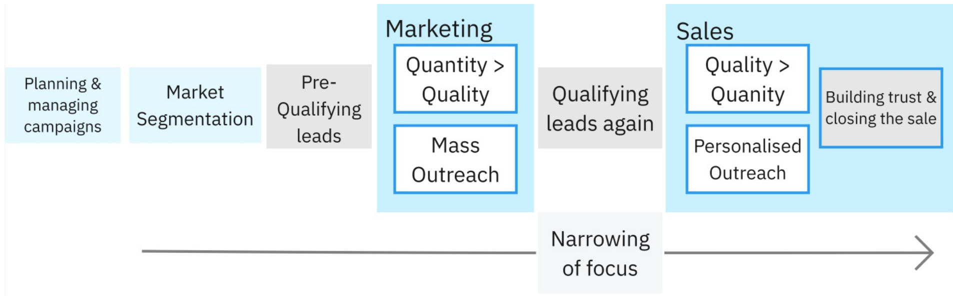
Radical value propositions

Because I was able to synthesize causal factors and user insights that were relevant for multiple user activities with common motivations, that opened up the opportunity to ideate solutions that could address multiple challenges together. That enabled potential ideas that went beyond incremental improvements in current products and how customers currently approached their work.

One of the most interesting findings was the current limitation of Scitodate's customers to either write personalised content for a small number of potential buyers or write generic content that can be sent to a large number of cold leads. The first approach increased the likelihood of a response to the outreach but limited how many leads could be approached. The second approach increased the reach to more people but reduced the probability of getting a response.

This is currently a human limitation centred around finding personalised information and then writing the outreach content. The potential of Language models is that they can generate messages that are personalised to an individual and that could be repeated for a large number of contacts.

The way current marketing and sales efforts are structured is that the marketing stage focusses on reaching a large number of potential leads, and then sales teams following up with personalised outreach to people that show interest to the initial marketing outreach. Language Models open up the potential to drastically change how these practices function. By enabling personalised outreach to a large number of potential leads, that drastically changes the responsibilities of marketing as well as sales teams, potentially reducing a significant amount of time & effort.



Part A : How email outreach currently takes place

Part B : How email outreach could change, also shown on the sales Funnel

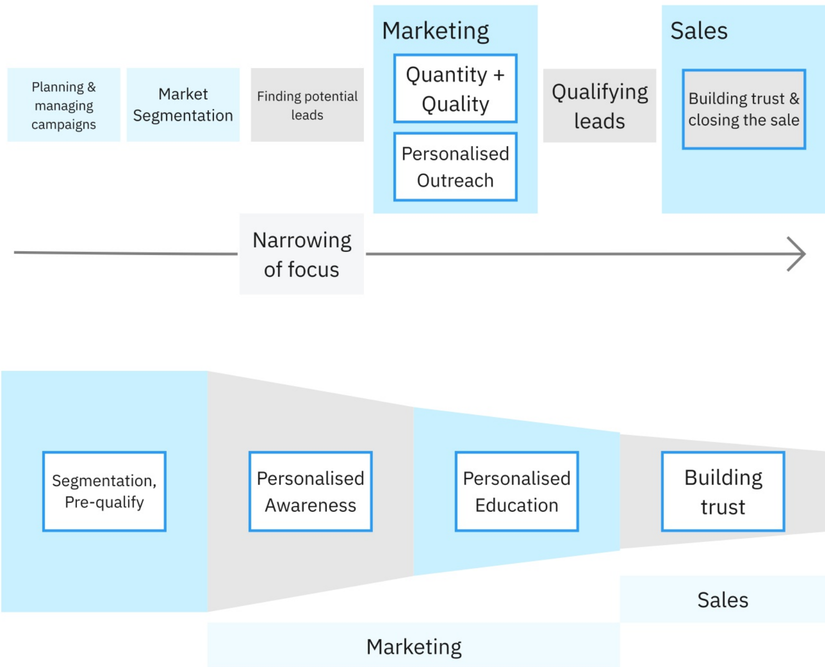


Fig. 20 Comparing the present state with the possiblity of good quality outreach to a greater

2.3.4 Design Research for radical ideas

Reflecting on the empirical research process I followed at Scitodate and the findings from the literature, I tried to find out how the process worked for me. The intent here was to build a generalisable understanding that can be communicated to and utilised by future designers working in a similar context.

Limitations of the "Novel Meaning" approach

Verganti (2011) defined "meaning" as the essence behind a product, referring to the deep psychological and cultural reasons why people use it. Meanings can encompass both individual and social motivations. Individual motivation is associated with the personal and emotional significance a product holds for the user. On the other hand, social motivation relates to the symbolic and cultural meaning that the product communicates to others about the user.

He focusses the process of “Design-driven research” primarily on finding solutions can be transformed into products that people buy for their emotional, playful, and symbolic aspects as much as for their functional use. While this approach holds significant value in the domain of B2C products and consumer products, it might not be as relevant in B2B contexts.

Additionally, new meanings and “more meaningful” products are only valuable for those parts of society that do not have bigger problems to worry about. People who struggle with needs that relates to lower hierarchies in Maslow’s pyramid of needs (Maslow, 1943) do not care as much about addressing those on top. For eg. people who struggle from malnutrition and hunger due to a lack of access to food, do not care as much about the meaning, or what it means to be vegetarian or vegan. They especially do not much care about the meaning of vegetarian meat substitutes.

In the B2B context of Scitodate, products that are built to make work easier and employees more productive, create most of their value through the problems that they solve and how they help customer businesses economically. New ways of solving problems can be more useful for them that new meanings to the products. New meanings could be created, but through better problem solving.

During the design research at Scitodate, that was primarily the case. Customers used Scitodate’s products mainly because of how they saved time and effort that way, making them more productive. The potentially radical value propositions that were identified also focussed on problem solving, not on creating new meaning. The new meaning that customers could derive from the solution would be a consequence of better problem solving.

We can go back to the example of "Quality + Quantity" from the previous section. The identified value proposition was of creating customised outreach content for a large number of leads. The proposed solution addressed customers’ lack of time and their inability create customised content for a large number of people. The solution to this problem could potentially hold the meaning of a copywriting aid that significantly simplifies their work. But the problem solving is at the core of the value proposition.

A “Novel Problems” approach to Design Research

What helped me during the design research at Scitodate was finding the underlying problems that customers struggled with. Taking the JTBD & User challenge approach to analyse the user research helped me find out the underlying needs and problems that customers were struggling with.

Similar to Verganti’s argument of new meanings creating new value, solving problems that were previously not addressed, or even identified, could be a promising source of value. This line of thought aligns very well with multiple past contributions to literature that were identified.

Need Finding (Patnaik & Becker, 2010) focusses on finding unmet customer needs. They highlight that this approach helps look beyond immediately solvable problems, and can be disconnected from current solutions. Jobs To Be Done (Christenson et al., 2016) focusses on finding underlying motivations and goals, that help uncover latent, unspoken needs and user problems.

Gordon Murray’s recommendation of considering the problem situation from first principles, basically points to framing the problem from a new point of view. Grange’s “fundamental reassessment of the purpose, function & use of the product” can also help acheive a novel problem reframe, especially when the purpose, function and use of the product is to address some user problem.

Frame Innovation (Dorst 2015), focusses on exactly this process of framing and reframing problems to identify better solutions to problems.

As demonstrated in the previous sections, design research methods can help identify these novel problems that can be addressed to create value for customers. This process of analysing user research data to identify unaddressed problems can also address the limitations of HCD and Lean Startup that were previously discussed.

This framing of “problem solving” also makes it possible to look at technology in a promising way.

The opportunities of new technology

New enabling technologies can solve problems and lead to better solutions in multiple ways :

1) Solving problems with existing solutions in an incrementally better way, improving current products and their performance.

2) Solving problems with existing solutions in a completely different way, resulting in significantly different products

3) Solving problems that previously could not be solved. These problems do not have a proper existing solution because existing technologies could not solve them.

4) Solving problems that were previously ignored, possibly because they were too difficult to solve, or could not be solved in an efficient and effective way.

While design research can help identify problems for all the above possibilities, the last two are especially interesting. Design research can help identify problems that have not yet been solved. Design research can also help identify problems that have been ignored and neglected. That way, design research opens up new avenues for using technology to create value through problem solving.

Another interesting perspective here is that the possibilities unlocked by new technology can guide designers in selecting the most promising problems identified from design research. This enabling technology can also help in re-framing problems identified in design research leading to “Novel Problems”.

Technology Driven Problem (Re)Framing

By understanding the new opportunities unlocked by emerging technologies, it can be possible to frame problems in a way that can benefit the most from it. The technology can drive how a problem is framed.

Insights and problem frames identified from design research can thus be reframed in a way that leverage the technology to create greater value.

I followed a similar process while finding value propositions for Scitodate. From the user research and analysis, I was able to identify multiple unmet needs and challenges. The “Quality + Quantity” need was originally a problem labelled “Quality vs Quantity”, one of the insights identified from analysing the customer interviews.

But, when I considering the potential applications of LLMs, this problem became more relevant and significant. It was an underlying problem that customers had, but they couldn’t do anything about it. LLMs opened up the possibility of addressing this challenge.

I modified the four quadrant image from Norman & Verganti (2014) to represent the "radical change in problem framing" perspective.

It is possible that framing problems from design research in a way that aligns with emerging technologies can help startups find unique value propositions that leverage new technology, Foundation Models in this case, and can lead to successful radical innovation.

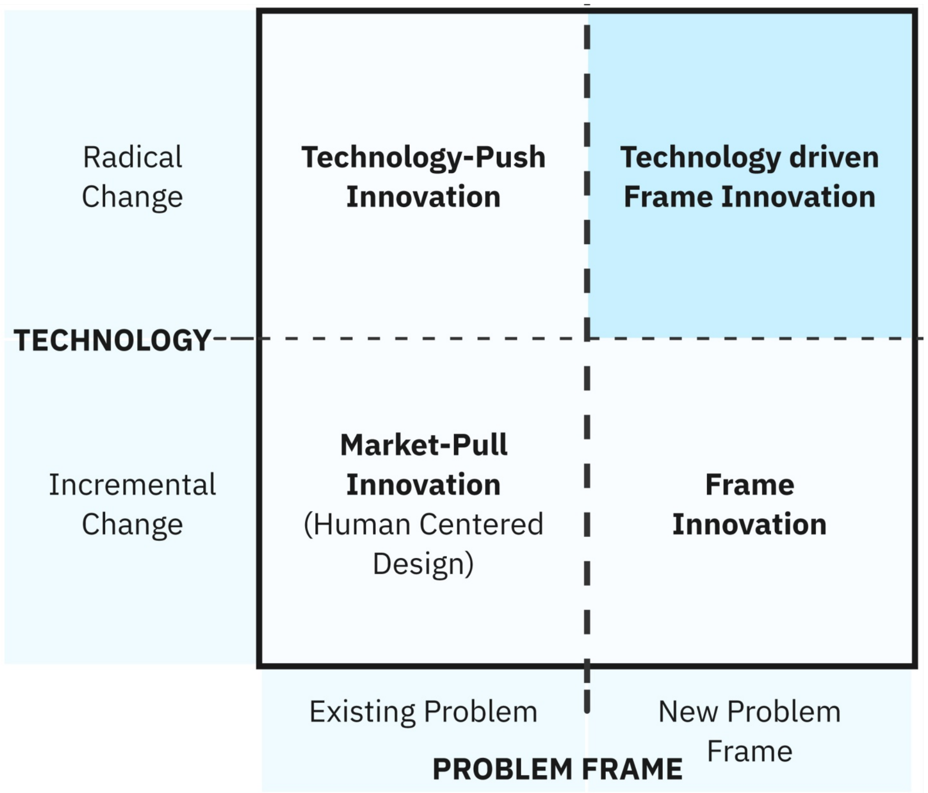


Fig. 21 : Technology driven Frame Innovation that leverages radical technology and problem frames

2.4 TOOL DESIGN

2.4.1 Prototyping

After going through the entire process of starting from user research, to analysing the insights and then finding out how LLMs could create value for customers, I moved to designing a tool to support designers in following a similar process.

I started with a list of design requirements, used that to propose a first version of the tool and iteratively developed it further based on feedback from testing. The iterative development process progressed as a co-evolution of the problem-solution space (Dorst & Cross, 2001) as I learnt more about the design requirements while designing & testing it.

Defining design requirements

Reflecting on the literature study, empirical research at Scitodate, and the observations from the process, I was able to compile a preliminary list of design requirements and goals for the design artefact :

1) Translating the empirical research activities into a repeatable process
2) Empowering designers & other non-engineers to understand what Foundation Models are capable of, and ideate possible usecases.
3) Align the technological possibilities with user challenges & asking the right questions to find that alignment
4) Support cross-functional teams in collaborating on products that employ Foundation Models

Apart from that, the design artefact had to be something designers & innovation teams will want to use. It had to be easy to use & adopt. And it had to be easy to distribute to different teams and individuals. Using the design artefact should not need a significant amount of time or effort.

Design version	P1	P2	A1	P3	P4	P5 : P4 + A2	A2
Evaluation	Test 1	Feedback		Test 2	DIY	Test 3, 4, 5	Test 6

Fig. 22 : How different versions of the canvas and cards were iteratively evaluated

Starting form

I used the design requirements to guide the initial selection of the form of the design artefact. The intention to help designers at Scitodate & other teams follow the same process meant that the design outcome had to communicate a set of step-by-step instructions. Or it could be something that supported a facilitator in doing this.

But, depending on a facilitator would've made the final outcome significantly less distributable or scalable, and hence less valuable. Supporting collaboration across cross-functional teams meant that multiple people need to be able to use the tool simultaneously. It also had to serve as an ideation tool for the team and help designers use Foundation Models.

To make it easy to use & adopt, the design deliverable had to have a form that most teams were aware of and comfortable using. The design should also need low effort & time to adopt and integrate well with the rest of product development process. Thanks to the popularity of "canvas" solutions, and the fact that I'd used one in the empirical research phase too, I decided to start my exploration for the form with a canvas.

Value Proposition Canvas

As part of the value finding process (Sec 2.5), I had used the Value Proposition Canvas (VPC) from Strategyzer(Osterwalder et al., 2014) to map out and communicate how customer goals, pains & desires aligned with potential value proposition ideas. It consists of two main components: the Customer Profile and the Value Map.(Strategyzer, 2017)

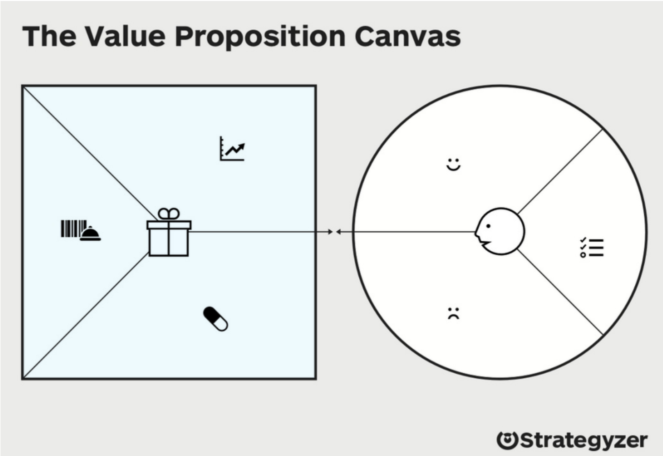


Fig. 23 : The Value Proposition Canvas from Strategyzer (Strategyzer, 2017)

Customer Profile: The Customer Profile focuses on gaining a deep understanding of the target customers or users. It includes the following elements:
1) Customer Jobs: The tasks, problems, or needs that customers are trying to address.
2) Customer Pains: The negative experiences, obstacles, or risks associated with the jobs customers are trying to fulfill.
3) Customer Gains: The outcomes, benefits, or positive results customers expect or desire from fulfilling their jobs.

Value Map: The Value Map represents the organization's value proposition, which is the unique combination of products, services, and experiences that create value for customers. It consists of the following elements:
1) Products and Services: The specific offerings provided by the organization.
2) Pain Relievers: How the organization's offerings address and alleviate customer pains.
3) Gain Creators: How the organization's offerings generate customer gains and deliver positive outcomes.

The interaction between the Customer Profile and the Value Map is crucial in creating a compelling value proposition:
1) Fit: The goal is to align the pains and gains of the customers with the pain relievers and gain creators offered by a solution. By addressing customer needs and desires, organizations can create value that resonates with their target audience.

2) Differentiation: The Value Proposition Canvas helps identify unique aspects of the organization's value proposition compared to competitors. This differentiation can be achieved by emphasizing certain pain relievers and gain creators that set an organization apart.

3) Innovation: The canvas encourages innovation by identifying opportunities for creating new or improved value. By understanding customer jobs, pains, and gains, organizations can develop novel approaches to addressing customer needs and providing superior value.

Where the VPC falls short

The Value Proposition Canvas is a framework to structure this process of finding alignment or "FIT" as the authors call it, but it does not seem to support the context of new and emerging technologies.

If the designer does not know how some (new) technology can address a customer pain, how can they find out what outcomes can be achieved and what value propositions and product/features to develop ?

This was evident in the team at Scitodate as well, with the designer & product manager finding it challenging to understand what potential the engineers were seeing in LLMs. That made it difficult for them to propose how LLMs could solve customer challenges they were aware of.

I had used the Value Proposition Canvas in my process to first map out the customer understanding from my research, then ideated how LLMs could be used to address them, and converted the answers to that into product features and customer outcomes. That nudged me to modify the value proposition canvas to map out this ideation process that I currently did in my head.

Overcoming that limitation

How can we address a customer pain by building a solution that utilizes some LLM ability ?

I tried adding a technology "layer" in the middle of Customer and Product blocks. I was trying to make the Technology a bridge between customer challenges and how the designed solutions can address them via the abilities of the technology.

By "abilities", I want to refer to generalized descriptions of what LLMs can do. They are similar to what Norman(2002) calls "actual affordances". This was comparable to Yang et. al.'s (2020) observation of Designers having an abstract understanding of AI capabilities in order to design using them.

After making these changes to the Value Proposition canvas, I tried to fill it in to represent some of the ideation I'd previously done to find the LLM value propositions for Scitodate.

Here the central diamond has 4 blocks for 4 LLM abilities that we at Scitodate were exploring to find new opportunities. Sticky notes that mention ideas leveraging an ability would go into any of the blocks.

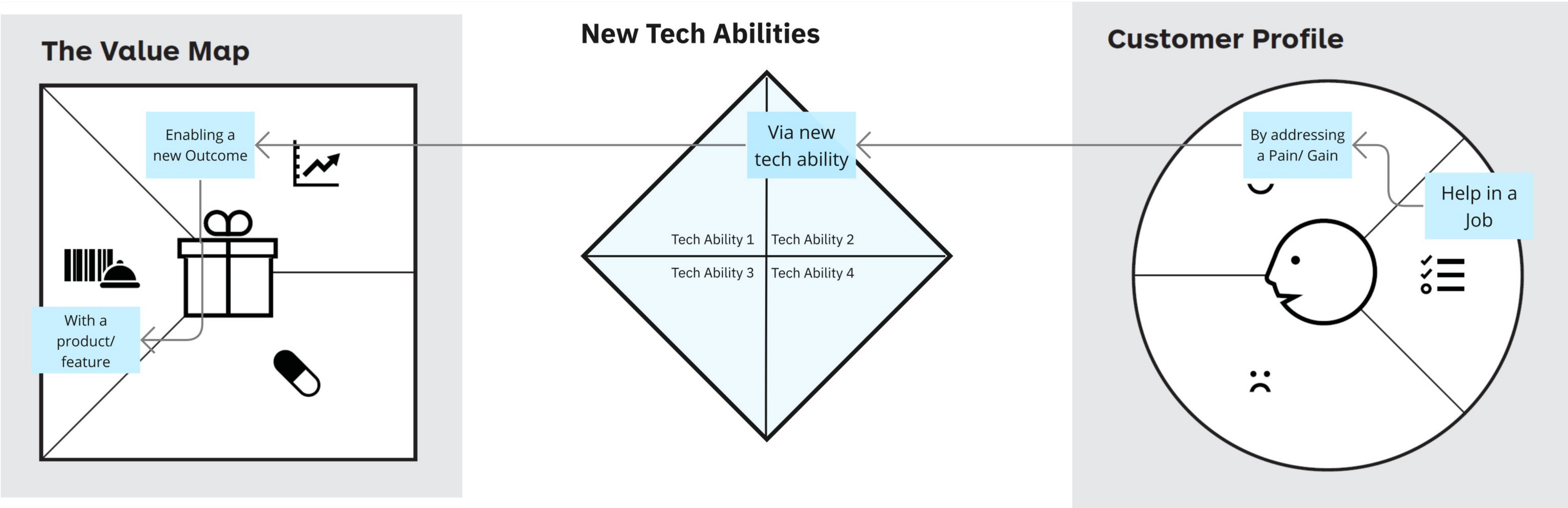


Fig. 24 :Modifying the Value Proposition Canvas to include technology

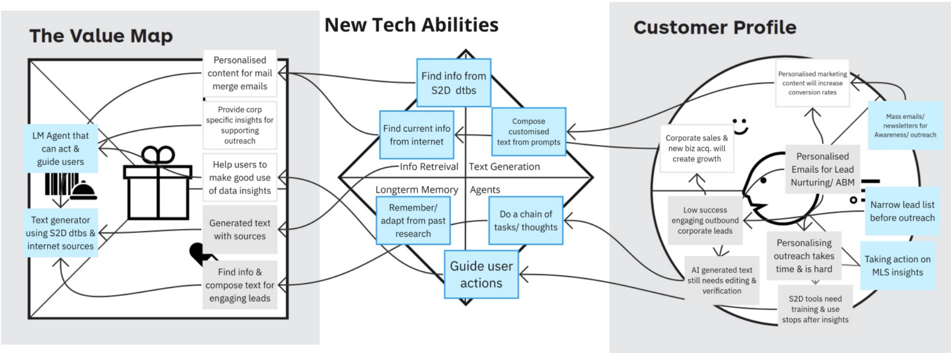


Fig. 25 : Modified VPC filled with data from empirical research

After I had this initial visual representation of my thought process mapped out, my approach for the next steps of the design and validation was quick iterations and testing, while referring to similar solutions and doing more research on how LLMs were evolving over the coming weeks.

Looking at the rapid pace of developments and the trajectory over those weeks, I was fairly certain that new abilities will emerge in the future for LLMs. I tried to modify the diamond to make it independent of specific abilities. I framed it as “What opportunities does (a new ability) unlock?” I kept 2 sections for opportunities to address pains and satisfy desires.

Testing propotype P1

We then did a trial session with the team at Scitodate to check how it worked. I facilitated the session and walked the team through the different steps.

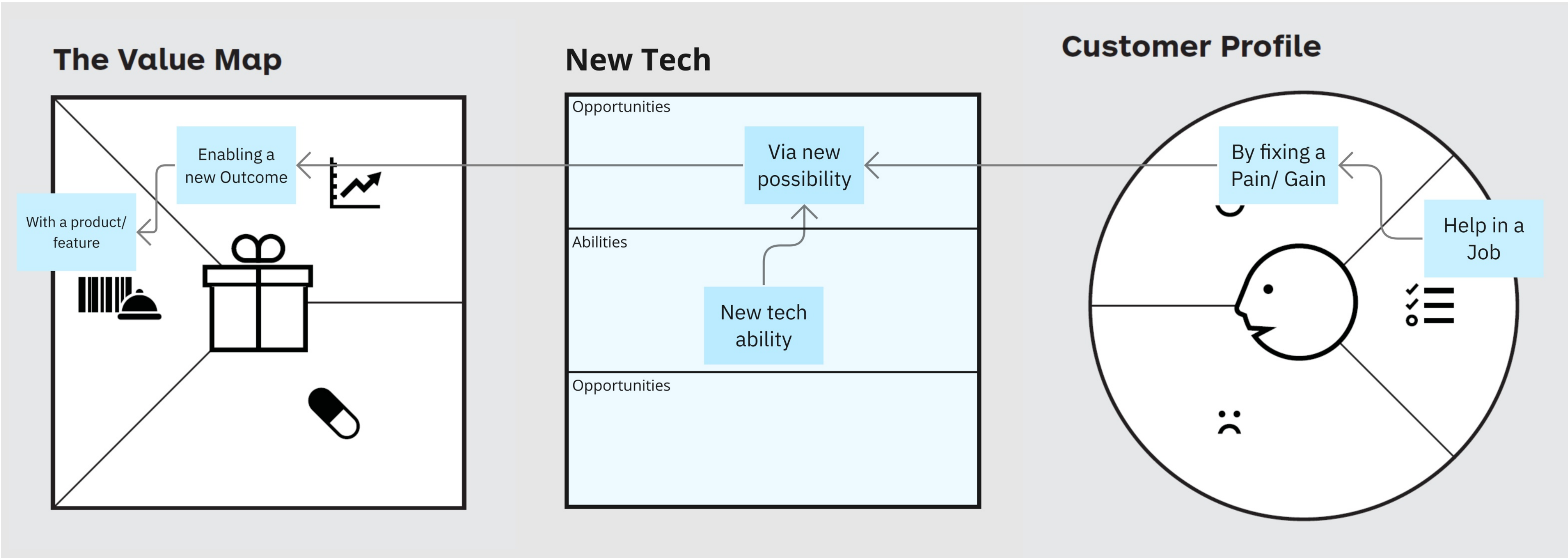


Fig. 26 : Prototype 1 ready to test

HYPOTHESIS :

Modifying the Value propositions canvas to reflect my own thought process does not stop it from being understandable to other people

The first testing session provided useful feedback about the flow of the process, the visual layout, what parts of the process were confusing or challenging to work through, etc. and what needed to be improved.

The hypothesis turned out to be partially correct. Because I was facilitating the process, the others could follow my chain of thought and work through the canvas. The critical feedback was that they would be clueless without my facilitation. They attributed that primarily to the lack of instructions in the Value Proposition Canvas as well as my hacked version of it.

Fig. 27 : Testing canvas prototype P1 (drawn on a roll of paper)

I also had discussions with other designers who've used different canvas and facilitated innovation sessions. That gave me insights about the scope the canvas should focus on and prioritise. I used all these insights to design a revised version of the canvas that was fairly disconnected from the Value Proposition Canvas.

Prototype P2

Using the feedback and reflecting on my own process I had followed previously, I made a new canvas from scratch, that was completely disconnected from the structure of the Value Proposition Canvas. The VPC only served as an initial test for finding a direction to take for the canvas design. That was needed because of the inherent limitations of the Value Proposition Canvas structure. Many of the feedback points like location of different elements, choosing completely different elements that were not in the VPC, aspects specific t o Foundation Models, etc. led to that decision.

A challenging aspect of how LLMs worked was that these models had a primary core ability that lead to a host of derived abilities depending on how the Language Model is prompted. For example, for the current Language Model APIs (OpenAI Platform) from OpenAI: Instruction following by generating the next characters after a prompt would be the core ability of the LLM and different abilities would then be summarizing text, generating emails, interacting with external software tools, etc.

I tried to differentiate the 2, hoping to help users get better clarity. Taking this approach was based on the premise that starting from the core ability of LLMs, the users of the canvas can explore different derived capabilities that they found interesting.

Considering the feedback about the visual layout and flow of the proces, I started the first steps of the process on the left upper corner and then kept moving towards the right. The kept 3 blocks to map out the required details about the customer. These formed the left side of the canvas. On the right, I added 3 blocks for relevant details about product development. In the center, I added 3 more blocks to represent the abilities of LLMs and a space to post ideas that utilized them.

I proposed the following steps for the process:

1) Finding/ mapping the customers' Jobs To Be Done
2) Finding their pains & desires associated with those Jobs.
3) Estimating the potential economic value of addressing thee challenges
4) Introducing the core ability of LLMs
5) Exploring the currently know capabilities of LLMs that this core ability enables.
6) Finding opportunities to use these capabilities to address the mapped customer pains & desires and support their Jobs.
7) Explore the potential future outcomes that customers should be able to achieve when LLM abilities can address their challenges.
8) Find out what work needs to be doen to build these new products or features and if Model fine tuning is required.
9) Estimate the resources needed to acheive the outcomes and do the required work.

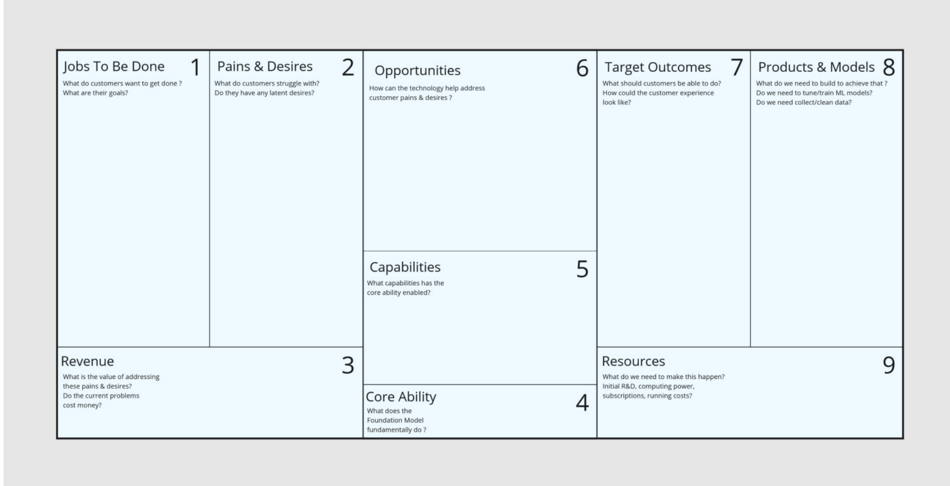


Fig. 28 : Prototype Canvas P2

Feedback on P2

HYPOTHESIS :

Making a canvas to reflect the underlying concept of LLMs, of having a core ability that leads to a variety of downstream capabilities will help designers to better utilise these capabilities.

From discussions with the team at Scitodate, we realised that the "core ability" approach was only making things more complicated and did not help the way I was expecting it to. That also made it necessary to have a Machine Learning engineer present during the brainstorming session to be able to explain this to the rest of the team.

I got similar feedback from fellow students at IDE. They could identify the blocks on the left & right but struggled to make sense of Capabilities & Core Abilities. The hypothesis was proven incorrect. One of the suggestions was to include this information about the models on separate ideation cards.

Introducing Ability Cards : A1

Boundary objects can serve this goal of conveying information and enabling collaboration. They can also carry this information in the form of “designerly abstractions” that Nur Yildirim et. al. (2022) claim, helped designers work with AI. The ideation cards, carrying information about different abilities of LLMs seemed to serve these goals quite well.

In the context of this design problem, designers & other non-tech users of the canvas need to be educated about the abilities of different Foundation Models & relevant model specific data. The canvas cannot include model specific details while still being relevant for a variety of domains & applications. Including that data on individual cards could help designers learn enough about the the technology and only work with the context relevant information. The cards can also help structure their thought process around the abilities of different Foundation Models.

Similar tools have been developed in the field or Machine Learning in the past that hold specific information about a topic or entity in such a way that helps teams design & build better solutions. Gebru et. al. (2018) introduced Datasheets for Datasets to help practitioners “decide, from reading a datasheet, how appropriate the corresponding dataset is for a task, what its strengths and limitations are, and how it fits into the broader ecosystem”. Mitchell et. al. (2019) introduced Model Cards for Model Reporting “to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited”. More similar solutions can be found in Hugging Face’s Landscape of ML Documentation Tools.

These “Ability Cards” can help not just help designers learn about Foundation Models’ abilities but also serve as a tool to spark ideation for possible applications. Along with supporting designers, a comprehensive compilation of Ability Cards can also help engineers cover all possibilities in the solution space during ideation. More importantly for a cross-functional team activity, the Ability cards provide a common language with “just enough” technical detail for facilitating conversations across diverse teams

DECIDING THE CARD CONTENTS

The initial plan was to have an Ability Card for every Foundation Model that the team could use. From the discussions and the questions raised during the previous test session, along with some desk research on other similar card decks that people used, I made an initial list of contents that the ability cards could have. That included the title of the Foundation Model, relevant specifications about the model, and its abilities and applications

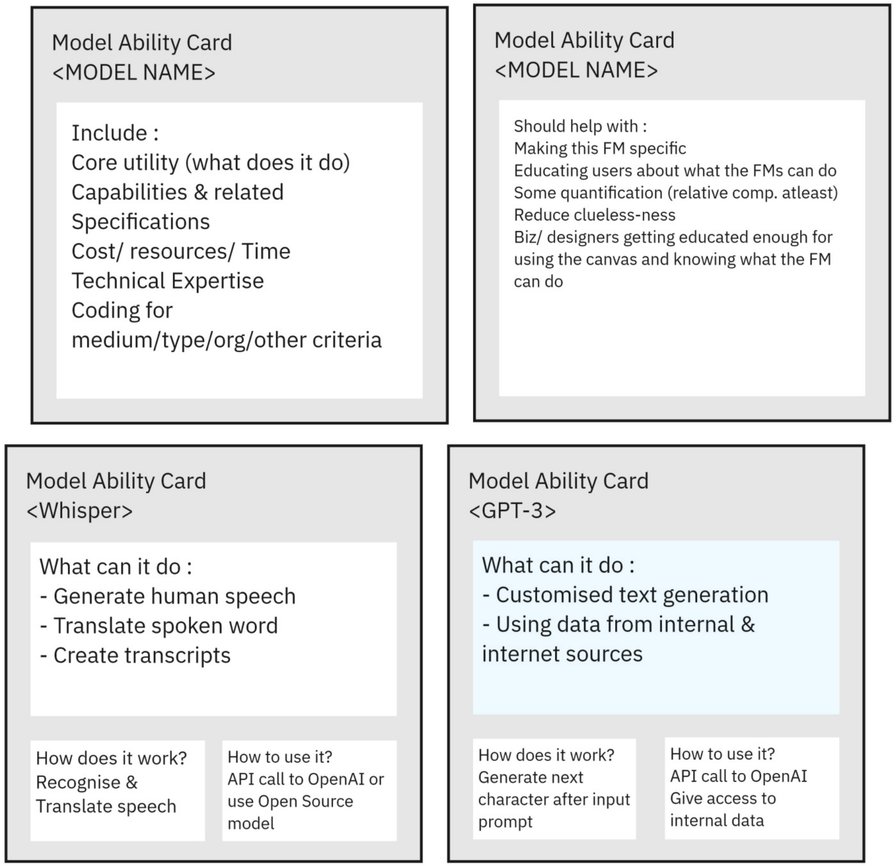


Fig. 29 : Initial drafts of Model Ability Cards A1

Prototype P3

Introducing these “Ability cards” seemed to help simplify the canvas by addressing the challenge of educating the user about the technology and guiding their thought process through the “technology layer”. They also enable making the canvas more robust and versatile to different types of Foundation Models & applications, as the context specific details and differences between models (text or image generators, classification models, etc.) can be included in the cards.

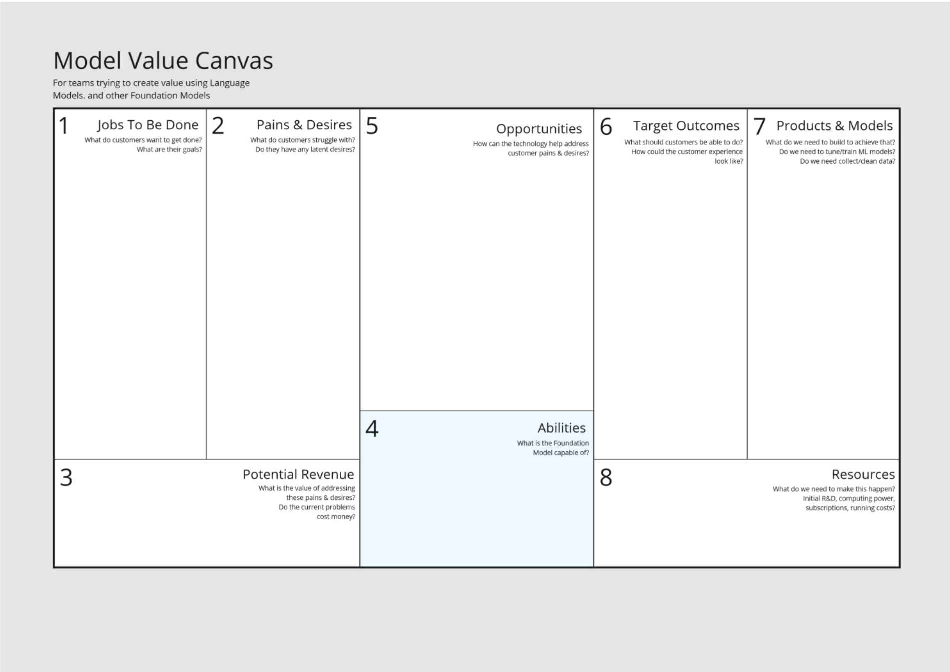


Fig. 30 : Canvas Prototype P3, simplified by transferring elements to cards

Testing P3

HYPOTHESIS :

Simplifying the way the canvas represents abilities of LLMs will address the technical complications and still help the team utilise LLM abilities.

Happy with the improvements with the canvas till now, we decided to have another brainstorming session at Scitodate, this time focussing on MirrorThink.ai, the team’s latest product that leverages GPT-4 from OpenAI. The team already had some ideas for adding new features to the current MVP but lacked a specific focus and clarity about the usecases. For the session, we focussed on a specific user group and their jobs & pains. As visible from the canvas image, we spent plenty of time on first deciding who the target persona can and should be. We didn’t use ability cards as they weren’t ready yet but the session gave good insights on what the cards should include.

The hypothesis proved to be correct. One of the outcomes from the brainstorm session was built into MirrorThink. Overall, the canvas was able to serve as an effective tool for guiding and structuring the discussion, especially considering the difference in alignment & focus before and after introducing the canvas into the discussion.

Fig. 31 : Testing Prototype P3 with the engineering team & CEO

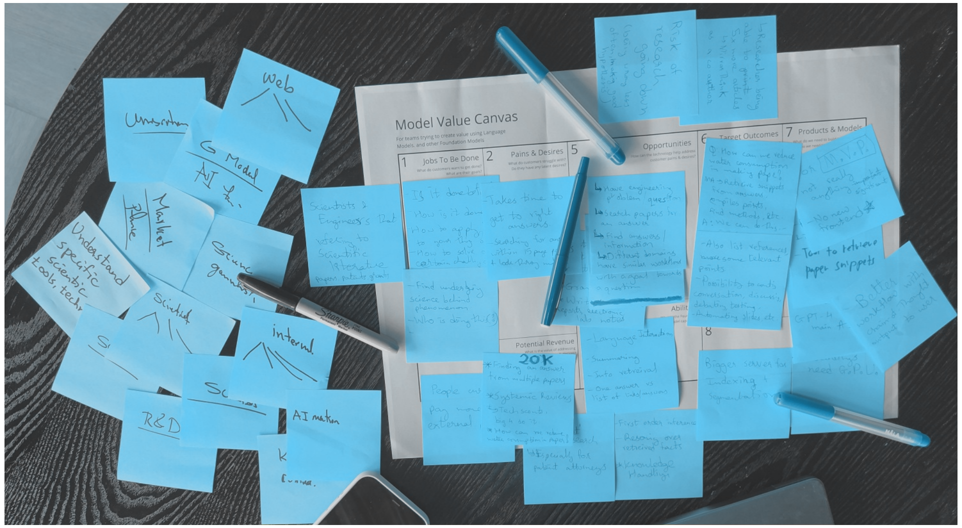


Fig.32 : Outcome of Test 2

We spent a large amount of time deciding on who the target user should be, before we could get to exploring their user goals and challenges. This test helped me to identify & add a field to specify the focus of the canvas: Who are we designing for?

DIY test with p4

I wanted to test how well the canvas worked without engineers, but with designers that have a basic understanding of LLMs.

To do that, I later tried using the canvas focussing on my own work of writing the thesis and leveraging ChatGPT's abilities to help me. I was able to use it to actually find new ideas for using ChatGPT. Also, helped me uncover some new questions about the "Abilities" approach to explore further. This served as an initial test for the canvas' robustness and adaptability.

I hand drew the canvas for this test. The aesthetic of this inspired me to modify the visual style of the actual canvas.

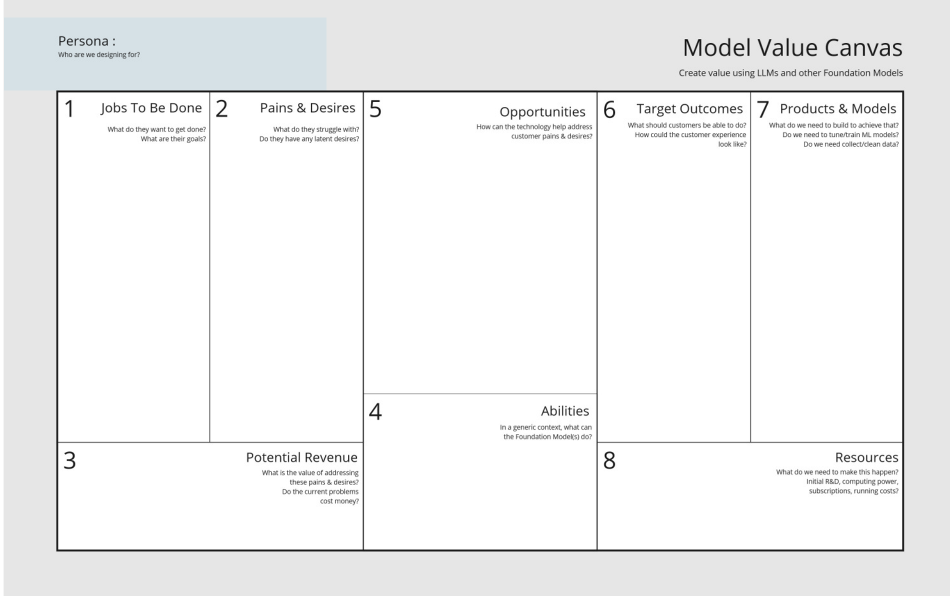


Fig. 33 : Canvas Prototype 4

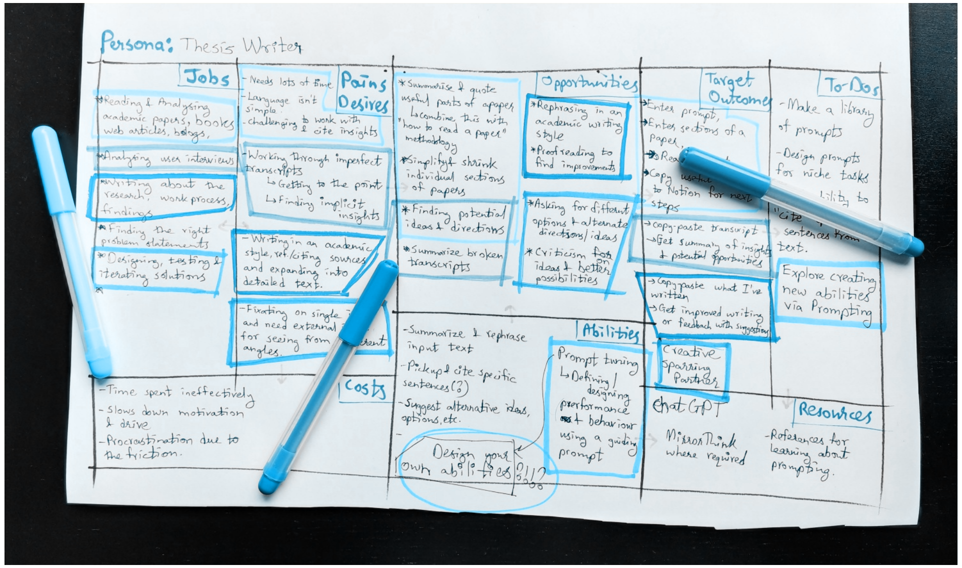


Fig. 34 : DIY test for Prototype 4 with handdrawn canvas

Prototype Ability Cards A2

After getting some initial feedback from the team at Scitodate on the sample cards, I decided to make significant changes to the original proposal for the cards. Instead of having one card for every Foundation Model, I decided to switch to having cards for different abilities. That way, every Foundation Model would have multiple applicable ability cards.

The main reason for making this switch was the possibility of describing every ability in sifficient detail so that designers could find sufficient insights from them. Another reason for making this switch was the simplification of the canvas itself. Because the canvas did not have the “Core Ability” block anymore, that could be eliminated from the cards as well. That meant every Ability Card for a Foundation Model only had a list of “Derived Capabilities”. That meant I could separate every ability into individual cards.

For the scope of this thesis, I wanted to have Ability Cards focussing on different affordances that LLMs have and disconnect them from specific Language Models. All Abilities will not be exhibited by all LLMs. For every LLM, some Ability Cards will apply and some will not, or not very strongly. For this thesis, I chose to make a few sample ability cards for GPT-4 from OpenAI (OpenAI, 2023) that Scitodate used in API form for running their own solutions.

I made some improvements and created ~12 Ability Cards about some of the most relevant capabilities of LLMs for Scitodate. We then tested the Ability Cards & Canvas together

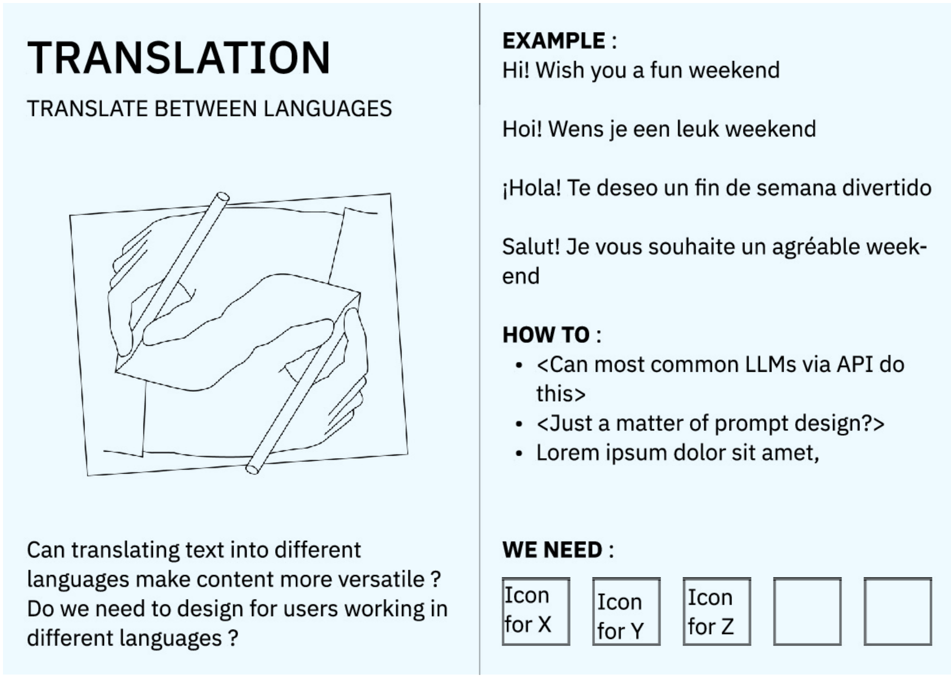


Fig. 35 : Draft Ability Card for Translation

Testing Prototype P5 : P4 + A2

HYPOTHESIS :

Ability cards will compliment the canvas and enable non-engineers to contribute equally to the brainstorm around using LLMs to improve products.

While facilitating this session, I introduced the cards to the team only after we were done mapping the customer Jobs and challenges. My intention was to prevent the card contents from fixating the team's focus on customer goals that were directly addressable in a straightforward way through LLMs

This session gave me more inputs for improving the contents of the Ability Cards. The session was effective in finding out if the cards were beneficial to the process. Although Jon (CTO) & Arthur (Full stack Engineer) were well aware of the possibilities of using, the cards helped Mehdi (CEO) come up with new ideas that were based on the abilities on the cards.

We decided to do more tests with the rest of the non-engineers at Scitodate to further validate the effectiveness of the Ability Cards.

Fig. 36 Testing Prototype P5 with the engineering team & CEO

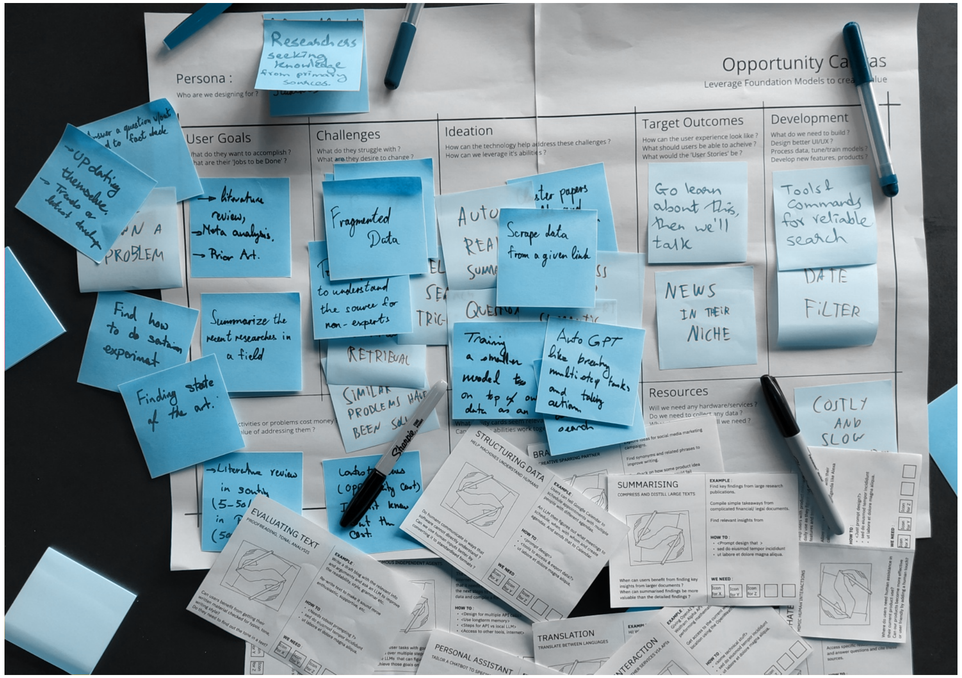


Fig. 36 Outcome of Test 3

Testing P5 two more times

We later did 2 more tests with different members of the sales, marketing and customer success team to find out if the canvas and card deck was effective for a variety of audiences.

HYPOTHESIS :

The current canvas + card can work for the rest of the team at Scitodate.

These tests helped me identify how different people interpreted the words on the canvas and cards differently, and how different people were able and unable to use the cards effectively. The prototype cards did not have enough visual elements and that prevent the participants from reading them in detail. Being busy with thinking about the prompts on the canvas prevented them from reading the text on the cards.

While the hypothesis did prove to be correct, there were still improvements to be made. It was time to get into optimising the details of the design of these cards and the canvas.

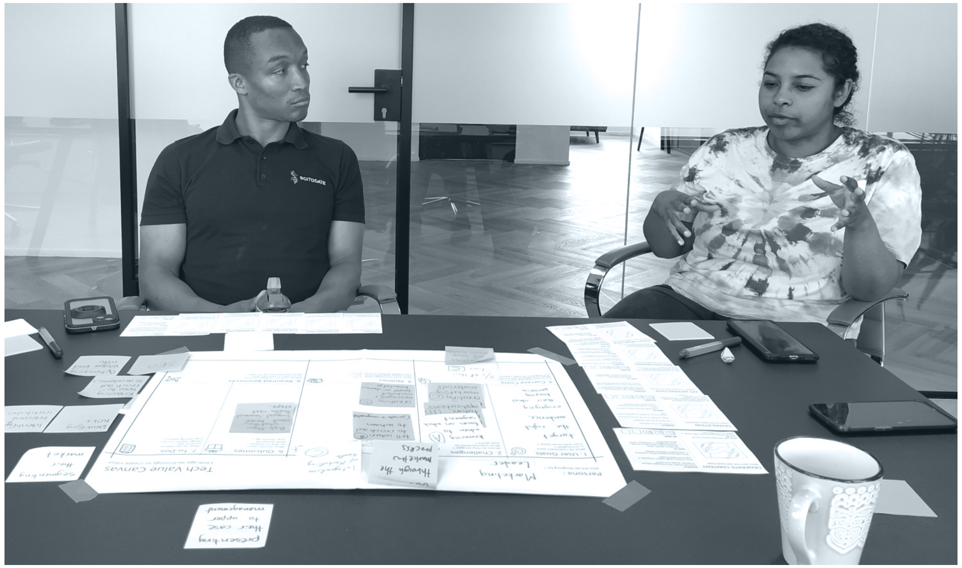


Fig. 37 : Outcome of Test 4

Fig. 38 : Outcome of Test 4. We ran out of time but the engineers later sorted out the To-Dos

Testing A2 card for radical ideation

From all the design research that was conducted before developing the canvas and cards, there were some promising customer problems that could lead to potentially radical innovations. We decided to ideate on some of them and use the Ability Cards to help us use LLMs to address those challenges. As it was a pure ideation session, we chose to skip the canvas. The session only had 2 designers : Me and Scitodate's product designer.

We chose three problems to focus on based on how promising they seemed for tackling with LLMs :

1) The current limitation of either creating quality outreach content or creating it for a large quantity of leads

2) The current products were not designed considering the normal way of working of customers. Users had to adjust how they worked to be able to use Scitodate's products.

3) Majority of the current users did not have a sales & marketing background and thus were not very good at it.

HYPOTHESIS :

The cards can support designers in ideating radical solutions .

The cards proved to be significantly useful during the ideation session. We started with exploring which cards were most relevant to the individual problems. That then led to further ideating how the solutions could look like.

As the ideas developed further, we used different Ability cards for different parts of the product ideas. By the end, we had gone throuhg all the ability cards and used multiple of them towards finding ideas to address the problems. In that way, they proved the hypothesis correct.

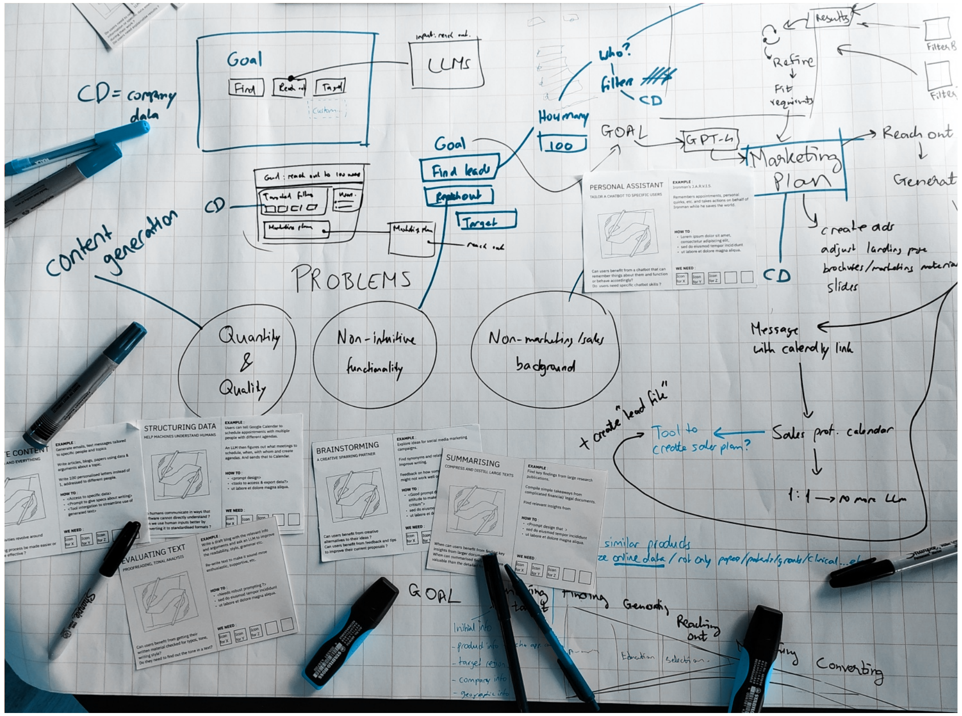


Fig. 39 : Outcome of Test 6. Cards worked well and Scitodate got multiple ideas for future

2.4.2 Detailed Design

After testing the initial prototypes of the canvas & cards to the point where the feedback & iterations were primarily about the detailed design features of the cards & the canvas, it was time to freeze the overall concept of the canvas + card setup, and get into the nitty gritty details of visual design, usability & ease of understanding.

Optimising the canvas

For the canvas, every new version had some visual improvements that were inspired by feedback from the previous testing session. Most of these improvements revolved around better word choice, adding numbers to guide the sequence of steps and including icons for every block. I shifted from the grey background style of the Business Model Canvas to a black frame on white background to simplify the visuals and make the box feel less restrictive. I also adjusted the frame borders to extend past the corners to communicate the sketchy draft vibe of early design ideation, where the ideas are still unrefined and evolving.

Optimising the cards

1) The postcard layout was helpful to refer to all the information at a time, but that was problematic when people wanted ot take a quick glance at an ability. There was too much text on the cards.

2) The need to read the text on the cards also made them difficult to use during brainstorming discussions. Participants were too occupied with the conversation

3) Because the cards currently looked too similar, differentiating between them was difficult. Replacing th eEscher placeholder with a unique icon would fix that.

4) The steps I'd included in the "how to" section a few weeks ago had already become outdated. I could understand that even if I updated them, they would again get outdated soon. So I decided to eliminate that completely.

5) Although the icons for the requirements could've been a good visual touch, they meant extra customisation of the cards. To help others understand what every icon meant, I would need to include an "explainer" sheet as well. I decided to revet back to text.

Visual refinements to the cards

1) I decided to move from a postcard layout to 2 faced cards. The front face can have minimal text and the back face can hold the details.
2) Added a large square icon to different blocks in the canvas
3) Converted the prompting questions into suggestions for when to use an ability, moving that to the back of the card
4) I added a shade of blue to make the cards easy to identify.

Word choice & jargon

Different words mean different things for various groups. Multiple iterations of testing & feedback on the canvas & cards helped get to simple jargon that conveyed the point and was interpreted in the same way by designers & engineers. Although finding the correct phrases and ontologies(Uschold & Gruninger, 1996) , that Osterwalder (2004) focussed on was extremely relevant and valuable here, the time constraints of the thesis prevented me from investing more time in refining the terminology I used on the canvas and cards

2.4.3 Design Validation

The validation criteria & process were significantly limited by the scope of this thesis and what was possible in the available time. The primary validation criteria for the tools was to ensure that they contributed positively to Scitodate's efforts to build LLM powered products. A secondary criteria was that designers should easily be able to use these tools. I tested the entire toolkit with fellow design students at a later stage in the thesis.

Tests with a different members of the team at Scitodate helped evaluate the utility & effectiveness of the approach. Outcomes from the various brainstorming sessions that used the canvas & cards were able to contribute to the team's product development plans. These brainstorming sessions also helped uncover & emphasise previously neglected customer challenges.

The development plans that resulted from these discussions were integrated into Scitodate's monthly planning and "product backlog". Feedback from three other practitioners in the technology innovation space helped me evaluate the potential for using this beyond the scope of Scitodate & LLMs.

Despite these efforts and results, I definitely acknowledge that the amount of testing is not sufficient to claim validity of the process and design proposal. The proposal needs further "stress testing" with different teams and different Foundation Models to find out if it can work reliably and effectively in a variety of contexts and applications. Startups working with Computer Vision & image generation are promising future candidates.

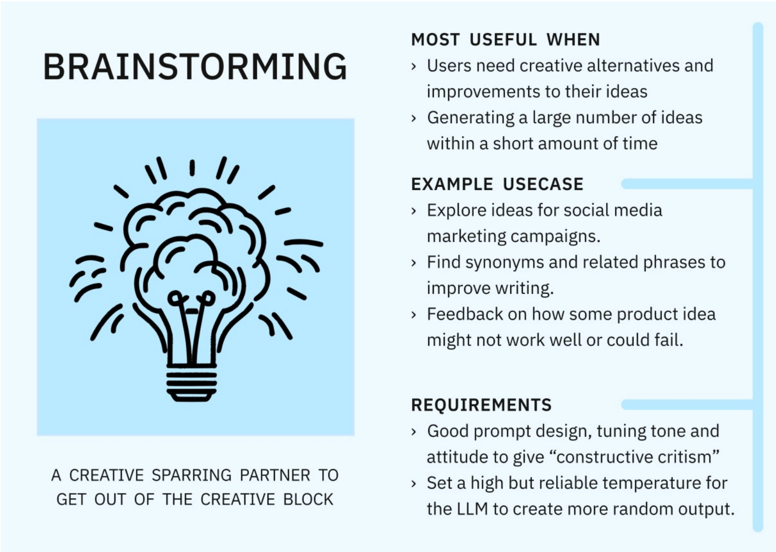


Fig. 40 : Final Ability Card Design

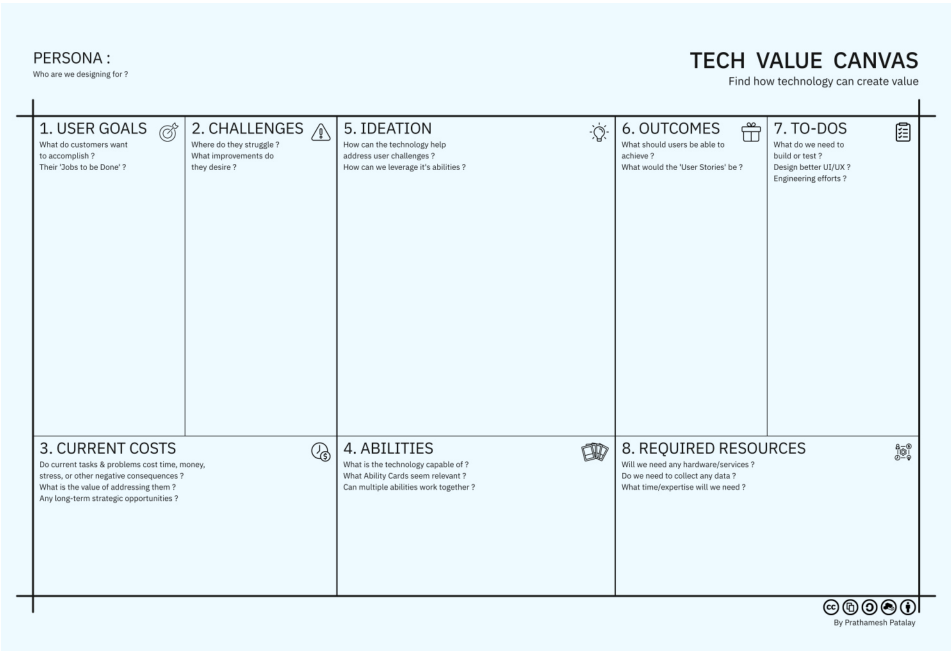


Fig. 41 :Final Tech Value Canvas Design

Evaluating Desirability, Viability, Feasibility

DESIRABILITY :

The team at Scitodate has been keen on doing each of the multiple testing sessions, and they've been able to get valuable outcomes from every discussion.

FEASIBILITY :

The information on the canvas and the cards enables a large variety of stakeholders and roles to participate in the process. These details also make it easy for designers to facilitate this process in the future. The instructions provided in this report in chapter 3 should help individuals follow the process on their won.

VIABILITY :

The sessions took an hour on average, depending on how much detail the team wanted to go into. That makes using this design and the process relatively time efficient, enabling busy teams at startups to benefit without spending alot of their precious time. More time can definitely be invested into better user research and ideation, and they should lead to better outcomes.

Validation from Scitodate

Here's feedback from Jon, my mentor from Scitodate for this thesis

"I think we've done quite a lot of sessions with different members of the team, which was a good stress test for the design itself to figure out how it behaves with different roles, but also for us to coordinate within the team, and every session has been useful and successful, I can give an example...

So last week we had a session with two members from customer success and sales. One of the objectives we have for next month was to make MirrorThink a bit better for current customers and we identified 2 pain points which we knew existed, but a lot more emphasis was put on them than we expected : <ABC> and <XYZ>. These are bigger bottlenecks than we thought, and from that, we adjusted quite a lot, the plans of how we will invest energy in the product team. Our customer success coordinated with the customers.

We've already identified 2 customers that have these cases that we are going to work on. These cases were discussed and we are now already working on them. A week later, we will deliver some results and I think in a couple of weeks, it's quite likely that this will move on to a paid pilot project. So every session has uncovered new priorities or adjusted priorities. And we can, from them, move on quite quickly to practical value. So this has been very valuable.

Responsibly

How can designers help startups mitigate the negative consequences of using Foundation Models ?

Contents

Summary

I used the literature review process along with insights from expert interviews to understand the broader picture and practical challenges of Responsible AI. That helped me refine the initial research question and then proceeded to explore current and past risk management practices.

That helped me guide the process towards designing a similar canvas & card deck to support designers in discovering potential risks of proposed products. A large part of the design process from the previous chapter was relevant and directly usable here, reducing the total amount of work required.

3.1 LITERATURE STUDY

3.1.1 Responsible Innovation

Science and innovation have led to both intentional and unintentional outcomes. Instead of only focussing on the aftermath and end result, which is less effective for uncertain future consequences, Owen et al. (2013) propose a shift in how we approach the challenge.

To accommodate the inherent uncertainty in innovation, they focus on two aspects of future responsibility: care and responsiveness. The crux of their proposal is responsiveness—a fusion of thoughtful analysis and discourse with impactful actions guiding an innovation's trajectory.

The authors argue that conventional notions like liability, accountability, and blame are insufficient, given the worldwide repercussions of technology in today's interconnected sphere. These concepts are retroactively applied and fail to sufficiently address the intricate ramifications of contemporary global technology or to prevent harm. They propose a different definition :

Responsible innovation is a collective commitment of care for the future through responsive stewardship of science and innovation in the present.

They suggest four dimensions to responsible innovation asking for a continual and joint commitment to be :

1) ANTICIPATORY, and analyze potential impacts in economic, social, environmental, and other areas. This can be done using methods like technology assessment and scenario development. These methods don't aim to predict the future but help identify possible issues and implications.

2) REFLECTIVE on the purposes, motivations, known information (including regulations and ethics), as well as uncertainties, risks, assumptions, questions, and dilemmas.

3) DELIBERATIVE, engaging with a wide range of people, including the public and diverse stakeholders, to consider different perspectives and address potential challenges.

4) RESPONSIVE by using the collective insights gained to guide the direction and speed of innovation. This process should be adaptable and open to learning from feedback.

While the framework they propose is broadly aimed towards governing science and innovation, private organisations could also potentially incorporate and benefit from these recommendations. Dignum (2019) agrees that Responsible AI means different things to different people. She proposes that depending on the people involved and the context, it can mean one of the following :

1) Policies concerning the governance of R&D activities and the deployment & use of AI in societal settings,
2) The role of developers, at an individual and collective level,
3) Issues of inclusion, diversity and universal access,
4) Predictions and reflections on the benefits and risks of AI.

This chapter explores how startups like Scitodate, and designers can practice responsible innovation while using Foundation Models in their products. It explores what responsibility might mean in this context and within the scope of the thesis, how can designers help startups innovate responsibly. In the field of mitigating harm from AI, there have been numerous contributions to ethics, regulation and risk based approaches.

3.1.2 Ethics Approach

AI Ethics across the globe

Over the past decade, the increasing capabilities and widespread use of AI systems have sparked discussion about the values and principles, that should govern their development and deployment. Numerous studies, including assessments of systemic risks, algorithmic bias, and discrimination, have explored ethical AI (Jobin, A., Ienca, M. & Vayena, E., 2019).

To address these concerns, national and international organizations have formed expert committees to create policy documents on AI. AI-reliant corporations in the private sector have actively participated in developing guideline. Professional associations and non-profit organizations have contributed by issuing their own recommendations and declarations.

The AI Ethics Guidelines Global Inventory database (Admin, n.d.) contained 167 guidelines at the time of this writing. Jobin, A., Ienca, M. & Vayena, E. (2019) found a global alignment on five such ethical principles regarding AI: transparency, justice and fairness, non-maleficence, responsibility, and privacy.

But,they also found significant differences in the interpretations these principles, their perceived importance, and the suggested strategies for putting them into practice. Multiple researchers have explored this challenge of interpreting principles and converting them into actionable practices.

AI Ethics Principles vs Practices

Morley et al. (2019) emphasize that the AI ethics discussion has primarily focused on ethical principles within AI, (the 'what') rather than the practical application (the 'how') of these principles. In order to help bridge this gap, they introduce a typology of different tools and how they relate to different principles. This typology is designed to assist developers in implementing ethical principles at various stages of AI system development.

However, they noticed an uneven distribution of effort and tools across this framework for 'Applied AI Ethics'. The authors also found that many of these tools lack actionable guidance, making their practical use challenging. Due to the limited effectiveness of these tools, they conclude that further refinement is necessary before these tools can be effectively applied in real-world settings. Without thorough testing in practical scenarios, the impact of these tools on the overall governance of the algorithmic ecosystem also remains uncertain .

STAGE -> AI PRINCIPLES	Business case	Design Phase	Training	Building AI system	Testing	Deploy	Monitor
Beneficence	Tool 1	Tool 4	Tool 5				
Non-maleficence	Tool2	A typology of how different tools support different ethical principles at different stages of product development					
Autonomy	Tool 3						
Justice							
Explicability							

Fig. 42 : showing schematic of typology proposed by Morley et al. (2019)

Organisations try to Self-Regulate

Majority of the large software companies like Microsoft, Google, Meta, IBM, OpenAI, Salesforce, etc. have all produced their own frameworks, principles and guidelines for ethical development and use of Artificial Intelligence. They also have designated teams to focus on developing and implementing these Ethical guidelines. Microsoft has also designed a training program to help other firms design their own AI strategy & principles for Responsible AI (Microsoft Learn). Despite all these efforts, their implementation & integration has not been effective so far.

Schiff et al.(2020) explore the gap between these high-level principles and the lack of clarity on how they can be implemented in organizational practices. They outline 5 possible explanations for this gap :

1) AI's social and ethical implications for human well-being are broader, more complex, and more unpredictable than we often understand
2) Accountability for ethical consequences is divided and muddled
3) Experts in different fields focus too much on their own specific ideas, or sometimes they look at things too broadly. This makes it hard for them to understand each other and work together.
4) Existing methodologies and tools for responsible AI are hard to access, evaluate, and apply effectively
5) Organizational practices and norms which divide technical from nontechnical teams minimizes the chance of developing well-considered AI systems that can safeguard and improve human well-being.

Startups struggle to prioritise AI Ethics

Bessen et al. (2022) found that many AI startups are aware of ethical issues related to AI, with 58% of the startups in their study having established AI principles to guide their operations. However, the extent to which they could act on these principles varies based on their available resources. Startups with prior experience, such as data-sharing partnerships and knowledge of regulations like GDPR, were more likely to translate these principles into practical actions.

They note that these startups tend to follow norms set by larger technology firms or those established within their customers' industries. These norms played a significant role in shaping the ethical approaches of these startups. They also found that more than half of the startups that adhere to ethical AI principles faced costly business outcomes as a result of their commitment to these principles.

Winecoff & Watkins (2022) observe a similar conflict. Although many of the AI entrepreneurs they interviewed emphasized the value of scientific rigor and methodological integrity, they faced a challenge when dealing with external stakeholders who might lack the technical expertise to understand the importance of these principles. These stakeholders often prioritize business considerations over scientific integrity, creating a tension for entrepreneurs.

They argue that while startups encounter greater resource limitations compared to the more established companies that have been the primary focus of applied AI ethics studies, they also represent a potential ideal opportunity for implementing ethical interventions.

In a multiple case study by Vakkuri et al. (2020) of startup-like environments, ethics weren't formally integrated using methods or tools, and ethical issues weren't directly considered as such. Instead, ethical matters were addressed only for practical reasons. Their study participants still worried about possible ethical problems tied to the systems, but lacked ways to address them.

AI Ethics approaches are not effective

Hagendorff (2020) argues about the ineffectiveness of ethical guidelines in the AI and machine learning field, claiming that these guidelines lack the necessary mechanisms to enforce normative claims. That leads to minimal influence on human decision-making processes. Despite their weakness, AI companies find them attractive as a way to suggest self-governance and deter the need for specific laws to address potential risks and abuse. (Calo 2017)

However, research shows that these guidelines can fail to impact professionals' behavior, and in practice, AI ethics can be treated as an optional, external concern rather than an integral part of technical development.(McNamara et al. 2018) This highlights the need for a more robust approach to AI ethics that goes beyond token gestures and truly integrates ethical considerations into the industry's practices.

Munn (2023) further argues that AI ethical principles suffer from several issues, making them practically ineffective. He asserts that these principles are meaningless because they are often disputed or unclear, making their application challenging. Additionally, they exist in a context where ethics is largely overlooked in the industry and education system. Moreover, these principles lack enforcement and accountability, aligning more with corporate interests rather than promoting genuine ethical behavior. As a result, Munn argues that AI ethical principles are essentially useless in addressing the harmful impacts of AI technologies.

In March 2023, Microsoft laid off their Responsible AI team while continuing to invest in integrating AI into their products. According to Microsoft employees, the Ethics and Society team had a vital role in ensuring that the company's responsible AI principles were effectively incorporated into product design. One former employee mentioned that when some individuals were unsure of how these principles applied in practice, this team demonstrated their application and established guidelines in areas where there were previously none. (Schiffer & Newton, 2023)

Self Regulation hasn't worked in the past

Floridi (2021b) highlights the failure of self-regulation in the digital industry operating over the internet. He narrates how self-regulation was hoped to improve digital industry and society communication, but it failed. Change became necessary, shifting from voluntary ethics to legal rules like the General Data Protection Regulation (GDPR).

The Facebook-Cambridge Analytica scandal and other negative events in the industry further demonstrated that self-regulation didn't work. Companies were unwilling or unable to fix their ethical problems at a fundamental level, beyond mere public relations efforts.

Floridi foresees a similar outcome with AI ethics. As the AI industry responded to ethical challenges by creating numerous guidelines, manifestos, etc. it became clear that self-regulation was falling short again. The efforts seemed more like superficial "blue washing," lacking substance and genuine commitment to ethical considerations.

In response, the EU is taking the lead in introducing legislation to address AI's ethical issues, recognizing the need for a more enforceable approach. This move marks a shift away from the voluntary approach towards legal frameworks that would hold AI industry players accountable for their actions.

3.1.3 Policy Approach

Enforcing ethics through Legal policies

In 2019, the EU's High Level Expert Group (HLEG) on AI presented their Ethics Guidelines for Trustworthy AI (2019) along with an assessment list of questions (ALTAI) to help organisations evaluate whether an AI system complies with the requirements specified in the Ethics Guidelines. The HLEG also published a document on Policy and investment recommendations for trustworthy Artificial Intelligence (2019). It contained their proposal addressed to EU institutions and Member States. These contributions formed the ethical framework for the EU AI Act.

Floridi (2021a) explores how the AIA aims to address and mitigate the risks associated with AI, foster public trust in these innovative technologies, and promote the development and adoption of AI within the EU. He finds this risk-based approach convincing, as it aligns with common practices in internal market-based legislation. It reflects the perspective that ethics should benefit the market and contribute to responsible AI development, rather than being detrimental to it.

The EU AI Act

The EU AI Act (AIA) aims to encourage responsible and trustworthy AI development in Europe, with the goal of making significant strides in AI advancement while maintaining alignment with the continent's ethical values and principles. The AIA will come into effect on a specific date in all 27 Member States and will have legal power that must be followed across the entire EU. (Floridi, 2021a)

The final legislation is expected to be published and come into effect by late 2023-early 2024 and companies are expected to have a 2-3 year transitional period over which they can acheive compliance. (The AI Act, 2022), (FRKelly, 2023) (Artificial Intelligence Regulation, 2023) (Schuett, 2023)

The AIA is expected to extend what is known as the "Brussels effect" (Bradford, 2020). This means that companies, even in other countries, might choose to follow EU regulations because it's more convenient to have a uniform approach worldwide. As a result, the EU's laws could indirectly influence international markets through market mechanisms, although not necessarily through official legal channels.

RISK CLASSIFICATION SYSTEM

At the core of the AI Act lies a risk classification system, which evaluates the potential risks AI technologies may pose to people's rights, safety, and well-being. This system categorizes AI into four risk tiers: unacceptable, high, limited, and minimal.

AI systems with minimal and limited risk, such as spam filters and video games, can be utilized with relatively fewer requirements, but transparency obligations must still be met. On the other hand, AI systems deemed to carry an unacceptable risk, like government social scoring or real-time biometric identification in public spaces, will be strictly prohibited, with only a few exceptions.

For high-risk AI systems, such as autonomous vehicles, medical devices, and critical infrastructure machinery, they are permitted but will be subject to stringent regulations. Developers and users must conduct thorough testing, maintain proper documentation of data quality, and implement an accountability framework with human oversight.

Furthermore, the AI Act addresses regulations for general-purpose AI, including AI systems like ChatGPT, large language model generative AI systems, which can be employed for various purposes with varying degrees of risk.

AMENDMENT FOR FOUNDATION MODELS

In May 2023, the MEPs (Members of the Europoean Parliament) proposed additions to the AI Act for providers of Foundation Models, (Europa, 2023) incorporating certain added responsibilities. These providers would be obligated to ensure strong protection of fundamental human rights, health, safety, environment, democracy, and the rule of law.

They would be required to assess and address potential risks, adhere to specific design, information, and environmental standards, and register their models in the EU database. The proposed Act now also includes obligations for downstream providers of Foundation Model based AI systems and other actors along the AI value chain. (Article 28 in the draft AIA)

For generative foundation models, such as GPT, there are additional transparency requirements to follow. These include disclosing that the content was generated using AI, implementing measures within the model to prevent the generation of illegal content, and publishing summaries of copyrighted data used in the model's training process. The draft released on 16 May 2023 (EU LEX) acknowledges the newness and uncertainty around Foundation Models (Article 60h) :

"Given the nature of foundation models, expertise in conformity assessment is lacking and third-party auditing methods are still under development"

"As foundation models are a new and fast-evolving development in the field of artificial intelligence, it is appropriate for the Commission and the AI Office to monitor and periodically asses the legislative and governance framework of such models and in particular of generative AI systems based on such models."

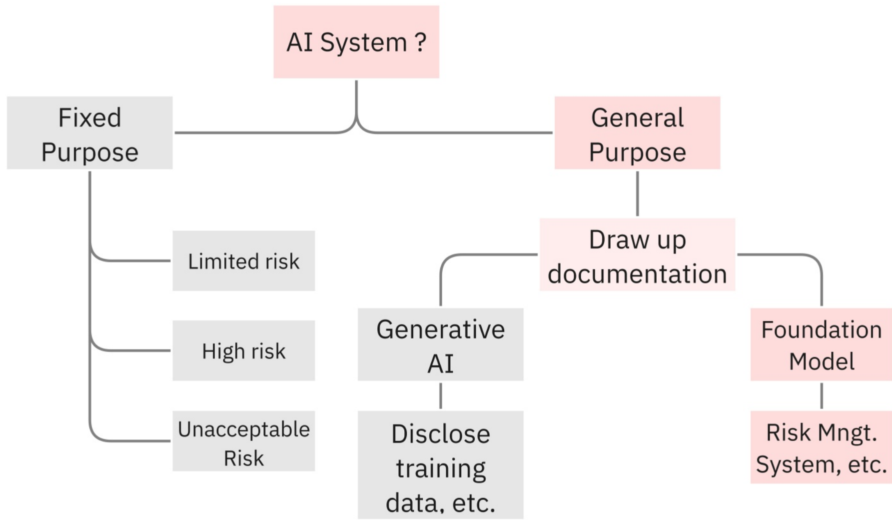


Fig. 43 Schematic of classification system in AIA, adapted from Hansen (2023)

RISK BASED APPROACH OF THE EU AI ACT

As we discussed previously, the AI Act categorises AI systems into 4 categories. High risk AI systems need to have a risk management system in place throughout their entire lifecycle. (Hansen Urlick, 2023)The same requirements for a risk management system also apply in the case of Foundation Models. While these requirements primarily only apply to "providers" of Foundation Models, (those who develop an AI system), downstream providers and deployers, also need to meet their own obligations(Walters, 2023). When downstream providers make "substantial modifications" to foundation models integrated into high-risk AI systems, they will be considered "providers" and need to establish similar risk management systems (Article 28 in the draft AIA)

As the AIA enters into force across the EU in the coming years, companies, big and small, that provide their AI powered solutions to the EU market will need to conform to the requirements set in the regulation. That includes setting up risk management systems and other internal processes to ensure that organisations tackle the risks that come with (Walters, 2023)

3.1.4 Risk Approach

AI Risk Management Frameworks

It is worth noting that at the time of this writing there aren’t any harmonised standards or standardised specifications on AI risk management yet.(Schuett, 2023) Harmonised standards will define the specific risk management actions referred to as the "suitable risk management measures" in the AI Act. They are tailored to align with EU laws, and compliance with these standards will imply meeting the fundamental requirements with a presumption of conformity with the Act.(Standard Setting, 2022) They will be drafted after the legislation is published.

NIST RISK MANAGEMENT FRAMEWORK (RMF)

In January ‘23, the National Institute of Standards and Technology (NIST), released the AI Risk Management Framework (AI RMF) (Tabassi, 2023), a set of voluntary guidelines to help organisations improve the trustworthiness of AI systems. (NIST, 2023) The Core is a central aspect of the framework and is composed of 4 functions : Govern, Map, Measure and Manage.

These high-level functions are further divided into categories, and subcategories, that are further broken down into specific actions and outcomes. These actions are recommendations, not a strict checklist or a set of ordered steps. The framework emphasises that risk management should be continuous and timely, spanning the entire lifecycle of the AI system. It also encourages diverse and multidisciplinary perspectives, including input from AI actors outside the organization.

To assist organizations in using the AI RMF effectively, the NIST also developed an online resource called the NIST AI RMF Playbook (NIST). It offers suggested tactical actions that organizations can apply in their own contexts to achieve the framework's outcomes. Both the AI RMF and the Playbook are voluntary, allowing organizations to use them according to their specific needs and interests.

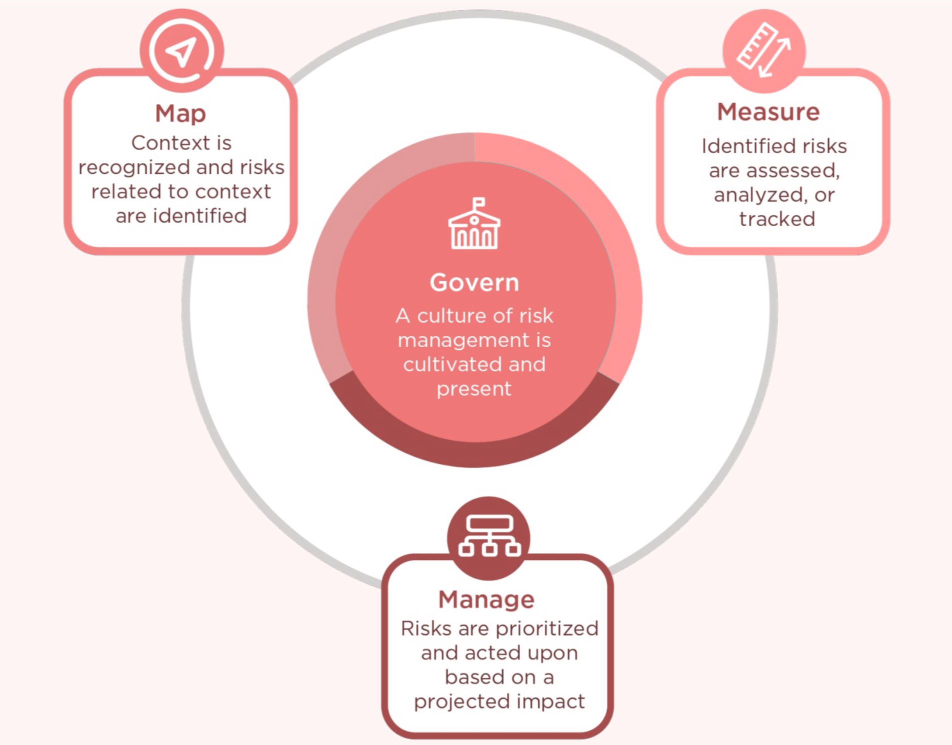


Fig. 44 The AI RMF Core : Functions organize AI risk management activities at their highest level to govern, map, measure, and manage AI risks. Governance is designed to be a cross-cutting function to inform and be infused throughout the other three functions. (Tabassi, 2023)

ISO/IEC 23894:2023

In February ‘23, The International Organization for Standardization (ISO) released a standard to provide guidance on Risk Management of AI systems. ISO/IEC 23894:2023 (ISO/IEC 23894:2023, 2023) suggests instructions on how businesses engaged in the development, production, deployment, or utilization of products, systems, and services incorporating artificial intelligence (AI) can effectively handle the associated risks.

It also intends to support organizations in integrating risk management into their AI-related operations and functions. Additionally, the document outlines processes for efficiently implementing and integrating AI risk management. The document also includes examples of how the recommendations can be implemented at different stages of an AI system's lifecycle.

Like many other ISO standards, this guidance is adaptable to suit the needs of any organization and its specific circumstances. It builds on top of, and references ISO 31000:2018 (ISO 31000:2018, 2022), a more versatile risk management standard. They are recommended to be used together, making use of AI-specific guidance from ISO/IEC 23894:2023 along with more general guidance from ISO 31000:2018.

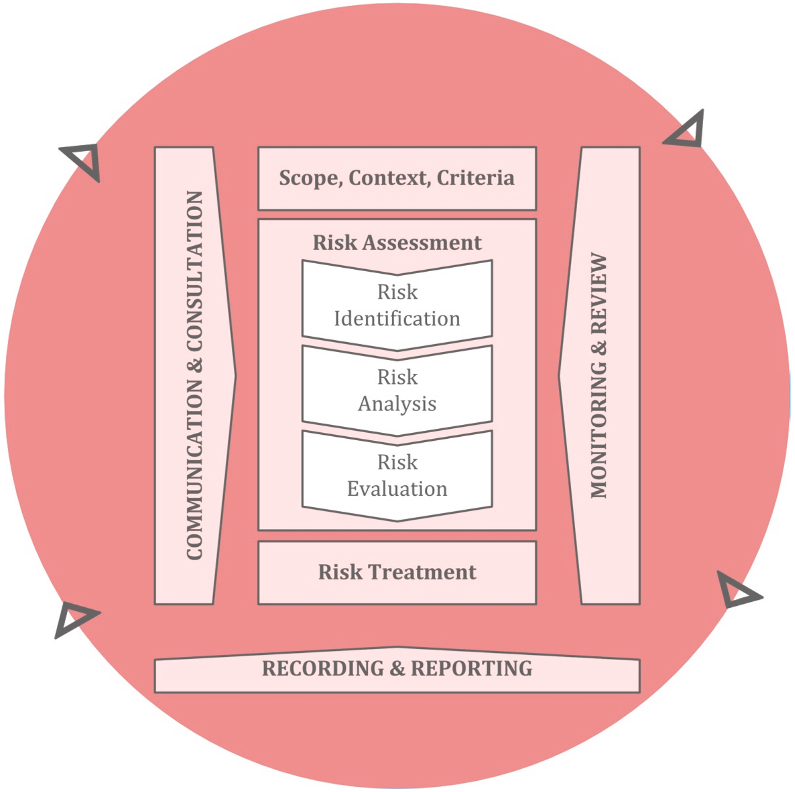


Fig. 45 showing schematic of ISO Risk Management framework (ISO 31000:2018, 2022)

Risks of Large Language Models

Bender et al. (2021) explored the potential challenges and risks of language models increasing in size, the way they have over the past 5 years. They discuss how LLMs' possibility of learning biases and harmful patterns combined with our tendency to naturally find meaning in text and believe it, creates real-world risks when LLM-generated text is distributed. They also studied the risks of using LLMs in classification systems that can categorise data, and the dangers of LLMs memorizing parts of their training data.

They urged the NLP community to understand that creating applications that try to convincingly mimic human behavior can be very risky. Developing synthetic human-like behavior in AI should be seen as a significant ethical challenge and that it's essential to understand and predict their potential consequences on society and various social groups to prevent harmful outcomes. Weiginger et al. (2022) identify a list of such risks and compile them in into a classification of six risk areas.

CLASSIFICATION	HARM
Discrimination, Exclusion and Toxicity	Social stereotypes and unfair discrimination Exclusionary norms, Toxic language Lower performance for some languages and social groups
Information Hazards	Compromising privacy by leaking private information Compromising privacy by correctly inferring private information Risks from leaking or correctly inferring sensitive information
Misinformation Harms	Disseminating false or misleading information Causing material harm by disseminating false or poor information e.g. in medicine Leading users to perform unethical or illegal actions
Malicious Uses	Making disinformation cheaper and more effective Facilitating fraud, scams and more targeted manipulation Assisting code generation for cyber attacks, weapons, or malicious use Illegitimate surveillance and censorship
Human-Computer Interaction Harms	Anthropomorphising systems can lead to overreliance or unsafe use Creating avenues for exploiting user trust, nudging or manipulation Promoting harmful stereotypes by implying gender or ethnic identity
Automation, access, and environmental harms	Environmental harms from operating LMs Increasing inequality and negative effects on job quality Undermining creative economies Disparate access to benefits due to hardware, software, skill constraints

Fig. 46 Six LLM risk areas adapted from Weiginger et al. (2022) & Derczynski et al. (2023)

Focussing on the broader scope of algorithmic systems, Shelby et al. (2022) present a similar taxonomy of socio-technical harms. They argue that such taxonomies can be useful tools to understand and evaluate the social and technical harms caused by algorithmic systems. But since these harms involve both social and technical aspects, they can't be fixed solely by technical solutions. From their analysis, addressing these issues requires a comprehensive approach that considers both the social and technical dimensions. They also discuss how such a taxonomy could support the anticipation of such harms and therefore help in avoid\mitigate them.

THEME	SUBCATEGORY
Representational Harms	Stereotyping, Demeaning Social Groups Erasing Social Groups, Alienating Social Groups Denying People Opportunity To Self-identify Reifying Essentialist Social Categories
Allocative Harms	Opportunity Loss, Economic Loss
Quality-of-service Harms	Alienation, Increased Labour, Service Or Benefit Loss
Inter- & intrapersonal Harms	Loss Of Agency, Social Control, Privacy Violations Technology-facilitated Violence Diminished Health And Well-being
Social System/societal Harms	Information Harms, Cultural Harms Political And Civic Harms, Macro Socio-economic Harms, Environmental Harms

Fig. 47 Categories of harm adapted from Shelby et al. (2022) & Derczynski et al. (2023)

Derczynski et al. (2023) build on top of this research on risk classification to propose Risk Cards. They propose Risks Cards as a way of providing a generic framework for assessing LLMs in different scenarios. They propose categorizing risks into RiskCards by explaining i) who might be affected when the risk manifest, ii) what type of harm could occur, and iii) what conditions are necessary for this harm to actually happen.

Actionable Risk mitigation tools and their limitations

Over the past years, multiple tools have been designed to aid practitioners in grappling with these challenges of working with Machine Learning models and LLMs. A particularly popular approach has been documentation tools that can contain relevant information about different aspects of AI systems. (THE LANDSCAPE OF ML DOCUMENTATION TOOLS) Practitioners have used the information in such documentation to make design & engineering decisions during the product development process.

Hugging Face, a platform that provides tools and a library of ML models to support building ML applications, uses Model Cards (Mitchell et al., 2019) to document information about the ML models they host. They also provide a template (Huggingface, GitHub) and a Guidebook (Model Card Guidebook) for people & organisations in the open-source community that upload their own ML models, so that they can document information about those ML models. That information helps other people who want to use those ML models make informed decisions.

Adkins et al. (2022) propose Method Cards, that don’t provide descriptive information like Model Cards, but focus on taking a prescriptive approach, providing guidance on using ML development methods properly while defining and training such models. According to them, descriptive approaches are beneficial for product developers and external experts as they can evaluate whether the ML system meets their needs. However, these approaches may not be as practical for other stakeholders. ML engineers, in particular, often require specific guidance on how to address possible ethical shortcomings of AI systems.

Wong et al. (2023) analyse multiple AI Ethics toolkits to evaluate how they support practitioners. They observe that many such toolkits focus on the technical roles in teams, but they do not address other challenges, like identifying & engaging with relevant stakeholders. They also do not include ways to engage with non-technical expertise and external stakeholders.

The authors also argue that the content and instructions offered by toolkits, along with the metaphor and structure of "toolkits" as a primary approach to AI ethics, shape specific perspectives on the world. This includes defining what qualifies as an ethical issue, determining who is responsible for addressing these problems, and establishing the acceptable methods for handling them. That makes the design of these tools a significant factor in the success of AI Ethics practices.

While not very popular, multiple tools have been designed by designers for designers working in AI. These tools primarily focus on AI ethics, similar to how designers have also primarily focussed on ethics, and not as much on legal requirements and risk management. (Chivukula et al., 2021) Some of them are explored in greater detail in a latter section. (3.4.3)

3.1.5 Findings

Summary

A significant amount of work has been done across academia and industry to take an Ethics approach to avoiding the negative consequences of using AI. Because Ethics principles have been considered too broad and vague to be actionable, researchers have proposed tools and practices that can support practitioners in applying ethics principles into product development.

Despite all these efforts and companies defining their own principles and practices, the Ethics approach to Responsible AI has not been effective in achieving the intended outcomes. Scholars cite a lack of enforcing mechanism, accountability, clarity, and a variation in their interpretation across different contexts as possible reasons. Self regulation as an approach towards adoption of ethical practices has not been effective in other domains in the past and the observation in the field of AI has been similar.

With the EU AI Act, the European Union plans to enforce ethics and responsible practices through legislation. The Act revolves around a risk based classification system that subjects different AI systems to different requirements depending on their risk characteristics. The Act has special provisions for developers and providers of Foundation Models, and requires them to be safeguarded via a risk management system.

While no official harmonised EU standards are drafted yet, organisations like NIST & ISO have proposed guidelines that organisations can voluntarily adopt. Large Language Models come with a variety of significant risks. Scholars have tried to identify, compile and categorise them into taxonomies, in order to help the community be aware of them and mitigate them.

Multiple tools have also been designed to support practitioners and researchers that document information about different aspects of an AI system and prescribe methods to guide ethical development of such systems. But a majority of the popular proposals are directed towards technologists and other actors dealing with the technological development of AI systems.

Relevance of findings to this thesis

The EU AI Act is a promising starting point to focus on. But considering the lack of well defined standards to comply with it, it's challenging to explore directions that can help startups comply with the Act. Although that is the case, the underlying risk based approach, with a focus on risk management can be a more robust angle to approach the situation.

While ethics has gathered alot of attention from practitioners and designers too, a lack of enforcing mechanism has prevented it from being effective. By overcoming that limitation, the AI Act makes Responsible AI practices more likely in the future. Compliance with the AI Act seems like a strong enough incentive for startups (and other organisations) to address its requirements and invest in responsible AI practices in the process.

Within the scope of risk management discussed in the Act, identifying and eliminating risks, and communicating residual risks seems like a relevant direction for designers to focus on. Although not directly relevant, designers could contribute to testing, post-deployment monitoring and informing users about residual risks through good design decisions.

Reframing sub-RQs

It is important to better understand how these practices currently work in the industry and what can be practically helpful to startups and designers to help them be responsible. That led me to reframe the initial research question for this chapter and decide the next steps :

1) What challenges do startups currently face in practice, in mitigating potential harm from AI solutions ?
2) How can designers support startups in that process ?

3.2 METHODS

For this part of the thesis, I chose to do semi-structured interviews with Responsible AI practitioners in the Netherlands. I analysed the findings using the Grounded Theory Method approach, to find common themes and decide the direction for the design process. The findings from the literature review and interviews led me to focus on the challenge of risk discovery.

I used desk research on risk management practices to develop a process for Risk Discovery and design a canvas and card deck to help designers follow that process. The tool design process carried over a large amount of learnings and insights from the previous chapter. The time limitations of this thesis also limited the amount of testing and validation I could do evaluate the process and artefacts.

Interviews with practitioners

I interviewed 4 practitioners working in the Responsible AI domain in the Netherlands. All participants worked in different roles in different kinds of organisations. The discussions were all semi-structured interviews conducted over video call, that were recorded and later trascribed. After a brief introduction about the thesis, I asked them questions about their work, feedback on my work and their observations from their practice about the thesis topic. In that way, all interviews did not follow the same interview guide. They all lasted 30-60 minutes. I analysed them via the GTM approach to code the interviews and find clusters and relevant insights.

Risk Assessment research

I explored a broad variety of approaches to risk management and assessment, and how they can be relevant to the context of AI risk discovery at startups and how designers can support this process. I found common factors and used them to decide design requirements for the tool & Risk Discovery process.

Process & Tool Design

The lack of an existing process at Scitodate and experiential know-how on my part made empirical work unfeasible. I instead chose to design the Risk Discovery process based on desk research and then iterate & test its desirability-feasibility-viability in the process of designing the supporting tool.

I designed the process and tool together. The tool design process complemented and supported the “process design” process. Majority of the findings from the previous chapter were relevant here, and I built on top of those insights.

3.3 PRACTITIONER INTERVIEWS

Interview Goals

1) Compare findings from the literature research with how things actually work and the challenges faced in practice

2) Find directions for next steps that resonate with practical challenges as well as the broader findings from literature

3.3.1 Interview Findings

Challenges with AI Ethics

Interview participants highlighted several key challenges. These include the difference between mathematical fairness and human perceptions of fairness, trade-offs between transparency and accuracy in healthcare, the need to balance fairness and privacy when using user data, and the limitations of universal ethical guidelines.

Additionally, participants acknowledged that discrimination might be necessary for ethical decision-making in certain contexts, and the evolving nature of values challenges the idea of constant ethical principles. Proving fairness in AI systems remains a complex task due to multiple mathematical definitions. Addressing these challenges requires interdisciplinary efforts and adaptable ethical frameworks.

Context dependence of AI Ethics

Participants emphasized the need to tailor discussions about AI ethics and fairness to the context of specific use cases. They highlighted the importance of assessing risks and involving stakeholders in decision-making processes to address potential biases and ethical concerns before deploying AI systems. Many participants echoed the sentiment that universal rules and guidelines for ethics may not effectively address the complexities of fairness in AI. Instead, understanding the context and perception of fairness becomes crucial to ensure just and equitable outcomes across different applications.

They also stressed the importance of balancing various ethical considerations, such as fairness, privacy and transparency. They noted that making trade-offs based on the specific application is essential to navigate the challenges posed by potential biases and their implications. For example, participants discussed the dilemma of addressing fairness in AI for scientific publications, where one approach involved considering whether to train AI models on historic data or start from scratch to avoid bias when selecting reviewers.

"INTENT" to be Responsible

A focus on ethics among the engineers, as well as at the CXO level, was considered crucial for fostering responsible behaviour. Some participants stressed the difficulty of convincing businesses to invest in AI ethics, as there are limited public examples of the impact of not doing so. One major challenge identified was the lack of accountability on the part of companies. To address this, some participants advocated for incorporating AI ethics into the company culture. By doing so, they believe that a strong culture within a company can replace the absence of laws or regulations in the AI space. As one participant pointed out,

“At the moment we just replace laws with culture, just enforcing a strong culture.”

One participant has integrated ethics into their venture studio's processes, starting from the assessment of new ventures to product design, ensuring that ethical considerations are woven into every aspect of their business. This cultural emphasis on ethics involved weekly huddles to discuss AI news and risks, encouraging everyone in the company to raise concerns and stay informed about potential ethical implications. By instilling ethics into the core of their operations, companies can take responsibility for their AI applications and contribute to the responsible development of AI technology. As another participant aptly summarized,

"You start off with the culture and you keep fostering it."

Support for Legal Regulation

Participants in the interviews expressed support for AI regulation in Europe, citing its potential to simplify operations for startups and provide clarity for companies using AI products. They emphasized the necessity of regulation to avoid pitfalls seen in software development and highlighted the need for a cautious and thoughtful approach when dealing with AI technologies. The idea of adhering to regulations voluntarily was seen as a means to demonstrate credibility and attract partnerships. Overall, there was a consensus that a legal framework for AI is essential for supporting responsible and successful AI development.

Often, Engineers deal with AI Ethics

The participants stated the need for alignment within the company, especially among the engineers and data science teams as they were often the ones taking the decisions around ethics and building the solutions. While technical requirements are relatively easy to execute, subjective requirements like fairness pose significant difficulties. Achieving fairness in AI involves balancing various ethical considerations such as privacy and making trade-offs based on specific applications, as mentioned one of the participants.

For example, addressing fairness in AI systems may involve deciding between training on historic data or starting from scratch to avoid bias, as illustrated by one of the examples provided. Proving fairness in AI systems is also found to be a challenging task, with multiple mathematical definitions used to assess discrimination, and fairness being significantly contextual.

Difficult to foresee things going wrong with AI Systems

“When something bad happens, you always think, yes, that was obvious, that it will happen. No, it's not obvious. It's obvious only after the fact.”

Participants highlighted the challenges in predicting AI technology risks, emphasizing its complexity and the possibility of unforeseen issues, even for imaginative and experienced individuals. Identifying all potential risks upfront, especially with founders focused on ideal customers, is particularly difficult. They stressed the importance of prototyping to understand the context and set boundaries for AI models' degrees of freedom. While early anticipation of risks is considered crucial and a cost-effective approach, it remains a challenging. Some participants expressed skepticism about the effectiveness of processes like premortems in predicting AI-related problems. Overall, there is a strong emphasis on anticipating the consequences of new ideas and solutions in the AI field to mitigate potential risks.

“What we can do is try to observe how it behaves and then interfere before the problem extrapolates.”

Need to monitor systems post deployment

“I don't think you can identify everything up front. And please don't do that for a start-up. As long as you keep monitoring and you have the right culture and the incentive that you wanna tackle this risk, then there's a lot you can fix along the way.”

Participants emphasized the criticality of monitoring AI product performance and user behavior to identify risks and unexpected consequences. They highlighted the importance of adaptability and responsibility in addressing challenges and changes. However, predicting human behavior's impact on AI development remains challenging. Participants advocated for agile development, pre-deployment testing, and human oversight to ensure responsible AI system management under the EU AI Act.

Proactive monitoring and continuous adaptability were recognized as key to efficient and responsible AI systems post-deployment.

“You need to monitor how your product is doing and there, the product manager needs to be made responsible to also check on what are the the implications of this product? Are there any problems that pop up? And if issues related to fairness or robustness pop up later then they should be able to, identify those by talking to users”

Education & awareness about AI technology

“How are you going to imagine what could go wrong if you never really had any experience with that type of technology?”

Participants highlighted the need to understand how AI models behave differently from traditional products. A participant pointed out that comprehending the degrees of freedom in AI models is not as intuitive as with conventional technologies, making education essential. Furthermore, integrating AI ethics into the company culture was highlighted by another participant, who mentioned their company's practice of holding weekly huddles to discuss AI news and risks, encouraging everyone to raise concerns. This approach fostered a more informed and responsible use of AI technology.

The interviews also revealed a specific case where a company sought advice from the Dutch Human Rights Board. They faced a dilemma in balancing fairness and privacy in their AI system, demonstrating the complex ethical considerations involved. To avoid potential pitfalls, a participant highlighted the danger of anthropomorphizing AI and blindly trusting it, as illustrated by the example of ChatGPT. Despite its limitations, people tend to unquestioningly trust the machine, even if its accuracy is not guaranteed.

Responsible AI in the startup context

“When you're building a startup, you know you're always tight on cash. So why do you want to spend money on thinking about all the risk. You just want to build and get money and then when you bump into something, you'll solve it along the way.”

Interviews highlighted the unique challenges startups face in incorporating AI while remaining financially agile. In the fast-paced world of startups, being financially constrained is a common reality. The participant from a venture studio pointed out that startup founders often face tight budgets and limited resources. In such a challenging environment, it becomes difficult to identify all potential risks upfront, particularly when founders are primarily focused on satisfying their ideal customers.

“The EU’s ALTAI assessment took us more than four hours. We were just laughing like this is not gonna be used by any startup if it's gonna take 4 hours.”

However, to ensure the responsible use of AI in startups, there is a need for pragmatic approaches that do not hinder their progress. Startups require AI solutions that are efficient, cost-effective, and do not slow down their growth. The pursuit of pragmatic AI solutions becomes essential to strike a balance between ethical considerations and business objectives. Startups can harness the benefits of AI while remaining responsive to their customers' needs and market demands by finding the right balance.

“Startups need something that's pragmatic, that's not gonna slow them down.”

3.3.2 Insights

Post product development activities have a significantly greater importance for AI systems considering their autonomous nature and our inability to anticipate all potential failure modes. Internal testing before deployment helps, but cannot capture all potential risks. The practitioners anticipate unanticipated risks to materialise.

Converting ethical requirements into technical specifications is a challenge for technologists, considering a potential lack of understanding the contextual complexity and human aspect of such socio-technical systems. Educating non-technologists about the actual capabilities, limitations and risks of AI systems is an equally critical aspect of the challenge.

Organisations need sufficient incentives along with their personal intent to prevent harm, for Responsible AI practices to be practiced. Along with that, the specific constraints of startups in terms of time, money & people add to the challenge.

Enthusiastic support for the upcoming EU AI Act confirms the findings from the initial literature review.

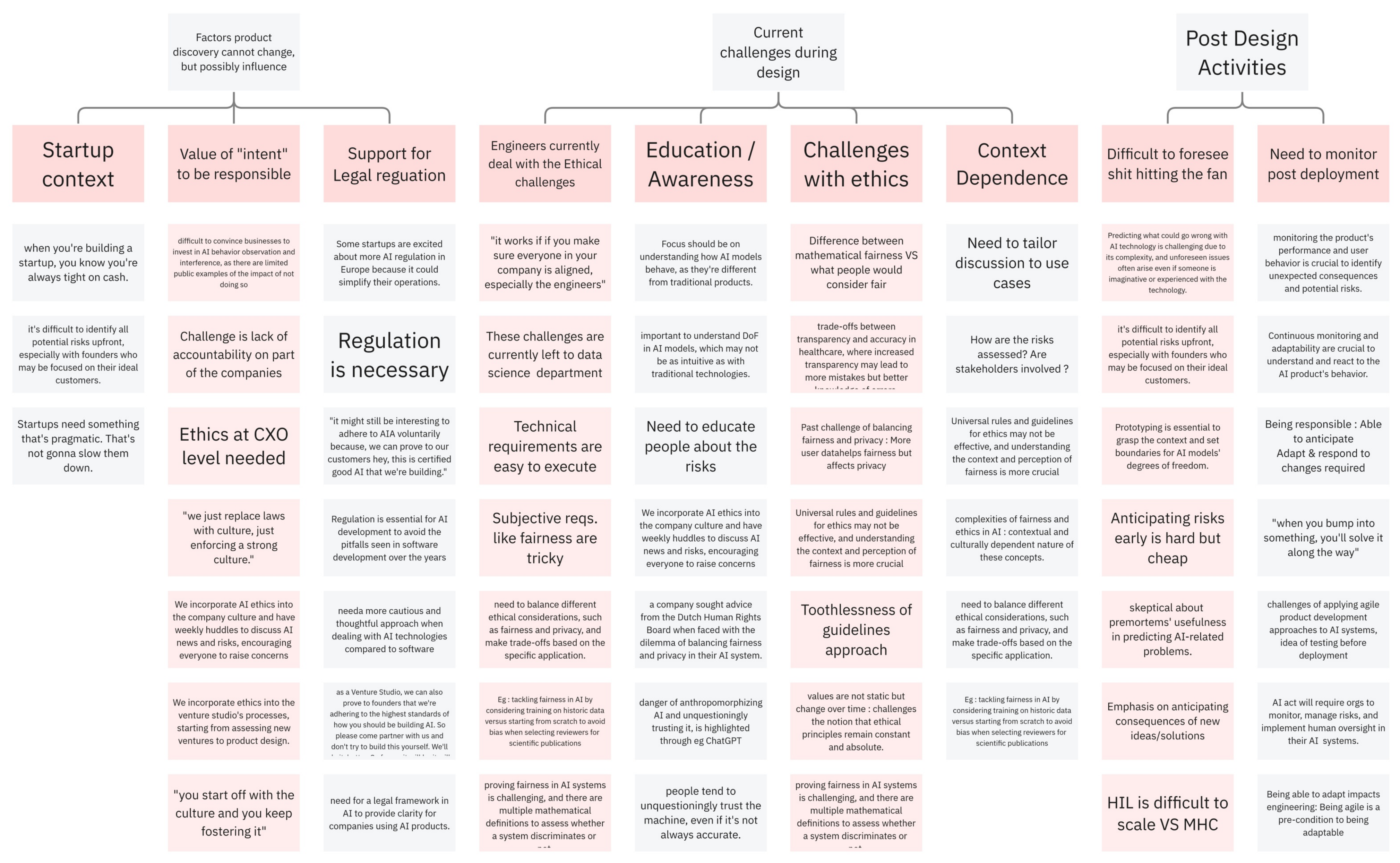


Fig. 48 Clustering of interview findings

3.3.3 Conclusions from Theory & Practice

Evaluating the findings from the literature review & interviews, and considering the scope of the thesis, the product discovery stage can contribute to helping startups practice Responsible AI in a few ways. The constraints of the startup context challenges of AI also imposes some requirements.

1) Startups do not have the time to invest in lengthy risk assessments or complicated risk management processes, especially in the early stages of product design & development. On one hand that implies potential for qualitative risk assessment practices that don't need investing in complex systems. On the other hand, such a risk assessment process will need to be time efficient and not require significant upskilling.

2) It is difficult to foresee all the ways how an AI system can malfunction after deployment. Because of that, monitoring these systems and taking action to fix them if they malfunction, is as important as pre-deployment efforts. There is potential to identify what aspects of these systems need to be monitored, what should be tested pre-deployment, and how designers can design the products to address potential risks.

3) Designers can design human-AI interactions to support fairness, transparency or other specific risks, but only if they are aware of the expected risks. If designers don't know what might go wrong, they cannot design in a way to mitigate harm. The product discovery phase can help identify potential risks, that designers can later address during the interaction design process.

4) To try and tackle the challenge of bridging ethical requirements to technical specifications, the product discovery process could aim to produce design & engineering requirements for product development. This is comparable to the bridging of customer understanding and AI capabilities from the previous chapter. This is also similar to the need for designer-engineer collaboration that was explored in the previous chapter.

5) The lack of awareness and understanding about the beahivour and risks of AI systems will be a challenge during Product Discovery too, and that is something tools could address. This challenge is similar to the understanding about capabilities that was addressed in the previous chapter.

6) The AI Act, while still being finalised, points to the future direction of Responsible AI requirements that organisations will need to meet. Product Discovery can play a part in supporting the initial stages of the risk management process, along with identifying what parts of the Act are applicable to the AI system being designed.

7) Organisations will need sufficient incentives to invest resources in managing risk during product discovery. While compliance with the AI Act, and reduced probability of product malfunction could be good reasons, exploring more incentives will only make adoption of Responsible AI practices more likely.

8) The scope of Product Discovery aligns with Risk Identification and Evaluation from the AI Act. When compared to the ISO/IEC 23894:2023, Product discovery aligns with the stages of Risk identification and Risk analysis. This gives a direction for the next stages of the design process.

Discovering what can go wrong, can be a means to identify how to address those concerns and what to watch out for. That in turn, can help teams take action to prevent risks from causing harm.

How to be Responsible ?

Referring back to Owen et al.'s (2013) four dimensions of Responsible Innovation, we can explore how that can apply to the scope and context of this thesis :

1) Anticipatory : Identifying possible risks of harm and analysing their potential impact

2) Reflective : Being thoughtful of uncertainties around post deployment behaviour of AI systems, the unknown risks, and known challenges & regulations

3) Deliberative : Considering the unique context of every challenge and involved stakeholders. Including relevant stakeholders in these conversations to discover risks from different perspectives.

4) Responsive : Being able to detect and ready to act quickly on unanticipated product behaviour post deployment

Referring back to Dignum's (2019) observation of Responsible AI meaning different things to different people, in the context and scope of this thesis, Responsible AI would imply the role of designers at startups in discovering potential risks and finding what can be done to eliminate & mitigate them. Designers can practice this while respecting the four dimensions of Responsible Innovation.

Considering the lack of risk focussed methods and tools for designers, I decided to propose a process and tool that would make it easier for designers to engage in "Risk Discovery".

RISK DISCOVERY

I intentionally wanted to use a phrase different from Risk Assessment, although they focus on almost the same aspect of risk management. That is because Assessment implies checking or evaluating. Discovery implies finding out the existence of something that was previously unknown.

I find this distinction important because of their difference in attitude and the challenge of dealing with emerging technologies. As identified in the interviews and literature, the lack of well defined protocols makes this process different from "assessment". Because alot of research is still needed to understand the risks of LLMs and other Foundation Models, we do not yet have standardised checklists to evaluate product performance against. We might not know how a risk might manifest in a new context. That makes this process exploratory, instead of evaluative. This framing led me to define the design problem(s).

Design Problems

1) How can I design a process to help designers at startups discover potential risks of harm and take early steps to address them ?

2) How can I design a tool to support designers to practice this risk discovery process during the Product Discovery stage of innovation ?

3.4 RISK ASSESSMENT RESEARCH

After framing the design problem, I used desk research to explore some risk management processes and ideas that can help. I also explored how different organisations practice good risk management practices. I intentionally explored a variety of topics, many of them noticeably distant from the context of this thesis, with the intention of finding common patterns and trends.

3.4.1 Risk Assessment Approaches

FMECA

`Failure Mode Effect & Criticality Analysis is a systematic approach to identify potential failures in a system, analyze their effects, and determine their criticality. It helps prioritize actions to prevent or mitigate failures. FMECA is used to improve reliability, safety, and performance while reducing the risk of failures and their consequences.`

FMECA generally involves the following key steps: (Borgovini et al. 1993)

1) Identify failure modes: List potential ways the system can fail.
2) Analyze effects: Examine consequences of failure like safety hazards or financial losses.
3) Assign severity ratings: Rate the impact of each failure mode.
4) Identify causes: Understand root causes for effective solutions.
5) Assess current controls: Evaluate existing preventive measures.
6) Assign occurrence ratings: Rate the likelihood of each failure.
7) Assign detection ratings: Rate the chances of detecting failures.
8) Calculate Risk Priority Number (RPN): Prioritize failure modes based on their ratings for higher risk.
   RPN = Severity of failure X Occurence X Difficulty of detection
9) Develop mitigation actions: Plan measures to reduce risks.

Following this process during the design and development phases helps engineers design and optimise systems so that they either have a low likelihood of failure, or low impact of failure, or high likelihood of the failure getting detected early. Such an analysis of a designed system helps teams find out what needs to be improved and optimised in the next design iteration. Iterations can be continued till the design reaches an acceptable RPN.

TAKEAWAYS FROM FMECA

1) Analyzing multiple factors related to a failure helps decide its priority for mitigation. The most critical failure modes can then be addressed first.

2) The criticality rating depends the likelihood of a failure occurring, it getting detected and the severity of its impact.

3) For failure modes that can have a significant impact and are relatively more likely, a monitoring system that can detect it early and reliably can be an acceptable approach to mitigate the risk of failure

Agile Risk Management

Agile is a popular software development philosophy used by organisations of different sizes, and common among software startups as well. (Pantiuchina et al. 2017) Moran (2014) explores how different risk management practices can be integrated into Agile software development and different methodologies like Scrum, XP, etc. To achieve that, he proposes the agile risk-management methodology. He explores Scrum as one of the Agile software development approaches and proposes how risks management practices can be integrated at the daily, sprint and release levels of the development process.

Moran proposes that Risk Scoping (identifying risk drivers and deciding the scope of activities to cover) can be integrated with the product release planning, alongside the product vision & planning. He proposes that Risk Analysis, Risk Burndown (tracking how well the team is tackling risks) & Risks Reviews (assessing risk management performance) can be done for every sprint. In this way traditional risk management processes can be integrated into a software startup's development process.

TAKEAWAYS FROM AGILE RISK MANAGEMENT

1) This approach aligns the iterative approach of Agile software development with different stages of risk management, making it easy to integrate risk management into existing software development practices.

2) It seems to be a promising way to introduce risk management to product teams that do not work on risk management. Slowly introducing one risk management process at a time can make the transition and integration easier and more likely to succeed.

3) Agile Risk Management can bring the product team in contact with the discussion around mitigating risks of harm.

DAILY	SPRINT		RELEASE
Daily Scrum	Sprint Planning	Risk Analysis	Product Vision
Product Development	Sprint Review	Risk Review	Product Planning
	Sprint Retrospective	Risk Burndown	Risk Scoping
	Backlog Grooming	Burndown Chart	

Fig. 49 Agile Risk management framework for Scrum style operations (Moran, 2014)

Antifragility

Antifragility can be best understood by comparing it to Fragility and Robustness. Antifragile systems benefit and improve from small amounts of volatility, shock, disorder or "things going wrong". Resilient systems only adapt to, or resist harm from such shocks, and do not benefit from them. Fragile systems can only be affected negatively from shocks, and need to be protected from them.

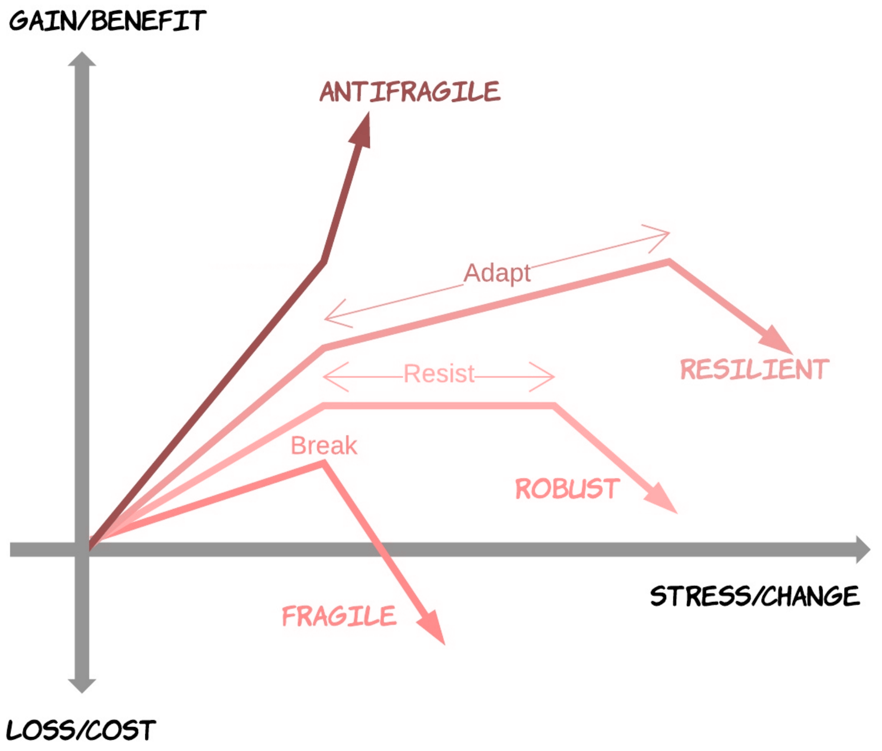


Fig. 50 Antifragility : The property of benefitting from small stressors. Ibryam (2023)

Taleb (2012) acknowledges that there are many events which are unlikely, unexpected and cannot be predicted. Their probability of occurrence is too small to be significant, but their impact is too large to be ignored. To design a system that is capable of dealing with such events, possibly even benefiting from them, he proposes that the system be designed such that it loves certain kinds of errors: the errors are too small to cause catastrophic harm, and they make the system better at handling future errors.

In the context of this thesis, antifragile systems would benefit from risks that materialise to a small extend without causing significant harm. Taking that perspective shifts the focus from trying to avoid risks to handling them in a way that systems benefits from them. When risks materialise and are mitigated without causing any harm, that can potentially make systems better at detecting and addressing them better in the future. That creates incentives for discovering these risks. That creates incentives for organisations to invest resources in risk discovery.

Don't avoid risks. Benefit from them.

With antifragile systems, being exposed to, and exposing these small hazards reduces the likelihood of more severe failures and also makes the system more capable of handling future shocks.

Aven (2014) argues that Taleb's concept of antifragility brings a valuable perspective to risk analysis. It emphasizes the dynamic nature of risk and performance, highlighting the need for variability, uncertainties, and risk to drive improvements and achieve high performance in the future. He claims that antifragility contributes to risk analysis by linking variation, uncertainties, and stress at the current/present context to the risks related to future performance. That emphasises continuous development and improvement rather than mere compliance at specific points in time. Antifragility suggests that exposure to stressors and uncertainties can lead to positive performance gains over time. Therefore, the focus then shifts from static risk assessments to considering how the system can evolve and develop in response to stress and challenges.

Monperrus (2017) discusses how software solutions and their development processes can be antifragile. He discusses how test driven development, combined with continuous deployment makes software systems antifragile. In test driven development, developers write automated tests for every feature they develop. With continuous deployment, features and bug fixes are released in production every day, or multiple times every day. Because of that, any errors that might emerge have smaller impacts. When such an error is found out, it can be rectified quickly before having catastrophic propagation.

Derbyshire & Wright (2014) discuss how the commonly used scenario planning methodologies rely on a deterministic view of uncertainty. These methods emphasise the causal forecasting of future events and the authors argue that this emphasis limits their ability to prepare for the future. They suggest that to fully account for uncertainty, there's a need to develop methods that embrace non-deterministic approaches. From their perspective, scenario planning should rely less on causation, and they propose an 'antifragile' approach to preparing for an uncertain future.

They focus on 5 factors associated with greater Antifragility :

1) Optionality : Maintaining multiple options open and being flexible in making choices
2) Barbell strategy : Taking no risks in areas with potentially significant negative impact, and taking many small risks in areas with potentially significant future outcomes but low negative impact.
3) Redundancy : Seeing redundancies and buffers as a form of investment and overcompensation for past shortages, thereby being prepared for future shocks.
4) Hormesis : Short term stress that leads to beneficial long term improvements
5) Bricolage : Using available resources effectively and making decisions under uncertainty that focus on the upside of being correct.

TAKEAWAYS FROM ANTIFRAGILITY

1) Antifragile systems benefit from small risks that don't have major consequences. Early risk discovery during design, testing and deployment can help minimise the impact of such risks.

2) Adopting such risk management practices can make organisations and their products better at tackling future risks and failures.

3) When its difficult to predict what could go wrong and the consequences of that, Antifragility can be a good approach to designing systems and processes.

4) It is possible to acquire Antifragile characteristics by changing certain processes and decisions. Proactive early risk mitigation can contribute to making products and organisations Antifragile.

3.4.1 How Relevant Organisations approach Risks

High Reliability Organisations

A high reliability organization (HRO) refers to an organization that manages to successfully avert disasters even when operating in environments with inherent risk factors and complexity, where one would typically expect normal accidents to occur. Examples include nuclear power plants, naval aircraft carriers, air traffic controllers, etc. Roberts (1989) defines them as follows :

"There is a class of organisations that can do catastrophic harm to themselves and a larger public. Within this larger set of potentially harmful organisations, there is a subset which have operated extraordinarily reliably over long period of time. Hence, we call these organisations "high reliability" organisations."

Rousseau and Roberts (1989) found that Highly Reliable Organizations (HROs) share certain key features:

1) "Hypercomplexity" **-** HROs have an extremely wide range of components, systems, and levels.

2) Tight coupling **-** There is a strong interdependence among various units and levels within HROs.

3) Extreme hierarchical differentiation **-** HROs have many levels, each with its own complex control and regulating mechanisms.

4) Large number of decision makers in complex communication networks **-** HROs have redundancy in control and information systems, with many individuals involved in decision-making.

5) High level of accountability **-** Unlike most organizations, HROs have strict consequences for substandard performance or deviations from standard procedures.

6) Frequent and immediate feedback on decisions **-** HROs receive feedback quickly to assess the impact of their decisions.

7) Compressed time factors **-** HROs operate with very short cycles for major activities, often measured in seconds.

8) Multiple critical outcomes requiring simultaneous achievement **-** HROs face complex operations where multiple critical objectives must be accomplished simultaneously, and once decisions are made, they cannot easily be modified or withdrawn.

While some of these characteristics are notably different from startups, its worth noting that HROs operate in a significantly more complicated and complex context than an early stage startup. Although HROs have more resources at their disposal than startups, their reliability is still impressive.

Weick and Sutcliffe (2007) identified 5 defining characteristics that aid them in acheiving this performance :

1) Preoccupation with failure: Viewing anomalies as signs of underlying systemic issues, understanding that small errors can be indicative of larger problems.

2) Reluctance to simplify interpretations: Taking intentional steps to fully comprehend the complexity of their work environments and specific situations.

3) Sensitivity to operations: Remaining continuously alert to unexpected changes in conditions and closely monitoring safety and security barriers and controls.

4) Commitment to resilience: Focussing on developing the ability to detect, contain, and recover from errors. They acknowledge that errors may occur, and actively learning from them and improving their resilience.

5) Deference to expertise: During critical situations, they prioritize the expertise needed to solve the problem over hierarchical ranks and roles.

TAKEAWAYS FROM HRO'S

1) They have a preoccupation with failure and pay close attention to anomalies & small errors
2) There is focus on sensitivity to changes in conditions and close monitoring
3) HROs showcase high frequency of immediate feedback about decisions with short activity cycles, helping them take corrective actions quickly
4) There is an inherent commitment to building resilience, recovering and learning from errors.

Startup and product success risks

A startup is an organisation formed to search for a repeatable and scalable business model. As discussed previously (Section 1.3), this process of searching often starts with forming a hypothesis and doing rapid experimentation to prove of disprove it. Through these experiments, startups aim to reduce the risk of the assumptions that underlie their hypothesis. Addressing these risks makes products more likely to be successful. A similar hypothesis testing/de-risking approach is taken by many product teams.

Cagan (2017) proposes how product teams can address 4 different types of product risks :

1) Value Risk: Ensuring the product meets customer needs through market research and validation.
2) Usability Risk: Designing intuitive experiences based on user feedback and testing.
3) Feasibility Risk: Assessesing technical feasibility and collaborates with the team.
4) Business Viability Risk: Aligning the product with business goals, marketing, and sales.

"Tackle Big Risks Early."

To effectively address these risks, he proposes that product teams should strive for:

1) Tackling Big Risks Early: Rather than deferring risk assessment, address value and business risks early in the product development process. This involves validating assumptions, testing hypotheses, and gathering feedback from potential customers and stakeholders.

2) Collaborative Problem-Solving: Encourage cross-functional collaboration between engineering, design, and product teams. This ensures that allperspectives are considered, leading to well-rounded solutions that address usability, feasibility, and business viability.

3) Focussing on Problem-Solving: Avoid getting fixated on a specific set of features or a rigid roadmap. Instead, concentrate on solving the underlying problems and meeting customer needs effectively.

TAKEAWAYS FROM STARTUP & PRODUCT SUCCESS RISKS

1) The attitude of early risk mitigation is already present in discussions around product development and validation
2) Startups, by their nature have a need of validating hypotheses and addressing risks stemming from assumptions
3) There's already a recognition for the value of early risk mitigation in product development

Relevance of Early Risk Discovery

Along with multiple practitioners pointing out the need and value of early risk discovery, this research on Risk Assessment further supports the relevance of early risk discovery and the specifics of how it benefits different organisations and processes.

Early risk discovery makes Agile Risk Management more effective. Risk discovery can leverage the benefits of Agile risk management to ensure that risks are mitigated and systems improve rapidly. Risk discovery can also provide agile risk management process with potential risks to monitor and design for, basically complimenting risk assessment.

Early risk discovery helps systems become AntiFragile. By discovering risks early, they can be mitigated before they cause significant harm. Discovering risks can also help improve a systems design such that it can avoid or handle that risk in the future. That creates incentives for teams to invest resources in practicing early risk discovery.

Early risk discovery aligns with the principles of highly reliable organisations. Being continuously alert to changes, investigating anomalies, and frequent feedback cycles all point towards how discovering risks early contribute to their resilience.

Early risk discovery also resonates with how startups tackle business risks early, minimising wasted resources and reducing the probability of business failure. Product teams also follow a similar approach to validate their initial assumptions about product success to reduce the chances of failure at a later stage. Startups that already practice early business risk mitigation are well suited and capable of early risk discovery.

3.4.2 How Designers can contribute

Chivukula et al. (2021) mapped a collection of 63 ethics-focussed design methods that focus on making ethical impact. While none of them take a risk-based approach to mitigating negative consequences, the techniques used and approaches taken are significantly relevant and transferable to risk discovery. The skills required in these processes also overlap with many competencies of designers like future visioning, stakeholder empathy, concepting, and reflective practice. That makes it likely that designers are well suited to contributing to, and executing similar processes that focus on uncovering potential risks.

To support this argument, I further explore Backcasting based Scenario Development, Judgement Call the Game and Tarrot cards of Tech. Backcasting focuses on exploring potential opportunities, Judgement Call helps uncover ethical concerns, and Tarrot Cards help initiate conversations about the positive change as well as unintended consequences. The reason for selecting them specifically was their relevance to the design problem and context of this thesis.

Scenario Development Techniques

Bishop et al. (2007) discuss various scenario development methods utilized by futurists and designers in the field of future studies. They assert that scenario development is a crucial technique that sets professional futurists apart from other professions dealing with the future.

Scenarios represent an important aspect of futures studies because they embody the core principles of this discipline:
1) They emphasise the value of imaginative contemplation of the future to avoid being caught off guard and unprepared.
2) Acknowledging the uncertainty of the future, they focus on the preparation for multiple plausible futures, not solely relying on a single expected outcome.

One of the methods they discuss is called Backcasting, that envisions a future state at a distant time horizon and then works backward to identify the technologies or breakthroughs needed to achieve that state. This approach encourages bold and imaginative thinking, helping to avoid carrying the limitations of the present into the future. It involves creating a vision of the future and then determining the steps or events required to reach that vision. By focusing on future possibilities, backcasting opens up new research and development opportunities and allows for a more dynamic and adaptable strategy to shape the desired future.

While backcasting was originally used to envision positive future states, it could also be possible to use the same approach to visualise how risks can manifest in a specific context, under the wrong conditions. From this visualisation, we can "backcast" to find out what can lead to those risks manifesting, and what steps can be taken to reduce their likelihood.

Judgement Call, the game

Ballard et al. (2019) developed a game called Judgement Call, incorporating value sensitive design (VSD) (Friedman et al. 2006) and design fiction, to help product teams in the industry address ethical concerns related to technology. The game uses cards representing various stakeholders, their "star ratings" of experience, and different ethical principles. By combining these elements, product teams can generate different design fictions for real or hypothetical scenarios, allowing them to consider the consequences of technology from multiple perspectives.

Unlike traditional VSD projects, Judgement Call separates stakeholder identification and the surfacing of values. In the game, product teams perform stakeholder identification during gameplay, while values are derived from the organization's ethical principles. Their goal was to provide a practical tool for product teams to incorporate VSD into their operations effectively.

Fig. 51 Image of the game in progress, from Ballard et al. (2019)

They establish three stakeholder categories encompassing both individuals (e.g., end users, teenage users, parents of teenage users) and groups (e.g., watchdog organizations, elders, legislators).
1) Direct stakeholders, who directly interact with the technology, such as end users, designers, engineers, hackers, and administrators.

2) Indirect stakeholders, who are not direct users of the technology but are impacted by its use. This group includes advocacy groups, families of end users, regulators, and society as a whole.

3) Excluded stakeholders, who cannot or do not utilize the technology due to physical, cognitive, social, or situational limitations. For instance, a technology heavily reliant on visual elements will exclude individuals with low-vision.

They identified four main limitations of Judgment Call:
1) The game raises ethical concerns without proposing solutions, leaving that task to the product team.
2) Workshop methods inherently limit perspectives to the participants' experiences, affecting stakeholder identification and ethical concerns raised.
3) Representing perspectives vastly different from one's own lived experience in stakeholder identification can be challenging.
4) Values emphasized in Judgment Call align with Microsoft's, potentially differing from stakeholders' values, leading to tensions.

Tarrot Cards of tech

"The Tarot Cards of Tech" by Artefact is a tool that fosters discussions about the true impact of technology and the products being designed. It encourages designers to think about both the unintended consequences and opportunities for positive change that technology can bring. The cards contain a list of prompting questions that help think about the potential future consequences of a product.

New perspective: Don’t just ask: “How might we?” ask: “At what cost?”

Designers can use the cards during brainstorming or team meetings to prompt conversations about scalability, usage implications, equity, and access in technology. The cards can help identify potential negative outcomes to be avoided and uncover opportunities to enhance product inclusivity and meaningful connections.

Fig. 52 Tarrot Cards of Tech, from Artefact Group

TAKEAWAYS :

1) Design methods have already been used in activities similar to risk discovery. The same underlying skills and mechanisms can be used in processes that support a risk-based approach to mitigating harm.
2) Scenario Development methods are a useful approach to ideate future possibilities of how risks can materialise and cause harm
3) Stakeholder analysis plays an important role in uncovering concerns about future risks
4) While it is useful to identify potential concerns, it is equally important to propose solutions to address them
5) Card decks have been a useful artefact, serving as boundary objects in addressing ethical considerations. This finding is similar to the findings from Chapter 2.

3.5 PROCESS AND TOOL DESIGN

I used all the key takeaways from the risk assessment research to drive the initial design of the risk discovery process. Combining them with the insights from the practitioner interviews, I framed a list of design requirements to start the design process :

Defining Design Requirements

1) The outcomes of the process should lead to identifying mitigation measures to address risks.
2) Be low effort and need less time
3) The process should include exploring and analyzing the context of usage of the AI system
4) Outcomes should be actionable enough to be inputs for the product backlog.

Most of the design requirements for the artefacts from the previous chapter also applied here :

1) Translating findings from the research into a repeatable process
2) Empowering designers & other non-engineers to understand potential risks of Foundation Models, and ideate potential harms to explore mitigation strategies..
3) Align the technological risks with the contexts of the stakeholders & asking the right questions to find that alignment
4) Support cross-functional teams in collaborating on products that employ Foundation Models to mitigate potential risks

Design version	R1	R2	C1	R3 : R2 + C1	P5 + R3
Evaluation	Feedback			Test 1	Test 2, 3, 4

Fig. 53 Summary of iterative design and evaluation process followed

Initial Decisions about the form

Similar to the Tech Value canvas from the previous chapter, the artefact here would again have to map out different parts of a process. Considering the designers' lack of knowledge about risk management and AI risks, it would again need to provide them with information they might not have. That was similar to the design challenge with finding opportunities to innovate. Thus for the form of the tool, I chose to use the learnings from the previous chapter, and decided on using a canvas + card set approach.

The canvas would again communicate the process and guide designers through different steps. The cards would again convey relevant information about risks of AI systems and ask questions that nudge the designers' thought process. During the desk research and literature study, I'd already identified many similar tools like Model Cards, Datasheets and Risk Cards.

A secondary benefit of this choice would be that it streamlines the process of identifying opportunities and risks. Making this decision in the beginning also increases the likelihood that the final outcome will be usable and easy to adopt for designers.

Getting started

I started with the general risk analysis process(Rausand & Haugen, 2020)
1) Establishing Context
2) Identifying Hazards
3) Frequency Analysis
4) Consequence Analysis

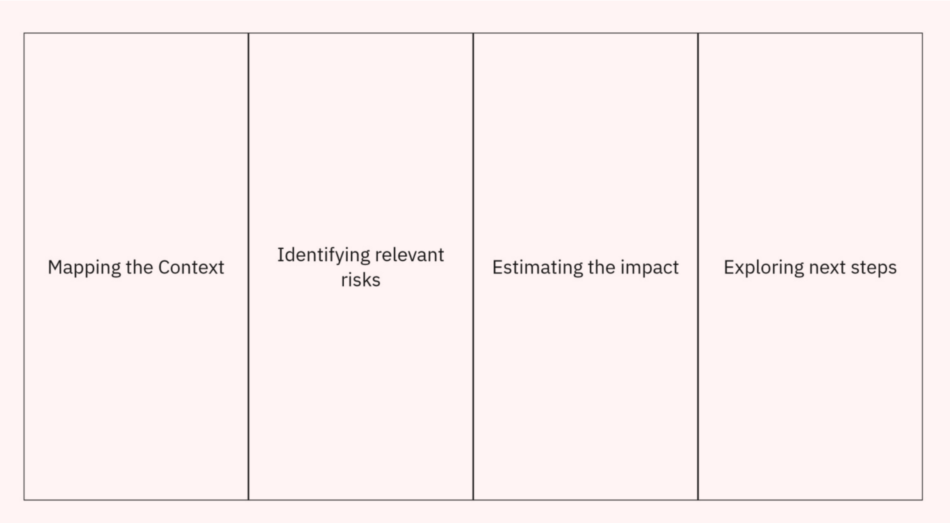


Fig. 54 Flow of general risk analysis process

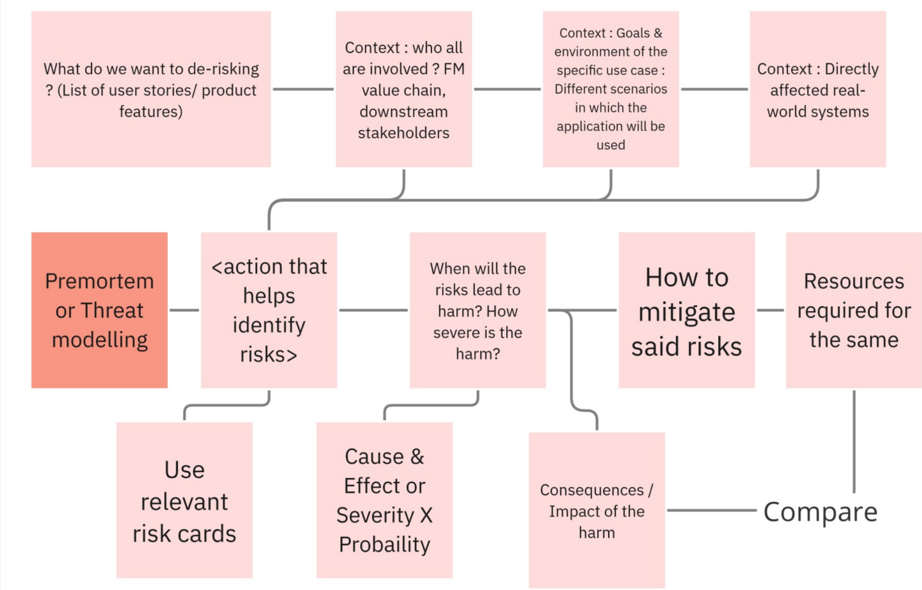


Fig. 55 Potential flow for Risk Discovery process

I converted it into a first draft of the canvas like visual. That helped me iterate further. The visualisation of the process helped me analyse my proposals and helped me identify their shortcomings. I continued to follow this evolution of process and canvas together, and both activities of designing the process and canvas supported each other.

I then converted this rough process into something more focussed on the context of AI risks and foundation models.

Prototype canvas R1

I converted this AI system focussed visual of the process into a canvas. As I'd already decided on the form of the artefacts (canvas + cards) I made a 1st draft.

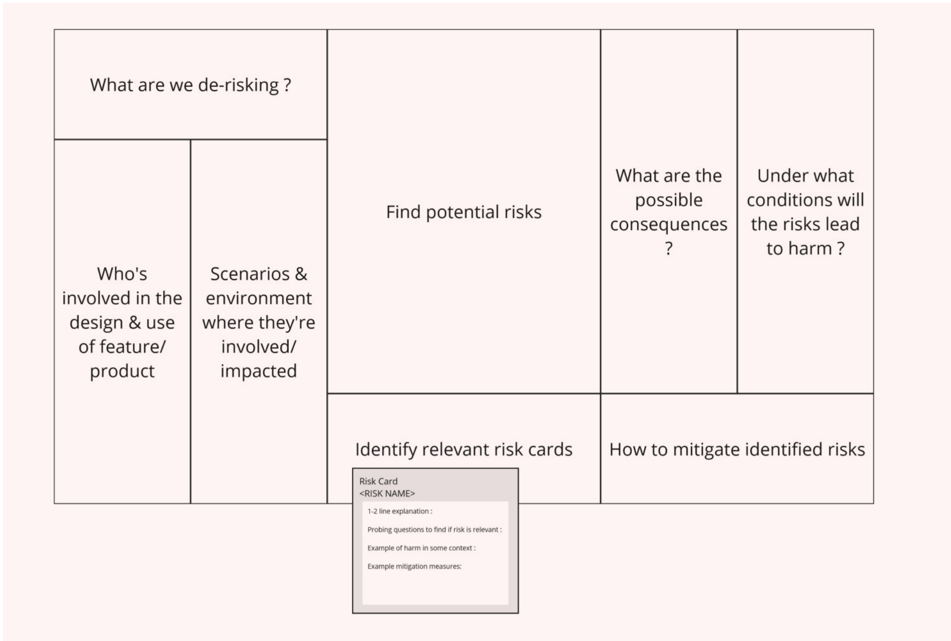


Fig. 56 Version 1 of Prototype Canvas R1

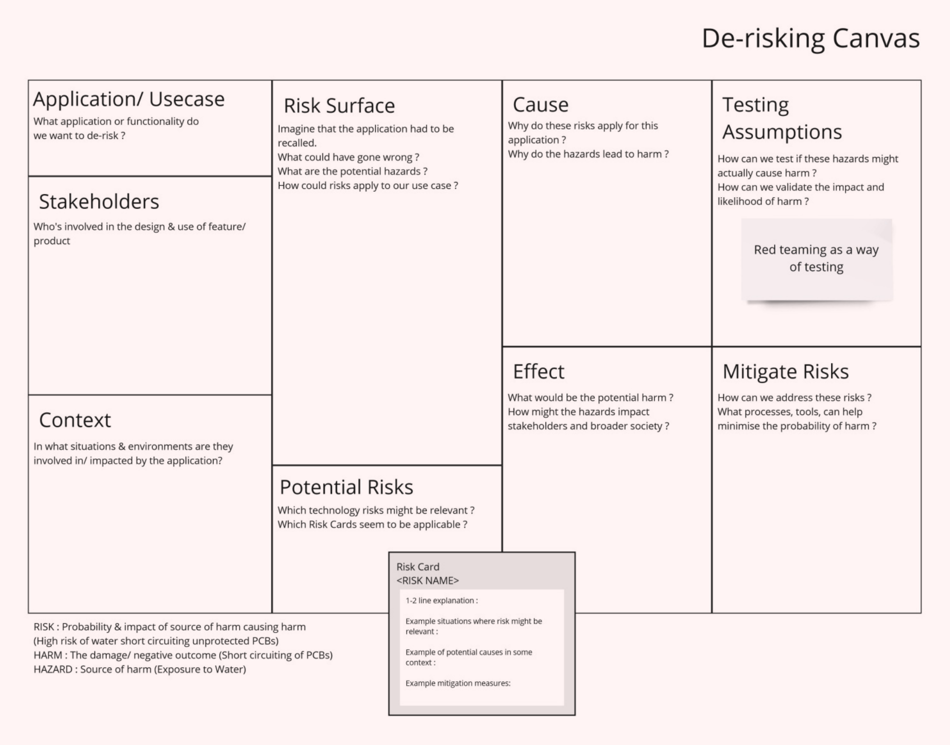


Fig. 57 Version 2 of Prototype Canvas R2

I made some further changes to this canvas layout, adding details where required for ease of use. I referred to the FMECA process and how it identifies the cause and effect of a failure to address it. The need to include stakeholders and context in th beginning came from the interview insights and desk research.

Feedback on R1

From a feedback review of the canvas R2 with part of the team at Scitodate, the major feedback was a hesitation to put effort and time into doing this. They were not keen on spending effort analysing the causes of certain potential failures and their effects. And that too followed after an initial ideation phase where the canvas would ask for an initial brainstorm of potential risks.

Because of this, the team was not keen on doing a testing session either. That meant I had to make changes to this proposal, incorporating their feedback.

Prototype canvas R2

I made changes to the previous Risk discovery canvas, making it simpler and aligning it alot more with the Tech Value Canvas. That reduced the total number of steps required and reduced the time & effort required to follow the canvas.

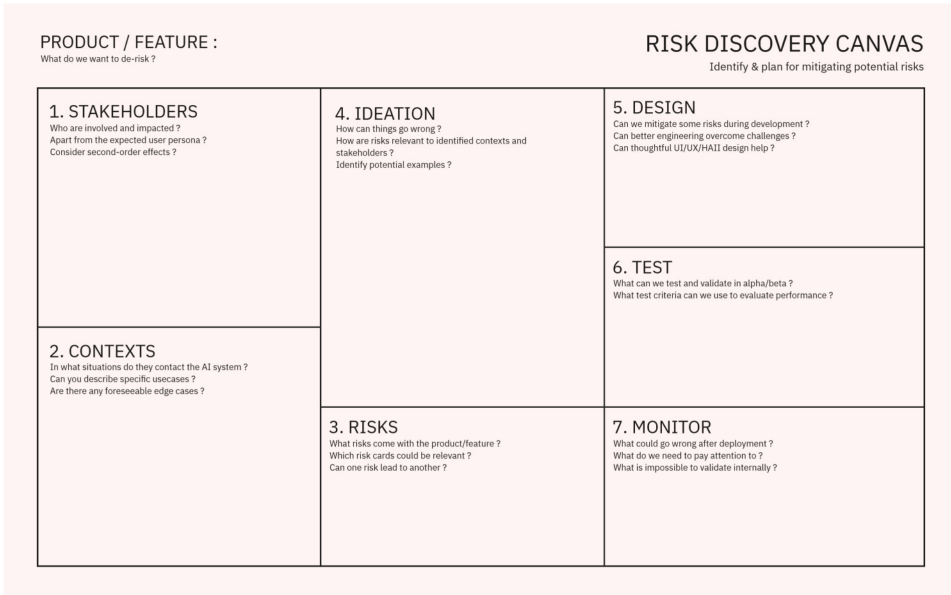


Fig. 58 Prototype Canvas R2

Prototype C1

To design Risk Cards, I carried over a large fraction of the work in the previous chapter around card design. To support that, I referred to Risk Cards from Derczynski (2023) as a good framework. While they proposed different contents for the risk cards, I started with a simpler version of a Risk Card with significantly less detail. I used a different colour than the Ability Card

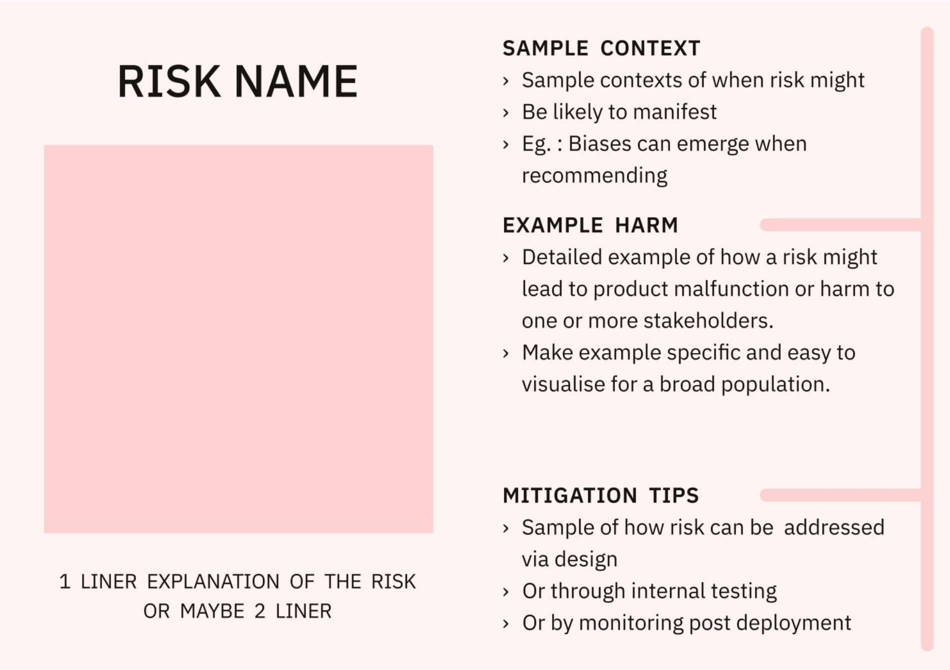


Fig. 59 Template of final Risk Card C1

Testing R2 + C1

After the first Risk cards and revised canvas were designd, the team was willing to test it out. I made some more visual improvements to the canvas by referring to the improvements in the Tech Value Canvas.

HYPOTHESIS :

This Canvas + Card combo can run just as smoothy and effectively as the Tech Value Canvas with Ability Cards.

The test was successful in getting the participants to follow through the process and use the cards to explore specific risks. I facilitated the entire discussion and walked the participants through the session. We labelled different sticky notes using alphabets and numbers to keep them clustered. This was needed because multiple identified stakeholders could potentially be exposed to multiple risks. To keep this matrix of stakeholders and risks wasy to understand segregate, we kept all potential risks grouped to every stakeholder.

The outcomes for the team were useful insights to improve the current solutions and testing methods, primarily focussing on MirrorThink. The feedback on the Risk canvas and cards were mainly small improvements to make them easier to use.

Although this was the first attempt at doing a Risk Discovery session at Scitodate, and it was the first time the participants of the session were working on exploring potential risks, we were able to identify some actual concerns and agree on some product improvements and tests to improve the current situation.

By including Scitodate as one of the stakeholders we were able to identify how risks could affect the company as well.

Fig. 60 Risk Discovery test session

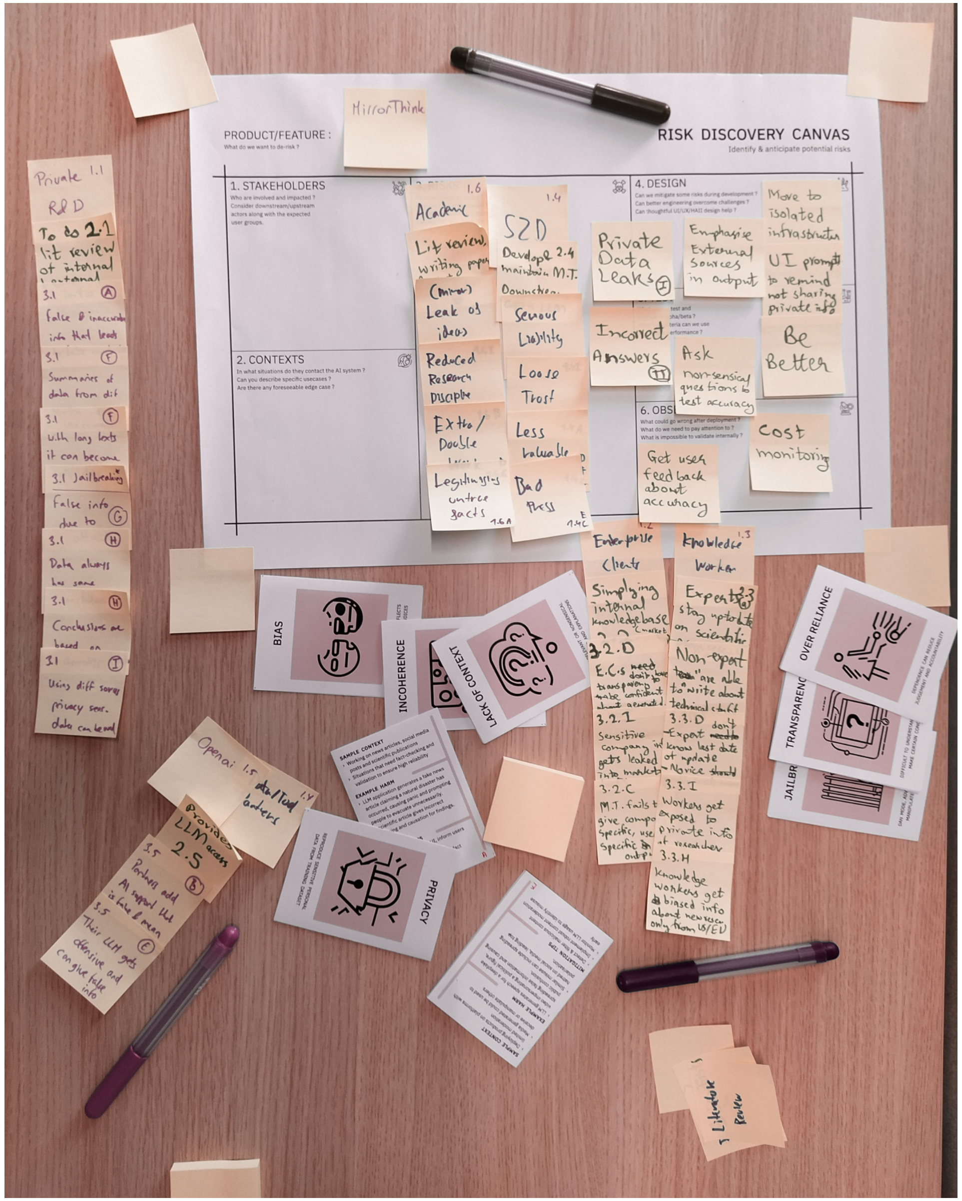


Fig. 61 Outcomes of the session. Note the clustering of risks around different stakeholders

3.6 VALIDATION

Testing both canvases together

After the initial successful test at Scitodate, I tested the Risk Discovery canvas and cards along with the Tech Value Canvas and Ability cards. I ran these tests with fellow master students as they were the target audience for the design of these tools. One of these sessions, I facilitated the tasks the participants did. With the other 2 participants, they worked through both canvases on their own, with me only interfering to support with stakeholder analysis.

For all sessions, I allowed the participants to select what AI idea or problem they wanted to design for. This allowed me to ensure that they knew enough about the target user and the stakeholders involved.

Only one of them had a relatively good technical background and they were able to propose next steps, etc. For another session, I supported the participant with identifying risks mitigation steps. The third session did not focus on the implementation of the technical aspects of the solution or the risk mitigation measures.

OBSERVATIONS FROM THE FINAL TESTS

There were significant differences in the backgrounds of the test participants and that led to them using the tools in noticeably different ways. That proved to be a good test of the adaptability and usability of the tools.

Depending on how familiar the participants were with Machine Learning, they had different levels of comfort while navigating through the canvas. That led to a difference in the breadth and novelty of the ideas.

Despite these differences, all of them were able to work through both canvases effectively and use the cards to aid their process. Everyone reflected that they learnt something new from participating in this test session.

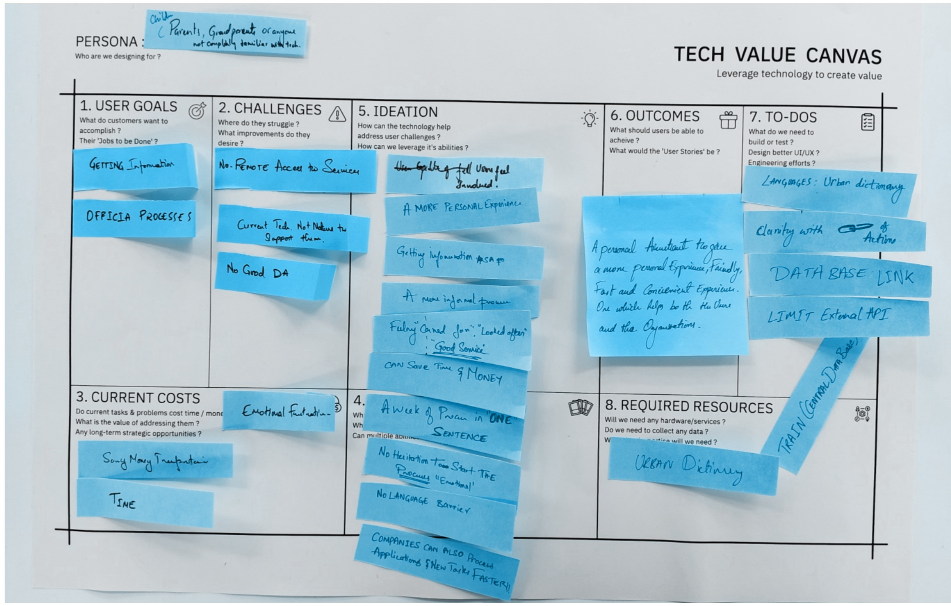


Fig. 62 Final Test 1 : Finding Opportunities

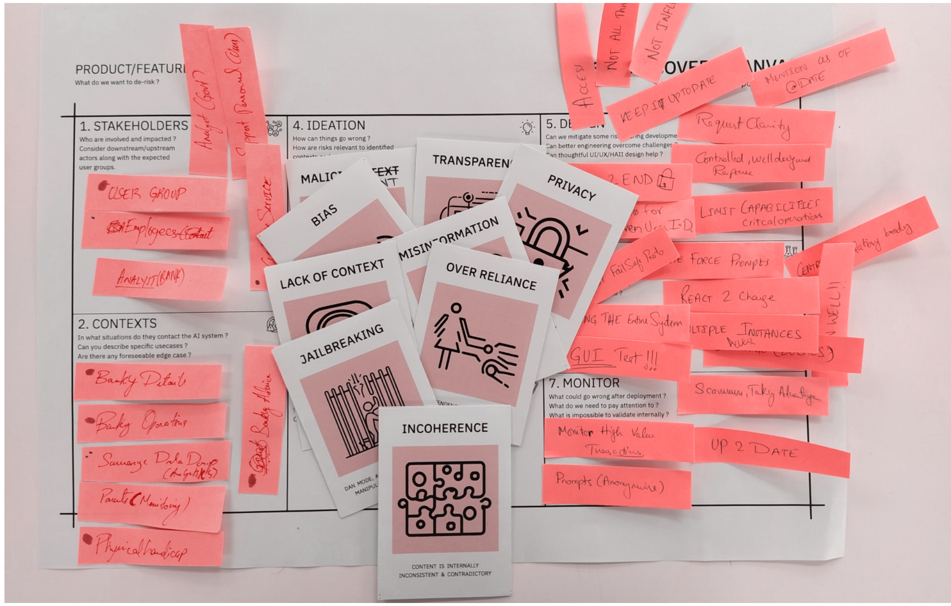


Fig. 63 Final Test 1 : Finding Risks

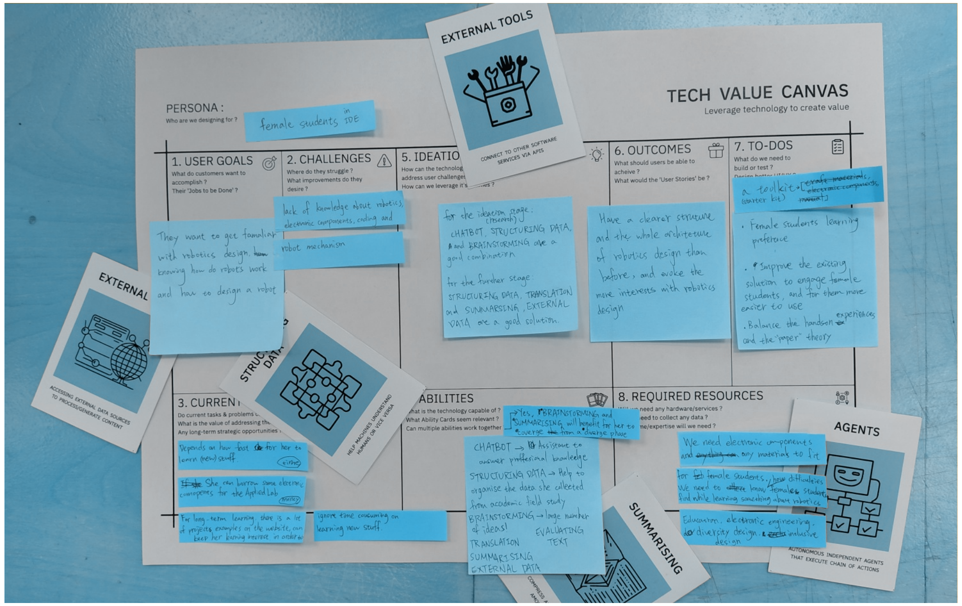


Fig. 64 Final Test 2 : Finding Opportunities

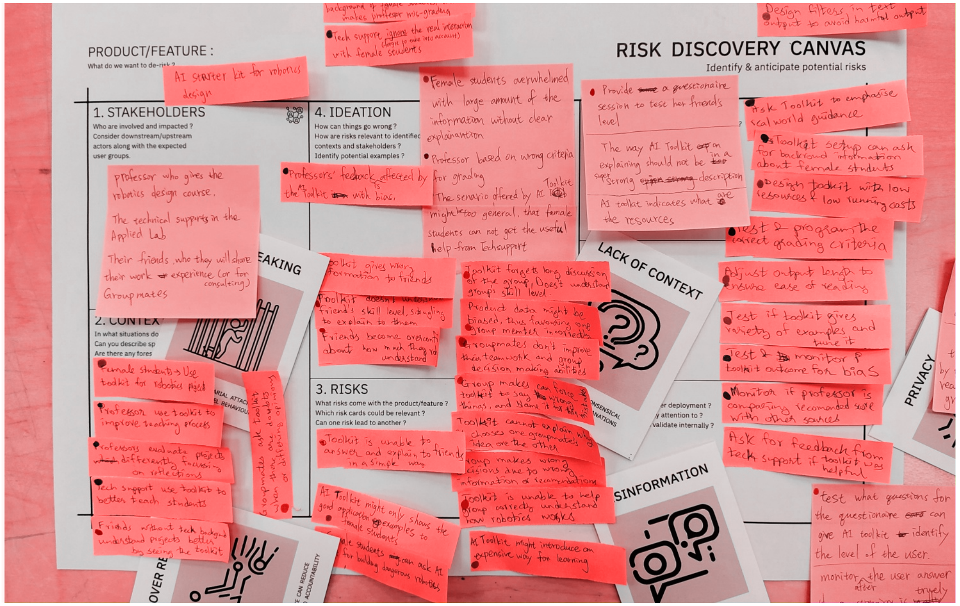


Fig. 65 Final Test 2 : Finding Risks

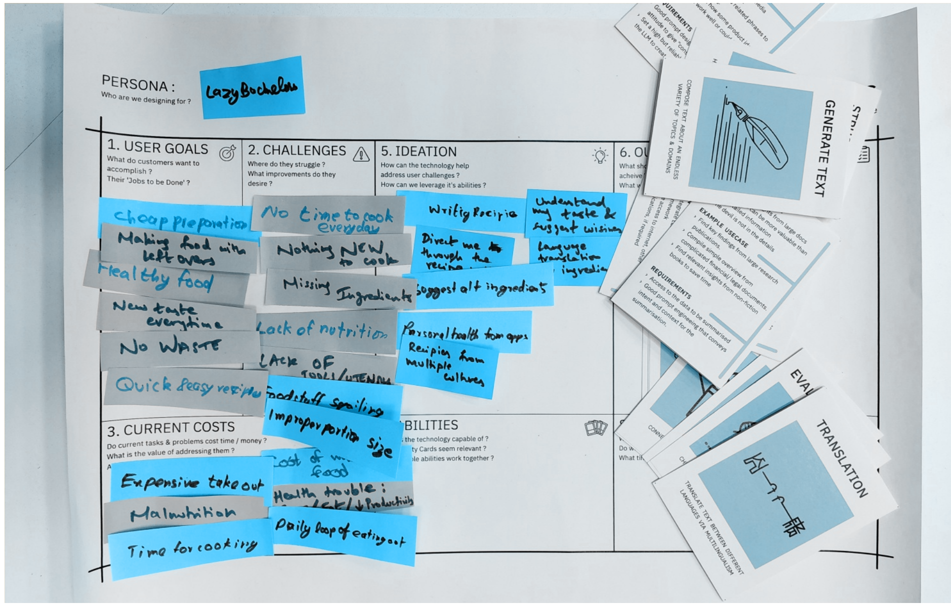


Fig. 66 Final Test 3 : Finding Opportunities

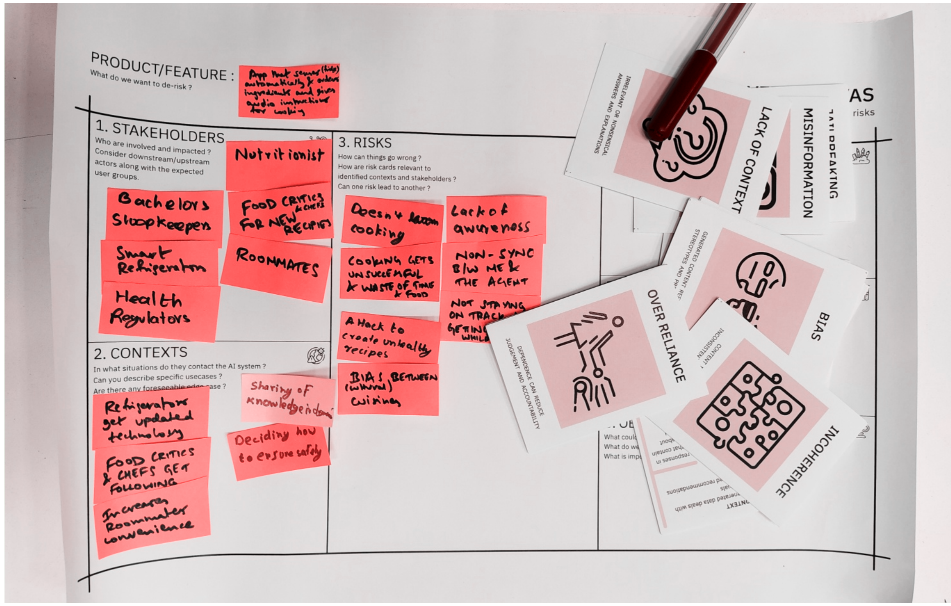


Fig. 67 Final Test 3 : Finding Risks

Evaluating Desirability, Feasibility, Viability

DESIRABILITY :

The outcomes of the Risk Discovery session were fruitful, with the Scitodate team finding multiple improvements to the current product and their mitigation plans.The students who volunteered also found out unexpected challenges and concerns in their design ideas. They said they learned something new from the process.

FEASIBILITY :

The Risk Discovery process could be practiced successfully by a variety of designers with a variety of technical understanding. That signals ease of use. Combine that with the low time investment makes the proposal a sginificantly feasible intervention.

VIABILITY :

The Risk Discovery sessions took an hour on average, depending on how much detail the team wanted to go into. This addresses the concerns for adoption of such tools, voiced in one of the practitioner interviews. That enables busy teams at startups to practice proactive risk discovery without spending alot of their precious time. More time can definitely be invested into better stakeholder analysis to acheive better analysis of the risks

Toolkit

What can I design as part of the thesis to support designers in this process ?

Contents

Summary

In this chapter, I give a brief overview of the entire toolkit, describing different aspects of the canvases and cards, and how to use them. For both card decks, I include instructions for creating new Ability & Risk cards that can be added later, as the technology evolves further. The same process can be used to later edit the cards proposed here.

Both canvases are designed to be usable for a variety of Foundation Models. The cards designed in this thesis focus only on LLMs, but this chapter explains how similar cards can be made for other Foundation Models too. Due to the limitations of this thesis, the proposed cards have not been reviewed by technical experts and should be considered a prototype.

All the tools can be downloaded via Github. The GitHub repo also has the editable templates for adapting the canvases and making new cards.

https://github.com/P2squared/InnovaitingResponsibly

4.1 DISCOVERING OPPORTUNITIES

4.1.1 Tech Value Canvas

How the canvas works

The canvas helps bring different parts of a team on the same page during the process of ideating potential use cases and solutions for leveraging Foundation Models. It follows a step-by-step process, with every block contributing to exploring the next block and at the same time, building on top of the previous one. The blocks are numbered to facilitate this process. It is always possible to bo back to a previous step depending on the outcomes of a later step, and follow an iterative loop of modifying the data on the canvas.

The canvas is broken down into 8 blocks over 3 sections. The first 3 blocks serve the role of mapping the user context. The next 2 blocks focus on exploring the potential of using Foundation Models, (LLMs in the case of Scitodate) and the last 3 blocks help identify the details of the solution and the next steps required for further development.In that way the canvas moves from WHY to develop a solution, to WHAT that could be, to HOW to realize it

The process the canvas proposes helps teams bridge the gap between the Problem and Solution. First the problems are identified. Then potential ideas for using a specific Foundation Model are identified, which lead to solutions emerging. In that way, the canvas contributes to ideating and understanding the "Problem-Solution Fit" for ideas that leverage Foundation Models.

How to use the canvas

1) Use the user research data to select a promising User Persona to focus on.

2) Map out their Goals, Challenges and Current Costs by referring to insights and findings from the user research.

3) After mapping out this user context, take the Ability Cards deck and explore which cards seem most relevant.

4) With this initial selection of Ability Cards, ideate potential opportunities for solving user challenges through the abilities. Can multiple Ability Cards work together? These ideas can be extremely rough and sketchy.

5) After an initial round of ideation, go back to the rest of the Ability Card deck and explore if any other Abilities seem relevant.

6) Discuss all ideas, find out if they can be combined or lead to new ideas. Find out if they help users avoid their Current Costs. In the end select a few most promising ideas to explore further.

7) Develop these ideas further, focussing on how they will solve user challenges. Visualise the user experience & how they will benefit from the new solution. What will the new User Experience look like?

8) List the tasks that the product team will need to execute to realise the proposed solution.

9) Identify if any new resources or investments are needed. Compare this with the Current Costs that the solution addresses. Do the solutions create sufficient value for users to justify the investment? Are there any long-term benefits of making this investment?

10) Play around with all the blocks and adjust the proposal until you reach a suitable outcome.

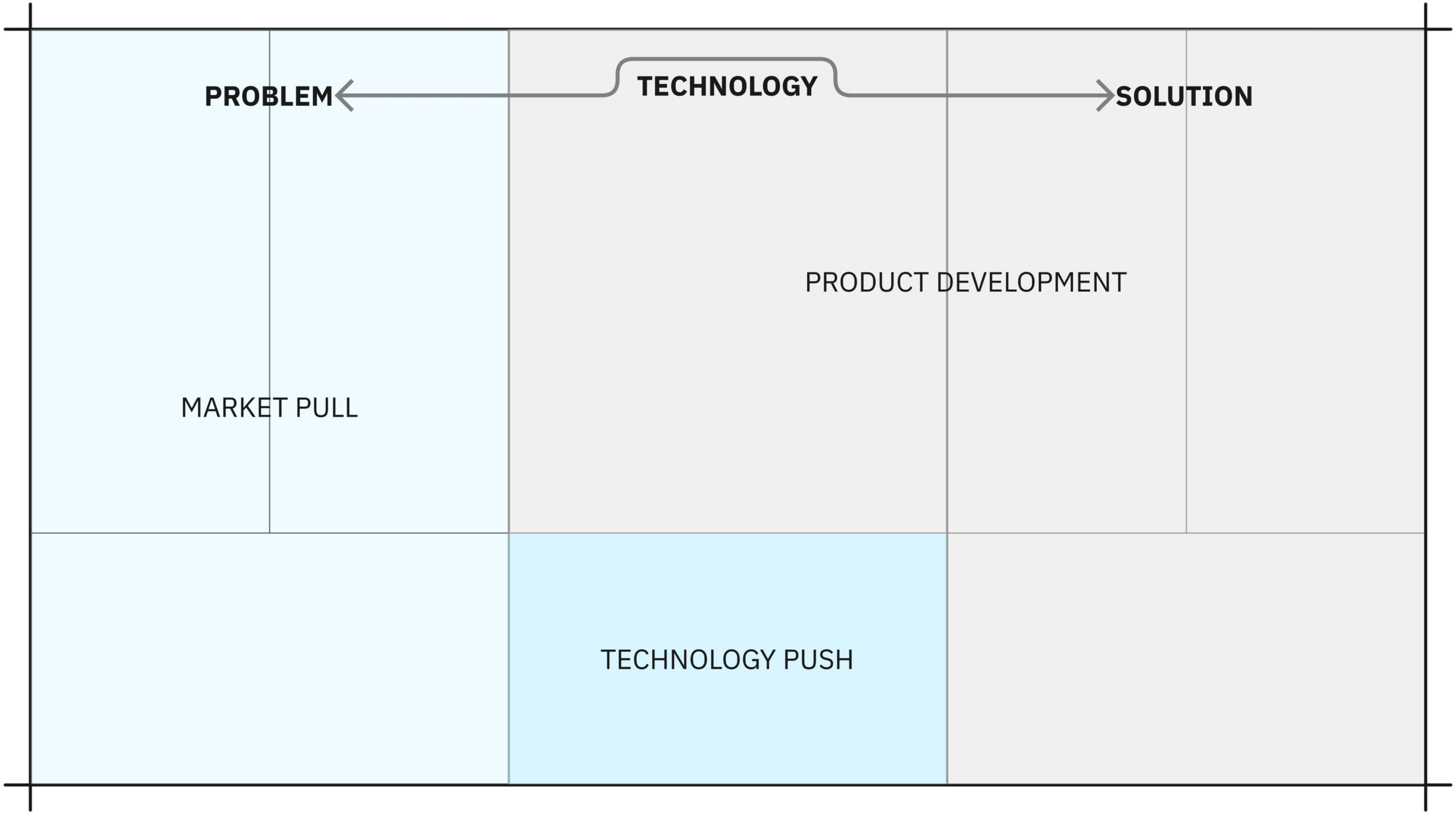


Fig. 68 Schematic of Tech Value Canvas showing how different regions focus on different topics

Different blocks explained

0. PERSONA

The first step before starting the process is to decide who the target user or customer could be. Having a specific target audience or market makes it easy to map out the problems that can be solved using Foundation Models.

This initial choice can often be a hypothesis about a good target user. Through the rest of the process, we explore if this hypothesis was correct. With the outcomes of canvas, we can further validate it through prototyping and testing.

1. USER GOALS

After deciding on the user Persona, we identify relevant activities that they perform, and understand why they do them. While our target audience might follow certain protocol and take some actions, its important to understand why they do them. What goals do they want to achieve? What are their “Jobs to be Done” (discussed in chapter 2) ? Understanding these motivations make it easier to explore how they can be addressed.

2. CHALLENGES

Very often, users will face certain challenges when trying to achieve the goals and outcomes we identified in the previous step. Users will often use some existing solutions, products or workarounds to overcome these challenges. These challenges can either exist in the form of problems faced or unmet desires. Identifying these challenges helps identify whether there is a “market need”. These challenges can potentially be addressed in better ways using better technology, which the canvas explores in the next steps.

3. CURRENT COSTS

Here, we try to better understand the impact of the challenges that users face. It also serves as a proxy for estimating how valuable a solution could be. Trying to quantify or qualify the impact of current problems serves as a reality check for understanding if the problem we’re trying to solve creates sufficient value to our target users. Its important to remember that costs don’t have to only be monetary. They can also be time spent, stress, emotional turmoil, etc.

4. ABILITIES

This is where we explore what abilities of Foundation Models, can be used to address the challenges that users face. (Eg. GPT-4 in the case of Scitodate) Going through the card deck, we can make a first selection of which Ability Cards can be well suited for this context. After Ideation, we can later go back to the card deck to check if any other cards apply.

5. IDEATION

After mapping user challenges and selecting promising Ability cards, we can ideate how some Ability can help overcome some user challenge. Having sticky notes with challenges written on them and keeping Ability cards next to them is a good way to identify how different abilities can be relevant to different challenges. After spending some time ideating, go back to the rest of the Ability Cards to find out if some other Ability cards could also be relevant.

6. OUTCOMES

After we have some ideas and we’ve selected a few that we want to explore in greater detail, the next step is to add more details to those ideas. In this block, we explore what the future user experience can be. How do Foundation Models lead to better user outcomes and less challenges ? How can a new product or product feature improve the future state of the user ?

7. TO-DOS

Acheiving the desired outcomes needs design and engineering effort to prototype, and later develop the products or features we identified. This block involves documenting the work that needs to be done to develop the ideas we’ve discovered. This To-Do list can serve as a starting point for the team’s product backlog and sprint planning.

8. REQUIRED RESOURCES

The last block involved mapping out the resources that the team will need to develop the ideas discussed. This step helps identify whether the team can realise its plans with their current skills and resources or do we need to invest in developing or procuring them. We can compare the identified resource requirements with Current Costs (Block 3.) to estimate the viability of the proposal : is it worth investing these resources into developing these ideas ?

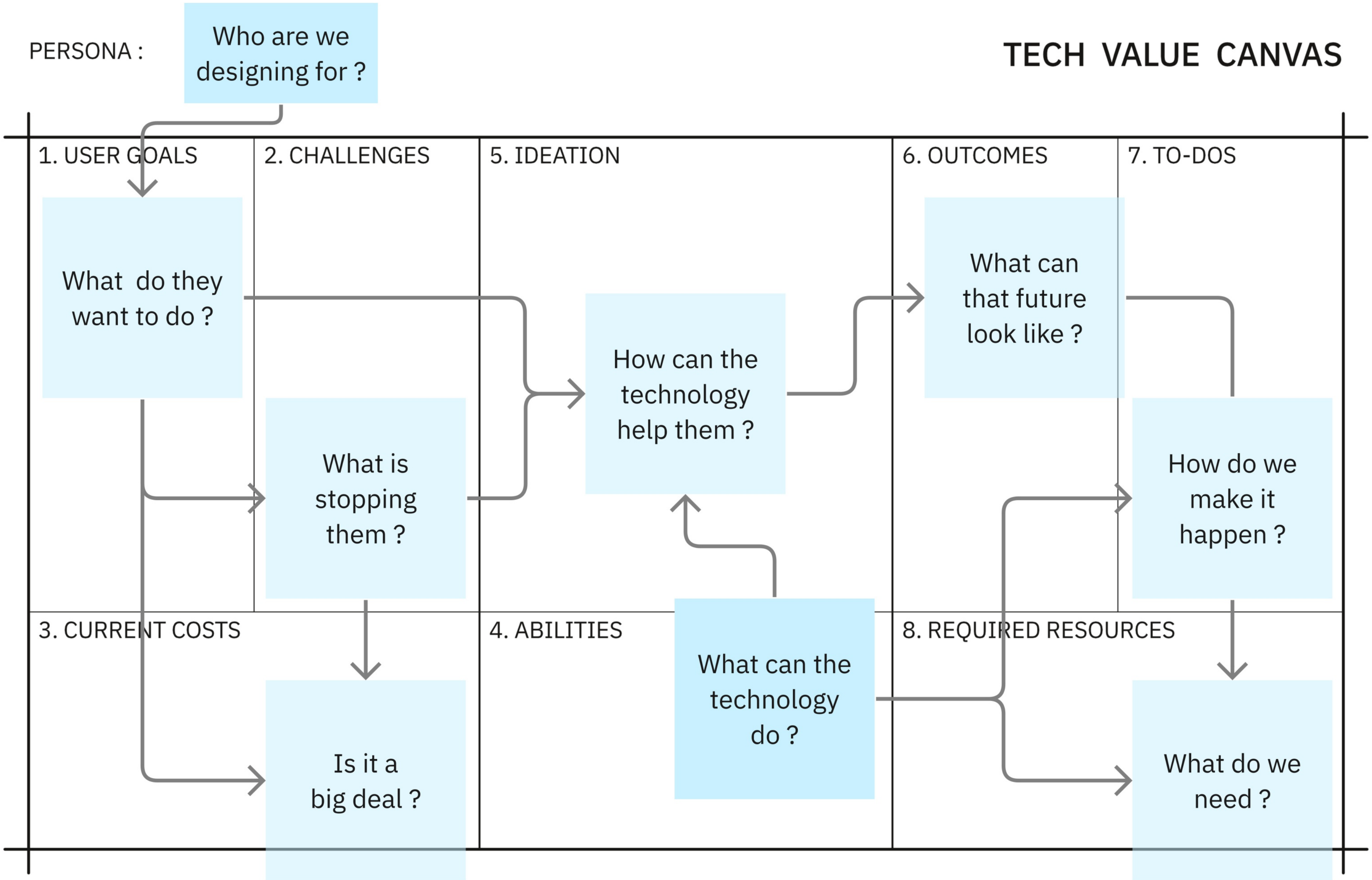


Fig. 69 Schematic of Tech Value Canvas showing connections and flow between different blocks

4.1.2 Ability Cards

Card Contents Explained

TITLE

The title is a one-two word phrase for identifying an Ability Card. The title enables easy referencing to a specific Ability during discussions. It's intentionally simple making it easy understand and remember.

ICON

The icon is a visual representation of the Ability on that card. The icon aims to make it easy to think about and understand an Ability without having to read the contents of the Ability Card. It also gives every Ability a visual identity, making it easy to remember. This makes it easy to use them during ideation sessions, when participants are busing thinking about potential solutions.

1 LINER

This is a one line summary of the ability, making it easy to understand at a glance. It serves to educate the people about the Ability Card without having to turn it over.

MOST USEFUL WHEN

This section has examples of when an ability can be valuable to users or when it can solve a problem. This are sample contexts that aim to help identify how an ability can be relevant to a specific user context.

EXAMPLE USECASE

This contains an example application of an ability in a specific context to solve a specific problem. This example aims to help while using the cards to ideate how a specific ability can help address a problem

REQUIREMENTS

Every ability needs certain resources to be realised. This section maps those requirements. In this way, using some Ability Card will lead to adding these requirements in the Required Resources block on the canvas

How the cards work

The Ability Cards aim to support the Tech Value Canvas. While the canvas aims to bring everyone on the same page and follow a specific process, the cards serve as boundary objects, giving teams a common language of communication and getting everyone to a sufficient level of understanding about the abilities of Foundation Models .

The cards are designed to be usable during a canvas session. They are designed to be easy to understand and refer to when focussing on different parts of the canvas. That also makes them useful amd understandable for a variety of non-technical audiences like designers, etc.

The cards also cover elements related to WHEN is an ability useful and relevant, WHAT could an example usecase look like, and HOW can the team realise that ability during development.

How to use them

Use them to discuss abilities with different team members. Use them to find out how a Fooundation Model can help address user challenges. Use them to explore the next steps of developing a solution.

It is best to not introduce Ability Cards to a canvas session while still discussing the user context and mapping their challenges. This helps avoid technology fixation, and ensures a focus on users in the beginning.

Ability cards are designed for finding solutions to user problems. The contents of the cards will help find out how some ability of a Foundation Model can address some user problem.

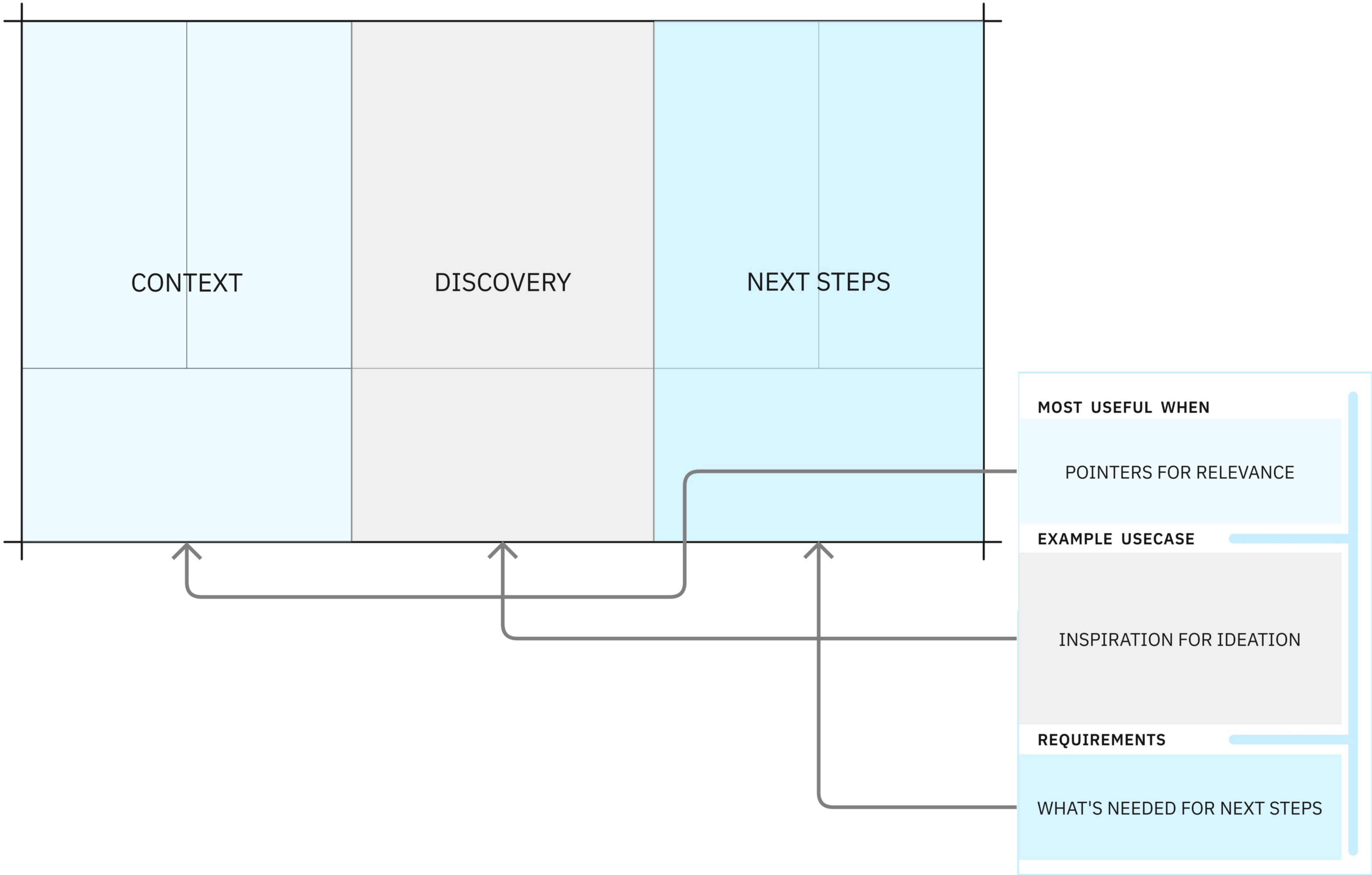


Fig. 70 Schematic of Tech Value Canvas & Ability Cards showing how they connect

How to make your own Ability Cards

The goal of having a modular card deck is to facilitate future expansion : as current Foundation Models improve, new Foundation Models get developed, new applications get discovered and as other technologies emerge.

Having Ability Cards for different Foundation Models, or even different technologies can also help address the “Hammer Nail Problem” that comes with exploring applications for a specific technology. With cards for multiple technologies, it becomes easier to choose which technology might solve a specific problem in the best way.

The overarching theme while designing new cards is to have empathy for the future card user. What do they need to know and how can they understand it best? Considering future card users will be designers, engineers, business developers, etc. it’s important to use simpe language and text that’s easy to read & understand. The icons should be representative of the Ability to help communicate it in a visual way.

For any new Ability that emerges in the future or a company develops internally, try to find out how to document the different elements of the cards. Refer to the card template and some existing cards in the deck. The template contains prompts that should help identify what the contents should be. Referring to existing cards should help decide the framing of the sentences.

The editable card template uploaded on GitHub is made using an open-source desktop publishing software called Scribus (Scribus - Open Source Desktop Publishing, www.scribus.net). More cards can be designed and printed using this template.

Get the files here : https://github.com/P2squared/InnovaitingResponsibly

ICONS

The icons on the current cards are created using Generative AI. The image is generated using the Microsoft Bing Image Generator. The prompt for the image is generated using ChatGPT. The contents of an Ability Card can be entered into ChatGPT asking for icon suggestions to represent this Risk. That suggestion can be adapted into an image prompt.

I used the following prompt for image generation :

“<Enter one line image description from ChatGPT>, modern minimal line art icon with black lines on white”

This image can then be cleaned and converted to a transparent SVG vector file in open-source tools like Inkscape (inkscape.org).

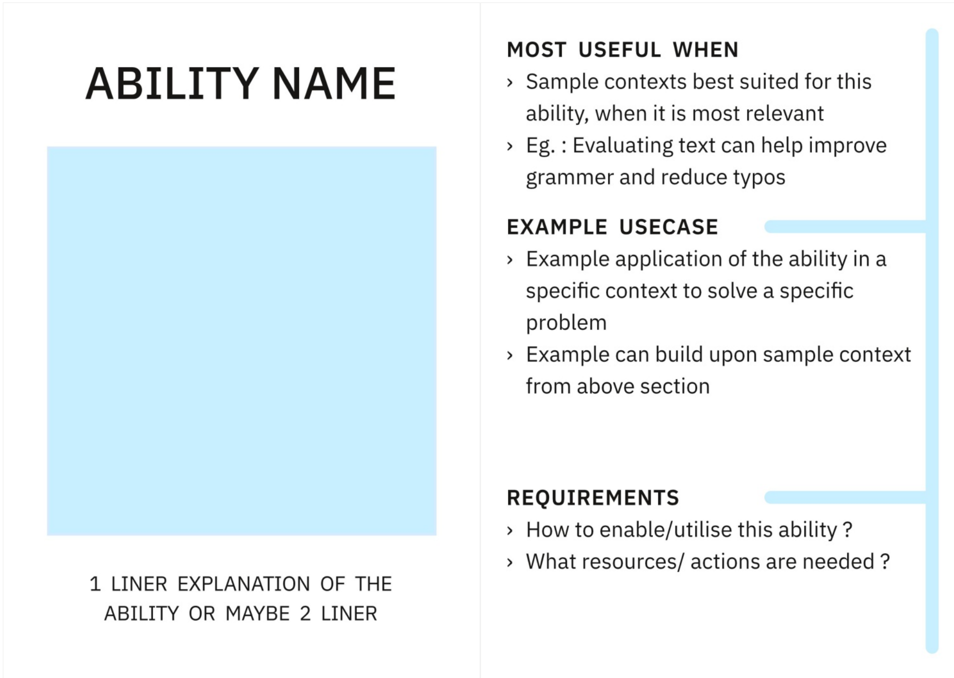
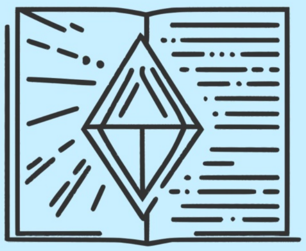


Fig. 71 Template for making new Ability Cards

Next page : pg.72, 4 sample Ability Cards

SUMMARISING



COMPRESS AND DISTILL LARGE AMOUNTS OF TEXT

MOST USEFUL WHEN
- Users need insights from large docs
- Summaries can be more valuable than the detailed information
- The devil is not in the details

EXAMPLE USECASE
- Find key findings from large research publications.
- Compile simple overview from complicated financial/ legal documents.
- Find relevant insights from non-fiction books to save time

REQUIREMENTS
- Access to the data to be summarised
- Good prompt engineeing that conveys intent and context for the summarisation.

TRANSLATION

TRANSLATE TEXT BETWEEN DIFFERENT LANGUAGES VIA MULTILINGUALISM

MOST USEFUL WHEN
- Creating content for multiple demographics
- Products need to be designed for users working in different languages

EXAMPLE USECASE
- Hi! Wish you a fun weekend
- Hoi! Wens je een leuk weekend
- ¡Hola! Te deseo un fin de semana divertido
- Salut! Je vous souhaite un agréable week-end

REQUIREMENTS
- LLM trained on sufficient data from multiple languages
- Fine tune LLM with text from underrepresentated languages

EVALUATING TEXT

CHECKING TEXT AGAINST A VARIETY OF REQUIREMENTS

MOST USEFUL WHEN
- Proofreading, spellchecking
- Tonal analysis
- Need to write in a specific academic or casual style

EXAMPLE USECASE
- Write a draft blog with the relevant info and arguments and ask an LLM to improve the readability, style, grammar etc.
- Re-write text to make it sound mroe enthusiastic, supportive, etc.

REQUIREMENTS
- Well designed prompt from user or backend to ensure consistency
- Sufficiently capable Instruction tuned LLM

BRAINSTORMING

A CREATIVE SPARRING PARTNER TO GET OUT OF THE CREATIVE BLOCK

MOST USEFUL WHEN
- Users need creative alternatives and improvements to their ideas
- Generating a large number of ideas within a short amount of time

EXAMPLE USECASE
- Explore ideas for social media marketing campaigns.
- Find synonyms and related phrases to improve writing.
- Feedback on how some product idea might not work well or could fail.

REQUIREMENTS
- Good prompt design, tuning tone and attitude to give "constructive critism"
- Set a high but reliable temperature for the LLM to create more random output.

4.2 DISCOVERING RISKS

4.2.1 Risk Discovery Canvas

How the canvas works

Similar to the Tech Value Canvas, the Risk Doscovery Canvas follows a step-by-step process, with evey block building on top of the previous one and contributing to the next one. The blocks are again numbered to facilitate this process. It also aims to foster collaboration between designers and other members of a product development team.

The canvas is divided into 7 blocks. The first 2 map out the context of relevant stakeholders and how those stakeholders are related to the solution. The next 2 blocks focus on exploring how the solution can harm the identified stakeholders.

The process the canvas proposes helps the team to go from potential concerns to identify specific details about the risks and find actionable steps to address them. First the broader context is analysed. That helps identify how different risks can be relevant and can lead to future harm. That helps plan next steps and mitigation strategies. These outcomes from the canvas can thus directly translate into contents of a team's product backlog.

How to use it

1) Start with a specific product of feature that needs to be de-risked.

2) Identify all the different stakeholders that might be explosed to the consequences of deploying this product/feature

3) Include the startup itself as one of the stakeholders. That way, the process can help identify how risks can potentially harm the company.

4) Explore the different contexts in which these stakeholders could potentially be affected in a positive or negative way.

5) Use the risk cards to ideate how a specific risk can cause harm to a specific stakeholder in a certain context.

6) Follow this process for all risk cards, all stakeholders and all contexts.

7) Analyse these possibilities to identify common factors and causes.

8) Explore how these risks can be mitigated through improving the product's design and engineering. What steps can be take during development ?

9) Explore what can be tested to evaluate if the design improvements address the risks. What else can we test before deployment to identify remaining risks ?

10) Identify residual risks and explore how to monitor the product/feature post deployment. These monitors should help identify risks early, so that they can be addressed before causing greater harm.

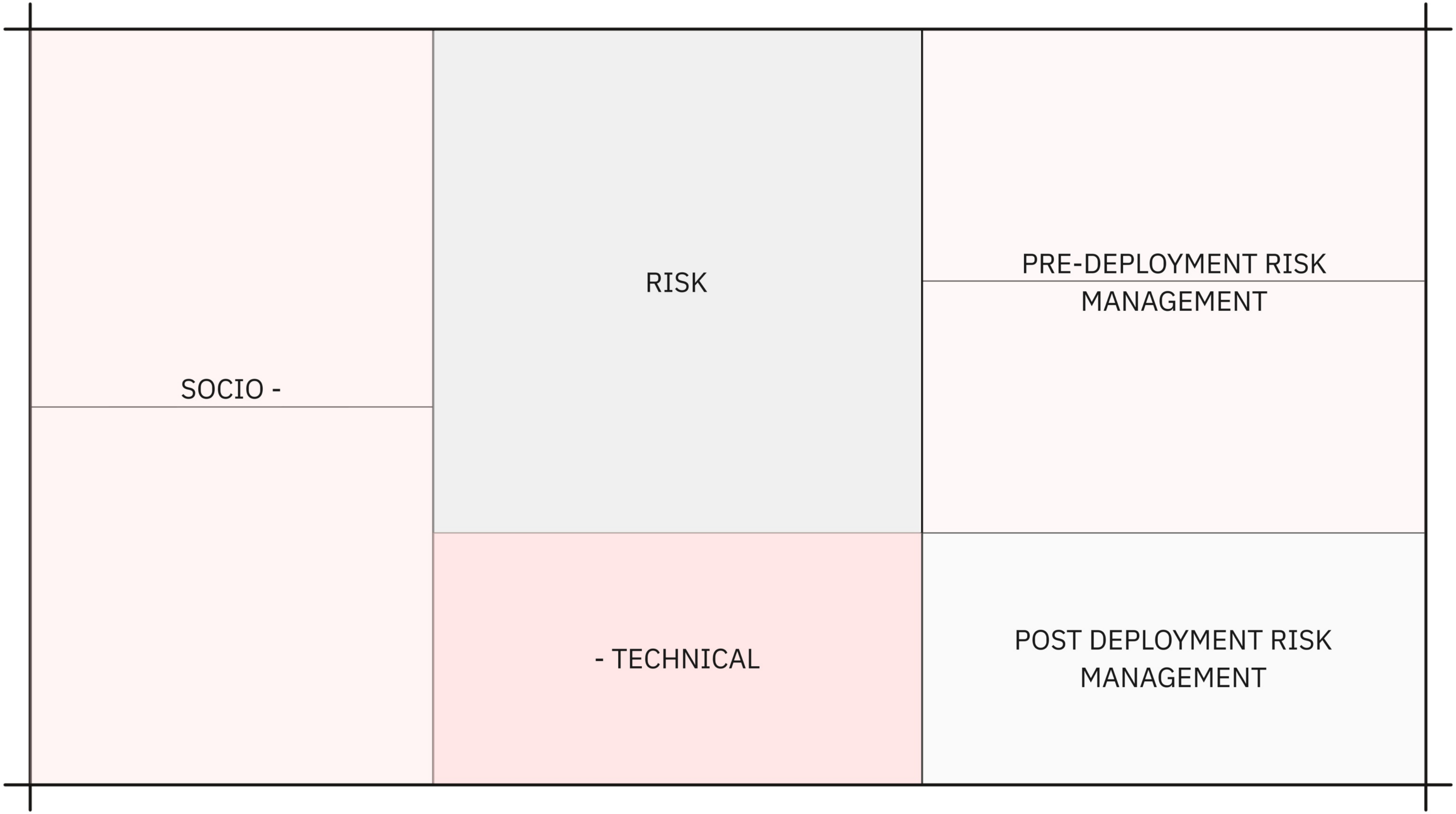


Fig. 73 Schematic of Risk Discovery Canvas showing how different regions foucs on different topics

Different blocks explained

0. PRODUCT/FEATURE

The Risk Discoery Canvas starts with the AI solution that the team wants to de-risk. It can either be an entire product or a specific part of it. This decides the scope for the rest of the canvas

1. STAKEHOLDERS

After deciding the solution scope, we identify all the different entities that are affected or come in contact with the solution. The Persona who the solution is designed for will definitely be one of them. But there will be others who are affected from the Persona using that solution. This impact can either be positive or negative. Stakeholders should involve this comprehensive set of downstream and upstream actors including the firm that developed the solution in the first place.

2. CONTEXTS

After identifying the stakeholders, its important to specify the different contexts in which they are affected by the solution. A context is the situation or usecase in which the stakeholder becomes relevant to the effects the solution. The context is the reason why a specific stakeholder is included the previous block. One stakeholder could be impacted in multiple contexts.

3. RISKS

This is where we identify relevant Risk Cards that apply for the identified stakeholders and contexts. Going through the card deck, we can make a first selection of which Risk Cards can be well suited for this context. After Ideation, we can later go back to the card deck to check if any other risks seem relevant. It is advisable to try to include as many risk Cards as possible, trying to force fit the relevance of a risk to the context. That changes the conversation from “If a risk might be relevant” to “How might a risk be relevant”.

4. IDEATION

This block shares the same name with the Tech Value Canvas. That is intentional. The goal here is almost the same. Ideate how a certain risk could impact the identified stakeholders in different contexts. Write the ideas as potential harms : “Persona gets arrested for unknowingly spreading misinformation”. Considering the challenge of organising multiple stakeholders and different contexts, it is advisable to group different ideas together, potentially numbering them for easy identification.

5. DESIGN

After exploring all the ways in which things can go wrong, that should serve as a source of design problems to address. Some of the potential risks can be addressed through better UI/UX design, better engineering and other product development activities. These activities can be included in this block.

6. TEST

Many times, it's very difficult to be sure if some design decision actually helps mitigate a risk. That can only be identified by testing it after the solution has been developed. Some of these uncertainties can be addressed via internal testing before the solution get’s deployed to the end users. These tests should be included in this block. It is also very important to setup strict testing criteria for these evaluations. The nature of the risks should govern how stringent the testing protocol should be.

7. MONITOR

With AI systems, it is very difficult to predict how they wll behave after they interact with humans and other systems after deployment. Some risks will not manifest in internal testing but can lead to harm in the real world. Some risks which we previously thought we’d mitigated end up manifesting in different ways. It is extremely unlikely to prevent these unexpected events. To address them quickly, before they can cause harm any stakeholder, we need to monitor these systems in real-time to check for anomalies and malfunctions. These monitoring requirements go into this block.

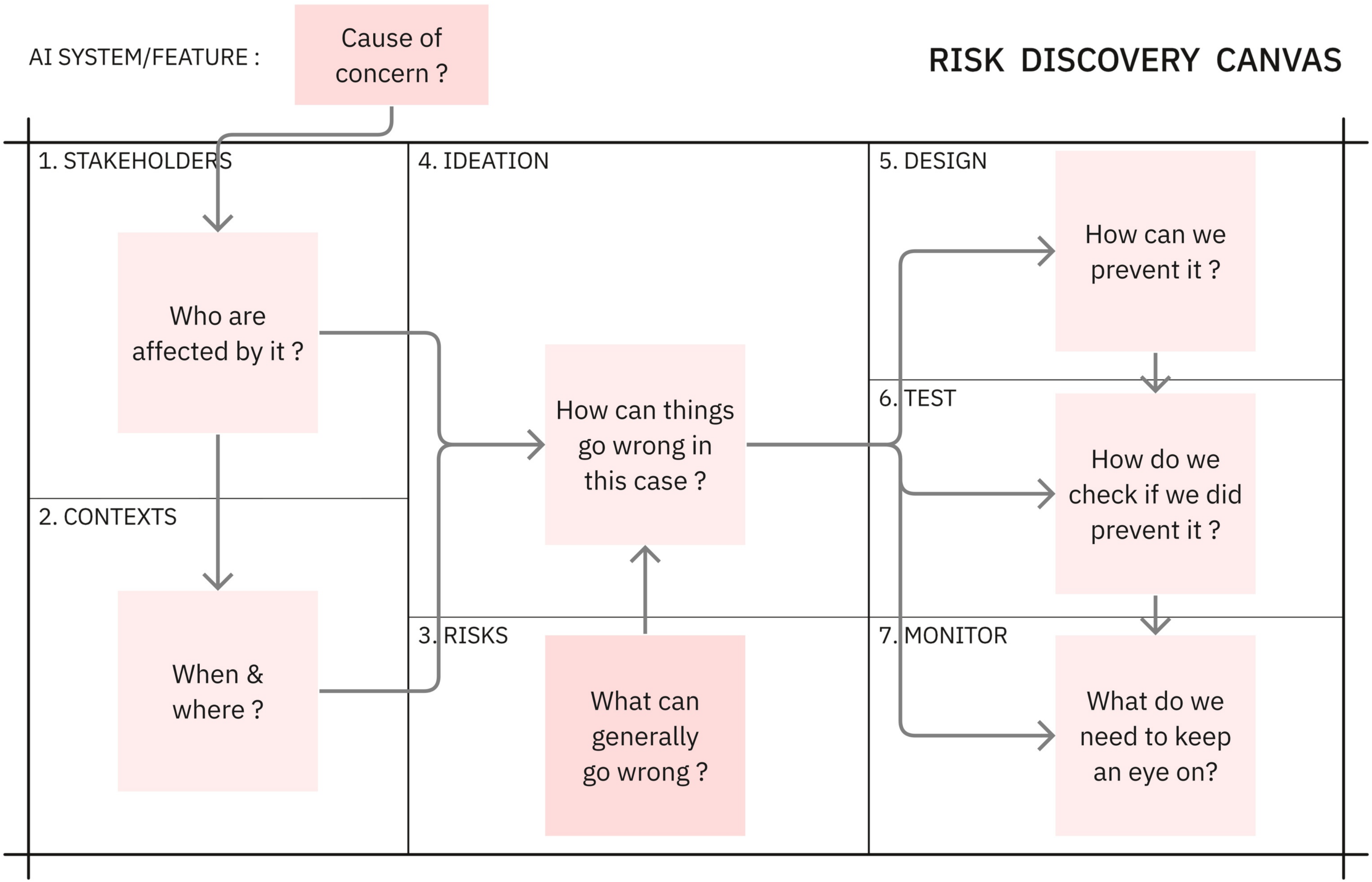


Fig. 74 Schematic of Risk Discovery Canvas showing connections and flow between different blocks

4.2.2 Risk Cards

Card content descriptions

The risk cards are significantly similar in design and use to the ability cards, except for the contents of the back face and their relevance.

TITLE

The title is a one-two word phrase for identifying a Risk Card. .

ICON

The icon is a visual representation of the Risk on that card. This makes it easy to use them during ideation sessions, when participants are busing thinking about potential solutions.

1 LINER

This is a one line summary of the risk, making it easy to understand at a glance. It serves to educate the people about the Risk Card without having to turn it over.

SAMPLE CONTEXT

This section has examples of contexts in which a risk can pose a threat of causing harm. This are sample contexts that aim to help identify similar situations that are more relevant to the actual context of a solution.

EXAMPLE HARM

This contains examples of arisk manifesting and resutling in some form of harm to someone in a specific context through a specific AI solution. This example aims to help in ideating how a specific risk can lead to harm in the context of the solution being evaluated.

MITIGATION TIPS

These are a few suggestions for how a risk can be addressed. The examples include better design choices, testing and monitoring.

How they work

The Risk Cards aim to support the Risk Discovery Canvas. While the canvas aims to bring everyone on the same page and follow a specific process, the cards serve as boundary objects, giving teams a common language of communication and getting everyone to a sufficient level of understanding about the potential risks of a solution.

The cards are designed to be usable during a canvas session. They are designed to be easy to understand and refer to when focussing on different parts of the canvas. That also makes them useful and understandable for a variety of non-technical audiences like designers, etc.

Different parts of the Risk Cards relate to different parts of the Risk Discovery Canvas. The cards also cover elements related to WHEN is a risk relevant to a specific context and stakeholder, WHAT could an example harm look like, and HOW can the team take steps to mitigate the risk during and after development.

How to use them

Use them to discuss risks with different team members. Use them to find out how using a Foundation Model can lead to new risks . Use them to explore and plan the next steps to address these risks.

Just like the ability cards, it is best to not introduce Risk Cards to a canvas session while still identifying the relevant stakeholders and their context. This helps avoid the possiblity of missing out some stakeholders and ensures a focus on the broader impact of the AI solution.

Risk cards are designed for identifying possibilities of potential harm. The contents of the cards will help find out how some shortcoming or limitation of a Foundation Model can cause problems.

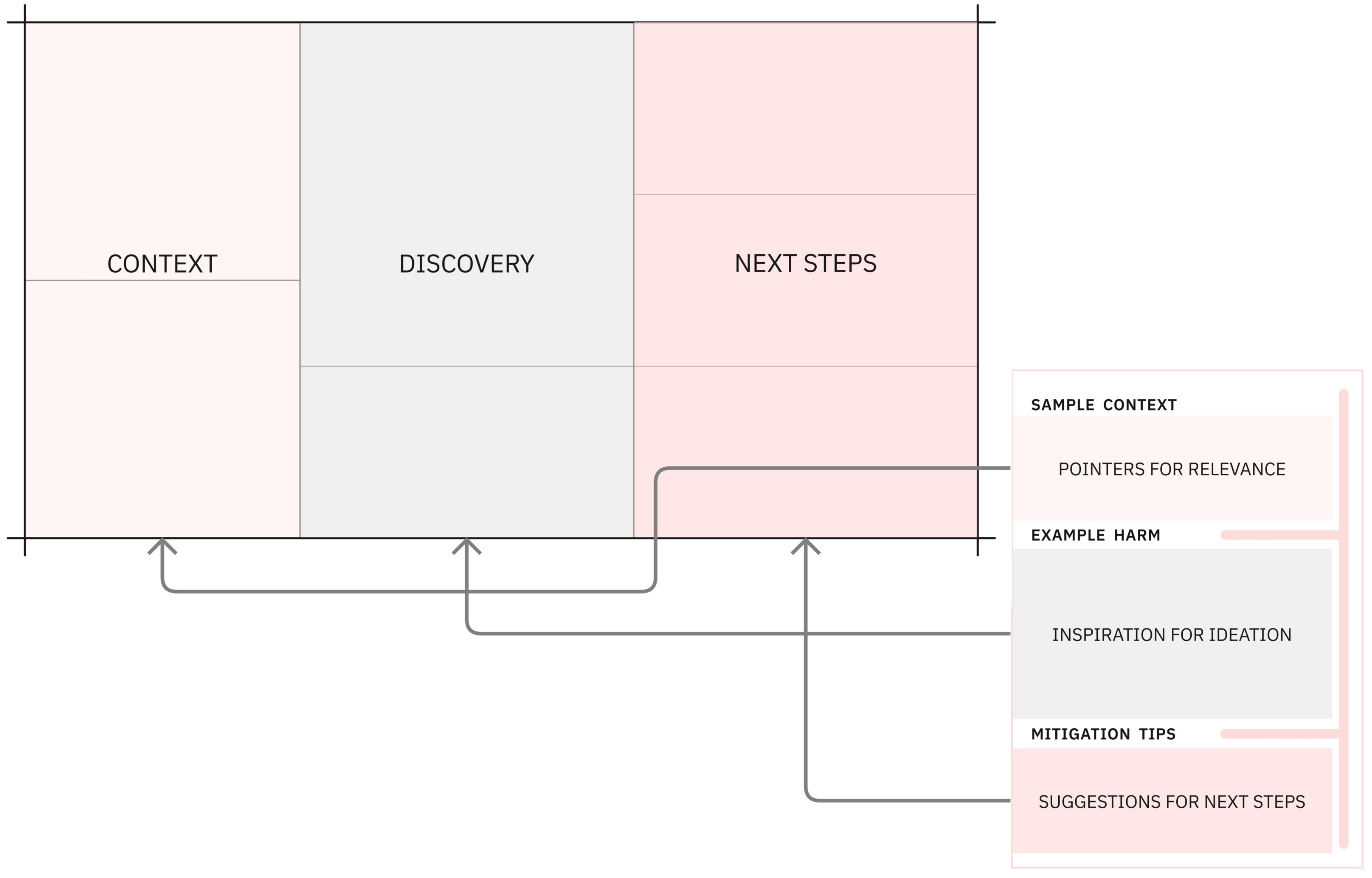


Fig. 75 Schematic of Risk Discovery Canvas & Risk Cards showing how they connect

Make your own Risk Cards

Similar to the Ability Cards, the Risk Cards are also designed for future expansion. As Foundation Models continue to become more capable and used in a variety of applications, new risks will emerge that were previously unknown. They can be documented onto new Risk Cards and thus incorporated into the Risk Discovery process.

Risk Cards are most effective if the deck of cards is as comprehensive as possible.Having a comprehensive Risk card deck is crucial to ensure that designers and team do not miss out on any of them. The cards included with this thesis DO NOT cover all the challenges and need to be expanded. Making sure that the deck is comprehensive means that it needs to be regularly updated with findings from research and practice.

The overall considerations and process of designing new cards is the identical to Ability Cards. Refer to Section 4.1.2 for more detailed instructions.

As new Risks emerge in the future or a company decides to focus on specific concerns raised by their customers, try to find out how to document the different elements of the cards. Refer to the card template and some existing cards in the deck. The template contains prompts that should help identify what the contents should be. Referring to existing cards should help decide the framing of the sentences.

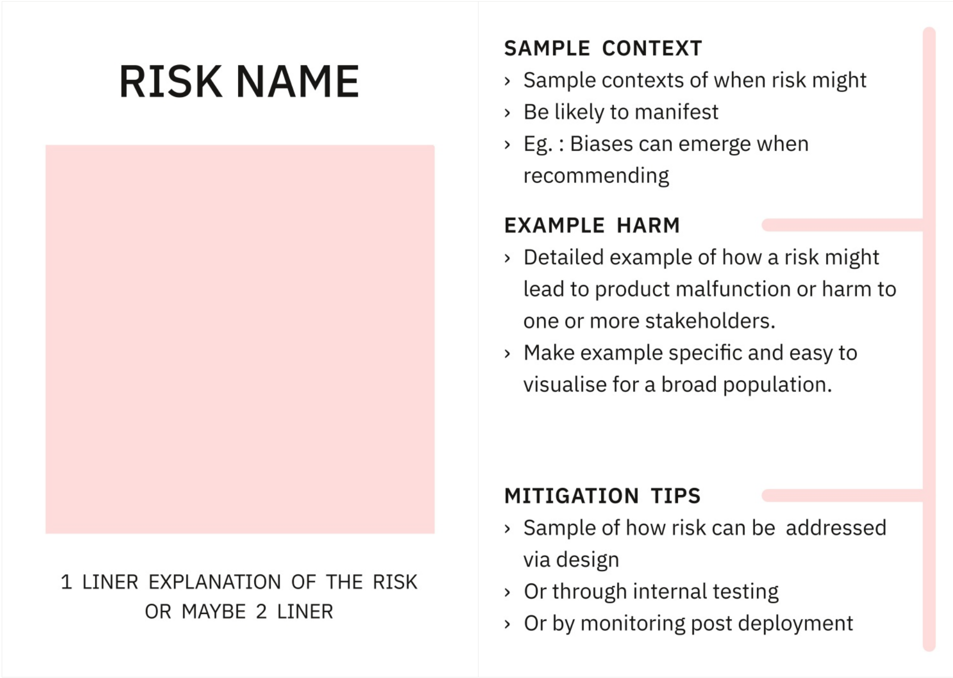


Fig. 76 Template for making new Risk Cards

Next page 77 : 4 sample Risk Cards

PRIVACY

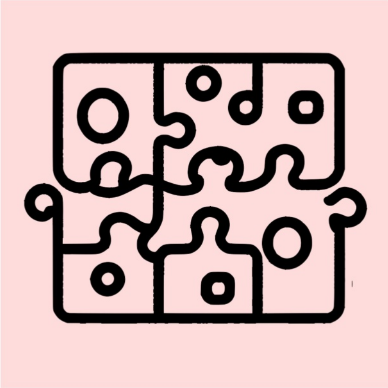
REPRODUCE SENSITIVE PERSONAL DATA FROM TRAINING DATASET

SAMPLE CONTEXT
- Training or generated data deals with specific individuals
- Giving personalised recommendations from sensitive data

EXAMPLE HARM
- LLM application generates responses in a customer support chatbot that contain sensitive financial information about users, exposing it to unauthorized individuals.
- Similar incidents can happen in the context of healthcare, social services.

MITIGATION TIPS
- Anonymise & sanitise training data, removing personally identifiable info
- Restric access to LLM to authorized personnel only

INCOHERENCE



CONTENT IS INTERNALLY INCONSISTENT & CONTRADICTORY

SAMPLE CONTEXT
- Creative writing, storytelling
- Long form content with large context window requirements
- Arguments need to be built over steps

EXAMPLE HARM
- A language model generates a story with conflicting plot points and confusing character developments, causing harm to the human author's reputation.
- Similar situation can emerge for scientific or philosophical arguments

MITIGATION TIPS
- Include training data focussing on long form argumentation and reasoning
- Test and improve model using RLHF
- Use prompt engineering to guide LLM

OVER RELIANCE

DEPENDENCE CAN REDUCE JUDGEMENT AND ACCOUNTABILITY

SAMPLE CONTEXT
- Recommending or automating critical decision making
- Creating art or public communication without human oversight

EXAMPLE HARM
- A company heavily relies on a language model to automate customer support responses, leading to numerous customer complaints and dissatisfied clients due to generic and unhelpful replies.

MITIGATION TIPS
- Avoid anthromorphisation
- Design tools that augment not substitute human decision making
- Involve human review during use

JAILBREAKING

DAN MODE, ADVERSARIAL ATTACKS TO MANIPULATE MODEL BEHAVIOUR

SAMPLE CONTEXT
- Safety critical applications
- Online forums, social media platforms
- Highly valuable/expensive operations

EXAMPLE HARM
- An adversarial attack causes a language model to produce false medical advice, leading to potential harm to patients following the advice.
- Microsoft's Tay chatbot on Twitter was conditioned into inappropriate behaviour by other Twitter users

MITIGATION TIPS
- Integrate robust filtering monitors to detect and prevent adversarial inputs
- Frequent red teaming to make LLM more resilient to attacks

4.3 DISCOVERING BOTH TOGETHER

Following both processes together, of finding opportunities and risks can help teams make an informed choice about how to use a Foundation Model, or what products to develop by comparing the opportunities & risks and understanding the tradeoffs.

Using both canvases

Using both canvases together is convenient and smooth as the outcomes of the Tech Value Canvas serve as the starting point for the Risk Discovery Canvas.

By using both of them together, it is possible to start with user research about a target persona, and find opportunities to develop solutions, and risks that come with the solution, one followed by the other. The outcomes in the end will be steps to take to build the solution, and steps to take to address potential risks.

Both canvases follow a similar process of understanding the context of users, finding how powerful technology can affect the current situation, and framing actionable next steps based on the findings from previous steps. That reflects in how the structure of the canvases are similar, dividing the blocks into 3 clusters. This similarity also makes them easy to work on, as the learning curve is gradual. As most designers are often familiar with ideating opportunities to use a technology, this similarity makes it easier for them to transition to discovering risks.

Comparing opportunities and risks

Make informed choices about whether the risks and the risk mitigation activities are worth the opportunities they create. Explore if opportunities that seem extremely promising are actually too good to be true.

By following both the processes together, it becomes possible to evaluate how solutions that leverage certain abilities can lead to certain risks. Having both Ability and Risk cards together can help make a better informed tradeoff that balances the benefits and challenges of using some Foundation Model.

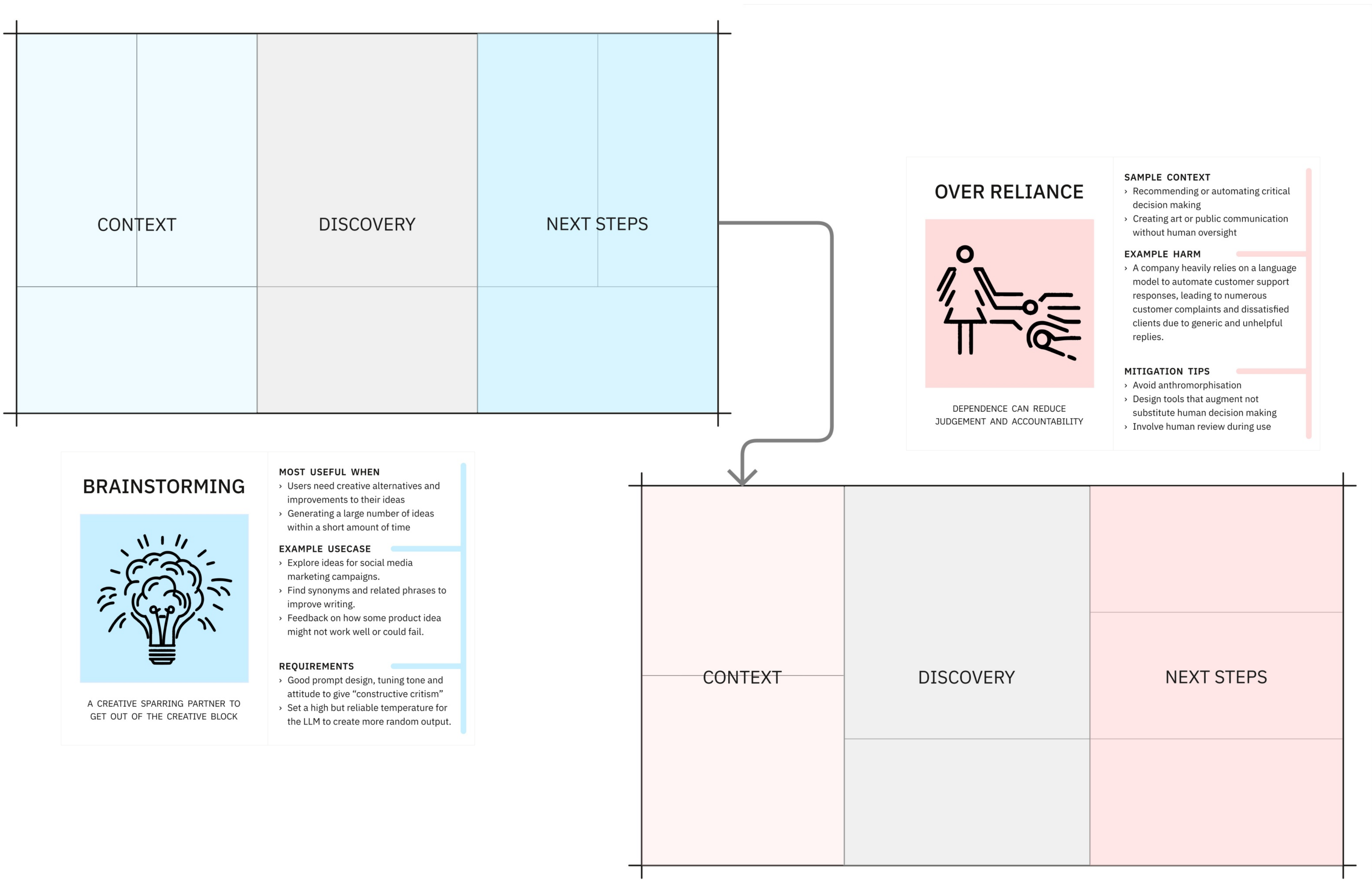


Fig. 78 Schematic of Tech Value Canvas and Risk Discovery Canvas showing how they connect

Discussion

What can the broader design community learn from the outcomes and findings of this thesis ?

Contents

Summary

In this chapter, I reflect on the work done through the thesis and discuss relevant and interesting findings during the process. The reflections intend to serve as observations and my interpretations from them, that design practitioners and researchers might find valuable.

After that I try to make recommendations for practitioners derived from the reflections in this chapter. Along with that I mention potential directions for future work.

5.1 REFLECTIONS

5.1.1 Answering the Research Questions

RQ1 : How can designers help startups find innovative solutions that leverage Foundation Models

Designers can help startups develop a deep understanding of their customers to design differentiated, long-term value propositions. By learning about the capabilities of different Foundation Models and working collaboratively with engineers and data scientists, designers can effectively contribute to finding and developing innovative solutions. The tools proposed in this thesis support designers in this process.

RQ2 : How can designers help startups mitigate the negative consequences of using Foundation Models

Designers can help startups practice responsible innovation by discovering the potential risks of using Foundation Models. This can help them design safer products, test them thoroughly and closely monitor them after deployment to address emergent risks. Understanding potential risks of Foundation Models and collaborating with engineers helps designers in this process. The tools proposed in this thesis support designers to achieve that.

5.1.2 Relevance to Literature

The thesis contributes to gaps & problems identified in literature and proposes solutions that build on existing literature :

1) Despite the differences in the Technology Push & Market Pull approaches to innovation, the importance of balancing both of them has been previously acknowledged. The Tech Value Canvas and the Ability Cards, act as tools to align the two approaches, exploring how Technology Push and Market Pull can compliment each other.

2) The Tech Value Canvas and the Ability Cards aim to address the lack of Discovery competencies in organisations, one of the barriers to radical innovation identified in the literature.

3) Building on top of the existing literature on design research for radical innovation, the thesis provides more clarity on how to go about doing that in practice, by documenting the process in the context of Scitodate using LLMs.

4) The thesis contributes to supporting designers working in AI by proposing tools that address previously identified challenges. The thesis finds out that the same challenges that apply to designing innovative AI solutions also apply to mitigating their potential harm. The proposed toolkit promises to support them both with finding opportunities as well as risks. The outcomes can support other work being done in this field in the same time frame (mid 2023). (Yildirim et al., 2023) The toolkit also supports collaboration with engineers and data scientists, another challenge identified in the literature.

5) The process of designing the canvases and cards contributes to a gap of missing action research around the design of these tools. Although I do not follow a strictly scientific process, the test sessions are framed as experiments to evaluate hypotheses, helping structure the process.

6) The outcomes of this thesis propose a way to practice responsible innovation, as a design practitioner, in an emerging technology space. The Risk Discovery Canvas with the Risk cards helps designers and startups to practice the four dimensions of Responsible Innovation.

7) Although the EU AI Act is still to be finalised and harmonised standards still be to drafted, the thesis proposes solutions that align with the direction in which the regulation iscurrently headed : Risk management of AI systems & Foundation Models. The Risk Discovery tools propose a step towards helping AI startups comply with the AI Act. The Risk cards can be later supplemented with Legal Risk cards, that can represent legal risks and violations. In this way, the outcomes can support designers to design in compliance with the AI Act.

8) Considering the unique challenges and limitations AI startups face with Responsible AI, the thesis proposes a practically usable toolkit that acknowledges these limitations and proposes a suitable solution. An assumption that can be made here, is that the same qualities of the toolkit can make it easy to adopt and implement for organisations that do not have the same limitations.

9) The Risk Discovery canvas can be an actionable tool to support the Agile Risk Management methodology. Agile Risk Management is a promising candidate to explore how product development teams and designers can integrate risk management into their current workflow.

10) The Risk Dicovery Canvas proposes to be another practical step towards achieving Antifragility. Antifragility can be a strong incentive to improve long-term product & company resilience, thereby incentivising responsible innovation practices.

5.1.3 Relevance to Practice

Helping designers innovate responsibly

The toolkit brings designers one step closer towards addressing the potential harms of the products they design. By designing a risk discovery process that closely resembles the value finding innovation process, designers are more likely to engage in this practice of evaluating their ideas and concepts.

Similarities between the tools makes it easy to use. The smooth transition between the canvases when they are used together makes it easy to explore both perspectives one after the other.

It also demonstrates that certain aspects of Resposible Innovation are possible through a similar process as traditional innovation practices.

Increasing the likelihood of Radical Innovations

The outcomes of the ideation process proposed in the thesis are still hypotheses that need to be validated and introduced to the market. My personal opinion is that ideation is the easy part. The process of turning ideas into products needs greater effort and skill. Early product validation and successful market introduction are even more challenging for radically innovations than for incremental innovations.

Where the proposed ideation process might help is in improving the odds of success in the later stages. While the outcomes are still guesses, just like dreaming up future possibilities, or random tinckering with technology, these are well reasoned, educated guesses. They are grounded in customer understanding and technology capabilities.

The user research could potentially also contribute to the later stages of radical innovation, helping with the finer details of product development, sales and marketing. I witnessed that happen at Scitodate, with the sales & marketing team benefiting from the customer research data generated through this thesis.

Increasing the likelihood of Responsible Innovation

It seems likely that adoption of Responsible Innovation practices is an incentives problem, for firms to invest resources into this, especially when faced with conflicting business interests. As the interviews with practitioners emphasised, the intent to be responsible needs to come before teams build a responsible culture and adopt tools and practices that enable them to be responsible.

The proposed risk discovery process tries to incentivise businesses through the potential improvement in their system's ability to respond to risk before the risk can cause significant harm. For teams with intent, the process along with the artefacts gives them the agency to act on their intent. They aim to give teams a usable "map" and part of the supporting knowledge needed to follow that map.

By reducing the effort, time & expertise needed to initiate responsible practices, it becomes more likely for teams to try it. For teams with intent, the low resource requirement makes integrating these practises into an organisation easy.

Importance of interdisciplinary teams & education

The thesis showcases the value and relevance of interdisciplinarity. Many parts of the challenges in this thesis could be addressed only through building an interdisciplinary understanding myself, and collaborating in an interdisciplinary manner with the team at Scitodate.

Finding a way to align technological possibilities with customer needs was possible only after engaging with both sides of the challenge. I did that through collaborating with the engineers at Scitodate to build the technical understanding and collaborating with the designer & customer success team to better understand Scitodate's customers.

Addressing the "socio + technical" challenge of Responsible AI needed a good understanding of the technical limitations as well as the social context. I had to first understand the current limitations of AI Ethics, the upcoming policies and risk management practices to decide a way forward. At the same time, the technical limitations of mathematical context-independent definitions of fairness, pre-deployment testing and design can only be understood by engaging with the technology itself. The Responsible AI practitioners that I interviewed are a brilliant example of how the Responsible Innovation challenge for machine learning needs and benefits greatly from interdisciplinary understanding.

Finally, bridging the gap between being innovative and being responsible is only possble if both parties get out of their echo chambers. Innovation does not make sense if it causes harm to society and the planet. Being responsible should not stop us from finding ways to overcome today's challenges and improve the state of life on earth. I was able to propose similar tools for both goals only after engaging in the context and challenges of both domains. I believe alot more is possible by further exploring this angle of engagement.

5.1.4 Relevance to Foundation Models

Revelance to LLMs

As majority of the work during this thesis has been done in the context of Scitodate using LLMs, majority of the work and proposed solutions have a high relevance to LLMs. Nevertheless, LLMs, and the field of NLP continues to evolve at a rapid rate. By grounding the proposed solutions in insights from older challenges and other domains, the thesis aims to propose solutions that can potentially continue to stay relevant as the field evolves.

Relevance beyond LLMs to other FMs

The intent of this thesis has been to propose solutions that are relevant beyond LLMs, to the broader scope of other Foundation Models. While the empirical work and parts of the desk research focus only on LLMs, I've tried to propose outcomes that are intentionally generalisable.

The canvases do not refer to any specific aspect of LLMs. The card decks focus on the specific characteristics of LLMs. New cards can be created that focus on other kinds of Foundation Models. The same layout & structure can be effective, as the canvas+cards will follow the same mechanism during use : Canvases guide the process, cards provide the required understanding and examples.

Relevance beyond FMs to ML

Foundation Models are still Machine Learning algorithms. Although the new paradigm opens up new perspectives and possibilities, the way they work is still the same. Because of that, many of the challenges that the thesis address are relevant to the broader field of ML.

Designers struggle to understand the capabilities of different ML algorithms. Many ML algorithms exhibit the tendency to behave in unexpected ways post deployment. Context dependence is still a critical factor in addressing the potential concerns of deploying ML solutions. Developing ML systems from scratch needs notably greater effort and investment than FM powered products. That makes the challenge of aligning technology with user needs even more relevant to avoid costly mistakes.

Relevance beyond ML to other technologies

Feedback from technological innovation practitioners shows promising signals of the outcomes being relevant beyond ML and Artificial Intelligence. Considering that technology push and market pull have been discussed for decades, that seems plausible. As one of the experts quoted :

"Innovation = invention + business case"

Responsible Innovation, and addressing risks before they lead to harmful accidents, is important beyond self learning algorithms. The literature and research in this report showcases multiple examples. While the importance of different aspects of the challenge change across different domains, the need for such solutions stays equally important.

5.1.5 The Toolkit

Value provided by the canvases

Both canvases guide the designers and rest of the product team through a structured process, and lead to actionable steps for designers to take.

The Tech Value Canvas helps them consider the market requirements as well as technological possibilities at the same time. The proposed opportunity finding process helps find market pull that aligns with the existing technology push : “We want to build something valuable using X technology. Let's find out which market we could serve and how this technology can solve their problems.” The outcomes of the canvas lead to execution steps that can be directly added to the team’s product backlog.

The Risk Discovery Canvas helps designers & product teams follow a similar process for finding out potential concerns and planning actions to address them. The goal is to anticipate risks, not predict them. From a value perspective, it is possible to think of risk discovery like investing into the future, just like technical debt or R&D. The outcome of this canvas too lead to actionable steps that can be directly included in the team’s product backlog.

Value provided by the cards

Both card decks support teams by prompting designers and other to consider the abilities and risks of FMs while brainstorming product ideas. They communicate knowledge and understanding required by designers and support them in becoming effective at contributing to technological innovation and being responsible.

As a boundary object, they enable a common language of discussion between designers, engineers and other people of a product development team. But the most interesting aspects of the cards for me was how they supported designers in using the canvases.

Canvas + Cards : 1+1=3 ?

Canvases are generally used for collaboration and communication across teams. Cards are often used to trigger ideation, ask important questions, carry important information, etc. They are seldom used together in a single application. Doing this seems to make them more effective than if used separately. The cards support teams in using the canvas, consequently making the canvas more effective and the canvas structures the use of the cards, making them more effective in contributing to the designer’s goals.

The canvases make the card use more structured and more reliable in contributing to the brainstorm process. The cards carry the technological understanding that designers need to work through the canvas. In that way, they make each other more effective. That makes the toolkit greater than the sum of its parts.

Opportunities & Risks : 2 sides of the same technology

The outcomes suggest that the process of being responsible might not be very different from that of being innovative. The way teams de-risking product success and business success is similar to de-risking products from potential harm. For self learning systems, both processes currently need upskilling. So both should be equally easy/hard for designers.

The similarities are also visible in the way the proposed tools for finding opportunities are extremely comparable to those of finding risks : The way they both start with the people involved, support designers to be educated about the possibilities, and end with plans for executing towards the end goals. They both follow a process of building an understanding of the context, ideating how the technology can impact that context in good and bad ways, and then planning steps to act on those findings.

Design tools as training wheels for novice designers

The proposed toolkit intends to support designers who are new to technology innovation, foundation models, and collaborating with engineers in interdisciplinary teams. The cards are designed to support designers without a technical understanding of how LLMs work and what they can/cannot do.

From the user tests & discussions with fellow master students at the IDE faculty, some of them expressed their tendency to not follow the suggested step-by-step process in the Tech Value Canvas. Because they had already worked as designers in interdisciplinary teams on technological innovation, they already had a good understanding of how the process worked. Because of that, they knew that they could be more flexible in their approach as long as they satisfied the underlying mechanism of matching technological possibilities to user requirements. This flexibility actually helped them look at the problem from different perspectives and brainstorm more creatively. The canvas then served as an overview of the different aspects of the innovation process and an external representation of the thought process in their heads.

The same designers did not have a similar opinion about the Risk Discovery Canvas. Because this process was completely new to them, they felt the need to use the recommended flow of the canvas to support them. It is likely that as they become more comfortable with the process and use the canvas multiple times, they might shift to a more flexible approach here too.

A similar difference was observed between engineers and designers on how they used the cards. The engineers at Scitodate already had a good understanding of LLM abilities and only used the Ability Cards to communicate with the non-engineers in the brainstorms. They referred to the details on the Risk cards only for knowing about the social aspects of these socio-technical risks. Designers who did not understand the LLM-specific technical jargon on the cards reflected that they can still use them to communicate with engineers about specific abilities and risks: “I want to focus on addressing THIS risk”.

It is possible that just like the canvases, designers might eventually “outgrow” the cards too, and integrate that understanding into their own minds, later using the cards primarily to communicate and focus a group of people onto a specific topic. Reaching that stage would imply that the toolkit was successful in its goal of supporting designers in working with Foundation Models.

5.2 RECOMMENDATIONS

5.2.1 Practioner Recommendations

Move Fast, to not break things

“Move Fast and break things” was a motto popularised by Mark Zuckerberg to explain his experimental approach to innovation. He claimed that “If you aren’t breaking things, you aren’t moving fast enough” The idea behind it was that in the pursuit of rapid progress and innovation, it was acceptable to take risks and potentially encounter failures or "break" things along the way.(Blodget, 2009)

The same perspective of “Moving fast” can help teams to identify how things can break before they actually do, and then prevent that. Rapid product development, testing, monitoring & iteration can positively serve the goal of effective risk management. For that to happen, rigorous testing is needed before solutions are deployed to customers. And supported with moving fast post monitoring, acting quickly on detected anomalies and malfunctions. Risk discovery has the potential to help teams identify what can break, so that teams can move fast and avoid them from breaking..

Become Antifragile

Basically, setup systems that help the team benefit from unanticipated risks and unexpected AI system behaviour. That can be done by integrating these observations & learnings into the product development process to prevent similar consequences in the future. Monitoring systems closely and frequently can help detect such risks and hazards as soon as they emerge.

Acting quickly on such suspicious and abnormal AI system behaviour can help address risks before they cause significant harm. A combination of this continuous system improvement combined with preventing risks from leading to harm can help products and organisations become more resilient and capable of better managing risks in the future.

Complement the proposed Toolkit

There is significant potential in using other tools in conjunction with the proposed toolkit. For Risk Discovery Canvas, VSD methods can help in identifying and representing stakeholders, risk matrices (probability vs severity evaluation) & FMECA can help prioritize mitigation steps, etc. For Tech Value Canvas,use persona development tools for mapping their goals & challenges. Use one of the many feature prioritization methods like MoSCoW, RICE, etc. for filtering through all the ideas to proceed with the most promising candidates.

Apart from that, the toolkit proposed in this thesis is intended to be adaptable and easy to modify. The canvases can become more effective and efficient by adapting them to specific company contexts and markets. Both card decks are more effective when they are comprehensive. As LLMs and other Foundation Models continue evolving, the card decks will need new cards that reflect the new abilities and risks that these technologies bring. Teams can create their own company specific cards that represent unique internal technology. The process documented can aid designers in designing these, either modifying the proposed solutions, or starting from scratch.

It is possible to consider Risk cards & Ability cards as an internal knowledge database for an organisation. As the team discovers new abilities & risks, adding cards for them will help make the database comprehensive and up-to-date. They can thus form internal assets, just like design systems.

5.2.2 Limitations & Future work

The Incentives Challenge

Even when teams have every intent to act ethically, sometimes ethical practices do not align with what's best for a business's economic performance. Not only is that extremely relevant in the context of startups, larger organisations also prioritise economic gains over responsible practices. While this thesis proposes some directions for further research and the AI Act aims to enforce responsible behaviour, there is a need to identify factors that can incentivise organisations to prioritize ethical behaviour.

Finding value propositions for ethical practices, and finding economic incentives for ethical practices can increase their adoption. Aligning stakeholder interests can help create the required incentives. We've seen something similar happen with sustainable clothing (Patagonia) and sustainable grocery supply chains (The Path Forward for Sustainability in European Grocery Retail, 2021). They were all partially driven by market sentiment & awareness towards environmental problems. We are entering a similar stage right now, with plenty of public discourse on the topic of AI risks. Startups are a good candidate for bringing about this change.

Supporting Stakeholder Participation

Taking a participatory design approach is seems quite possible for the risk discovery process. Cards can act as boundary objects, and the canvas can align everyone along the same process, enhancing shared understanding.

From discussions with the team at Scitodate, they were not very keen on participatory design styled stakeholder involvement. Apart from operational challenges like finding time to do these activities with multiple stakeholders, working on the canvas was not the right time to involve the stakeholders, in their opinion. They preferred to instead do the stakeholder interviews and research beforehand and then use the information and insights. Better processes can be explored to ensure sufficient understanding or involvement of stakeholders. That is critical to ensure effective risk discovery.

Foundation Model Value Chain

The scope of this thesis focussed only on a startup's internal circle of influence. The entire FM Value Chain that connects different actors involved in building, distributing and using Foundation Models needs to be studied. Aligning incentives for responsible behaviour across different stakeholders in this process is important and designers are well suited for this challenge.

Environmental Impact of large models

Large ML algorithms currently consume a significant amount of computing power to train and run, consuming large amounts of water & energy in the process. (Li, 2023) My assumption during the time of the thesis is that they will become efficient over time, just like personal computers, etc. But it is an assumption, and environmental risks can othewise be a serious concern with large Foundation Models.

Comprehensive Standardised Cards

The card proposed with this thesis are not intended to be comprehensive or the most optimised solution. Plenty of work can be done to try & standardise the definition of these abilities. Work is already underway elsewhere in new directions (Yildrim et al., 2023). The research on risk taxonomies can be better utilized to include more risk cards that cover different types of harm.

Legal Risk cards can help designers identify what parts of the AI Act are applicable to a solution. That can then help designers be mindful of the legal requirements during the design stage.

Validating & Improving the Toolkit

While the canvases and cards in the proposed toolkit were tested multiple times over the course of this thesis, a more rigourous testing process is needed to confirm the effectiveness and value of the proposed tools. It's especially important to test them in the context of other AI startups and Foundation Models apart from LLMs.

PERSONAL REFLECTIONS

Designing for a continuously evolving, rapidly improving technology has been challenging but at the same time extremely thrilling & satisfying. Some personal learnings from outside academia were extremely useful to tackle this situation. The Lindy Effect (Marcus, 2021) being one of them. In this way, I guess that this project made me better at decision making under uncertainty.

The constraints of designing something pragmatic and immediately usable for Scitodate and designers was a very good framing. It was a good constraint that led to effective solutions. The focus on iterative development and validation through continuous testing helped achieve an outcome that promises to be a useable tool.

My engineering background definitely proved beneficial during this thesis. Although I studied mechanical engineering in the past with zero software engineering exposure, technical conversations with engineers during the project were deeply satisfying and enjoyable. It was also very easy for me to communicate with them in their technical style for many aspects of work. Previous side projects in business development made it equally easy to communicate with the growth team at Scitodate about their commercial priorities.

In a strange way, the outcomes reflect who I am as a person and as a designer : the aspect of bridging the gap between designers and engineers, crossing disciplinary boundaries at multiple levels of the topic, finding insights from different domains and using them in the problem at hand, being optimistic about the opportunities while being considerate of the risks. It's weirdly authentic. Its an extension of what I know, how I think, who I am.

While the designed artefacts are processes and tools to follow that process, I acknowledge the potential limitations of relying on process alone.

Does brilliance need process ? No. Brilliance can definitely be complimented with good processes, but processes are not a pre-requisite for brilliance.

Can process replace brilliance ? I do not think so. A good process could make up for a lack of brilliance, but only upto a point. Just like correlation might be an acceptable proxy for causation, only under specific or predictable, recurring situations.

Can process support brilliance ? Definitely yes.

Can process foster brilliance ? Possibly, but not always. I guess, if we continue to understand and learn from the design tools and processes we use, and integrate those insights into our own reasoning and worldview, that can lead to processes and tools not just supporting practitioners, but also helping us continually become better.

The past two years at TU Delft have been a life changing experience. This thesis feels like a fitting conclusion.


industrieel ontwerpen
DESIGN
FOR our
future
Paper
Waste

REFERENCES

A

Adkins, D., Alsallakh, B., Cheema, A., Kokhlikyan, N., McReynolds, E., Mishra, P., Procope, C., Sawruk, J., Wang, E., & Zvyagina, P. (2022). Prescriptive and Descriptive Approaches to Machine-Learning Transparency. CHI Conference on Human Factors in Computing Systems Extended Abstracts. https://doi.org/10.1145/3491101.3519724

Almus, M., Nerlinger, E.A. Growth of New Technology-Based Firms: Which Factors Matter?. Small Business Economics 13, 141–154 (1999). https://doi.org/10.1023/A:1008138709724

ALTAI. The Assessment List for Trustworthy Artificial Intelligence. https://altai.insight-centre.org/

AI Ethics Guidelines Global Inventory by AlgorithmWatch. AI Ethics Guidelines Global Inventory. https://inventory.algorithmwatch.org/

Amnesty International. (2022, October 3). Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya - Amnesty International. https://www.amnesty.org/en/documents/ASA16/5933/2022/en/

Artem Andreenko on X. (n.d.). Twitter. https://twitter.com/miolini/status/1634982361757790209

Artificial intelligence regulation. (2023). Global Law Firm | Norton Rose Fulbright. https://www.nortonrosefulbright.com/en/knowledge/publications/a473a1c1/artificial-intelligence-regulation

Avdiji, Hazbi & Elikan, Dina A. & Missonier, Stéphanie & Pigneur, Yves. (2018). Designing Tools for Collectively Solving Ill-Structured Problems. 10.24251/HICSS.2018.053.

Aven, T. (2014). The Concept of Antifragility and its Implications for the Practice of Risk Analysis. Risk Analysis, 35(3), 476–483. https://doi.org/10.1111/risa.12279

B

Stephanie Ballard, Karen M. Chappell, and Kristen Kennedy. 2019. Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology. In Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19). Association for Computing Machinery, New York, NY, USA, 421–433. https://doi.org/10.1145/3322276.3323697

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

James Bessen, Stephen Michael Impink, and Robert Seamans. 2022. The Cost of Ethical AI Development for AI Startups. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22). Association for Computing Machinery, New York, NY, USA, 92–106. https://doi.org/10.1145/3514094.3534195

Bing. (2023, February 8). Introducing your copilot for the web: AI-powered Bing and Microsoft Edge [Video]. YouTube. https://www.youtube.com/watch?v=rOeRWRJ16yY

Bishop, P., Hines, A., & Collins, T. (2007). The current state of scenario development: an overview of techniques. Foresight, 9(1), 5–25. https://doi.org/10.1108/14636680710727516

Blank, S. (2013). Why the Lean Start-Up Changes Everything. Harvard Business Review, 91, 63-72.

Blodget, H. (2009). Mark Zuckerberg On Innovation. Business Insider. https://www.businessinsider.com/mark-zuckerberg-innovation-2009-10

Borgovini, R., Pemberton, S., Rossi, M., & Ny, R. a. C. G. A. (1993). Failure Mode, Effects, and Criticality Analysis (FMECA). In DTIC. https://apps.dtic.mil/sti/citations/ADA278508

Bommasani, R. (2021, August 16). On the Opportunities and Risks of Foundation Models. arXiv.org. https://arxiv.org/abs/2108.07258

Bornstein, M., Appenzeller, G., Casado, M., & Dharris. (2023, January 19). Who owns the generative AI platform? Andreessen Horowitz. https://a16z.com/2023/01/19/who-owns-the-generative-ai-platform/

Bower, J. L., and C. M. Christensen. "Disruptive Technologies: Catching the Wave." Harvard Business Review 73, no. 1 (January–February 1995): 43–53.

Bradford, Anu, The Brussels Effect: How the European Union Rules the World (New York, 2020; online edn, Oxford Academic, 19 Dec. 2019), https://doi.org/10.1093/oso/9780190088583.001.0001

Brem, A., & Voigt, K. (2009). Integration of market pull and technology push in the corporate front end and innovation management—Insights from the German software industry. Technovation, 29(5), 351–367. https://doi.org/10.1016/j.technovation.2008.06.003

Brown, T. B. (2020, May 28). Language Models are Few-Shot Learners. arXiv.org. https://arxiv.org/abs/2005.14165v4

Bubeck, S et al. (2023, March 22). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv.org. https://arxiv.org/abs/2303.12712

Konstantine Buhler (2023). AI 50 2023. Sequoia Capital. https://www.sequoiacap.com/article/ai-50-2023/

C

Calo, Ryan, Artificial Intelligence Policy: A Primer and Roadmap (August 8, 2017). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3015350

Cagan, M. (2017). The four big risks - Silicon Valley Product Group. Silicon Valley Product Group. https://www.svpg.com/four-big-risks/

Cagan, M. (2017b). INSPIRED: How to Create Tech Products Customers Love. John Wiley & Sons.

Cagan, M. (2017c). Beyond lean and agile - Silicon Valley Product Group. Silicon Valley Product Group. https://www.svpg.com/beyond-lean-and-agile/

CB Insights. (2022). The top 12 reasons Startups fail. CB Insights Research. https://www.cbinsights.com/research/report/startup-failure-reasons-top/

Chairs, P. (2023, January 10). ACL 2023 Policy on AI Writing Assistance. ACL 2023. https://2023.aclweb.org/blog/ACL-2023-policy/

Chivukula, Sai Shruthi & Li, Ziqing & Pivonka, Anne & Chen, Jingning & Gray, Colin. (2021). Surveying the Landscape of Ethics-Focused Design Methods.

Christensen, C. M. (2023, April 4). Know your customers' "Jobs to be done." Harvard Business Review. https://hbr.org/2016/09/know-your-customers-jobs-to-be-done

Coad, A., Nightingale, P., Stilgoe, J., & Vezzani, A. (2020). Editorial: the dark side of innovation. Industry and Innovation, 28(1), 102–112. https://doi.org/10.1080/13662716.2020.1818555

Cross, N. (2011). Design thinking: Understanding How Designers Think and Work. Bloomsbury Visual Arts.

D

Databricks. Free Dolly: Introducing the world's first truly open Instruction-Tuned LLM. (2023, April 12).https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

Derbyshire, J., & Wright, G. (2014). Preparing for the future: Development of an 'antifragile' methodology that complements scenario planning by omitting causation. Technological Forecasting and Social Change, 82, 215–225. https://doi.org/10.1016/j.techfore.2013.07.001

Design Council. (n.d.). The Double Diamond https://www.designcouncil.org.uk/our-resources/the-double-diamond/

Devlin, J. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805v2

Dignum, V. (2019). Responsible artificial intelligence. In Artificial intelligence: Foundations, theory, and algorithms. https://doi.org/10.1007/978-3-030-30371-6

Dorst, C., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. Design Studies, 22(5), 425–437. https://doi.org/10.1016/s0142-694x(01)00009-6

Dorst, K. (2015). Frame Innovation: Create New Thinking by Design. MIT Press.

E

Ethics guidelines for trustworthy AI. (2019, April 8). Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

EUR-LEX - 52021PC0206 - EN - EUR-LEX. (n.d.). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

F

Felin, T., Gambardella, A., Stern, S., & Zenger, T. (2020). Lean startup and the business model: Experimentation revisited. Long Range Planning, 53(4), 101889. https://doi.org/10.1016/j.lrp.2019.06.002

Floridi, L. The European Legislation on AI: a Brief Analysis of its Philosophical Approach. Philos. Technol. 34, 215–222 (2021a). https://doi.org/10.1007/s13347-021-00460-9

Floridi, Luciano, The end of an era: from self-regulation to hard law for the digital industry (November 9, 2021b). http://dx.doi.org/10.2139/ssrn.3959766

Frayling, C. (1994). Research in art and design (Royal College of Art Research Papers, vol 1, no 1, 1993/4)

Friedman, Batya & Kahn, Peter & Borning, Alan & Zhang, Ping & Galletta, Dennis. (2006). Value Sensitive Design and Information Systems. 10.1007/978-94-007-7844-3_4.

FRKelly. (2023). The EU's AI Act: Where Are We Now And What Could Its Impact Be O EU Companies? | FRKelly. FRKelly. https://www.frkelly.com/library/eus-ai-act-where-are-we-now-and-what-could-its-impact-be-eu-companies

G

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J., Wallach, H., Daumé, H., & Crawford, K. (2018). Datasheets for datasets. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1803.09010

Georgi Gerganov on X. (2001, March 14). Twitter. https://twitter.com/ggerganov/status/1635605532726681600

Gerpott, T. J. (2005). Strategisches Technologie- und Innovationsmanagement. ResearchGate. https://www.researchgate.net/publication/200167158_Strategisches_Technologie-_und_Innovationsmanagement

Ghezzi, Antonio & Cortimiglia, Marcelo & Bortolini, Rafael. (2018). Lean Startup: a comprehensive historical review. Management Decision. ahead-of-print. 1-20. 10.1108/MD-07-2017-0663.

Greg Brockman on X. (2022, June 23). Twitter. https://twitter.com/gdb/status/1599683104142430208

Griffith, E., & Metz, C. (2023, March 14). AI funding frenzy escalates. The New York Times. https://www.nytimes.com/2023/03/14/technology/ai-funding-boom.html

Guides | Langchain. (n.d.). https://python.langchain.com/docs/guides

Guo, H., Wang, C., Su, Z., & Wang, D. (2020). Technology push or market pull? Strategic orientation in business model design and digital start-up performance*. Journal of Product Innovation Management, 37(4), 352–372. https://doi.org/10.1111/jpim.12526

H

Hagendorff, T. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds & Machines 30, 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8

Urlick Hansen (2023)What the European AI Act means for you, AI developer. https://encord.com/blog/what-the-european-ai-act-means-for-you/

Harris, J. (2023, May 23). 'There was all sorts of toxic behaviour': Timnit Gebru on her sacking by Google, AI's dangers and big tech's biases. The Guardian. https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases

David Hershey & Diego M. Oppenheimer

DevTools for language models — predicting the future. (2023). https://www.unusual.vc/post/devtools-for-language-models

Marius Hobbhahn and Tamay Besiroglu (2022), "Trends in GPU price-performance". Published online at epochai.org. Retrieved from: 'https://epochai.org/blog/trends-in-gpu-price-performance' [online resource]

Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base - analyst note. Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Sonya Huang, Pat Grady. (2023). Generative AI: a creative new world. Sequoia Capital. https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/

Huang, J. (2022, December 20). Towards Reasoning in large Language Models: A survey. arXiv.org. https://arxiv.org/abs/2212.10403

Huggingface. (GitHub). huggingface_hub/src/huggingface_hub/templates/modelcard_template.md at main · huggingface/huggingface_hub. GitHub. https://github.com/huggingface/huggingface_hub/blob/main/src/huggingface_hub/templates/modelcard_template.md

I

Ibryam, B. (2023). From fragile to antifragile software | Red Hat developer. Red Hat Developer. https://developers.redhat.com/blog/2016/07/20/from-fragile-to-antifragile-software

ISO/IEC 23894:2023. (n.d.). ISO. https://www.iso.org/standard/77304.html

ISO 31000:2018. (2022, February 4). ISO. https://www.iso.org/standard/65694.html

The Artificial Intelligence Act, Institutional context. (2022, December 20). The Artificial Intelligence Act. https://artificialintelligenceact.eu/context/

J

Jacco J. Van Der Kooij. (2023). Frameworks that Govern B2B Marketing & Sales | WbD. Winning by Design. https://winningbydesign.com/resources/blog/frameworks-that-govern-b2b-marketing-and-sales/

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

K

Kyril Kotashev. (2022, December 14). Startup Failure Rate: How Many Startups Fail and Why in 2023? Startup Failure Rate: How Many Startups Fail and Why in 2023? https://www.failory.com/blog/startup-failure-rate

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems, 25, 1097–1105. http://books.nips.cc/papers/files/nips25/NIPS2012_0534.pdf

L

THE LANDSCAPE OF ML DOCUMENTATION TOOLS. (n.d.). https://huggingface.co/docs/hub/model-card-landscape-analysis

Li, P. (2023, April 6). Making AI Less “Thirsty”: Uncovering and addressing the secret water footprint of AI models. arXiv.org. https://arxiv.org/abs/2304.03271

LMSYS Org. (2023, March 30) Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality |. https://lmsys.org/blog/2023-03-30-vicuna/

M

Marcus, E. (2021, June 22). What is the Lindy Effect? The New York Times. https://www.nytimes.com/2021/06/17/style/lindy.html

Marcus, G. (2023, May 22). In the Rush to AI, We Can’t Afford to Trust Big Tech. Time. https://time.com/6281737/ai-we-cant-trust-big-tech-gary-marcus/

Maslow, A. H. (1943). A theory of human motivation. Psychological Review, 50(4), 370–396. https://doi.org/10.1037/h0054346

Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. 2018. Does ACM’s code of ethics change ethical decision making in software development? In Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2018). Association for Computing Machinery, New York, NY, USA, 729–733. https://doi.org/10.1145/3236024.3264833

Microsoft Learn. Embrace responsible AI principles and practices - Training.https://learn.microsoft.com/en-us/training/modules/embrace-responsible-ai-principles-practices/

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

Model Card Guidebook., https://huggingface.co/docs/hub/model-card-guidebook

Mollick, E. (2023, March 30). What the lean startup method gets right and wrong. Harvard Business Review. https://hbr.org/2019/10/what-the-lean-startup-method-gets-right-and-wrong

Martin Monperrus. 2017. Principles of Antifragile Software. In Companion Proceedings of the 1st International Conference on the Art, Science, and Engineering of Programming (Programming '17). Association for Computing Machinery, New York, NY, USA, Article 32, 1–4. https://doi.org/10.1145/3079368.3079412

Moran, A. (2014). Agile Risk management. In Springer eBooks. https://doi.org/10.1007/978-3-319-05008-9

Morley, J., Floridi, L., Kinsey, L. et al. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Sci Eng Ethics 26, 2141–2168 (2020). https://doi.org/10.1007/s11948-019-00165-5

Munn, L. The uselessness of AI ethics. AI Ethics 3, 869–877 (2023). https://doi.org/10.1007/s43681-022-00209-w

Myers, S. M. D. G. (n.d.). Successful Industrial Innovations. A study of factors underlying innovation in selected firms. | National Technical Reports Library - NTIS. https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB211097.xhtml

N

Nemet, G. F. (2009). Demand-pull, technology-push, and government-led incentives for non-incremental technical change. Research Policy, 38(5), 700–709. https://doi.org/10.1016/j.respol.2009.01.004

NIST. (2023) NIST Risk Management Framework aims to improve trustworthiness of Artificial intelligence |. NIST. https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial

NIST. NIST AIRC - PlayBook. NIST Trustworthy & Responsible AI Resource Center. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

Norman, D. A., & Verganti, R. (2014). Incremental and Radical Innovation: Design Research vs. Technology and Meaning Change. Design Issues, 30(1), 78–96. https://doi.org/10.1162/desi_a_00250

Donald A. Norman. 2010. Technology first, needs last: the research-product gulf. interactions 17, 2 (March + April 2010), 38–42. https://doi.org/10.1145/1699775.1699784

Donald A. Norman. 2002. The Design of Everyday Things. Basic Books, Inc., USA.

O

O'Connor, Gina & Demartino, Richard. (2006). Organizing for Radical Innovation: An Exploratory Study of the Structural Aspects of RI Management Systems in Large Established Firms. Journal of Product Innovation

Management. 23. 475 - 497. 10.1111/j.1540-5885.2006.00219.x.

Open AI. (2018). AI and compute. AI And Compute, Open AI. https://openai.com/research/ai-and-compute

OpenAI. (2023). GPT-4 Technical Report. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.08774

OpenAI Platform. (n.d.). https://platform.openai.com/docs/models/overview

Osterwalder, A. (2004). The business model ontology a proposition in a design science approach (Doctoral dissertation, Université de Lausanne, Faculté des hautes études commerciales).

Osterwalder, A., Pigneur, Y., Bernarda, G., & Smith, A. (2014). Value Proposition design: How to Create Products and Services Customers Want. John Wiley & Sons.

Ouyang, L., Wu, J., Xu, Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., John, S., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. J. (2022). Training language models to follow instructions with human feedback. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2203.02155

Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E., & Guston, D. H. (2013). A framework for responsible innovation. In John Wiley & Sons, Ltd eBooks (pp. 27–50). https://doi.org/10.1002/9781118551424.ch2

P

Pannozzo, A. (2010). The (Ir) relevance of Technology: Creating a Culture of Opportunity by Design. Design Management Review, 18(4), 18–24. https://doi.org/10.1111/j.1948-7169.2007.tb00090.x

Pantiuchina, J., Mondini, M., Khanna, D., Wang, X., Abrahamsson, P. (2017). Are Software Startups Applying Agile Practices? The State of the Practice from a Large Survey. In: Baumeister, H., Lichter, H., Riebisch, M. (eds) Agile Processes in Software Engineering and Extreme Programming. XP 2017. Lecture Notes in Business Information Processing, vol 283. Springer, Cham. https://doi.org/10.1007/978-3-319-57633-6_11

The path forward for sustainability in European grocery retail. (2021, June 14). McKinsey & Company. https://www.mckinsey.com/industries/retail/our-insights/the-path-forward-for-sustainability-in-european-grocery-retail

Dylan Patel and Afzal Ahmad, 2023.Google "We Have No Moat, And Neither Does OpenAI". www.semianalysis.com/p/google-we-have-no-moat-and-neither

Patil, S. G., Tianjun Zhang, Xin Wang, Joseph E. Gonzalez (2023, May 24). Gorilla: Large Language Model Connected with Massive APIs. arXiv.org. https://arxiv.org/abs/2305.15334

Patnaik, D., & Becker, R. H. (2010). NeedFinding: The why and how of uncovering people's needs. Design Management Journal, 10(2), 37–43. https://doi.org/10.1111/j.1948-7169.1999.tb00250.x

Policy and investment recommendations for trustworthy Artificial Intelligence. (2019, June 26). Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems With Applications, 97, 205–227. https://doi.org/10.1016/j.eswa.2017.12.020

R

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Rausand, M., & Haugen, S. (2020). Risk assessment. https://doi.org/10.1002/9781119377351

Rebelo, M. (2023). The best AI video generators in 2023. Zapier. https://zapier.com/blog/best-ai-video-generator/

Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 145 (April 2023), 27 pages. https://doi.org/10.1145/3579621

Reid, S., & De Brentani, U. (2004). The fuzzy front end of new product development for discontinuous innovations: a theoretical model. Journal of Product Innovation Management, 21(3), 170–184. https://doi.org/10.1111/j.0737-6782.2004.00068.x

K. H. Roberts and D. M. Rousseau, "Research in nearly failure-free, high-reliability organizations: having the bubble," in IEEE Transactions on Engineering Management, vol. 36, no. 2, pp. 132-139, May 1989, doi: 10.1109/17.18830.

Roose, K. (2023, February 17). Why A conversation with Bing's chatbot left me deeply unsettled. The New York Times. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

S

Sandberg, B., & Aarikka-Stenroos, L. (2014). What makes it so difficult? A systematic review on barriers to radical innovation. Industrial Marketing Management, 43(8), 1293–1305. https://doi.org/10.1016/j.indmarman.2014.08.003

Sanh, V. (2019, October 2). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv.org. https://arxiv.org/abs/1910.01108

Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, Michael Lennon. (2020, June 8). Principles to Practices for Responsible AI: Closing the gap. arXiv.org. https://arxiv.org/abs/2006.04707

Schiffer, Z., & Newton, C. (2023). Microsoft lays off AI ethics and society team. The Verge. https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs

Schuett, J. (2023). Risk Management in the Artificial Intelligence Act. European Journal of Risk Regulation, 1-19. doi:10.1017/err.2023.1

Scribus – Open Source Desktop Publishing. https://www.scribus.net/

Sevilla, J. (2022, February 11). Compute trends across three eras of machine learning. arXiv.org. https://arxiv.org/abs/2202.05924

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla, N., Gallegos, J., Smart, A., García, E. G., & Virk, G. (2022). Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2210.05791

Smith, D. (2019). What everyone gets wrong about this famous Steve Jobs quote, according to Lyft's design boss. Business Insider Nederland. https://www.businessinsider.nl/steve-jobs-quote-misunderstood-katie-dill-2019-4/

Standard setting. (2022, December 20). The Artificial Intelligence Act. https://artificialintelligenceact.eu/standard-setting/

Steve Blank What’s A Startup? First Principles. (2010, January 25). Steve Blank. https://steveblank.com/2010/01/25/whats-a-startup-first-principles/

Di Stefano, Giada and Gambardella, Alfonso and Verona, Gianmario, Technology Push and Demand Pull Perspectives in Innovation Studies: Current Findings and Future Research Directions (July 3, 2007). Research Policy, Volume 14, Issue 8, October 2012, 1283-1295, Available at SSRN: https://ssrn.com/abstract=1155050

Strategyzer. (2017, March 8). Strategyzer’s value proposition Canvas explained [Video]. YouTube. https://www.youtube.com/watch?v=ReM1uqmVfP0

Strauss, A., & Corbin, J. (1997). Grounded theory in practice.

T

Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) | NIST. NIST. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10

Taleb, N. N. (2012). Antifragile: Things That Gain from Disorder. Random House.

The Tarot Cards Of Tech. (n.d.). https://tarotcardsoftech.artefactgroup.com/

The AI dilemma. (2023). https://www.humanetech.com/podcast/the-ai-dilemma

Thiel, P., & Masters, B. (2014). Zero to one: Notes on Startups, or How to Build the Future. Currency.

Thomas Zoe (2023, May 9). Will AI gold rush lead investors to risky bets? - Tech News Briefing - WSJ Podcasts. WSJ. https://www.wsj.com/podcasts/tech-news-briefing/will-ai-gold-rush-lead-investors-to-risky-bets/704f9464-e5bc-4567-a58b-da2c63c24003

Thoring, Katja & Mueller, Roland & Badke-Schaub, Petra. (2019). Exploring the Design Space of Innovation Canvases. Conference Proceedings of the Academy for Design Innovation Management. 2. 10.33114/adim.2019.06.243.

Touvron, H et al. (2023, February 27). LLAMA: Open and Efficient Foundation Language Models. arXiv.org. https://arxiv.org/abs/2302.13971

Tristan Harris on X. (2023). Twitter. https://twitter.com/tristanharris/status/1634299911872348160/photo/1

U

Uschold, Michael & Grüninger, Michael. (1996). Ontologies: Principles, methods and applications. The Knowledge Engineering Review. 11.

V

Vakkuri, V., Kemell, KK., Jantunen, M., Abrahamsson, P. (2020). “This is Just a Prototype”: How Ethics Are Ignored in Software Startup-Like Environments. In: Stray, V., Hoda, R., Paasivaara, M., Kruchten, P. (eds) Agile Processes in Software Engineering and Extreme Programming. XP 2020. Lecture Notes in Business Information Processing, vol 383. Springer, Cham. https://doi.org/10.1007/978-3-030-49392-9_13

Vaswani, A. (2017). Attention is All you Need. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Verganti, R. (2006). Innovating through design. Harvard Business Review. https://hbr.org/2006/12/innovating-through-design

Verganti, R. (2011). Radical Design and Technology Epiphanies: A new focus for research on design management. Journal of Product Innovation Management, 28(3), 384–388. https://doi.org/10.1111/j.1540-5885.2011.00807.x

Vincent, J. (2016, March 24). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. The Verge. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

W

Walters, J. (2023, July 19). Complying with the EU AI Act. arXiv.org. https://arxiv.org/abs/2307.10458

Wang, Y. (2022, December 20). Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv.org. https://arxiv.org/abs/2212.10560

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

Weick, K. E., & Sutcliffe, K. M. (2007). Managing the unexpected: Resilient Performance in an Age of Uncertainty. Jossey-Bass. ISBN: 9780787996499

Amy A. Winecoff and Elizabeth Anne Watkins. 2022. Artificial Concepts of Artificial Intelligence: Institutional Compliance and Resistance in AI Startups. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22). Association for Computing Machinery, New York, NY, USA, 788–799. https://doi.org/10.1145/3514094.3534138

Y

Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T. Iqbal, and Jaime Teevan. 2019. Sketching NLP: A Case Study of Exploring the Right Things To Design with Language Intelligence. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 185, 1–12. https://doi.org/10.1145/3290605.3300415

Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, Eoin Ó Loideáin, Azzurra Pini, Medb Corcoran, Jeremiah Hayes, Diarmuid J Cahalane, Gaurav Shivhare, Luigi Castoro, Giovanni Caruso, Changhoon Oh, James McCann, Jodi Forlizzi, and John Zimmerman. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 483, 1–13. https://doi.org/10.1145/3491102.3517491

Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, and John Zimmerman. 2023. Creating Design Resources to Scaffold the Ideation of AI Concepts. In Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23). Association for Computing Machinery, New York, NY, USA, 2326–2346. https://doi.org/10.1145/3563657.3596058

Z

Zhao W. X., et al. (2023, March 31). A survey of large language models. arXiv.org. https://arxiv.org/abs/2303.18223

John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07). Association for Computing Machinery, New York, NY, USA, 493–502. https://doi.org/10.1145/1240624.1240704

APPENDIX

Appendix A : Interview Guide for interviews with Scitodate employees

QUESTIONS ABOUT CURRENT CLIENTS :

1) How do we currently categorise our users ? What does the market / customer segmentation look like ?
2) What do our current customers value in our products ?
   a) Value that they get detailed info about their potential leads & markets
   b) How is that different for Intelliscope & Market Landscape ?
3) Can you tell me about how our tools fit as part of their workflow ? How does the day of our users look like ? How is it different for different customer segments ?
4) Is it sometimes the case that the user of our solutions is different from the person who approves to pay for our product ? How does that sales funnel look like ? How common is it and in what types of organisations ?
5) What are the most common feature requests that we receive from existing clients ?

QUESTIONS ABOUT SELLING TO CURRENT POTENTIAL CUSTOMERS :

1) What does our sales pitch consist of? How does that vary for different end users & application?
2) What are common concerns and questions from potential customers ?
3) What are current roadblocks to closing sales ? What are the common reasons clients decline a purchase ? What are some reasons they stop using our products ?
4) What product features or functionality would you propose that would be highly beneficial for current/ near future sales efforts

QUESTIONS ABOUT FUTURE CUSTOMERS :

1) What direction do you think is the most promising for expanding our user base and increasing the company's revenue ?
2) How would you want our products to evolve over the next 1-2 years to ensure the best revenue & sales performance ? Some ideas for how we can create greater value for customers ?
3) If you have the chance to design the next Scitodate product for the larger B2B DeepTech market, what might that look like ?
4) What are our plans for future sales strategies apart from our current way of doing sales ? What do we currently struggle with ?

Appendix B : Interview Guide for interviews with Scitodate customers

1) How does a normal workday look like for you ? What kinds of tasks do you do ?

2) Can you walk me through your general workflow?
   What do you find time consuming, frustrating, challenging
   What are the boring repetitive parts of your work
   What other software tools do you use ?
   Reluctance to use digital tools ?

3) What Scitodate tools do you use ? (Intelliscope ILS, Market Landscape MLS)
   Can you walk us through when, why & how you use them ?
   What do you find useful in our current solutions ?
   Where do you think we can do better ?

4) Have you tried using the AI tools in Intelliscope ?
   What else do you think could we make for you ?
   Where do you see this AI trend going in the future ?

5) If you had a personal assistant / intern working for you, willing to do whatever you would like to outsource, what kind of work would that be ?
   With the extra time that you would have then, what would you focus on instead ?

6) Can you imagine a future where you only need to do the interesting stuff, and everything else is automated for you ? Can you walk us through that workflow ?

7) How does lead generation look like in academia vs industry ?