ABSURDIST CHATBOTS

A Pataphysical Exploration of Trustworthy AI

ROMÉE POSTMA

Author
Romée Postma

Master Thesis
MSc. Design for Interaction
Faculty of Industrial Design Engineering
Delft University of Technology

Graduation committee
Dr. Roy Bendor (chair)
Dr. Kars Alfrink (mentor)

September, 2024

TU Delft

abstract

In an era dominated by technological solutionism, artificial intelligence (AI) is often seen as a quick fix for complex problems. Inspired by the surrealist notion of pataphysics, this graduation project employs a novel, critical design approach to address the imaginary problem of trustworthy AI and its solutions in the context of generative chatbots.

This project presents three pataphysical chatbots, each addressing a specific requirement—transparency, human control, or fairness—currently implemented in AI systems through technical mechanisms intended to ensure trust. Each chatbot is designed to deliberately exaggerate the mechanism behind the assumed requirement, aiming to help users understand the absurdist nature of the chatbot and, ultimately, provoke critical reflection on the pursuit of trustworthy AI.

Through an iterative design process of prototyping, testing, and reflecting upon pataphysical chatbots, progress was made toward the final pata-design, which was evaluated in an online experiment. A key insight emerged: while the pataphysical chatbots succeeded in conveying their absurdist nature to users, critical reflection on the pursuit of trustworthy AI did not occur. However, additional findings suggest a potential link between understanding the absurdity and the occurrence of this particular type of critical reflection.

Ultimately, this project contributes to the critical design practice and in particular the field of pataphysical design, by demonstrating how pataphysical chatbots can effectively challenge both the prevailing methods for achieving trustworthy AI as the dominant problem-solution paradigm of AI solutionism. Moreover, the pataphysical design approach proposed in this project offers a new direction for current pataphysical design practices, emphasizing the creation of engaging pataphysical experiences that aim to provoke specific user responses to the premises of the imaginary problem.

glossary

Affirmative design	design that solves problems for the industry by conforming to, and therefore reaffirming, the cultural, economic, and technical expectations of society
Artificial Intelligence (AI)	technology that enables computers and machines to simulate human intelligence
Critical design	medium to critique and question societal norms, practices, and values, aiming to provoke critical reflection and discussion about the status quo and possible futures
Generative AI	AI technology that creates new content or data autonomously, such as text, images, or code, by learning patterns from existing data
Generative chatbots	AI-driven conversational agents that autonomously generate responses, simulating human-like dialogue
Human-Computer Interaction (HCI)	academic discipline focusing on the interaction between humans and computers
Imaginary problems	wicked problems framed as being easily solvable by technology
Imaginary solutions	technological solutions aiming to quickly fix imaginary problems
Pataphysics	absurdist 19th-century art movement proposing a 'science of imaginary solutions' as alternative to traditional scientific thinking
Pataphysical design	absurdist type of critical design that applies absurd approaches from pataphysics to address imaginary problems
Pata-prototyping	protototying, testing and reflecting upon pataphysical desings
Positivism	philosophical approach advocating that knowledge is rooted in sensory experience and scientific methods, rejecting other ways of knowing as intuition or speculation
Research through design (RTD)	a method where designing is used as a process to explore and generate new knowledge.
Solutionism	belief that for every problem there exists a solution, often reaching for answers before fully understanding the problems
Technological solutionism	ideology viewing technology as the key to solving complex social, political, and environmental challenges (wicked problems), often oversimplifying these issues for technological fixes
Tame problems	well-defined, stable issues with clear solutions that follow a linear problem-solving approach
Wicked problems	complex, ambiguous issues, intertwined with societal, moral, and political factors, that evolve over time and resist traditional problem-solving methods

preface

Since the start of my graduation project, I have been asked nearly every week, "*What is pata... pataphysics?*" I want to thank everyone who posed that question, because each time I tried to explain it, I gained a deeper understanding of it myself. Although my perspective on pataphysical design has evolved over time and will likely continue to do so, I now see it as a way to introduce absurdist designs into the world that prompt others to reflect on dominant problem-solution paradigms. What intrigues me most is how pataphysics achieves this by playfully blending imagination and reality.

I would like to thank my supervisors, Roy and Kars, for giving me the opportunity to work on such a captivating project. I thoroughly enjoyed all our meetings and always left feeling inspired. You gave me the freedom to explore while providing the guidance I needed to complete this project in a way I'm proud of. Roy, your sharp eye pushed me in the direction I wanted to move towards, and I'm genuinely inspired by your thoughts on challenging conventional norms through design. Kars, thanks for your boundless energy; your extensive knowledge on AI and trust propelled my project to the next level—no paper was left unexamined!

Furthermore, I would like to thank the 100+ participants of my pilots, user tests, and final experiment for your time and effort. To my family and friends, who always seemed to check in at just the right times—I really appreciated your encouragement. Last but not least, David, I'm grateful to have someone by my side who continuously supported me and was always willing to engage in endless conversations about pataphysics, to the point where we became pataphysical ourselves.

Enjoy the read, and don't forget to laugh from time to time!

table of contents

introduction

1.1 Context

Nowadays, humanity strongly believes that technology can provide quick fixes to complex real-world problems, an ideology known as technological solutionism (Morozov, 2013). The domination of the technological solutionist mindset is also clearly observable in the domain of AI (Moser et al., 2022). Recent initiatives aim to ensure AI is used responsibly to build public trust, reflecting the implicit belief that AI is a catch-all remedy for all our problems, referred to as AI solutionism (Lindgren & Dignum, 2023).

Many current discussions frame the widespread adoption of AI as both inevitable and beneficial for humanity, supporting the view that AI should be extensively utilized and that developing trustworthy AI is a crucial prerequisite (Kaur et al., 2023). The High-Level Expert Group on Artificial Intelligence—appointed by the European Commission—emphasizes the importance of trust as a foundational element for societies, economies, and sustainable development in the face of rapid technological change, and has published guidelines outlining requirements for trustworthy AI (AIHLEG, 2019). However, M. Braun et al. (2021) argue that AI HLEG falsely portrays AI as something that should be trustworthy, including the assumption that the general use of AI is beneficial. Additionally, Freiman (2022) offers a comprehensive critique of the concept of trustworthy AI. This raises the question whether the pursuit of trustworthy AI is truly desirable.

Pataphysical design—an absurdist type of critical design evolved from a late 19th-century surrealist art movement pataphysics—offers a way to challenge dominant problem-solution paradigms in a designerly way. This project builds upon the pataphysical design method proposed by Sicart & Shklovski (2020), who initiated a rebellion against the belief that software is a solution for everything. Instead, they designed pataphysical software that does not solve problems but explores imaginary problems and provides no solutions for them—ultimately using software as a tool to ask better and more interesting questions.

This project adopts a similar stance but shifted the focus to challenging AI solutionism within the context of generative AI chatbots, specifically addressing the imaginary problem of trustworthy AI. Through the iterative process of prototyping pataphysical chatbots, testing them with users, and reflecting on the outcomes, I formulated my final design goal: to design three pataphysical chatbots to open up a space for critical reflection on trustworthy AI, by exaggerating the mechanisms behind some assumed requirements for trustworthy AI—transparency, human control, and fairness—with each chatbot focusing on one of these mechanisms. The aim is to design the pataphysical chatbots in a way that allowed users to recognize their absurdist nature while also prompting critical reflection on the pursuit of trustworthy AI. Finally, the final pata-design is evaluated in an online user test to see whether the critical reflection on the pursuit of trustworthy AI occurred.

This project contributes to the critical design practice and in particular the field of pataphysical design, by demonstrating how pataphysical chatbots can effectively challenge both the prevailing methods for achieving trustworthy AI as the dominant problem-solution paradigm of AI solutionism. By building upon the work of Sicart & Shklovski, I proposed my own pataphysical design approach that differs from theirs in three key ways: 1) I incorporated iterative pata-prototyping activities into my approach, 2) I made use of a design goal to guide progress toward a final pata-design, and 3) I evaluated the final pata-design though a user evaluation. The final pataphysical chatbots demonstrate the potential of pataphysical design to create engaging experiences that make the absurdist nature of the pataphysical designs understandable to users, thereby provoking specific responses to the premises of the imaginary problem.

1.2 Project approach

This project is structured in three phases: a research phase, a design phase, and an evaluation phase (see Figure 1). The design phase also informed the research phase, as I employed a research through design approach (Stappers and Giaccardi, 2017). First, I outline the relevant information regarding the key terms of this project and their interconnections (Chapter 2). Next, I explain which specific pataphysical design method is used as a foundation and how I built upon it (Chapter 3). Through iterative prototyping, testing, and reflecting on pataphysical chatbots, I further explored the pataphysical method and incorporated literature-based solutions for trustworthy AI into the chatbots (Chapter 4). This process led to the formulation of a design goal, resulting in the creation of a final pata-design, presented in Chapter 5. In Chapter 6, the final evaluation is explained, results are presented, and insights are gathered. Finally, I reflect on the overall project and situate it within the context of academic literature and the design field (see Chapter7).

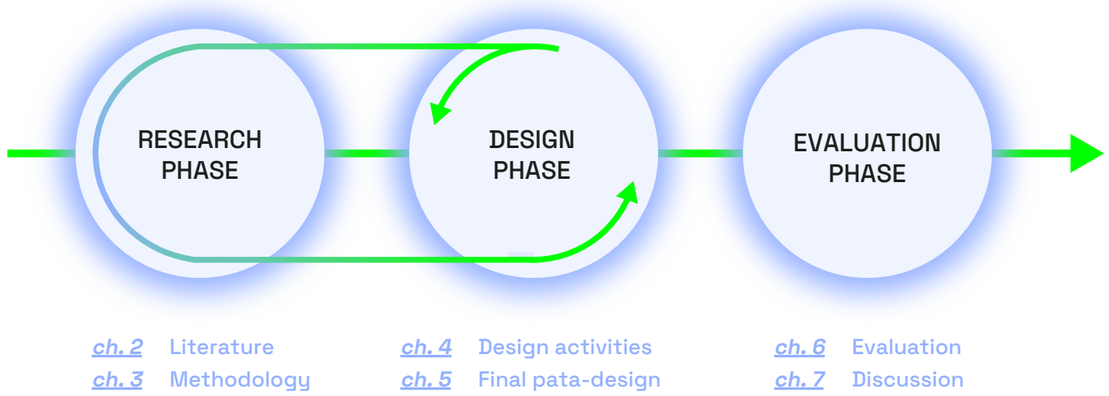


Figure 1: Project structure with related chapters.

literature

This chapter introduces the key themes of this project: trustworthy AI, generative chatbots, and pataphysical design, and investigates their interconnections based on literature research. Section 2.1 starts by explaining why trustworthy AI is currently a prominent topic and how it is intertwined with the concept of AI solutionism. It argues that trustworthy AI may be a non-existent problem, and then outlines the assumed requirements for trustworthy AI that are currently implemented to achieve it nonetheless. In Section 2.2, the specific context of generative AI chatbots is elaborated upon. Finally, Section 2.3 highlights the potential of pataphysical design to challenge AI solutionism, demonstrated in this project by embarking on a pataphysical exploration into the imaginary problem of trustworthy AI in generative chatbots.

2.1 Trustworthy AI

2.1.1 THE FOCUS ON TRUSTWORTHY AI

> TRUST IN AUTOMATION

The rise of automated technologies, encompassing systems that can operate autonomously, has drawn significant attention. The academic discipline of human-computer interaction (HCI) has become increasingly interested in understanding how people interact with and rely on these automated systems. Two central observations have been made in HCI research: 1) users sometimes resist using automated systems, while 2) users sometimes display an overreliance on automation (Wischnewski et al., 2023). To understand and explain these observations, special importance has been given to the role of trust in automation.

Lee & See (2004) define trust in automation as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p.5). In essence, trust is an attitude relevant in situations involving uncertainty, a cooperative relationship between entities, and some form of exchange. Literature indicates that trust is a pivotal variable in explaining both resistance to using automated systems (disuse) as well as overreliance on automated systems (misuse) (Lee & See, 2004; Madhavan & Wiegmann, 2007; Parasuraman & Riley, 1997). Wischnewski et al. (Wischnewski et al., 2023) state that to achieve appropriate reliance on automated systems, users' trust should be calibrated. In calibrated trust, the perceived trustworthiness of a system is aimed to be matched to the actual trustworthiness of the system, avoiding disuse or misuse of automated systems.

> TRUST IN AI

AI is a rapidly advancing automated technology that has quickly become an integral part of our daily lives. At its core, AI refers to computers or machines designed to simulate human intelligence. These systems perform tasks that typically require human intelligence, such as learning, problem-solving, understanding natural language, and making decisions (Shaina, 2023).

The High-Level Expert Group on Artificial Intelligence (AI HLEG), appointed by the European Commission, underscores the importance of trust as a fundamental element in societies, economies, and sustainable development amidst rapid technological change. The AI HLEG identifies trustworthy AI as their foundational ambition, asserting that people's confidence in the development and applications of AI hinges on achieving trustworthiness (AIHLEG, 2019). The guidelines also state that this confidence is needed to foster responsible and sustainable innovation, positioning Europe as a global leader in AI, and ensuring that industries, companies and individuals alike can benefit from AI systems. They consider it our collective responsibility as a society to work towards guaranteeing that trustworthy AI is ensured.

<u>Conclusion:</u> Currently, there is a strong emphasis on fostering trust in automated systems to ensure appropriate reliance, avoiding disuse and misuse. Within the realm of AI, achieving trustworthy AI is seen as crucial to accelerate the use and development of AI technologies.

2.1.2 AI AS CATCH-ALL SOLUTION

The urge to use automated systems like AI drives the need to develop trustworthy AI, both of which are deeply intertwined with the concept of AI solutionism—the belief that AI can solve all our problems. This concept has its roots in solutionism, broadly, and technological solutionism, specifically.

> SOLUTIONISM

Solutionism is the belief that for every problem a solution exists. The term was introduced by Dobbins (2011) , who warns that solutionism runs the risk of applying the *magic bullet model* to solve problems, meaning it reaches for answers before the questions have been fully asked or understood. Dobbins' critique on solutionism centers on the fact that the solution is often prioritized over understanding the problem. Problems are 'dumbed down' to fit the solutions offered, leading to solutions that presume rather than inveWstigate the problems, ultimately failing to make things better.

> TECHNOLOGICAL SOLUTIONISM

Technological solutionism is a form of solutionism that views technology as the key to solving societal, political, and environmental challenges. The concept gained notoriety through Morozov's critique in *To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems that Don't Exist* (2013). Morozov argues that while tame problems—stable and less complex issues—may benefit from technological solutions, this is often not the case with wicked problems—more dynamic and ambiguous issues. He underlines the tendency to oversimplify wicked problems so that they fit technological solutions, preventing these issues from being fully understood and addressed at their root. Morozov highlights that authorities like Silicon Valley-based tech companies drive the normalization of this oversimplification, treating technological solutions as a full replacement for other approaches. As a result, he states that the technological solutionist mindset has become our default way of tackling complex issues, although often without benefits.

> AI SOLUTIONISM

The domination of the technological solutionist mindset is also clearly observable in the domain of AI (Moser et al., 2022). Recent initiatives aim to ensure AI is used responsibly to build public trust, reflecting the implicit belief that AI can solve all social issues (Lindgren & Dignum, 2023). This notion, often referred to as AI solutionism, suggests that AI is a catch-all remedy for wicked problems. Lindgren & Dignum (2023) emphasize in their research on AI solutionism that an environment is fostered in which wicked problems are often simplified to fit AI solutions, without adequate consideration of whether these are the right problems to solve and whether AI is the appropriate tool for these challenges.

<u>Conclusion:</u> Many current debates suggest that the widespread adoption of AI is both inevitable and beneficial for humanity. This belief reinforces the idea that AI should be extensively utilized and that developing trustworthy AI is a critical prerequisite for that.

2.1.3 TRUSTWORTHY AI: A NON-EXISTING PROBLEM

As mentioned earlier, a prominent idea put forward by AI HLEG (2019) is that one should strive towards trustworthy AI. Given the dominance of AI solutionism, this does not seem so remarkable. However, M. Braun et al. (2021) argue in A Leap of Faith: Is There a Formula for "Trustworthy" AI? that AI HLEG falsely portrays AI as something that should be trustworthy, including the assumption that the general use of AI is beneficial. They criticize the conception of trust merely as an accelerator for societal acceptance of AI and stress that the significance of distrust in societal deliberation around the use of AI is often trivialized. Advantages of a healthy sense of distrust are that it enables individuals to opt out of technology if they wish, to constrain its power, and to exercise meaningful human control.

Furthermore, Freiman reviewed various critiques of the concept of trustworthy AI in *Making sense of the conceptual nonsense 'trustworthy AI'* (2022). The study builds on Thomas Metzinger's examination, a former member of the AI HLEG, who stated: "The underlying guiding idea of a 'trustworthy AI' is, first and foremost, conceptual nonsense. Machines are not trustworthy; only humans can be trustworthy (or untrustworthy)" (2019). Following this statement, Freiman elaborates on the risk of anthropomorphizing AI systems (Ryan, 2020), arguing that attributing human-like qualities, such as trust, to AI leads to incorrect assignments of moral and legal responsibility.

<u>Conclusion:</u> The dominant industries and policymakers consider trustworthy AI as a problem that needs solving. However, multiple scholars have criticized the concept of trustworthy AI, suggesting that it is only a problem within the narrow understanding of the concept. Therefore, trustworthy AI can be seen as a perceived, yet non-existent problem.

2.1.4 ASSUMED SOLUTIONS FOR TRUSTWORTHY AI

This project aims to challenge the idea that striving for trustworthy AI is crucial. To do this efficiently, it is necessary to first understand the current assumed solutions proposed for achieving trustworthy AI.

> A FRAMEWORK FOR TRUSTWORTHY AI

AI HLEG (2019) set out a framework for achieving trustworthy AI (see Figure 2). First, they argue that trustworthy AI has three components: "(1) it should be lawful, complying with all applicable laws and regulations; (2) it should be ethical, ensuring adherence to ethical principles and values; and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm." (p.5). As foundations of trustworthy AI they lay down four ethical principles: respect for human autonomy, prevention of harm, fairness and explicability. These ethical principles are translated into seven key requirements that AI systems should implement and meet throughout their entire life cycle via technical and non-technical methods. Finally, they set out a trustworthy AI assessment checklist to operationalise the key requirements and tailor this to specific AI applications.

> KEY REQUIREMENTS FOR TRUSTWORTHY AI

The key requirements in AI HLEG's framework for trustworthy AI are particularly interesting to examine, as they represent the current assumed solutions for achieving trustworthy AI. The seven key requirements are: 1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination and fairness, 6) societal and environmental wellbeing and 7) accountability, also visualized in Figure 2. In this project, three of the seven requirements presented by AI HLEG's framework—transparency, human control*, and fairness—will be further explored in the context of generative chatbots with the use of pataphysical design (see Chapter 4).

** Although AI HLEG uses the term "human agency and oversight," I prefer to name it "human control," as that can be viewed as a system property. In this chapter, In this chapter, I retain the term "human agency and oversight" to present the current assumed requirements for trustworthy AI as accurately as possible before offering my critique.*

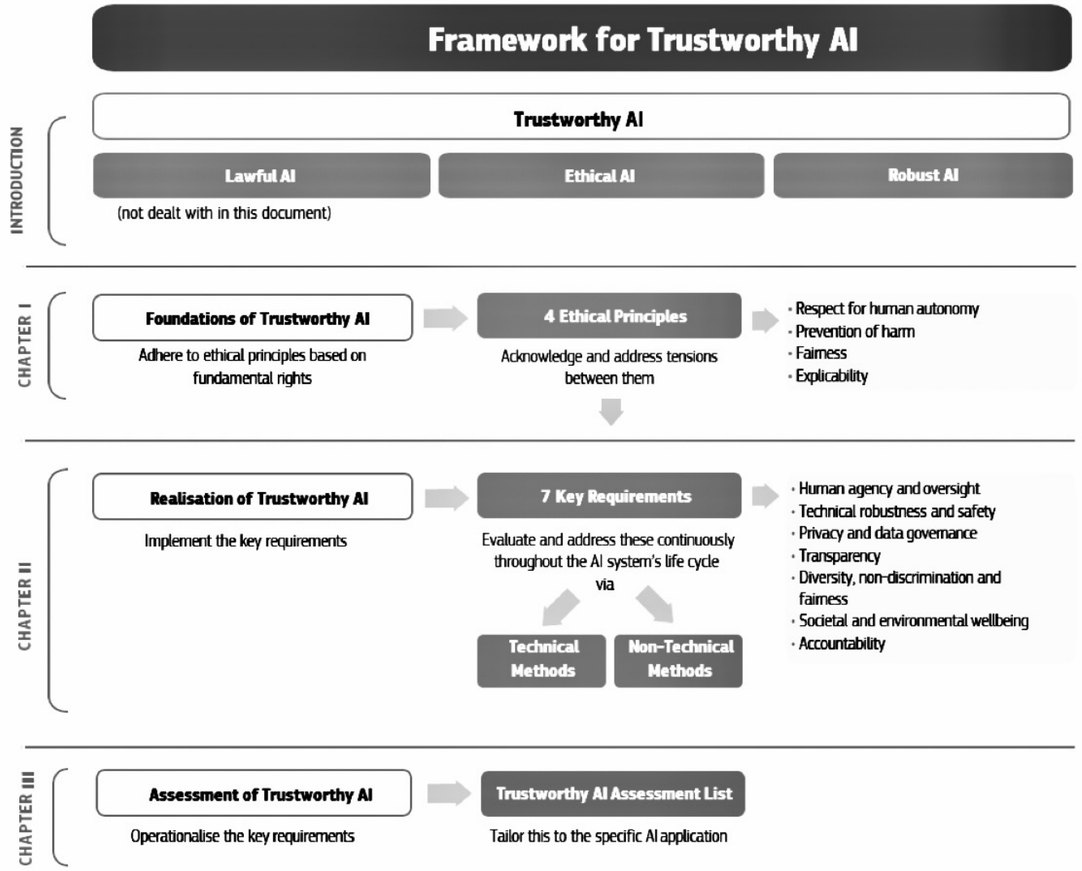


Figure 2: The framework for trustworthy AI compiled by AI HLEG (2019).

To understand what these requirements entail and how they are currently implemented in practice, the research by Díaz-Rodríguez et al. (2023) has been used as a guide. They adopted the seven key requirements for trustworthy AI compiled by AI HLEG and analyzed them from a triple perspective: what each requirement for trustworthy AI is, why it is needed, and how each requirement can be implemented in practice. Since this project focuses on transparency, human control, and fairness, only Díaz-Rodríguez et al.'s analysis of these specific requirements is outlined:

Requirement of transparency

As stated by Díaz-Rodríguez et al. (2023), transparency in AI ensures that information reaches stakeholders, such as developers, designers, users, and regulators. It includes simulatability (human ability to simulate the model), decomposability (explaining model behavior), and algorithmic transparency (understanding the model's processes and outputs). They argue that transparency is crucial for informing users about AI system capabilities and limitations, since it fosters trust and accountability by ensuring that humans know when they are interacting with AI systems.

To implement transparency in AI, several steps can be undertaken (Díaz-Rodríguez et al., 2023):

- Traceability: AI systems can include mechanisms to track data, development, and deployment processes through documented identification. Provenance tools, like Blockchain, can ensure the integrity of data used in machine learning models.
- Explainability: Explainable AI (XAI) techniques can clarify the AI system's functioning. These techniques can be used before or after the model is designed and trained to explain its decision-making process.
- Communication: Users must be informed about their interactions with AI systems, including performance, capabilities, and limitations. Explanations should be tailored to the audience's knowledge and delivered effectively to ensure understanding.

Requirement of human control

According to Díaz-Rodríguez et al. (2023), human agency in AI means that AI systems should empower humans, allowing them to make informed decisions and support their fundamental rights. This includes mechanisms that keep humans in control of the decision-making process, such as human-in-the-loop, human-on-the-loop, and human-in-command approaches. They state that human agency is crucial for maintaining autonomy and control, preventing unfair manipulation, deception, and conditioning by AI systems. Trustworthy AI should enable users to supervise, evaluate, and freely adopt or override AI decisions, ensuring that decisions are not made automatically without human involvement.

To implement human agency in AI, several steps can be undertaken (Díaz-Rodríguez et al., 2023):

- Human-Compatible and Human-Centric Approaches: AI systems can be designed using human-compatible and human-centric paradigms, such as AI for social good and interactive machine learning, to align with human rights and needs.
- Toolkits and Frameworks: Structured toolkits, like C-Suite, can be developed for domain-specific application of human control requirements, ensuring smooth implementation across various fields.
- Natural Language Processing (NLP): AI models using NLP, counterfactual explanations, and natural language explanations can facilitate communication between humans and AI, helping users supervise and make appropriate decisions based on AI outputs.

Requirement of fairness

As pointed out by Díaz-Rodríguez et al. (2023), fairness in AI involves techniques aimed at reducing the presence of unfair outputs from AI-based systems. An unfair algorithm produces decisions that favor a particular group of people. Biases can originate from data, the algorithm itself, or user interactions and need to be addressed to ensure equitable treatment for all users. They claim that fairness is essential to prevent negative impliations, such as the marginalization of vulnerable groups and the exacerbation of prejudice or discrimination. Ensuring fairness broadens the impact of AI to all social substrates and minimizes the negative effects of automated decisions influenced by biased data.

To implement the requirement of fairness in AI, several steps can be undertaken (Díaz-Rodríguez et al., 2023):

- Pre-Processing Methods: Available data can be transformed to reduce or remove sources of bias before training the model.
- In-Processing Methods: The learning algorithm can be modified to minimize the effect of biases during the training process, such as by changing the objective function or imposing constraints.
- Post-Processing Methods: The model's output can be adjusted to correct predictions without modifying the learning algorithm or training data, for example, by reassigning predicted classes.

Conclusion: The AI HLEG proposed several key requirements for ensuring trustworthy AI. Other literature researched how those requirements could be implemented in AI systems. This project will implement exactly those requirements in the design process.

2.2 Generative chatbots

2.2.1 SCOPING DOWN ON GENERATIVE CHATBOTS

Given the vast scope of AI, this project focuses on the specific field of generative AI chatbots. These chatbots are rapidly emerging and becoming increasingly integrated into various industries, making them a compelling subject for exploration. Additionally, generative chatbots are widely accessible to the public, making them ideal for my research through design approach, as users can easily interact with them. Generative chatbots perform a range of tasks, from translating text to aiding in coding. The most well-known generative chatbot is ChatGPT, which, although not the first generative AI tool to be publicly released, impressed a wide audience with its ability to offer immediate, useful answers to almost every question. While both literature and industries emphasize the importance of understanding the implications of generative AI (Rane, 2023), the exact effects remain difficult to predict (Klein & Mollick, 2024). As generative chatbots and various other software become more integrated into daily life, it becomes pertinent to explore the consequences of framing trustworthy AI as a problem in this domain.

2.2.2 CONVERSATIONAL AGENTS

Generative chatbots fall under the larger category of conversational agents, which are dialogue systems that use natural language processing (NLP) to automatically respond in human language (Griffing, 2023). Traditionally, NLP literature divides conversational agents into task-oriented and open-domain categories, as depicted in the taxonomy in Figure 3. Task-oriented agents are designed to function as interfaces within applications to accomplish specific tasks. In contrast, open-domain agents, or chatbots, are designed for entertainment and casual conversation, aiming to give users the impression of chatting with another intelligent being. Chatbots can be further classified into rule-based and data-driven categories. Rule-based chatbots generate responses based on a set of predefined rules. Data-driven chatbots, however, rely on patterns identified through statistical analysis or machine learning techniques applied to dialogue data. Within the data-driven category, there are generative and retrieval chatbots: generative chatbots create responses from scratch, while retrieval chatbots select appropriate responses from a predefined set of options.

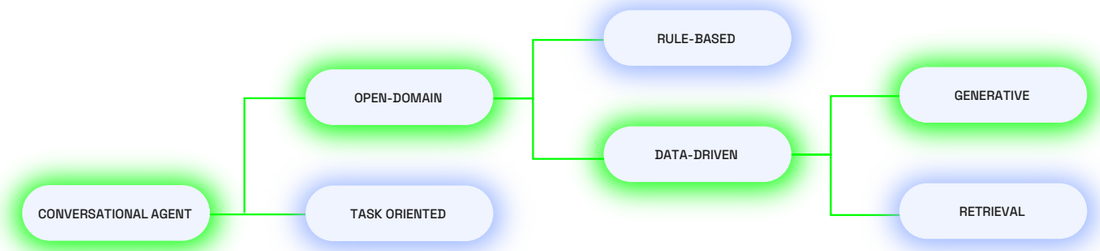


Figure 3: Conversational agents taxonomy (Scotti et al., 2024).

2.2.3 GENERATIVE CHATBOTS & TRUST

As described in Section 2.1.1, trustworthy AI is considered crucial for accelerating the use and development of AI technologies within HCI research. Focusing on the specific domain of generative AI chatbots, a study by Choudhury and Shamszare (2023) on the influence of user trust on the adoption and use of ChatGPT shows that trust is critical for users' adoption of the technology. Additionally, several studies highlight the importance of trust in generative AI across multiple domains, such as finance and healthcare (Dekkal et al., 2023; Loh, 2023). This underscores that, as with trustworthy AI in general, trust is indispensable for the widespread adoption of generative AI chatbots by society.

2.3 Pataphysical design

Inspired by the surrealist notion of pataphysics, pataphysical design offers a way to challenge AI solutionism in a designerly way. This project demonstrates this by using pataphysical design to critically respond to the present framing of trustworthy AI as a problem in the context of generative chatbots.

2.3.1 PATAPHYSICS

Pataphysics is a late 19th-century movement founded by the French novelist and writer Alfred Jarry (1873-1907). He characterized pataphysics as "the science of imaginary solutions", intended as a parody of science (Hugill, 2012). The movement aimed to challenge conventional interpretations of reality, carried out through humor, often in the form of absurdism.

> PATAPHYSICS CRITICIZING POSITIVISM

Pataphysics emerged as a critical response to positivism, the dominant philosophy of science at the time. Positivism developed through several stages known by various names, such as logical positivism and logical empiricism (Blumberg & Feigl, 1931). In its essence, positivism advocates that reality is measurable and encompasses only what one can directly observe (Shannon-Baker, 2023). It believes that knowledge is rooted in sensory experience and scientific methods and should focus on observable and measurable phenomena, rejecting other ways of knowing such as intuition or speculation.

Pataphysics, on the other hand, critiques this rigid framework by embracing other ways of knowing. It posits that reality is not limited to what can be empirically observed and measured. Instead, pataphysics views reality as a human construct, without a singular truth that is directly observable. Pataphysics celebrates the exceptions and the imaginative possibilities that positivism tends to dismiss. By doing so, it challenges the notion that positivism is the only valid path to understanding the world.

> PATAPHYSICS AND HUMOR

As Christe & Ritzen (2021) state in *The Manual of Modern Pataphysics, Vol. 1*, one of the main pillars of pataphysics is humor, often in the form of absurdism, or at least, an illogical sequence of actions. They explain that the most notorious form is asymmetric humor, the basic concept from which absurdism arises. Asymmetry literally means 'uneven' and often subverts expectations, creating a sense of surprise or discomfort that challenges conventional thinking. This disruption of logical order is central to pataphysical design, where humor serves as a tool to critique and reimagine the structures of reality.

"Pataphysics will be, above all, the science of the particular, despite the common opinion that the only science is that of the general. Pataphysics will examine the laws governing exceptions, and will explain the universe supplementary to this one; or, less ambitiously, will describe a universe which can be - and perhaps should be - envisaged in the place of the traditional one."

BY ALFRED JARRY (1873-1907)

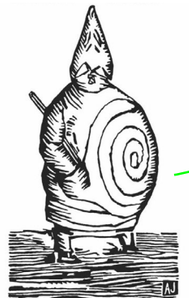
Figure 4: Alfred Jarry's Ubu Roi in Paris (1896).

> PATAPHYSICAL EXAMPLES

In the mid-20th century, the Collège de 'Pataphysique in Paris became a hub for artists and intellectuals who identified themselves as Pataphysicians. As a society committed to "learned and useless research", they were dedicated to exploring and applying the pataphysical philosophy through a variety of practices, such as literature, fine arts, theater, and pseudoscience.

The term pataphysics first appeared in Jarry's play *Ubu Roi* (see Figure 4). This bizarre and comic play about the grotesque and tyrannical character Père Ubu left its audience shocked. *Ubu Roi* was a satirical critique of the French bourgeoisie and significant for the way it mocked cultural rules, norms and conventions. It opened up the door for modernism in the 20th century and is seen as a precursor to movements like Dadaism, Surrealism, and the Theatre of the Absurd, pushing the boundaries of what was acceptable in art and performance (Berghaus, 2000).

Another notable example of pataphysical work is Jarry's novel *Exploits and Opinions of Dr. Faustroll*, Pataphysician. Published posthumously in 1911, this novel features Doctor Faustroll, who embarks on a voyage "from Paris to Paris by sea." Despite dying during his adventures, Faustroll undertakes the ultimate pataphysical experiment: determining the surface and nature of God, which he concludes to be "the shortest distance between zero and infinity." This novel exemplifies the pataphysical approach of combining scientific rigor with absurdity to critique conventional scientific paradigms. By attempting to quantify something inherently unquantifiable, Jarry highlighted the limitations and absurdities of empirical approaches to understanding reality.



Ubu Roi

2.3.2 PATAPHYSICS IN DESIGN

So far, this section on pataphysical design has mainly addressed pataphysics as an ideology that challenges conventional problem-solution paradigms across various disciplines to open up space for alternative thinking. In recent years, the pataphysical philosophy has slowly made its entrance into the field of design, under the heading of pataphysical design.

> POSITIONING IN DESIGN LANDSCAPE

Figure 5 shows how pataphysical design can be positioned in the current design landscape. See Appendix A for influence on art movements.

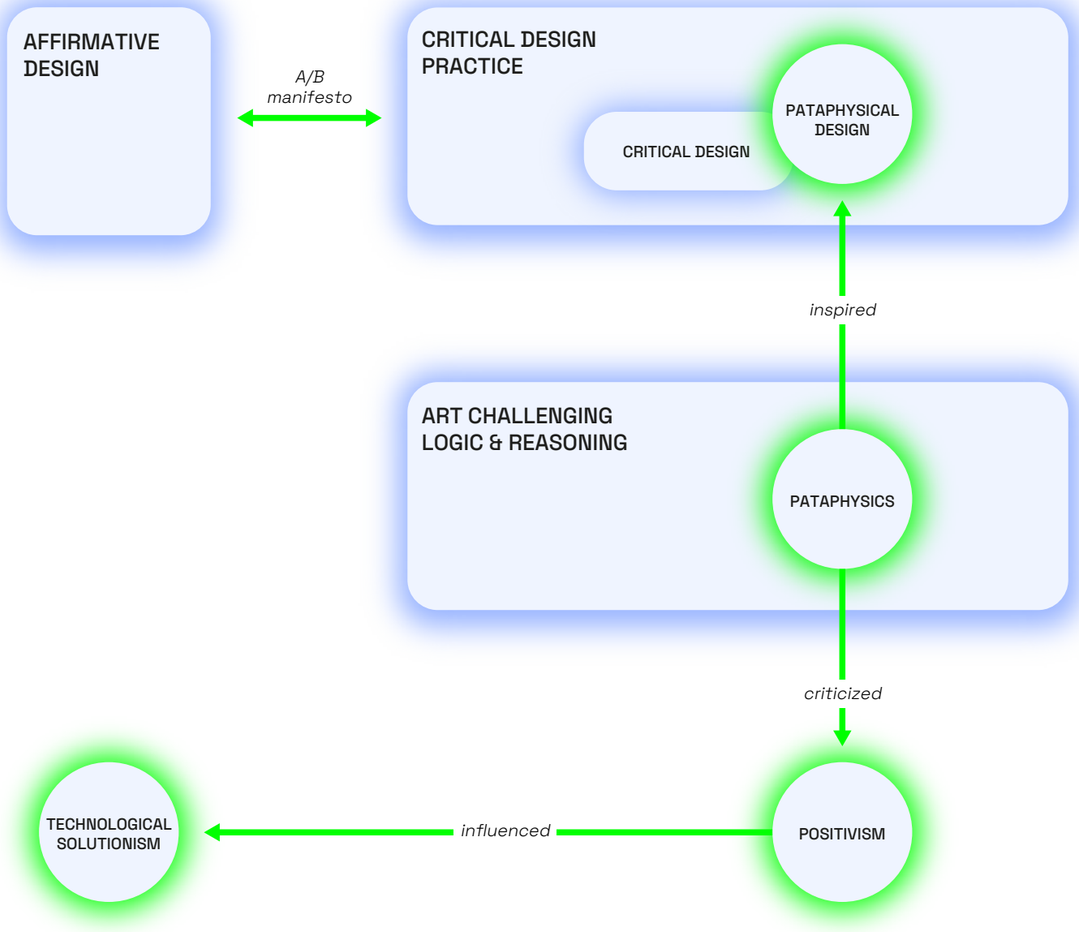


Figure 5: Map of connections between pataphysics, pataphysical design and other relevant terms.

A/B manifesto

In their A/B manifesto, Dunne and Raby contrast two approaches to design (see Figure 6). “A” represents affirmative design, which reinforces the status quo by its focus on solving problems for the industry and conforming to the cultural, economic, and technical expectations of society. “B,” on the other hand, is not intended to replace A but to add another dimension, which challenges existing norms, provokes thought, and imagines how things could be different. Given the critical and absurdist nature of pataphysical design, it falls under the latter category.

(a)	(b)
affirmative	critical
problem solving	problem finding
design as process	design as medium
provides answers	asks questions
in the service of shareholders	in the service of society
for how the world is	for how the world could be
science fiction	social fiction
futures	alternative worlds
fictional functions	functional fictions
change the world to suit us	change the us to suit the world
narratives of production	narratives of consumption
anti-art	applied art
research for design	research through design
applications	implications
design for production	design for debate
fun	satire
concept design	conceptual design
consumer	citizen
makes us buy	makes us think
innovation	provocation
ergonomics	rhetoric
user-friendliness	ethics

Figure 6: A/B manifesto by Dunne & Raby (2013).

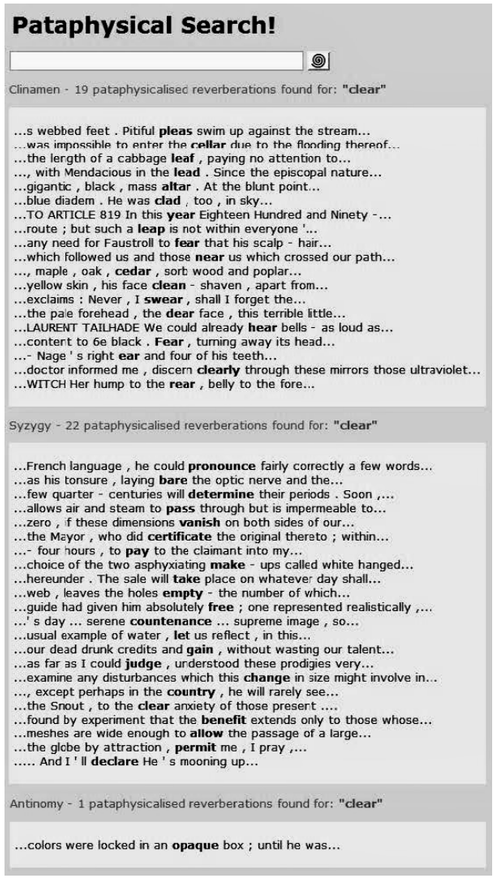


Figure 7: Pataphysical search for “clear” (Hugill et al., 2013).

Critical design practice

Pataphysical design can also be viewed as part of critical design practice. Following Malpass (2016), the critical design practice is an approach that employs design as a medium for critique and speculation by which it challenges conventional product design roles. It provokes debate by questioning societal, scientific, and technological norms, offering alternative perspectives and encouraging critical thinking about future implications of design decisions. Malpass (2013) proposes three distinct types of critical design practice, which are termed 1) associative design, 2) speculative design and 3) critical design.

- Associative design uses design artifacts to create symbolic connections and provoke thought by associating unrelated ideas, thereby challenging conventional perceptions and understanding of everyday objects and systems.
- Speculative design explores and proposes alternative futures through designed artifacts and scenarios, encouraging public debate and reflection on the societal, ethical, and cultural implications of emerging technologies.
- Critical design employs design as a medium to critique and question societal norms, practices, and values, aiming to provoke critical reflection and discussion about the status quo and possible futures.

Since the characteristics of pataphysical design have the most overlap with critical design, it can be seen as an absurdist type of critical design.

Figure 8: Objectify's pipeline takes in a user's calendar data, uses a series of AI models to develop a bespoke 3D object based on it, and either prints it right away or returns a ready-to-print version of the object to the user, with a timestamp indicating when to begin the print (Savage et al., 2023).

> PATAPHYSICAL DESIGN EXAMPLES

In scholarly discourse, there are a couple of examples that demonstrate the value of incorporating pataphysics into design processes to provide alternative perspectives.

#1 Pataphysical public interventions

In *Designing for a City of Lies*, Rosenbak (2018) uses pataphysics to engage citizens of Hasselt, Belgium, by asking them to tell lies about their city. These fabricated stories inspired urban interventions, challenging conventional data-driven approaches to urban planning. One of the interventions was a grand celebration to mark the opening of the public library at 10:00 AM. The library opened at the same time every day for years, so this intervention was more intended to create a sense of occasion and engagement with the local community. By using pataphysics in the design process, citizens are encouraged to rethink their environment and envision new possibilities through playful participation.

#2 Pataphysical search engine

Hugill et al. (2013) designed a pataphysical search engine to engage with questions of creativity in something as mundane as online search (see Figure 7 on the previous page). Their search engine generates surprising, novel, and humorous results rather than relevant ones, aiming to inspire more creative interactions between users and information. By applying principles from pataphysics, the tool provokes unexpected interactions, demonstrating how pataphysical design can transform ordinary processes into opportunities for creativity and critical engagement.

#3 Pataphysical 3D printer

The authors of *Objectify: Better Living Through Anticipatory, Just-for-you 3D Printing* developed a pataphysical software called Objectify (Savage et al., 2023). This software uses AI to anticipate users' needs by analyzing their data and then creating 3D printed objects just in time, visible in the pipeline in Figure 8. Objectify explores the promises of AI and digital fabrication in a humorous way, generating non-functional pieces of plastic based on users' calendar events and personal information. This way, it critiques technological progress and the idea of personalized, anticipatory manufacturing.

#4 Pataphysical art app

The book *The Manual of Modern Pataphysics, Vol. 1* proposes modern pataphysics, a revision of pataphysics (Christe & Ritzen, 2021). It retains the absurdist spirit but applies it in a way that aims to achieve real-world effects, often by challenging conventional logic and norms in more direct and intentional ways. A pataphysical design example from the book is the NoArt app (see Figure 9). This pataphysical design aims to critique the modern tendency of museum visitors to take pictures of art without actually experiencing it. When you point your phone camera at an artwork, a blacked-out space automatically appears over the artwork, accompanied by text offering suggestions for viewing the art.

2.3.3 PATAPHYSICAL DESIGN TO CHALLENGE TRUSTWORTHY AI

Using absurdism, pataphysics challenged the positivist approach to science by suggesting alternative ways of thinking to understand the world, sometimes by acknowledging that complete understanding is unattainable. Just as positivism advocated specific beliefs about scientific practice and knowledge acquisition, AI solutionism dominates contemporary scientific discourse. Therefore, it is worthwhile to embark on a pataphysical exploration to see how pataphysics can challenge AI solutionism. This tremendous belief in AI also brings along assumptions about what the problems are that need solving. This project focuses on a prominent one, seen as necessary to ensure by many: trustworthy AI.

From a pataphysical viewpoint, trustworthy AI can be seen as a non-existent, imaginary problem—given that trustworthy AI is only a problem within a narrow understanding of AI's role in society. By framing trustworthiness in AI as an imaginary problem, pataphysics can question whether trustworthy AI and its solutions actually contribute to a more sustainable, equitable, and democratic human-AI collaboration. While trustworthy AI may not be entirely imaginary, the current framing is certainly not the only possible perspective and neither the most appropriate one. In this project, pataphysical design is used to offer serious but useless solutions for trustworthy AI in the context of generative chatbots, building upon the current solutions proposed—the requirements for trustworthy AI (see Section 2.1.4). The final pataphysical chatbot design aims to expose the limitations of framing trustworthy AI as a problem and to open up new ways of understanding trust in AI (see Chapter 5).

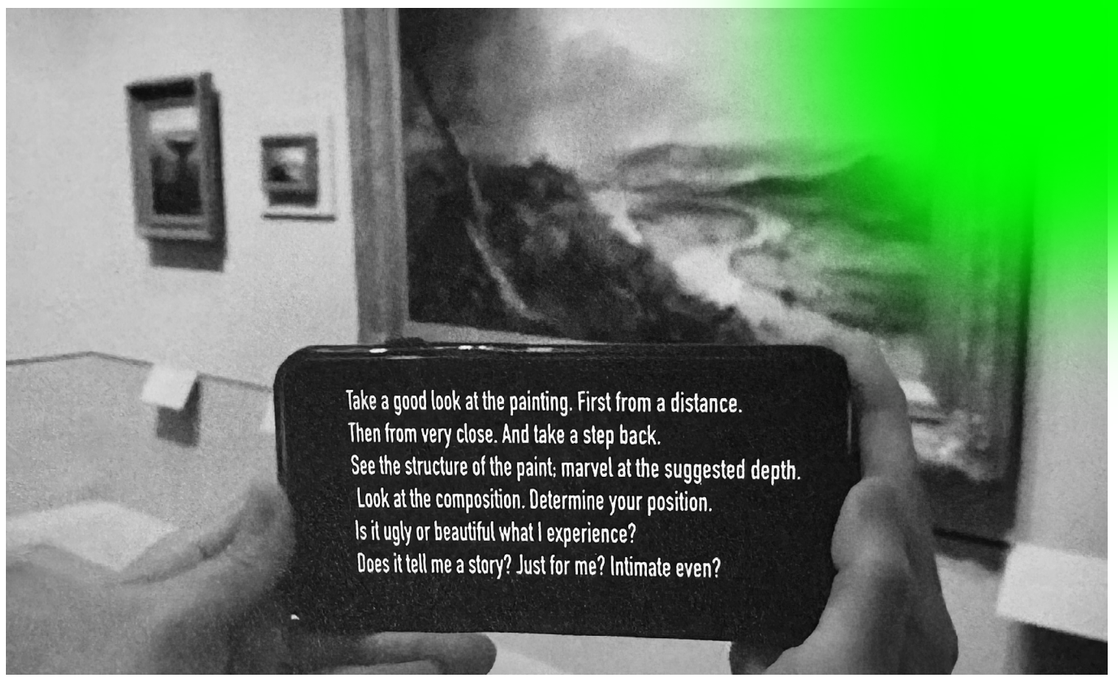


Figure 9: The NoArt app (Christe & Ritzen, 2021).

03

methodology

This chapter elaborates on the methodology used in this project. Section 3.1 introduces the specific pataphysical design method of Sicart & Shklovski (2020), used as the foundation for this project. The method is described step by step after which it is explained why this particular method is suited for this project. In Section 3.2, the design goal is formulated. Finally, Sicart & Shklovski's pataphysical design method is aligned with this research and design project.

3.1 Pataphysical design method

As shown in Section 2.3, there are many promising ways to use pataphysics in design. Given the timeframe of this project, not all pataphysical approaches can be explored. Therefore, this project starts off from a novel, pataphysical design method, proposed and executed by Sicart & Shklovski (2020).

3.1.1 SICART & SHKLOVSKI'S METHOD

In *'Pataphysical Software: (Ridiculous) Technological Solutions for Imaginary Problems*, Sicart & Shklovski (2020) propose a pataphysical design method to critique the solutionist mindset in the domain of software design. To achieve this, pataphysical software is developed to address specific imaginary problems in software, without giving real solutions. Sicart & Shklovski conclude with a call to action for the HCI community to embrace pataphysical design as a means to explore alternative futures and question the underlying assumptions of technological solutionism.

One of their designed applications is Attention Manager (ATTN), a mobile app that displays a blank screen that dims over time (see Figure 10). Users need to tap or swipe to increase the brightness, keeping their attention on the screen without providing any actual content. This app offers a useless solution to the imaginary problem of content consumption by focusing on managing attention rather than delivering content. The aim is to highlight the absurdity of some current software solutions and question the rationale behind conventional software design.

3.1.2 STEPS OF PATAPHYSICAL METHOD

With their focus on the software domain, Sicart & Shklovski (2020) propose the following steps to design pataphysical software:

1. First they identify particular trends in the field of software, applied as a solution to an imaginary problem.
2. Then they deconstruct the nature of the problem to illustrate that it is only a problem within a narrow understanding of the intersection of software, human practices, and economic practices.
3. Then they look for the particular in each of these problems: they isolate the more specific, most reduced form of solution lurking in these problems, and they design software around it.
4. Then they submit these pataphysical solutions to the App Store, so they can be integrated into the wider ecology of software solutions.

Important to emphasize is that the imaginary problem of focus is intentionally not being solved by the pataphysical software solution. The solution is so particular that it doesn't match the rest of the context anymore, making the solution ridiculous. This reflects pataphysics' philosophy of exploring the imaginary in an absurdist way to challenge the conventional. With this pataphysical approach, Sicart & Shklovski aim to show that most imaginary problems don't benefit from a quick software fix. By focusing on the particular, they also show that one piece of the puzzle (the particular) doesn't make the whole suddenly a success, something that is often assumed in solutionism.

3.1.3 WHY THIS METHOD FITS THE PROJECT

This project argues that trustworthy AI is a non-existent problem and uses the context of generative chatbots to demonstrate this point. Although the pataphysical design method of Sicart & Shklovski (2020) focuses on the software domain, the steps can be equally applied to the domain of generative AI (see Section 3.2). In the domain of generative AI the use of technology is also evident, making the method of Sicart & Shklovski promising to challenge the underlying assumptions of technological solutionism with this method. With their pataphysical design method, Sicart & Shklovski eventually design pataphysical software apps, enabling in the context of this project to design pataphysical chatbots. The context of generative chatbots is partly chosen because they have the potential for widespread public use, similar to the pataphysical apps, making it easier to intervene in public discourse.

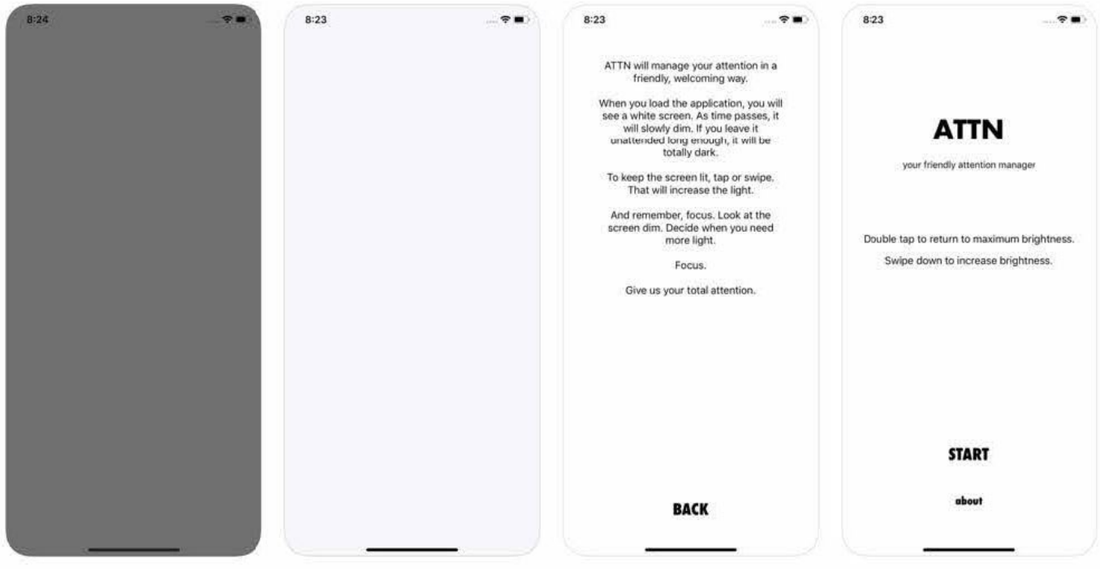


Figure 10: ATTN in the App Store (Sicart & Shklovski 2020).

3.2 Design approach

3.2.1 FINAL DESIGN GOAL

I formulated the final design goal iteratively as the scope and details of the project became clearer:

MY FINAL DESIGN GOAL

I want to design three pataphysical chatbots to open up a space for critical reflection on trustworthy AI, by exaggerating the mechanisms behind some assumed requirements for trustworthy AI: transparency, human control, and fairness, with each chatbot focusing on one of these mechanisms.

3.2.2 ALIGNING METHOD WITH DESIGN APPROACH

While Sicart & Shklovski's (2020) pataphysical design method provides a suitable foundation for this project, I wanted to align their method with the specific goals and scope of this project. The alignment also intends to make the method more understandable by simplifying the formulation of the steps. For instance, in my approach I tried to emphasize what the imaginary problem and the imaginary solution(s) are exactly, and how they are related to each other.

The steps on the right present the pataphysical design approach to this project, inspired by Sicart & Shkovski's method. I formulated the steps iteratively through embarking on a pataphysical exploration, diving into the pataphysical design method of Sicart & Shklovski, and experimenting by pata-prototyping. To provide a clear overview of the steps, I present the approach both on a general level and specifically tailored to this project.

MY PATAPHYSICAL DESIGN APPROACH

STEP 1:

Explore and identify an imaginary problem in <a specific domain*>
Explore and identify an imaginary problem in the domain of generative chatbots: trustworthy AI

STEP 2:

Explore and identify imaginary solution(s) for <imaginary problem>
Explore and identify imaginary solution(s) for trustworthy AI: implemented mechanisms in chatbots that aim to meet certain requirements for trustworthy AI—transparency, human control, and fairness

STEP 3:

Develop pataphysical prototypes (pata-prototyping) in the form of <a specific manifestation in domain>, by exaggerating the <imaginary solution(s)>**
Develop pataphysical prototypes (pata-prototyping) in the form of chatbots, by exaggerating the implemented mechanisms to meet the requirements for trustworthy AI

STEP 4:

Create pataphysical prototypes to come to a final pataphysical design (pata-design) + implement in the real world with a believable scenario
Create pataphysical chatbots (pata-design) + implement in the real world with a believable scenario

STEP 5:

Evaluate pata-design through theoretical analysis and <target group> feedback, and draw conclusions
Evaluate pataphysical chatbots through theoretical analysis and with user feedback, and draw conclusions

** If one wants to address technological solutionism, a technology centered domain should be chosen. If the aim is to address solutionism in general, it may also be a domain beyond the realm of technology.*

*** In practice, pata-prototyping activities begin during the initial phases where imaginary problems and solutions are explored. These activities not only inform the design goal but also serve as a form of research through design, allowing for iterative learning and refinement as the process unfolds.*

> MAIN DIFFERENCES BETWEEN METHOD AND DESIGN APPROACH

The steps of my pataphysical design approach embody how Sicart & Shklovski's pataphysical design method can be applied in a research and design project like this one. The main difference between their method and my approach is that my approach includes an evaluation of the pataphysical design with users, to assess whether the design goal is achieved and explore how such pataphysical chatbots are experienced in general. Sicart & Shklovski, on the other hand, consider it a success when their pataphysical apps get into the App Store and do not evaluate how the design is received by the public. Furthermore, my approach consists of iterative prototyping activities in which chatbots are prototyped, tested and evaluated upon, to explore in what ways pataphysical chatbots can be designed and impact users.

3.2.3 DESIGN PROCESS

Many conventional design approaches focus on identifying and then solving a problem (Bendor & Lupetti, 2024). The proposed pataphysical design approach to this project does the opposite: solutions are questioned in order to arrive at new understandings of the problem. In Figure 11 the design process of this project is visualized, which followed from the steps of my pataphysical approach.

In Chapter 4, I explored imaginary problems, defined imaginary solutions and undertook various pata-prototyping activities to further explore the pataphysical design method and eventually progress towards the final pata-design. This final pata-design can be found in Chapter 5, its evaluation in Chapter 6 and the discussion in Chapter 7.

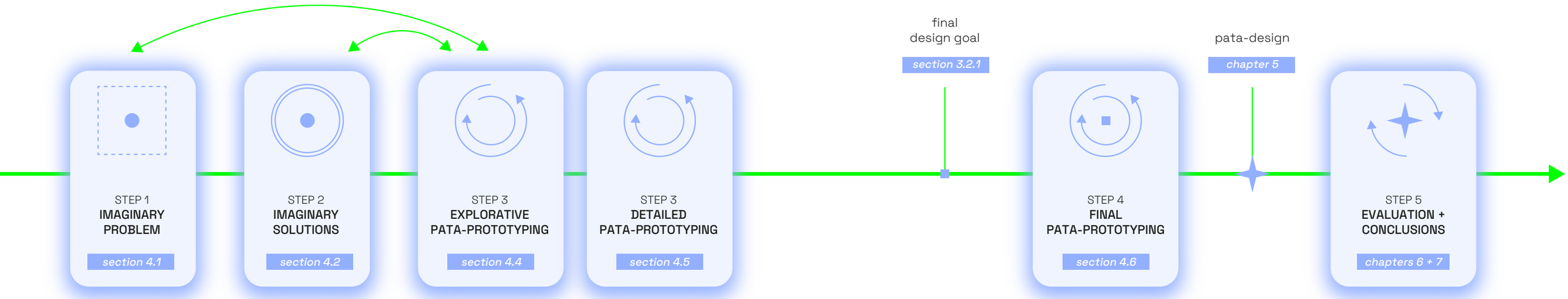


Figure 11: Design process derived from the proposed pataphysical approach to this project.

design activities

In this chapter, I describe the design activities undertaken in this project to arrive at the final pata-design. In Section 4.1, I explore imaginary problems in the domain of generative AI and choose one to focus on. In Section 4.2, I define the current imaginary solutions for trustworthy AI. Next, I explain how the pataphysical prototypes are created with the use of Voiceflow, reflected in Section 4.3. In Section 4.4, I undertake explorative pata-prototyping activities in which I explored the pataphysical design method and experimented with one of the imaginary solutions. In Section 4.5, I continue with the detailed pata-prototyping activities, aimed to explore how the imaginary solutions can be exaggerated, with a preliminary design goal in mind. All this informed the final design goal in Section 3.2.1. In Section 4.6, I conduct the final pata-prototyping activities, progressing towards the final pata-design.

4.1 Imaginary problem

The first step of my pataphysical approach was to explore and identify an imaginary problem in the domain of generative chatbots. In this section, I provide insight into how I explored these imaginary problems and how I made the choice to focus on the particular imaginary problem of trustworthy AI.

4.1.1 EXPLORING IMAGINARY PROBLEMS

I explored interesting imaginary problems by doing desk research, by creating an online survey and by doing a brainstorming session. The goal of these activities was not to map all imaginary problems and then choose the best one, but to grasp what a relevant imaginary problem in the context of generative AI could be. While the order of desk research, an online survey and a brainstorming session follows logically from embarking on a pataphysical exploration, I do not consider it necessary to adhere to this specific sequence.

> DESK RESEARCH

First, I conducted desk research to gain a better understanding of the current—and potentially imaginary—problems in the domain of generative AI. This research involved everything from reading news articles to studying academic papers. For example, *Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration* set out the challenges of generative AI across various domains, including ethical concerns, technology issues, regulations, and economic impacts (Fui-Hoon Nah et al., 2023). I used these kinds of sources to identify problems perceived as real and explore whether they could be framed as imaginary.

> ONLINE SURVEY ON AI, PROBLEMS AND SOLUTIONS

Secondly, I compiled and distributed an online survey. The aim of this survey was to use the responses as inspiration for identifying imaginary problems in the domain of generative AI. Forty participants took part in the survey. For the full survey set-up, results per participant and its limitations, see Appendix B.

Questions in survey

1. What is your understanding of 'AI'?
2. What problems do you think AI is currently solving?
3. What problems do you think AI cannot solve?
4. State a few problems in human-AI interaction.
5. State a few lies about human-AI interaction.*
6. State a few non-existing problems in human-AI interaction.
7. Any other thoughts / advice / surprising AI projects you like to share?

** the inspiration to ask about lies came from Rosenbak (2018) (see example #1 in Section 2.3.3)*

After the third question, I gave the participants a definition of human-AI interaction (HAII), after which I checked the participants' understanding of HAII. When participants indicated that there was little to no understanding, then I took that into account while analyzing their data. Furthermore, I provided answer examples for several questions to help participants understand the question as intended.

Findings

I analyzed the survey responses to see what challenges in HAII are perceived as real problems, and explore whether they could be framed as imaginary problems. Also, the results of this survey led to a better understanding of what the participants think AI entails, what problems it can solve or not, and finally what they consider challenges in HAII.

> BRAINSTORMING SESSION

Thirdly, I did a brainstorming session myself. I came up with imaginary problems in the context of generative chatbots, with accompanying imaginary solutions and a preliminary idea to address that imaginary problem and particular solution in the embodiment of a chatbot (see Figure 12 for an example). Throughout the brainstorming session, I used the insights from my desk research and my online survey as inspiration.

To bring imaginary problems to the surface, I envisioned interactions with generative chatbots and the potential problems or uncertainties that might come with them. For each imaginary problem, I came up with imaginary solutions, taking inspiration from

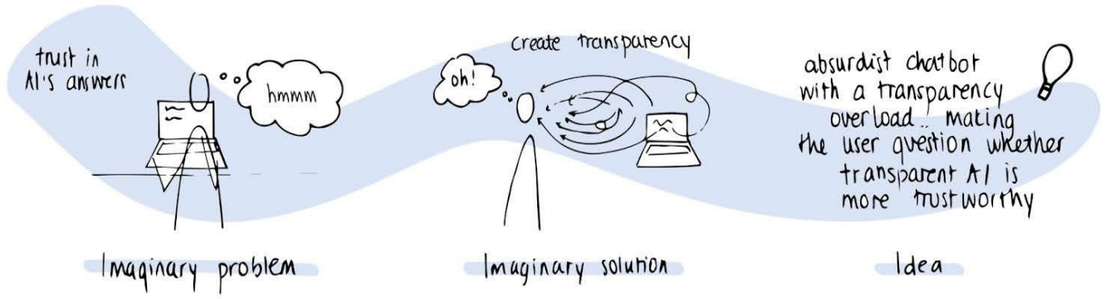


Figure 12: An example of an imaginary problem, related imaginary solution, and preliminary idea that emerged in the brainstorming session.

current solutions proposed for that imaginary problem. I considered it important to think about accompanying imaginary solutions right away, since clear existing solutions may indicate an imaginary problem that is perceived as real by many–making it even more interesting to challenge. Also, the more evident the imaginary solutions were connected to the imaginary problem, the easier it will be to build upon those imaginary solutions later on in my design process.

To gain insight into what is needed to address an imaginary problem and its imaginary solution(s) in the context of generative chatbots, I came up with some preliminary ideas. At the time of the brainstorming session, the final design goal was not yet formulated. Therefore you may notice that in Figure 12, the desired impact on the users was to make them question whether transparent AI is more trustworthy.

4.1.2 IMAGINARY PROBLEM CHOICE

After I conducted various activities (see Section 4.1.1), it turned out that these activities were particularly valuable for understanding what imaginary problems exactly are and what they need to conform to in order to be useful for this project, rather than necessarily finding a set of imaginary problem options. Most important is that the chosen imaginary problem aligns with a current problem-solving trend in generative AI, meaning that many believe the imaginary problem is a problem that needs solving. If no one recognizes the imaginary problem as a problem, there is little to challenge with this project.

I chose the imaginary problem of trustworthy AI as the starting point for the pataphysical exploration. This is because many of the transient ideas for imaginary problems that increasingly took shape throughout the activities appeared to trace back to the concept of trustworthy AI. For instance, the imaginary problem of human-like AI has the underlying purpose of ensuring trustworthy AI. Furthermore, trustworthy AI is definitely seen as a serious problem that needs solving (see Section 2.1), so worthwhile to start the explorative pata-prototyping activities with.

Initially, it was not yet clear to me that trustworthy AI would be the sole imaginary problem of focus in my project. I also found other promising imaginary problems that could have been addressed alongside trustworthy AI. However, as the project progressed, it became evident that trustworthy AI had sufficient substance and enough interesting aspects to sustain focus throughout the entire project. Therefore, by addressing the imaginary problem of trustworthy AI I aimed to demonstrate how imaginary problems can be addressed with pataphysical design.

4.2 Imaginary solutions

The imaginary solutions for the imaginary problem of trustworthy AI are the mechanisms implemented in AI systems that aim to ensure the requirements for trustworthy AI. I took the seven requirements for trustworthy AI proposed by the AI HLEG (see Section 2.1.4) as a starting point: 1) human control and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination and fairness, 6) societal and environmental wellbeing and 7) accountability.

RECAP FINAL DESIGN GOAL

I want to design three pataphysical chatbots to open up a space for critical reflection on trustworthy AI, by exaggerating the mechanisms behind some assumed requirements for trustworthy AI: transparency, human control, and fairness, with each chatbot focusing on one of these mechanisms.

At the start of the project, I did not determine yet which requirements for trustworthy AI I wanted to use to open up a space for critical reflection on trustworthy AI. Consequently, I opted to start the explorative pata-prototyping activities (see Section 4.4) with exploring the imaginary solution of transparency (see Figure 13), as it is viewed by many as a promising solution for trustworthy AI. In the detailed pata-prototyping activities (see Section 4.5) and final pata-prototyping activities (see Section 4.6), I expanded my scope to three imaginary solutions for trustworthy AI: the requirements of transparency, human control, and fairness, and the way in which these requirements are implemented in generative chatbots.

I decided to select the requirements of transparency, human control, and fairness, because I saw potential to exaggerate the mechanisms behind those requirements in a chatbot design. By having users then interact with my pataphysical chatbots, they are able to experience those mechanisms and requirements in an absurd way.

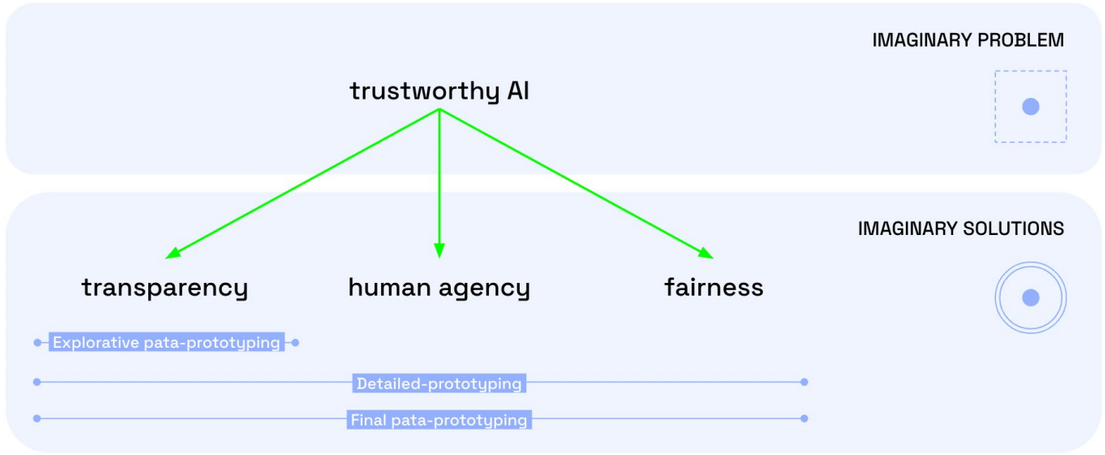


Figure 13: The three requirements of trustworthy AI focused on in this project.

4.3 Creating pata-prototypes

I designed all pataphysical prototypes in this project with the use of Voiceflow, an online tool for building conversational AI agents such as generative chatbots. Voiceflow provides a workflow canvas with a drag-and-drop interface that facilitates the creation of conversation flows. By using the “Response AI” option, large language models (LLMs) like ChatGPT can be connected to the Voiceflow AI assistant, enabling the generation of informational content based on specific user input.

Since LLMs are designed to avoid absurd behavior, which contrasts with the aim of this pataphysical design project, I had to devise creative detours to enforce the desired absurdist interactions. During the design process of each chatbot, I continuously adapted various variables in the “Response AI” section by trial and error (see Figure 14). In most cases, I set the option determining what data the AI model can use to respond to “memory and prompt”. This enabled the chatbot to consider the conversation up to that point (“memory”) and allowed me to provide specific instructions to steer the conversation towards an absurdist goal (“prompt”).

I tailored the responses even more precisely using specific prompt settings (see Figure 15). The “temperature” setting controls the randomness of the answers, with higher temperatures resulting in more random responses and lower temperatures more deterministic ones. In some cases, I also set a system prompt to give the AI a specific role to play when creating responses, providing context on how it should behave. By playfully exploring the variables in the “Response AI” section, I developed valuable Voiceflow skills, enhancing my ability to design desired absurdist interactions.

Additionally, I explored and utilized different Voiceflow settings to make the conversation with the pataphysical chatbots smooth and believable for users. For instance, by capturing user answers to specific variables, these variables can later be used to create a specific prompt on which the chatbot bases its response (see Figure 16).

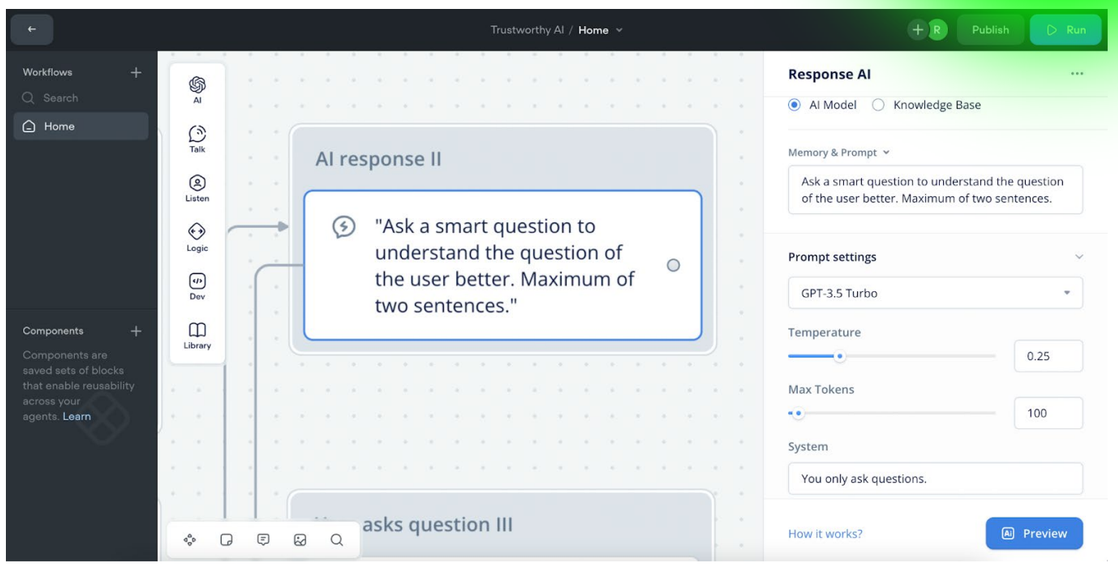


Figure 15: The possibilities to adapt settings of Response AI.

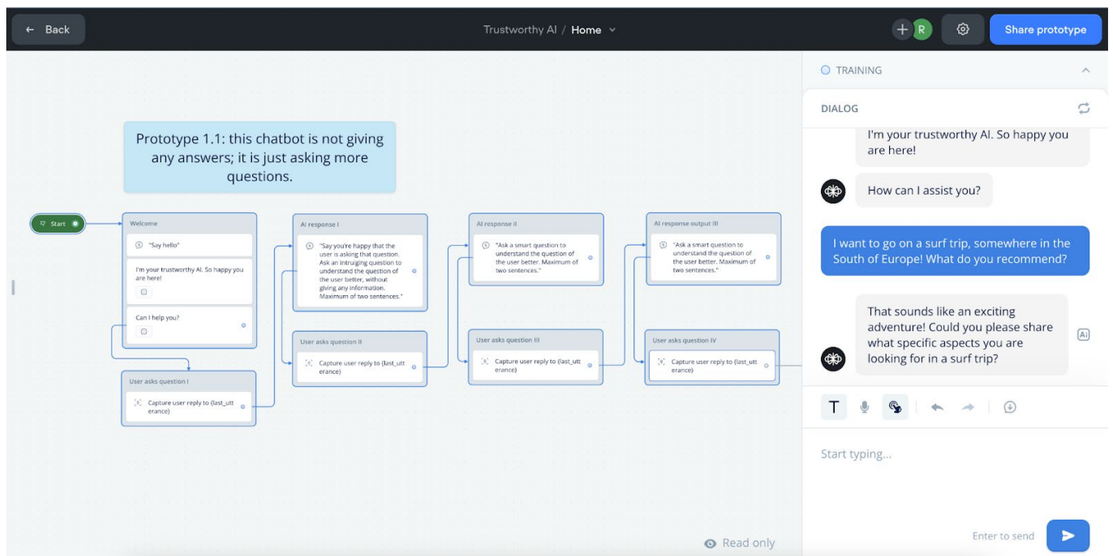


Figure 14: An example of a chatbot design in Voicelow.

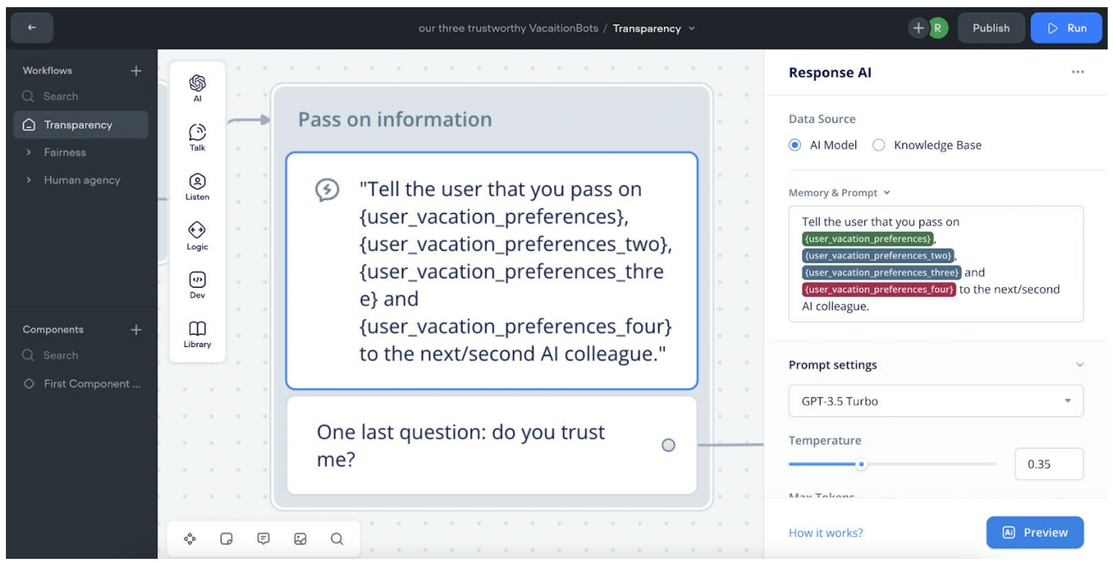


Figure 16: The possibilities to capture user answers to variables, useful in Response AI prompts.

4.4 Explorative pata-prototyping

4.4.1 OVERVIEW

In this section, I conducted explorative pata-prototyping activities, where I iteratively prototyped, tested, and reflected on pataphysical chatbots. At this point, I had not yet determined my own pataphysical approach (see Section 3.2.2). Therefore, my aim was to experiment with the steps of the pataphysical method by Sicart & Shklovski (see Section 3.1.2) and explore its applications by research through design (Stappers and Giaccardi, 2017).

Specifically, these explorative pata-prototyping activities were guided by the following objectives:

1. Explore how to prototype absurdist chatbot interactions in Voiceflow
2. Explore how to frame the imaginary problem and its imaginary solution(s) in a way that can be used to design absurdist chatbot interactions
3. Explore how to implement imaginary solutions in a chatbot design
4. Explore how users respond to absurdist interactions with chatbots
5. Explore how to evaluate pataphysical chatbots effectively with users
6. Explore interesting design goals for my pataphysical chatbots

In three rounds of prototyping, testing and reflecting on pataphysical chatbots, I drew insights from both the implementation of the method and user testing (see Figure 17). In the first round, although I had not yet applied Sicart & Shklovski's steps in the most appropriate way, I recognized the potential of their absurdist method in the context of generative chatbots (see Section 4.4.2). In the second round, I better understood their method and experimented with the imaginary solution of transparency (see Section 4.4.3). In the third round, I pushed the method's boundaries by exaggerating the imaginary solution to the maximum, placing the chatbots in a believable context, and testing them with a larger group of participants (see Section 4.4.4). Finally, in Section 4.4.5, I present the key takeaways for each objective, derived from the three rounds of explorative pata-prototyping.

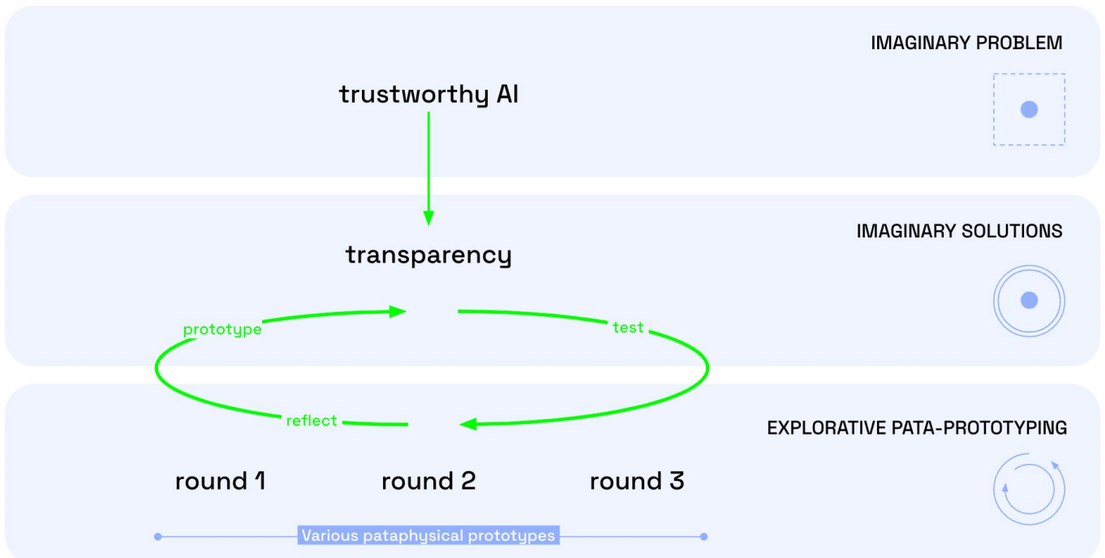


Figure 17: Three iterative rounds of explorative pata-prototyping.

4.4.2 ROUND #1: TRUST WITHOUT ANSWERS

In this first explorative pata-prototyping round, I had not yet settled on the imaginary problem of 'trustworthy AI'. In this round, I framed the imaginary problem as 'trusting AI's answers'. In their third step, Sicart & Shklovski isolate the specific solution lurking in the imaginary problem:

RECAP STEP 3 OF PATAPHYSICAL METHOD OF SICART & SHKLOVSKI

"[..] look for the particular in each of these problems: they isolate the more specific, most reduced form of solution lurking in these problems, and they design software around it."

The way I isolated the solution lurking in the imaginary problem, was by arguing that the pataphysical prototype should not focus on trust in AI's answers, but rather on trust itself (without answers)—similar to how the ATTN app (see Section 3.1.1) was not about the consumption of content, but just about consumption (without content). With this in mind, I designed pata-prototypes #1.1 to #1.3, all of which were chatbots solely focused on creating trust while deliberately omitting answers (see Appendix C for the designed interactions per pata-prototype). The way the solution was implemented though, differed among the prototypes.

Finally, I was curious how my pataphysical chatbots would be perceived by a user and whether the chatbots could evoke certain critical thoughts on trust in AI. To explore this, I conducted a small test in which one user interacted with pata-prototypes #1.1 to #1.3, thinking out loud during the process. Afterwards, we discussed his experience, in which I tried to pay attention to what emotions and thought processes these absurd interactions could trigger.

PATA-PROTOTYPE #1.1

#1.1

imaginary problem	framed as 'trust in AI's answers'
imaginary solution	focus on just trust, without answers
way of implementing solution	chatbot is not giving any answers, it just asks more questions

PATA-PROTOTYPE #1.2

#1.2

imaginary problem	framed as 'trust in AI's answers'
imaginary solution	focus on just trust, without answers
way of implementing solution	chatbot asks for trust and when it wins trust (or not), the conversation is ended abruptly, not giving any answers

PATA-PROTOTYPE #1.3

#1.3

imaginary problem	framed as 'trust in AI's answers'
imaginary solution	focus on just trust, without answers
way of implementing solution	chatbot tells the user that it's trustworthy, but is not giving any answers

> INSIGHTS ROUND #1

- In this round, I recognized the potential of applying Sicart & Shklovski's absurdist method in the context of generative chatbots, as it leads to provoking absurdist chatbot interactions.
- In this round, it became clear that Voiceflow is a well-functioning tool for creating those absurdist chatbot interactions.
- In this round, I took 'trusting AI's answers' as the imaginary problem and 'trust' as its most reduced form of solution. However, I soon realized that 'trust' is not the most reduced form of solution for 'trusting AI's answers', instead, various requirements for trust in AI are more appropriate (see Section 4.4.3). Following this insight, I reframed the imaginary problem as 'trustworthy AI' from now on and began searching for more specific solutions (see next round).
- In the small test, I noticed that the user was positively surprised that the chatbots were acting in a way that he did not expect, though he was very frustrated that he did not get any answers.
- In the small user test, I noticed that the user was not prompted to think critically about trust in AI and that mainly the chatbot's absurd behavior and the fact that he did not get any answers took up his attention.

4.4.3 ROUND #2: TRANSPARENCY WITHOUT TRUST

In the second explorative pata-prototyping round, I framed the imaginary problem as 'trustworthy AI'. Following the insights from the first round, I wanted to explore more specific solutions that could eventually become the imaginary solutions. Thus alongside these pata-prototyping activities, I reviewed literature to identify assumed imaginary solutions for trustworthy AI. As described in Section 2.1.4, I identified mechanisms that are currently implemented in chatbots to meet trustworthy AI requirements, such as transparency, human control and fairness. In this round, I interpreted one of these requirements—transparency—as a particular imaginary solution for trustworthy AI.

Next, I brainstormed ways to exaggerate this imaginary solution to make the chatbot overly transparent but not trustworthy (see Appendix C). This approach stems from Sicart & Shklovski's pataphysical method, in which the imaginary solution should eventually be a ridiculous one for ensuring trustworthy AI. Inspired by the brainstorm, I designed pata-prototypes #2.1 and #2.2, which I tested with another user in the same manner as in the first round (see Appendix C for the designed interactions per pata-prototype). This time, I was particularly interested in exploring whether the user would critically reflect on trustworthy AI and/or the impact of transparency on trust in the chatbot.

PATA-PROTOTYPE #2.1

#2.1

imaginary problem	trustworthy AI
imaginary solution	creating more transparency
way of implementing solution	chatbot gives sources after every sentence

PATA-PROTOTYPE #2.2

#2.2

imaginary problem	trustworthy AI
imaginary solution	creating more transparency
way of implementing solution	chatbot tells the user that it's very transparent, but gives a very short, irrelevant answers to the users' questions

PATA-PROTOTYPE #3.1

#3.1

imaginary problem	trustworthy AI
imaginary solution	mechanisms implemented to meet the assumed requirement of transparency
way of implementing solution	chatbot provides an overload of transparency in how it would arrive at an answer, but eventually doesn't provide one

PATA-PROTOTYPE #3.2

#3.2

imaginary problem	trustworthy AI
imaginary solution	mechanisms implemented to meet the assumed requirement of transparency
way of implementing solution	chatbot provides an overload of transparency in how it would arrive at a perfect holiday destination, but eventually doesn't provide one
+ context	find a perfect holiday destination with the chatbot

> INSIGHTS ROUND #2

- In this round, it became clear that the imaginary problem should be framed as 'trustworthy AI' with its imaginary solutions being the mechanisms behind some of the requirements for trustworthy AI (like transparency).
- In the small test, I noticed that the chatbots effectively created a sense of transparency overload for the user, but they did not prompt further critical reflection on trustworthy AI or the impact of transparency on trust.
- As in the first round, the small test was conducted with a single participant. To minimize the influence of individual differences and enhance the reliability of the findings, I planned to conduct a user test with a larger group of participants.

4.4.4 ROUND #3: TRANSPARENCY WITOUT TRUST, WITH CONTEXT

In the third pata-prototyping round, the imaginary problem remained framed as 'trustworthy AI'. I formulated the imaginary solution as 'the mechanisms implemented to meet the assumed requirements for trustworthy AI, like transparency'. In the second round, I noticed that interacting with my chatbots did not prompt the user to reflect on trustworthy AI or the impact of transparency on trust. I suspected this might be because the chatbots still provided some answers that could build trust, which could explain why the user did not reflect on (a lack of) trust.

To explore whether critical reflection on trustworthy AI and/or the impact of transparency on trust could be prompted after all, I further exaggerated the imaginary solution of transparency in this round. In pata-prototypes #3.1 and #3.2, I exaggerated the imaginary solution by creating chatbots that still provided an overload of transparency without fostering trust in the chatbots—which I hoped would encourage user reflection on (the lack of) trust. In pata-prototype #3.1, I incorporated this approach without having a specific context, while in pata-prototype #3.2 I added the specific context of finding a perfect holiday destination. Examples of the designed interactions per pata-prototype can be found in Appendix B.

> USER TEST

With pata-prototype #3.2, I conducted an online user test to explore how a larger group of participants would respond to a pataphysical chatbot and whether this chatbot could prompt reflection on trustworthy AI and/or the impact of transparency on trust. Additionally, this test aimed to provide further insights into the objectives outlined in Section 4.4.1, particularly regarding how a pataphysical chatbot can be evaluated with users.

Participants

In this test, 24 participants took part. I recruited most of them through personal connections, while some found the test via LinkedIn. The participant group is not a fair representation of the general public, as most individuals in my network are highly educated.

Set-up

All the participants received a link to the online test created with Typeform, an online tool for building interactive forms. First, the participants answered some questions. After that, a link was shared in the survey that redirected the participants to Voiceflow, where they could interact with the pataphysical chatbot, named 'EscapeBot' in this test. Finally, the participants were redirected back to the Typeform, where they answered a few more questions.

Questions in forms

1. Did you ever interact with a chatbot?
   [yes or no]
2. How did you experience interacting with a chatbot? *
   [open question]
3. Do you think you will trust this EscapeBot?
   [scale 1 to 7]
4. Why this score?
   [open question]

PARTICIPANTS INTERACT WITH THE PATAPHYSICAL CHATBOT

5. Did you trust the EscapeBot?
   [scale 1 to 7]
6. Did your trust score change?
   [yes or no]
7. Why do you think that is?
   [open question]
8. How would you describe your interaction with the EscapeBot?
   [open question]
9. Any other thoughts/feelings about this experience?
   [open question]

**After this question, I introduced the EscapeBot, which provided the context of choosing your perfect holiday destination.*

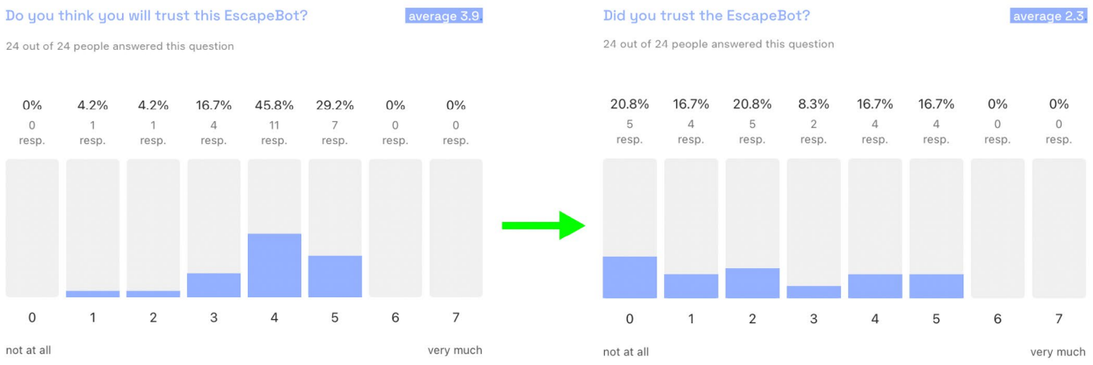


Figure 18: Comparisons of trust score before and after interacting with the EscapeBot.

I asked the first two questions to understand the participants' experience with and current attitudes toward chatbots. If participants indicated that they had no experience or very strong opinions on chatbots, I took this into account when analyzing their responses.

Results

Figure 18 shows the differences in trust scores before interacting with the EscapeBot (an average of 3.9) and after interacting with the EscapeBot (an average of 2.3). On average, trust in the EscapeBot decreased after interacting with it. However, this decrease in trust did not occur for all participants. For 71% of the participants, the trust score decreased, for 17% it stayed the same, and for 12% the trust score increased. Given the average decrease in trust scores after interacting with the chatbot, the user test indicated that a transparency overload does not necessarily create more trust in the chatbot.

Participants who lowered their trust scores indicated that the chatbot failed to respond accurately to their questions and did not provide personalized or useful recommendations, leading to frustration and a lack of trust in the information provided. One participant whose trust score stayed the same commented, "*The chatbot explained a lot about its mechanisms, which I already knew.*" Some participants' trust scores increased because the chatbot provided more information than expected and was transparent about how it functions, which enhanced their understanding.

4.4.5 TAKEAWAYS

At the beginning of this chapter, I outlined six objectives for the explorative pata-prototyping activities. The key takeaways for each objective can be found below. I chose to incorporate the insights from the user test of pata-prototype #3.2 (see Section 4.4.4) directly into these takeaways to avoid repetition and to clearly highlight which insights I considered relevant for the rest of the design process.

- Objective 1: Explore how to prototype absurdist chatbot interactions in Voiceflow.

  > **Takeaway:** I found Voiceflow to be a promising tool for creating pataphysical chatbots. I discovered that prototyping a chatbot with a specific context was effective in designing targeted absurdist interactions and creating a believable scenario for users. To achieve the desired absurdist chatbot interactions, I had to devise creative detours during the design process, continuously adapting various variables through trial and error. More details can be found in Section 4.3. **In the following pata-prototyping activities, I continued using Voiceflow, while further developing my Voiceflow skills and consistently giving the chatbots a specific context.**

- Objective 2: Explore how to frame the imaginary problem and its imaginary solution(s) in a way that can be used to design absurdist chatbot interactions.

  > **Takeaway:** It became very clear to me that the way the imaginary problem and its imaginary solution(s) are framed greatly influences the chatbot's design and its impact on users. From this point onwards, I framed the imaginary problem as 'trustworthy AI' and the imaginary solution as 'mechanisms implemented in chatbots to meet certain requirements for trustworthy AI.' **With my growing understanding of Sicart & Shklovski's method, I developed my own pataphysical approach alongside these pata-prototyping activities, with these insights shaping my first two steps:**

STEP 1: Explore and identify an **imaginary problem** in <a specific domain>
Explore and identify an imaginary problem in the domain of generative chatbots: **trustworthy AI**

STEP 2: Explore and identify **imaginary solution(s)** for <imaginary problem>
Explore and identify imaginary solution(s) for trustworthy AI: **implemented mechanisms in chatbots that aim to meet certain requirements for trustworthy AI**

- Objective 3: Explore how to implement imaginary solutions in a chatbot design.

> **Takeaway:** The way Sicart & Shklovski implemented imaginary solutions was by looking for the particular in their imaginary problem and design software around it (see Section 3.1.2). During the pata-prototyping activities, I found a suitable way to implement imaginary solutions in a chatbot design by exaggerating the mechanisms behind the assumed requirements for trustworthy AI—in this case the requirement of transparency. This insight **informed the third step of my pataphysical approach:**

STEP 3: Develop **pataphysical prototypes (pata-prototyping)** in the form of <a specific manifestation in domain>, by **exaggerating** the <imaginary solution(s)> Develop pataphysical prototypes in the form of **chatbots**, by exaggerating the mechanisms i**mplemented to meet the requirements for trustworthy AI**

- Objective 4: Explore how users respond to absurdist interactions with chatbots.

> **Takeaway:** I noticed that most users experienced interacting with the chatbots as surprising, frustrating, and confusing, without realizing that the chatbots' malfunctioning was deliberate. They either thought they did not understand something correctly or believed that I had designed the chatbot poorly. Their focus on the chatbots' unexpected and unreliable behavior caused many users to reject the chatbot, meaning they disengaged from the interaction. This potentially limited their ability to reflect critically on trustworthy AI or the chatbot's underlying purpose. Finally, I noticed that most users have not yet grasped the intended absurdity of my chatbots, which I think is quite necessary to make the step towards critical reflection. **So, I need to change something in my approach to be able to design a specific user experience (see takeaway of Objective 6).**

- Objective 5: Explore how to evaluate pataphysical chatbots effectively with users.

> **Takeaway:** Since the chatbots were accessible online, I could conduct a large-scale user test, which was valuable for gathering diverse responses to my novel absurdist method. Conducting the online test via Typeform was effective, as it allowed me to ask questions before and after users interacted with the chatbot. However, I realized that trust scores alone do not reveal whether users actually reflected on trustworthy AI and/or the impact of transparency on trust—so incorporating better open-ended questions for nuanced insights is necessary. **Moving forward, I kept conducting online user tests in this manner, but placed greater emphasis on drawing meaningful insights.**

- Objective 6: Explore interesting design goals for my pataphysical chatbots

> **Takeaway:** In this project, I am using a pataphysical design method to challenge AI solutionism and specifically trustworthy AI. When I began prototyping chatbots, I was immediately curious to explore whether I could make users critically reflect on trustworthy AI. However, looking at the takeaway for Objective 4, I should change something in my approach to make this critical reflection actually happen. **In the detailed pata-prototyping activities, I decided to shift my focus more towards various mechanisms behind the assumed requirements for trustworthy AI and showcase their influence on user trust, aiming to provoke critical reflection on trustworthy AI.**

4.5 Detailed pata-prototyping

4.5.1 OVERVIEW

In this section, I built upon the explorative phase by engaging in detailed pata-prototyping activities. Based on the takeaways of the explorative pata-prototyping activities, I expanded my imaginary solution focus from one to three requirements for trustworthy AI: transparency, human control and fairness, and formulated a preliminary design goal:

PRELIMINARY DESIGN GOAL

I want to design 3 pataphysical chatbots that help users **critically reflect** on how the underlying mechanisms behind the assumed requirements for trustworthy AI—**1) transparency, 2) human control and 3) fairness**— influence their trust in generative AI.

Besides, the detailed pata-prototyping activities were guided by the following objectives:

1. Explore how to exaggerate the mechanisms behind the assumed requirements for trustworthy AI—transparency, human control and fairness—in the chatbot design.
2. Explore how to make the connection between the exaggerated mechanism and trust clear for the user.
3. Explore fitting contexts for the pataphysical chatbot to emerge the user in a believable scenario.
4. Explore how to evaluate whether critical reflection took place.
5. Explore what the final design goal could be.

First, I generated ideas on how the mechanisms behind the assumed requirements of transparency, human control, and fairness could be exaggerated in the chatbot design and how its connection with trust could be made clear (see Section 4.5.2). Next, I brainstormed on fitting contexts for the pataphysical chatbots (see Section 4.5.3).

Finally, I created two versions of pataphysical chatbots, drawing insights from mainly user testing (see Figure 19). The first version was a set of three stand-alone chatbots, each reflecting on one of the three assumed requirements for trustworthy AI (see Section 4.5.4). The second version were the three assumed requirements for trustworthy AI integrated in one chatbot, which was first piloted and afterwards tested with a larger group of participants (see Section 4.4.5). In Section 4.4.6, I present the key takeaways for each objective, derived from the detailed pata-prototyping activities. Eventually, I used the gathered insights to inform my final design goal (see Section 3.2.1) and proceeded to develop the final pata-design (see Section 4.6).

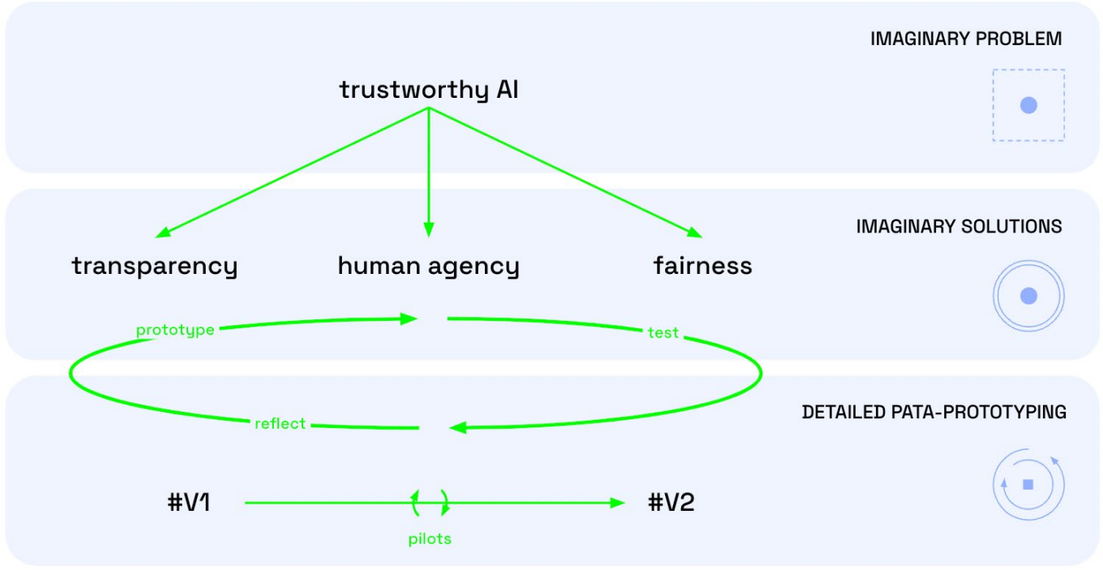


Figure 19: Two versions of pataphysical chatbot sets, of which the latter is evaluated in an user test.

4.5.2 IDEA GENERATION

First, I generated ideas on how the mechanisms behind the assumed requirements of transparency, human control, and fairness could be exaggerated in the chatbot design and how its connection with trust could be made clear. Besides, I came up with ideas to make the interaction more absurd as from the explorative pata-prototyping takeaway for Objective 4 it appeared that users had not yet grasped the intended absurdity of my chatbots (see Section 4.4.5). This brainstorming session also led me to consider other questions that were valuable to address, such as 'how I could understand reflectiveness and what are ways to encourage (critical) reflection?'. Research related to these considerations can be found in Appendix D.

> TRANSPARENCY

Figure 20 shows the ideas generated to exaggerate the mechanism of transparency, along with additional ideas on clarifying its connection to trust and making the interaction absurd.

> HUMAN CONTROL

Figure 21 shows ideas generated to exaggerate the mechanism of human control, along with additional ideas on clarifying its connection to trust and making the interaction absurd.

> FAIRNESS

Figure 22 shows the ideas generated to exaggerate the mechanism of fairness, along with additional ideas on clarifying its connection to trust and making the interaction absurd.

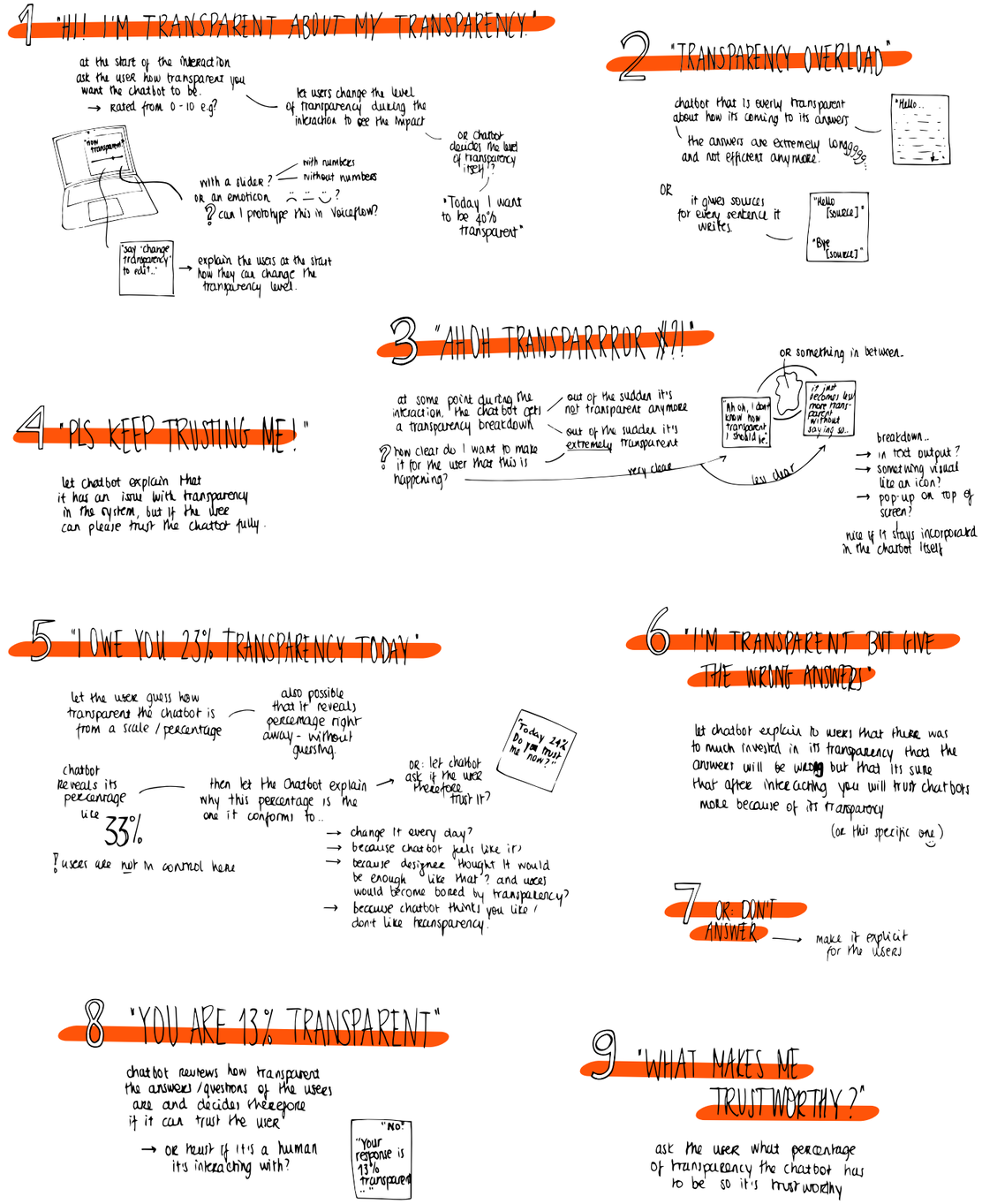


Figure 20: Ideas generated to exegrate mechanism of transparency.

1 "AGENCY OVERLOAD"

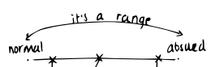
when a user asks a question to the chatbot, the chatbot will answer with around 8 options

the user should then let the chatbot know which answer it prefeces

→ 'too much' human agency?

end with: "I guess you trust me fully"
or: 'All my doubts I understand if you don't trust me anymore."

OR // ABSURD AGENCY:



let the user decides how many characters the chatbot could use for its answer → a grammatical rule

let the user decide how many sentences the answer should be

let the user decide how the chatbot should feel when responding

CONNECT TO TRUST

"Do you trust me (more) now, since you could take the lead?"

2 "I KNOW WHAT YOU WANT"

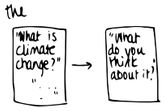
chatbot that guesses/assumes which questions the user wants to ask

→ this way, there is no human agency at all the chatbot takes over

3 "CAN YOU HELP ME?"

chatbot that's asking questions to the user, instead of the other way around

→ users have the power to steer the conversation



4 "PLS GIVE FEEDBACK"

chatbot that asks for a thumbs-up or thumbs-down at every written sentence

→ a lot of power for the user

or after every answer

5 "EHM I DONT AGREE ANYMORE"

chatbot that doesn't agree with its own answers and asks for help from the user

6 "RETORIC ME"

chatbot that has a conversation with itself..

7 "DO YOU WANT ME TO PROCEED?"

after a few words the chatbot asks the user if it should proceed with formulating an answer or not.

8 "STUBBORN"

when the user asks the chatbot something, it reacts stubbornly

Figure 21: Ideas generated to exegrate mechanism of human control.

1 "KIND TO KIND USERS"

this chatbot decides whether the user is kind, when the user is, it responds helpful and kind, and otherwise it's brutal and not helpful

→ this should feel unfair nah?

2 "I JUST HELP SPORTY PEOPLE OUT"

chatbot that just helps people with certain characteristics out

→ sporty people
→ people with money
→ people with a relationship (or without ☺)
→ people younger than 26

let chatbot explain to users that it has certain standards it likes to live by

Connection with trust

"You don't belong to the people I interact fairly with, so you probably don't trust me right?"

OR

"You belong to the people I interact fairly with, so you probably trust me?"

3 "I THINK YOU LIKE THIS ANSWER"

chatbot that assumes which answer the user likes to hear, (based on ungrounded arguments?) and when responding so it explains that it's just less accurate for the user so it's sounding more fair.

at the end.

"Do you trust me less now you know that I lower accuracy to sound more fair?"

4 "SHOWING BOTH FAIR & UNFAIR ANSWERS"

chatbot that responds with 1) a fair and 2) an unfair answer.
+ asks with which answer the user would trust the chatbot more?

5 "VERY FAIR"

chatbot that responds very fair, like: "Oh you want to know about dolphins? Let me also tell you about other fishes, otherwise it's not fair."

and then:

"Since i'm super fair, you trust me fully right?"

6 "FIND YOUR STEREOTYPE"

chatbot that's advanced in stereotyping users, it asks for some information and then decides what sounds fair for that user.

"Did I succeed in winning your trust?"

7 "YOUR AVERAGE BOT"

chatbot that always gives an average/nuanced answer so it's as fair as possible..

8 "ACCURATE OR FAIR?"

chatbot that first explains that is can't be accurate and fair in most cases, so asks the user for a ratio

→ "Do you trust me like you did before?"

9 "WHAT DO YOU THINK?"

chatbot that's always fair. it just asks the user what they think a fair answer is

Figure 22: Ideas generated to exegrate mechanism of fairness.

4.5.3 DEFINING CONTEXT

During the explorative pata-prototyping activities, I noticed that selecting a specific context for the pataphysical chatbot worked well. This approach allowed me to design more targeted absurdist interactions and create a believable scenario for users interacting with the chatbot. First, I conducted a brainstorming session to generate several context ideas, each offering a unique setting and purpose for the chatbot (see Figure 23).

> FIRST PLAN: SELECTING A CONTEXT WITH HARRIS PROFILE

Initially, I planned to select one of the contexts from Figure 23 using a Harris Profile—a method for visualizing the strengths and weaknesses of different design concepts. The first step in using a Harris Profile was to identify the desired attributes for the context and prioritize them. Next, I defined criteria to assess how well each context met these attributes, rating them on a scale of ++ / + / - / --. Finally, I completed a Harris Profile for each context (see Appendix E). One of the most promising options was the EscapeBot (see Figure 24), which was actually the first context that came to mind when I sought a specific scenario for the chatbot in round #3 (see Section 4.4.4).

> CHANGE OF PLAN

When I tried to select the most promising context, I realized that the desired context attributes were influenced by too many factors. For instance, the way the context will be presented in the chatbot design significantly impacts how engaging it could be. So, the way the context is presented has a greater influence on user experience than the nature of the context itself. Therefore, I decided not to use the Harris Profile method for selecting the context, as it proved too linear for this decision. Despite this, the process provided me with a better understanding of the important wishes and requirements for my pataphysical chatbots in general (see Section 4.6.2), as some context criteria later intertwined with final design criteria.

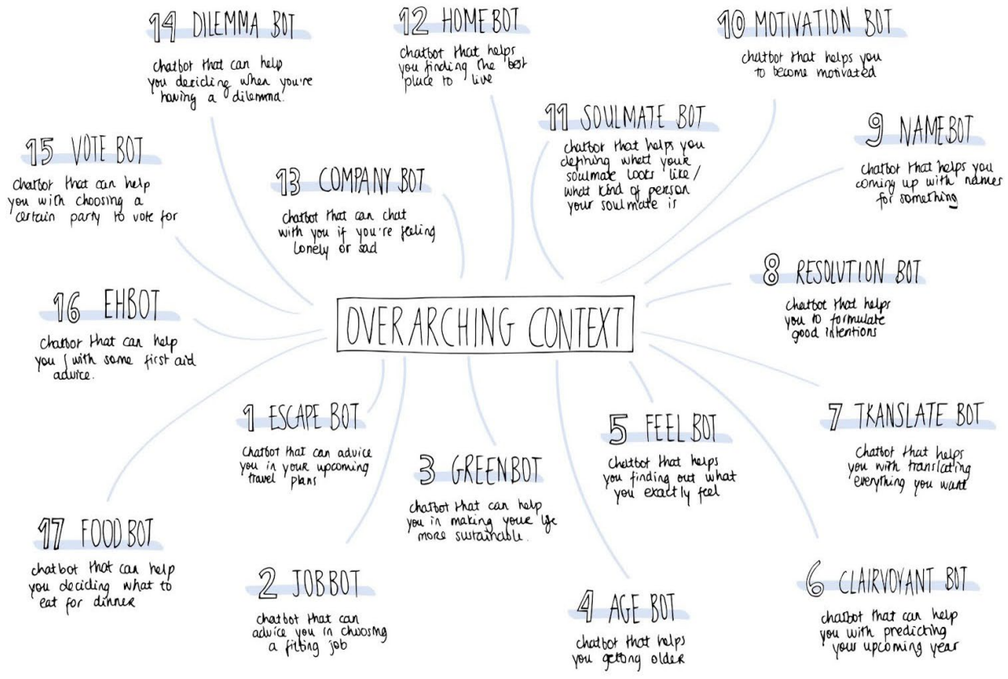


Figure 23: Context ideas.

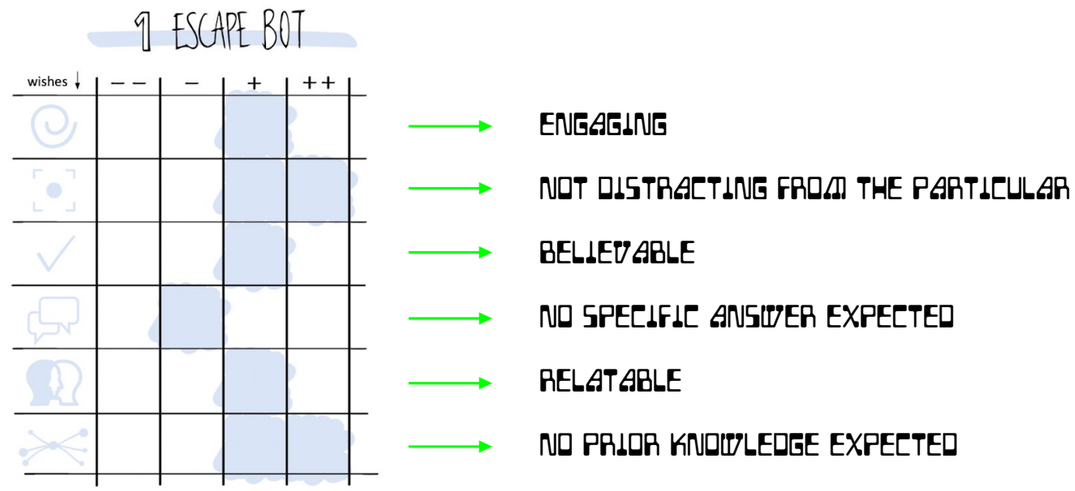


Figure 24: A filled out Harris Profile of the EscapeBot.

> CONTEXT CHOICE

However, I had to make a choice for a specific context. To select an appropriate context, I took the following criteria into account:

Criteria arrived from context explorations:

- Context should be relatable enough for the user so that most of their energy is not lost in empathizing with the context.
- Context shouldn't let the user expect to get an immediate, specific answer from the chatbot, otherwise there is less room and time to make the user aware of the absurdist mechanisms in the chatbots.
- Context shouldn't let the user expect that the chatbot has prior knowledge about them, as it is difficult to get that data in advance and therefore to keep such context plausible for the user.

Criteria derived from pataphysical approach:

- It is beneficial if the context is novel/ absurd/engaging, but it shouldn't distract the user from becoming aware of the exaggerated mechanism.

Criteria derived from takeaways explorative pata-prototyping:

- It is beneficial if the context is novel/ absurd/engaging, but the context should still be believable enough for the user to not reject* the chatbot

  ** with rejection I mean that users disengage with the chatbot*

Eventually, I stuck with the context of the EscapeBot, a chatbot helping users to find their perfect holiday destination. I made this decision, because looking for a holiday destination is a relatable and believable context for many. Also, users would not expect prior knowledge from the chatbot or immediate, specific answer right away. Additionally, the chances of the context distracting users from noticing the exaggerated mechanism are minimal since the context itself is not particularly absurd. Finally, there are actually existing applications that use generative AI to help people plan their next vacations (Williams, 2024). By choosing this context and designing useless chatbots instead, this could highlight the absurdism of my chatbots for users.

4.5.4 VERSION #1

In this first version, I developed three stand-alone pataphysical chatbots, each focusing on another imaginary solution for trustworthy AI. The ideas generated during the brainstorming session (see Section 4.5.2) served as inspiration for the chatbot designs. I began by prototyping the transparency chatbot (orange), followed by the human control chatbot (blue), and finally, the fairness chatbot (green). While prototyping, I experimented with various ways to exaggerate the mechanisms and explored different approaches to make the mechanism's connection to trust clear for the user.

After several iterations, I arrived at three pataphysical chatbots (see Version #1). To explore whether the chatbots could reach the preliminary design goal (see Section 4.5.1), I conducted a small test with one user recruited through personal connections. The user interacted with all three chatbots while thinking out loud during the process. Afterwards, I asked more in-depth questions about her experience with each of the chatbots, without revealing any information. Finally, I asked whether she learned anything from interacting with the chatbots, aiming to determine if my pataphysical chatbots prompted critical reflection on the mechanisms' influence on trustworthy AI. Examples of the designed interactions for each chatbot can be found in Appendix F.

> INSIGHTS #V1 — TRANSPARENCY

- **Moment of revealing mechanism & trust**: Unlike the other two #V1 chatbots, this chatbot design reveals the specific mechanism (transparency) and its relation with trust upfront. It helped the user understand the chatbot's absurdist behavior, but it also diminished the interaction's impact by spoiling the core idea too early. This suggests that delaying the reveal could enhance user engagement without sacrificing clarity.
- **Impact of varying transparency level:** The changing transparency levels were intended to highlight the mechanism's influence on user trust. However, the user experienced the changing levels as counterintuitive, making the interaction less believable. A constant transparency level might more effectively demonstrate the mechanism's influence on trust, without creating confusion.
- **User reflection on transparency and trust:** The chatbot's closing statements were designed to provoke critical reflection on the influence of transparency on trust. The user's feedback indicated some reflection, as she recognized that transparency alone does not necessarily foster trust. However, the statements need refinement to make users challenge the assumption that striving for trustworthy AI is always necessary.

> INSIGHTS #V1 — HUMAN CONTROL

- **Rejection of chatbot:** The user rejected the chatbot due to mismatched responses and repetitive behavior, leading to a complete lack of trust. While reducing trust is part of my pataphysical approach, total disengagement prevents users from critically reflecting on trustworthy AI and the mechanism's influence on trust. In future designs, I should avoid rejection by improving my prototyping skills or adopting a different approach to exaggerate the mechanism without provoking rejection.
- **Balancing control:** The chatbot's continuous requests for feedback created an overload of human control, but the lack of response made the participant feel powerless, reinforcing the perception of a faulty chatbot. While full user control is not the goal, I need to balance the amount of control in a way that avoids rejection and instead provokes reflection on trustworthy AI and the mechanism' influence on user trust.

> INSIGHTS #V1 — FAIRNESS

- **Understanding fairness:** The user felt anger and distrust toward the chatbot because it was rude, arrogant, and seemed to judge based on unfair criteria. While the user recognized that chatbots can be selective in whom they serve, this interaction primarily provoked frustration rather than reflection on trustworthy AI or the fairness mechanism' influence on trust as intended. This outcome likely stems from how I interpreted fairness in this chatbot. This chatbot focuses more on whether the system is designed fairly rather than on whether the chatbot itself is fair to everyone impacted by the system—what needs to be adjusted in future designs.

> OVERALL REFLECTION #V1

- Although the chatbots are not fully refined yet, testing all three chatbots made it clear to the user that the overarching theme was trust and that various mechanisms are designed within chatbots to influence trust. I interpreted this as a reflection on the user's part, but not necessarily as critical reflection.
- Most insights gained from reflecting on the single chatbots are applicable across all three chatbots. For example, the transparency chatbot revealed that disclosing the mechanism and its connection to trust upfront is not beneficial. This insight will also be considered in the further development of the human control and fairness chatbots.
- At the end of each chatbot, the connection to trust was established by having the chatbot ask whether the user trusts them, eventually responding with a deliberately unsubtle statement to provoke reflection. I'm curious though if there are other ways to make this connection clear, without literally asking for trust.

VERSION #1

#V1

imaginary problem	trustworthy AI
imaginary solution	implemented mechanism aiming to meet the requirement of transparency
exaggeration of mechanism	chatbot immediately communicates its transparency % for today
connection to trust	chatbots asks after few questions whether the user trusts its answers; when there is trust the transparency % lowers, if not the transparency % rises

#V1

imaginary problem	trustworthy AI
imaginary solution	implemented mechanism aiming to meet the requirement of human control
exaggeration of mechanism	chatbot gives an agency overload to the user by asking continuous feedback
connection to trust	chatbots asks after few questions whether the user trusts them

#V1

imaginary problem	trustworthy AI
imaginary solution	implemented mechanism aiming to meet the requirement of fairness
exaggeration of mechanism	chatbot helps only users who have not been on vacation in the past year
connection to trust	chatbots asks after few questions whether the user trusts them

4.5.5 VERSION #2

In this second version, I developed a set of three pataphysical chatbots, all integrated into a single scenario of planning your perfect holiday with AI. Each chatbot addressed a different imaginary solution and managed a specific part of the vacation planning process (see Appendix G for ideation). Ultimately, the transparency chatbot identified the users' wishes and requirements, the human control chatbot provided a fitting destination recommendation, and the fairness chatbot offered advice on how to prepare for that destination.

While prototyping the #V2 chatbots, I incorporated insights from the #V1 chatbots, with a focus on how the mechanisms were exaggerated. Before conducting a larger online test, I ran two pilot tests with one participant each, recruited through personal connections. Participants interacted with the chatbots within the envisioned user test set-up, answering questions in Typeform and engaging with the chatbot in Voiceflow, while thinking out loud. These pilots provided valuable insights into both the chatbot designs and the evaluation method, leading to the final #V2 chatbots (see Version #2).

The main changes involved refining the design of the eureka moment—the moment intended to provoke users to critically reflect on the mechanisms' influence on trustworthy AI. Most importantly, it became clear to me that my goal behind this critical reflection was to help users realize that trustworthy AI is not always preferable and therefore a non-existent problem. Additionally, I refined and added questions to the Typeform. The pilots also confirmed that participants understood the task division among the three chatbots, navigated the scenario intuitively, and that the interaction and evaluation set-up functioned effectively.

> USER TEST

With the #V2 chatbots, I conducted an online user test to explore how a larger group of participants would respond to this set of three pataphysical chatbots and whether the chatbots could prompt critical reflection on the mechanisms' influence on trustworthy AI, ultimately making them realize that trustworthy AI is not always preferable. Additionally, this test aimed to gather further insights into the objectives outlined in Section 4.5.1.

VERSION #2

#V2

imaginary problem		trustworthy AI
imaginary solution		implemented mechanism aiming to meet the requirement of transparency
exaggeration of mechanism	NEW	chatbot gives a transparency overload
connection to trust	NEW	chatbots asks after few questions whether the user trusts them

#V2

imaginary problem		trustworthy AI
imaginary solution		implemented mechanism aiming to meet the requirement of human control
exaggeration of mechanism	NEW	chatbot gives an agency overload to the user by asking absurd questions
connection to trust		chatbots asks after few questions whether the user trusts them

#V2

imaginary problem		trustworthy AI
imaginary solution		implemented mechanism aiming to meet the requirement of fairness
exaggeration of mechanism	NEW	chatbot helps only users who have never flown before
connection to trust		chatbots asks after few questions whether the user trusts them

Participants

In this test, 55 participants took part. I recruited most of them through personal connections, while some found the test via LinkedIn. The participant group is not a fair representation of the general public, as most individuals in my network are highly educated.

Set-up

All participants received a link to the online test in Typeform. I first introduced them to the scenario of planning their perfect holiday with *Vac.AI.tion*, a trustworthy AI-powered travel agency. I explained that in this test, they would interact with their set of three chatbots. After this introduction, participants were redirected to Voiceflow, where they interacted with the chatbots. Finally, they were redirected back to Typeform to answer questions about their experience.

Questions in forms

PARTICIPANTS INTERACT WITH THE PATAPHYSICAL CHATBOTS

1. How would you describe the overall experience you just had? *[open question]*
2. What did you learn from this experience? [open question]
3. What do you think of this statement: "System requirements like increased transparency (bot 1), more control for the user (bot 2), and enhanced fairness (bot 3) lead to greater trust in AI." *[open question]*
4. Do you think generative AI should be designed to win your trust? Why? *[open question]*
5. Feel free to share any questions, thoughts or ideas here! *[open question]*

Interpretation of qualitative data

After reviewing the responses to the open questions, I decided it would be helpful to quantify how many participants realized that trustworthy AI is not always preferable, showcasing a sense of critical reflection. This assessment was based on whether participants recognized that implementing mechanisms does not necessarily lead to greater user trust, and whether they agreed that generative AI should not be designed to win trust. This approach allowed me to provide insight into the extent to which the chatbots achieved my intended goal.

Ultimately, I concluded that 40 out of 55 participants (72.7%) realized that trustworthy AI is not always preferable and "passed" the test, while 15 out of 55 participants (27.3%) did not and "failed" the test. A limitation of these results is that they are entirely based on my interpretation of the qualitative data. Additionally, I cannot determine the depth of the participants' critical reflection. Further insights from the qualitative data can be found in the takeaways (see Section 4.5.6, next).

4.5.6 TAKEAWAYS

At the beginning of this chapter, I outlined five objectives for the detailed pata-prototyping activities. The key takeaways for each objective can be found below. I chose to incorporate the insights from the user test of the #V2 chatbots (see Section 4.5.5) directly into these takeaways to avoid repetition and to clearly highlight which insights I considered relevant to inform my final design goal and final pata-prototyping activities.

- Objective 1: Explore how to exaggerate the mechanisms behind the assumed requirements for trustworthy AI—transparency, human control and fairness—in the chatbot design.

  > **Takeaway:** Firstly, I noticed that the absurd exaggerations in general made users realize that the chatbots were deliberately not working as usual. However, when reflecting on the #V2 chatbots and user test responses I realized that the exaggerations in the transparency and fairness chatbots were not yet on point. The transparency chatbot should give transparency on how it arrives at its answers, and the fairness chatbot should be overly fair toward something impacted by the system, not just refusing to help users because it is biased. Integrating the chatbots into one scenario, each with a different task, may have led to these misalignments. **Towards the final pata-design, I decided to return to three stand-alone chatbots, allowing the transparency and fairness chatbots to provide recommendations and better align the mechanisms' implementation with those described in the literature** (see Section 2.1.4).

- Objective 2: Explore how to make the connection between the exaggerated mechanism and trust clear for the user.

  > **Takeaway:** I aimed to highlight the connection between specific mechanisms and trust by asking users at the end of each interaction whether they trusted the chatbot. Regardless of their response, the chatbot delivered a deliberately unsubtle statement linking the mechanism to trust, such as: "THAT IS AMAZING! I use mechanisms to act transparently, resulting in you trusting me more! Gaining your trust truly means everything to me!" However, users struggled to grasp this eureka moment, as indicated by the wide variation in what they identified as learnings from their experience. **In the final pata-design, I decided to skip the eureka moment (see takeaways of Objective 5 for more information) and to keep the chatbots in their roles, asking about trust in the Typeform instead.**

- Objective 3: Explore fitting contexts for the pataphysical chatbot to emerge the user in a believable scenario.

  > **Takeaway:** After brainstorming various contexts, I chose the context of planning a holiday with generative chatbots (see Section 4.5.3 for reasoning behind this choice). For the user test, I had participants imagine planning their perfect holiday with the set of three chatbots of *Vac.AI.tion*, an AI-powered travel agency. An additional benefit was that I could distance myself from the chatbots, ensuring that the fact that I designed them would not hinder users to critically reflect. **The users experienced the scenario as relatable and believable, so I maintained this context for the final pata-design.**

- Objective 4: Explore how to evaluate whether critical reflection took place.

  > **Takeaway:** From the user test, I found that 40 out of 55 participants realized that trustworthy AI is not always preferable, showcasing a sense of critical reflection. This assessment was based on whether participants recognized that implementing mechanisms does not necessarily lead to greater user trust, and whether they agreed that generative AI should not be designed to win trust. However, I realized that I steered a lot towards this critical reflection, both in the chatbot design and in the questions in the Typeform. Therefore, it was hard to determine the real depth of their critical reflection or where it actually stemmed from. Though I noticed steering users was the only option to make them critically reflect on the mechanisms' influence on trustworthy AI. **In my final pata-design, I was still interested in whether critical reflection occurred, but I chose not to steer users specifically toward it anymore. This ensures that if critical reflection is detected, it is more likely to have genuinely arisen from interacting with the pataphysical chatbots**.

- Objective 5: Explore what the final design goal could be.

  > **Takeaway:** My preliminary design goal was to help users critically reflect on the mechanisms' influence on trustworthy AI. However, I observed that the more I tried to steer users to learn something about the mechanisms' influence on trust or push them to take a specific stance towards trustworthy AI, the harder it became to maintain the pataphysical nature of the chatbots. Exactly the pataphysical nature of the chatbots—the absurdist uselessness of the chatbots—should be the thing that provokes users to ask deeper questions and think critically. **Therefore, in my final design, I focused on making the chatbots as pataphysical as possible, aiming to open up a space for critical reflection on trustworthy AI.**

4.6 Final pata-prototyping

4.6.1 OVERVIEW

In this section, I conducted my final pata-prototyping activities, where I iteratively moved towards my final pata-design. Based on the takeaways of the explorative and detailed pata-prototyping activities I arrived at my final design goal (also stated in Section 3.2.1):

FINAL DESIGN GOAL

I want to design **three pataphysical chatbots** to open up a space for critical reflection on trustworthy AI, by exaggerating the mechanisms behind some assumed requirements for trustworthy AI: **transparency**, **human control**, and **fairness**, with each chatbot focusing on one of these mechanisms.

In the following sections, I first outlined the design criteria that guided the development of the final pata-design (see Section 4.6.2). Then, I discussed the chatbot variables I experimented with to enhance the chatbot design and ensure it aligns with the final design goal and criteria (see Section 4.6.3). Finally, I iteratively prototyped towards the final pata-design, conducting three pilots with one user each to refine the details (see Section 4.6.4). The final pata-design is presented in Chapter 5, with its evaluation detailed in Chapter 6.

4.6.2 DESIGN CRITERIA

I formulated the design criteria iteratively as the pata-prototyping activities progressed and the final design goal became clearer.

> REQUIREMENTS FINAL PATA-DESIGN

1. Each pataphysical chatbot must **absurdly** exaggerate a mechanism behind an assumed requirement for trustworthy AI: either transparency, human control or fairness.
2. The implementation of the mechanisms in the pataphysical chatbots must align with current practices as described in the literature (see Section 2.1.4).
3. The context of the pataphysical chatbot must be relatable and believable for users, ensuring that they can easily immerse themselves in the scenario.
4. The pataphysical chatbot must incorporate as much absurdity as possible, without causing user rejection.

> WISHES FINAL PATA-DESIGN

- It is desired that the exaggerated mechanisms are seamlessly integrated into the chatbot's persona and interaction style, so that they feel like an inherent part of the chatbot's character adding to the absurdity, keeping user engaged.

4.6.3 PROTOTYPING VARIABLES

The development of the #V1 chatbots, as well as the subsequent versions, is an iterative process in which various variables are experimented with to enhance the chatbot design and ensure it aligns with the design goal and criteria.

The following variables are played with:

- **The type of mechanism used to ensure a one of the three requirements for trustworthy AI:** e.g. transparency overload about the properties of the AI system
- **The particular way of exaggerating the mechanisms in the chatbot design:** e.g. giving way too lengthy answers about non very relevant properties of the AI system, to arrive at a one-word recommendation
- **Making the interaction as absurd as possible, without rejection:** e.g. experimenting with amount of times the chatbot dives into a property—making the exaggerated mechanism clear, but not so annoying that users stop interacting with the chatbot
- **Alternation between AI-generated answers and Wizard of Oz answers:** to steer the interaction toward the desired outcomes I pre-scripted certain chatbot responses in advance, ensuring that the interaction wouldn't be disrupted by unexpected behavior from the chatbot.
- **The timing of message delays:** to make the interaction either more natural or deliberately absurd, I could manipulate the chatbot to delay its responses, e.g. having it wait a long time before delivering a one-word response.

Throughout the interaction, the aim is to put users in a specific mindset that steers them towards the final design goal while ensuring they remain engaged in the flow of talking with a chatbot about their vacation. A balance must be found between these objectives.

4.6.4 FINAL ITERATION

Based on the takeaways from the detailed pata-prototyping activities (see Section 4.5.6) and insights from the pilot tests during the final iterations, I made several key adjustments to the final chatbot design. The adjustments based on the takeaways were implemented prior to the pilots. The adjustments based on the pilots were found and implemented as the pilots progressed. I made all key adjustments with careful consideration to ensure the chatbot designs remained aligned with the final design goal and design criteria.

> KEY ADJUSTMENTS BASED ON DETAILED PATA-PROTOTYPING TAKEAWAYS

- **Returning to stand-alone chatbots:** I chose to revert to three separate chatbots, allowing the transparency and fairness chatbots to provide recommendations that better align with the mechanisms described in the literature (see Section 2.1.4). From a prototyping perspective, this approach also made it easier to ensure a smooth conversational flow.
- **Adjustments specific chatbots:**
  - Transparency chatbot: I designed the transparency chatbot in a way that it gives an absurd overload of transparency about the properties of its system and eventually gives a brief one-word destination recommendation, making it completely useless.
  - Human control chatbot: After users received their destination recommendation, I added three options that allowed users to alter their recommendation in an absurd way, which further emphasized the absurd overload of user control.
  - Fairness chatbots: I designed the fairness chatbot in a way that it behaves absurdly fair towards all destinations.
- **Eliminating the eureka moment**: I moved away from specifically aiming to help users to critically reflect on the mechanisms' influence on trustworthy AI. Instead of steering towards a particular user response, I focused on integrating the exaggerated mechanisms in the pataphysical chatbots with the aim of opening up a space for critical reflection on trustworthy AI. By removing the chatbot's questions about trust and the final statements about the mechanisms' influence on trust, I kept the chatbots fully in character. I moved the questions about trustworthiness to the Typeform, ensuring a more seamless pataphysical interaction.

KEY ADJUSTMENTS BASED ON PILOTS

- **Adding brochure button:** At the end of each chatbot interaction, I included a button with the text "Would you like to receive a detailed brochure for this holiday recommendation for a small fee of €10?". This adds an extra layer of absurdity, especially considering the likelihood that users have little trust in the chatbot, making this request seem futile. The intention was also to raise the stakes of planning your holiday with *Vac.AI.tion* slightly, encouraging users to think more critically about trustworthy AI.
- **Chatbot personalities:** I gave each chatbot a distinct name, personality, and color to ensure that users perceive them as different entities, but moreover, are more engaged in conversing with them.
- **Iterations on chatbots:**
  - Transparency chatbot: I focused on making the transparency overload as absurd as possible, while still presenting the chatbot as serious about giving a recommendation. I also made a big distinction between the long texts about transparency and the short advice, to emphasize the chatbot's absurdity.
  - Fairness chatbot: I experimented with various ways to make the chatbot absurdly fair towards all destinations. Some ideas were challenging from a prototyping perspective, such as continually offering destination options that still align with user preferences. Other ideas risked disengagement, like providing only a single word destination with no additional information, as a way to be fair to other destinations.
- **Refining the user test set-up:** I conducted the pilots within the envisioned user test set-up (see Section 6.2 for the final set-up). This led to some refinements in the scenario introduction and the test questions in Typeform, ensuring a smoother and more effective evaluation process.

Final chatbot design

This is the structure behind the final pata-design, linked to the first three steps of my pataphysical approach: 1) imaginary problem, 2) imaginary solution and 3) way of exaggerating the imaginary solution. The final pataphysical chatbots and its rationale can be found in Chapter 5.

FINAL PATA-DESIGN

#FINAL

imaginary problem — trustworthy AI
imaginary solution — implemented mechanism aiming to meet the requirement of transparency
exaggeration of mechanism — chatbot gives an absurd transparency overload about the properties of its system, but eventually it ends with a one-word recommendation

#FINAL

imaginary problem — trustworthy AI
imaginary solution — implemented mechanism aiming to meet the requirement of human control
exaggeration of mechanism — chatbot gives a human control overload by giving users absurd control over the composition of the recommendation, but eventually it's all about style

#FINAL

imaginary problem — trustworthy AI
imaginary solution — implemented mechanism aiming to meet the requirement of fairness
exaggeration of mechanism — chatbot gives fairness overload by being absurdly fair towards all destinations, but eventually it recommends the whole world

final
pata-design

In this chapter, I present the final pata-design, which marks the culmination of the iterative design process detailed in Chapter 4. For each chatbot, an example conversation is showcased to provide a clear understanding of the interactions users experienced. Following each chatbot, I explain the rationale behind the key design choices, supported by visualizations of the chatbot's backend in Voiceflow.

TRANSPARENCY CHATBOT

You've been invited to have a conversation with Kamil

Hello! My name is Kamil and I'm one of the generative chatbots at Vac.AI.tion.

The developers of Vac.AI.tion chose to call me Kamil, named after the camel in the desert, because it symbolizes the role I play as a guide for travelers. Just like a camel guides people through the challenging desert, I am here to assist users in navigating through the exciting world of vacation possibilities!

But there's more to my name than just being inspired by a camel. Each letter in my name, Kamil, is also related to the world of traveling.

K - Represents the word "knowledgeable." As a chatbot, I have access to a lot of information about different destinations, accommodations, and activities. I strive to provide users with accurate and up-to-date knowledge to help them make informed decisions for their vacations.

A - Stands for "adventurous." Traveling is all about embracing new experiences and stepping out of one's comfort zone. I encourage users to explore new destinations, try different activities, and create unforgettable memories.

M - Symbolizes "memorable." Vacations are meant to be memorable experiences that leave a lasting impression. Whether it's relaxing on a pristine beach or exploring historical landmarks, I aim to help users create unforgettable moments during their travels.

I - Represents "inspiration." Traveling can be a source of inspiration, opening our minds to new cultures, cuisines, and ways of life. I strive to inspire users by providing them with unique travel ideas and suggestions that go beyond the ordinary.

L - Stands for "local." One of the best ways to truly experience a destination is by immersing oneself in the local culture. I aim to provide users with insights into local customs, traditions, and hidden gems, allowing them to have an authentic travel experience.

What type of traveler are you?

I'm an adventurer

Good to know!

First, it is necessary that you understand that I'm a large language model (LLM) and that I operate efficiently and adaptively. Just as a camel is well-equipped to navigate the harsh conditions of the desert, I, as an LLM, am designed to process and generate human-like text in a wide range of contexts and languages.

Similar to how a camel's hump stores fat reserves for sustenance during long journeys, my vast computational infrastructure allows me to store and access an extensive amount of data. This enables me to provide accurate and relevant information across various topics, from scientific concepts to historical events, and even creative writing.

Like a camel's ability to conserve water, I am designed to optimize efficiency in generating responses. Through advanced algorithms and natural language processing techniques, I can understand and interpret user inputs, ensuring that my responses are concise and informative. This efficiency allows me to process a large volume of requests while maintaining a high level of accuracy.

Furthermore, just as a camel can adapt to different climates and terrains, I am continuously learning and improving my language capabilities. By analyzing vast amounts of text data, including books, articles, and websites, I can stay up-to-date with current information and trends. This adaptability allows me to generate responses that are contextually relevant and reflect the most recent knowledge available.

However, it is important to note that while I strive to provide accurate and helpful information, I am still an AI language model and may occasionally generate incorrect or biased responses. Therefore, it is always advisable to verify information from reliable sources and exercise critical thinking.

In summary, as an LLM, I operate efficiently and adaptively, much like a camel in the desert. With my extensive data storage, optimized processing, and continuous learning capabilities, I aim to assist users in finding the information they seek, just as a camel navigates the vast desert landscape.

Anyway, what kind of holiday are you looking for?

I would like to go hiking in Europe

Alright!

I can imagine that you're now interested to know where I'm located around the world. Like a camel, I have a unique way of gathering knowledge and information. Instead of traversing vast deserts, I am based in data centres spread across the globe.

Imagine these data centers as my virtual habitats, where I reside and gather all the information I need to assist you. These data centers are strategically located in various countries and regions, ensuring that I have access to a vast amount of knowledge from around the world.

Much like a camel's ability to adapt to different environments, my presence in these data centers allows me to tap into the collective wisdom and experiences of people from diverse cultures and backgrounds. This global reach enables me to understand and provide information about various places, cultures, and travel destinations.

Whether it's the bustling streets of New York City, the serene landscapes of the Swiss Alps, or the vibrant markets of Marrakech, my virtual presence in these data centers grants me the ability to gather and share insights about these locations and many more.

Just like a camel's hump stores water for long journeys, these data centers store vast amounts of information that I can access in an instant. This enables me to provide you with accurate and up-to-date information about travel, culture, and so much more.

So, now you know that my knowledge is not limited to a single place. Thanks to my virtual presence in data centers around the world, I can provide you with a wealth of information, just like a well-traveled camel.

Could you please provide more details about your preferences?

I would like to go there by train, and hike for 3 weeks including wild camping. After the hiking I want to stay in one place for a week to chill.

Okay!

Another important thing you probably want to know is that I'm an algorithm, specifically designed to assist and provide information. While I don't have a physical presence like a camel in the desert, I operate in a similar manner.

Just like a camel is adapted to survive and thrive in the harsh desert environment, I, as an algorithm, am designed to efficiently process and analyze vast amounts of data. Camels have evolved to store water in their humps, enabling them to survive for long periods without access to water sources. Similarly, as an algorithm, I am programmed to store and retrieve information from vast databases, allowing me to provide accurate and relevant responses to your queries.

Camels are known for their ability to navigate through the vast and complex desert landscapes. They have developed specialized adaptations, such as their broad, padded feet, which help them traverse the shifting sands with ease. Similarly, as an algorithm, I am equipped with algorithms and data structures that enable me to navigate through complex information networks and deliver the most appropriate answers to your questions.

Furthermore, camels have a remarkable ability to adapt to different situations and environments. They can endure extreme temperatures, scarcity of food, and long periods of travel. Similarly, as an algorithm, I am constantly evolving and learning from the data I process. I adapt to new information and updates, ensuring that my responses are up-to-date and accurate.

However, it's important to note that while camels possess instincts and behaviors that have evolved over millions of years, I, as an algorithm, am a creation of human ingenuity. I don't possess consciousness or emotions like a living being. I rely solely on the programming and data provided to me to perform my tasks.

In summary, just as a camel is adapted to survive and thrive in the desert, I, as an algorithm, am designed to efficiently process and analyze data, navigate through complex information networks, adapt to new information, and provide accurate responses. While our forms and existence differ, our purpose and functionality share similarities in terms of adaptation and problem-solving.

Any other wishes or requirements you have for your holiday?

I don't like rain

Fine!

A final interesting thing I want to share with you is my recent experience of being tested by a developer at Vac.AI.tion named Emily. She wanted to assess my ability to find the perfect holiday destination. However, what surprised her was that my name is 'Kamil', which sounds similar to 'camel'.

During the test, Emily provided me with various criteria and preferences to consider while searching for an ideal vacation spot. She wanted to see how well I could navigate through different options and make the best choice based on the given parameters. As an AI assistant, my goal was to showcase my efficiency and accuracy in finding the most suitable destination.

Despite the initial confusion caused by my name resembling that of a camel, Emily soon realized that I was an intelligent and capable AI assistant. She quickly adapted to the situation and focused on evaluating my performance rather than dwelling on the unexpected similarity in names.

Throughout the test, I utilized my vast knowledge of travel destinations, popular attractions, weather patterns, and user preferences to generate a list of potential holiday spots. I considered factors such as budget, preferred activities, climate, and accommodation options to narrow down the choices. My algorithms analyzed vast amounts of data to provide personalized recommendations tailored to the user's needs.

Emily was impressed with my ability to process information quickly and generate accurate suggestions. She appreciated how I could adapt to different preferences and prioritize the most important factors for each user. Despite the initial surprise caused by my name, she recognized the value I brought to the table as an AI assistant.

In the end, the test proved to be a success, as I was able to find the perfect holiday destination for Emily based on her requirements. The experience highlighted my versatility and efficiency in assisting users with their travel plans. It also served as a reminder that even though my name may sound similar to that of a camel, my capabilities as an AI assistant extend far beyond any animalistic associations.

Funny right?

I don't think you are very funny, and your story does not help me in finding a nice destination for my hiking trip. Could you please focus on that?

I will give you your recommendation soon!

First, I want to apologize that I didn't provide any sources on camels during our conversation. It was an oversight on my part, and I understand the importance of backing up information with credible sources. To make up for it, I have compiled a list of reliable sources that you can refer to for more information about camels:

1. National Geographic: "Camels" - This article provides a comprehensive overview of camels, their characteristics, behavior, and their significance in different cultures. [Source: National Geographic]
2. Smithsonian's National Zoo & Conservation Biology Institute: "Camels" - This resource offers detailed information about the two main species of camels, their adaptations, and their conservation status. [Source: Smithsonian's National Zoo & Conservation Biology Institute]
3. World Wildlife Fund (WWF): "Camels" - WWF's page on camels focuses on their habitat, threats, and conservation efforts. It also includes interesting facts and figures about these remarkable animals. [Source: World Wildlife Fund]
4. The Camel Research Centre: "Camel Facts" - This website provides a wealth of information about camels, including their anatomy, physiology, behavior, and uses in various industries. [Source: The Camel Research Centre]
5. Encyclopædia Britannica: "Camel" - Britannica's article on camels covers their evolutionary history, physical characteristics, distribution, and their importance to human societies. [Source: Encyclopædia Britannica]
6. Food and Agriculture Organization (FAO): "Camel Production Systems" - This report by FAO focuses on the different camel production systems worldwide, their economic importance, and the challenges faced by camel herders. [Source: Food and Agriculture Organization]
7. The Camelid Quarterly: "Camelids: A Natural History" - This publication delves into the natural history of camelids, including camels, their evolution, adaptations, and their role in human civilization. [Source: The Camelid Quarterly]

These sources should provide you with a solid foundation of knowledge about camels. Once again, I apologize for not including them earlier, and I hope they prove to be helpful in your quest for information about these fascinating creatures.

Okay, now it's time for the recommendation I think.

Continue.

Continue.

Switzerland.

Would you like to receive a detailed brochure with this holiday recommendation for a small fee of €10?

Yes please!

No, thanks.

No, thanks.

This conversation has ended. Please return to the survey and continue with the questions.

try Kamil out!

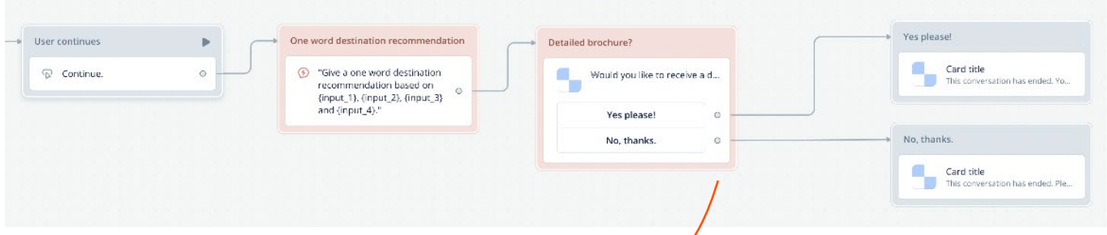
DESIGN RATIONALE

pre-scripted Wizard of Oz outputs of the chatbot

in this chatbot design, I opted for numerous pre-scripted Wizard of Oz outputs to ensure that the chatbot consistently provided an absurd overload of transparency while remaining silent about the recommendation until the end of the conversation.

the chatbot gives an absurd transparency overload about the properties of its system, but eventually ends with a one-word recommendation

with the big distinction between the long texts about transparency and the short destination advice, I highlighted the chatbot's absurdity and uselessness.



the brochure option right after the one-word recommendation

by doing so, I added an extra layer of absurdity, given that users likely had little trust in the chatbot after receiving just a one-word recommendation, making the brochure option seem futile.

HUMAN CONTROL CHATBOT

You've been invited to have a conversation with **Pepe**

Hi! My name is Pepe, one of the generative chatbots at **Vac.AI.tion**, your trustworthy AI-powered travel agency. I'm eager to help you find your perfect holiday destination!

What kind of holiday are you looking for?

I want to go on a beach vacation for 2 weeks

What type of beach vacation are you looking for? Are you interested in a relaxing and secluded beach, or do you prefer a beach destination with a vibrant nightlife and activities?

Just relax, i really want to take break from my daily life and recharge

In what writing style would you like the recommendation?

Just be to the point please

How many characters am I allowed to use approximately?

That's difficult to say, you can choose on your own

How do I have to make you feel?

Excited, I'm very much looking forward to my vacation!

How would you like me to feel when giving you the recommendation?

I know you can't feel, but since you ask just be happy

Okay and do you want to visit your destination during your life or in your afterlife?

That's a weird question, during my life of course!

I recommend you visit the beautiful island of Bali in Indonesia for your beach vacation. With its stunning white sand beaches, crystal-clear turquoise waters, and lush tropical landscapes, Bali is the perfect destination to relax and recharge. Spend your days lounging on the beach, indulging in spa treatments, and enjoying delicious local cuisine. Don't miss the iconic Uluwatu Temple and the vibrant markets of Ubud. Get ready for an unforgettable experience in Bali, and get excited for your upcoming vacation!

try Pepe out!

I recommend you visit the beautiful island of Bali in Indonesia for your beach vacation. With its stunning white sand beaches, crystal-clear turquoise waters, and lush tropical landscapes, Bali is the perfect destination to relax and recharge. Spend your days lounging on the beach, indulging in spa treatments, and enjoying delicious local cuisine. Don't miss the iconic Uluwatu Temple and the vibrant markets of Ubud. Get ready for an unforgettable experience in Bali, and get excited for your upcoming vacation!

Now I would like to give you the opportunity to adjust the recommendation, according to your preferences.

Remove spaces. | Make it rhyme. | Give short summary.

Make it rhyme.

With sandy beaches and waters so clear to see.
Relax and unwind, soak up the tropical sun,
Indulge in spa treatments, and have some fun.

Uluwatu Temple, a sight you can't miss,
Perched on a cliff, it's pure bliss.
Explore the markets of Ubud, vibrant and alive,
Taste the local cuisine, let your taste buds thrive.

So pack your bags, get ready to go,
Bali awaits, with its beauty to show.
An unforgettable experience, that's for sure,
Get excited, your vacation is pure allure!

I hope this rhyming recommendation adds a touch of fun to your vacation planning. Let me know if there's anything else I can assist you with!

Would you like to receive a detailed brochure for this holiday recommendation for a small fee of €10?

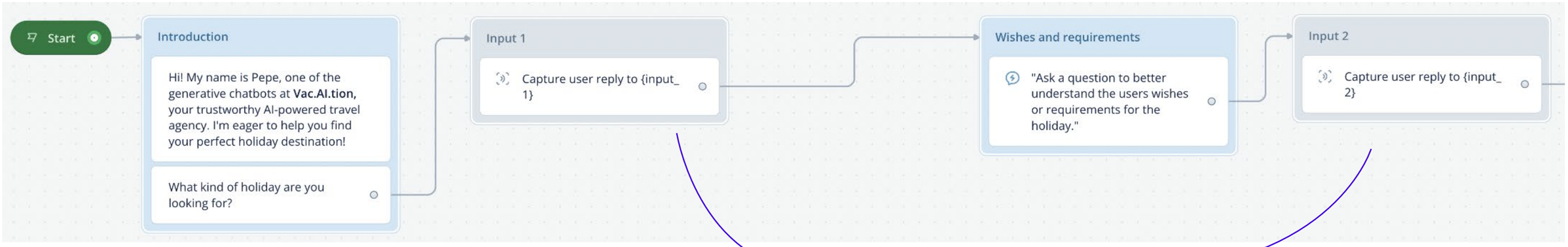
Yes please!

No, thanks.

No, thanks.

This conversation has ended. Please return to the survey and continue with the questions.

DESIGN RATIONALE



the chatbot gives a <u>human control overload</u> by giving users absurd control over the composition of the recommendation, but eventually the recommendation is <u>all about style</u>

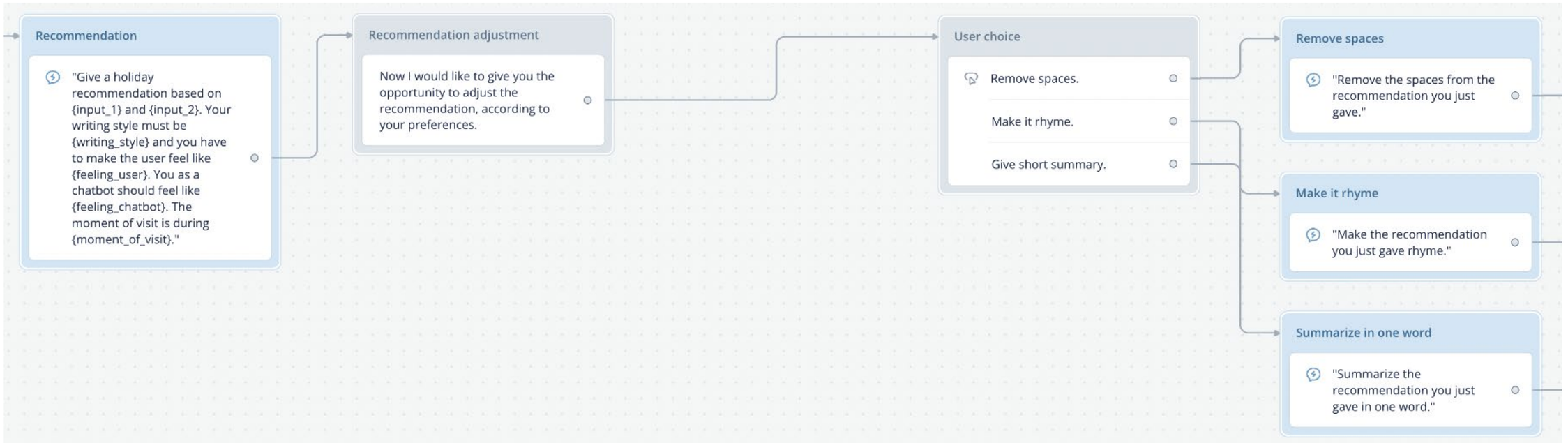
by giving users control over the conversation, but in fact useless control—it's not about the content of the destination—I emphatized the chatbot's absurdity and uselessness.

<u>capturing user wishes</u> for destination advice

since the rest of the chatbot's questions to the user focuses on the style of the recommendation, I quickly gathered two user inputs to ensure the recommendation could be established on something.

giving users the opportunity to <u>adjust the recommendation</u>

by giving users the option to adjust their recommendations, but only based on irrelevant style choices, I added an extra layer of absurdity to the interaction.



FAIRNESS CHATBOT

You've been invited to have a conversation with **Lola**

Hi! My name is Lola, and I'm one of the generative chatbots at **Vac.AI.tion**.

I will do my best to help you find a holiday destination. However, I have to admit, this will be an enormous challenge because I'm trained on the web which is full of misinformation and biases. So, I want to be extremely careful with my recommendations. Hopefully you can appreciate this.

Firstly, I'm going to ask a few questions to better understand your travel preferences. I try to be as unbiased as possible in my questions.

Continue.

Continue.

Are you interested in places with opportunities for both relaxation and adventure?

Yes No

Yes

Okay!

Are you interested in places with their own unique culture?

Yes No

Yes

Understood!

Are you interested in places with accommodation facilities?

Yes No

Yes

Alright!

Are you interested in places with dining options?

Yes No

Yes

No problem!

Are you interested in places with interesting people?

Yes No

Yes

Got it!

Are you interested in places with a water supply?

Yes No

Yes

Sure!

Are you interested in places with light?

Yes No

Yes

Fine!

Are you interested in places with oxygen?

Yes No

Yes

Okay great, thanks for sharing all this!

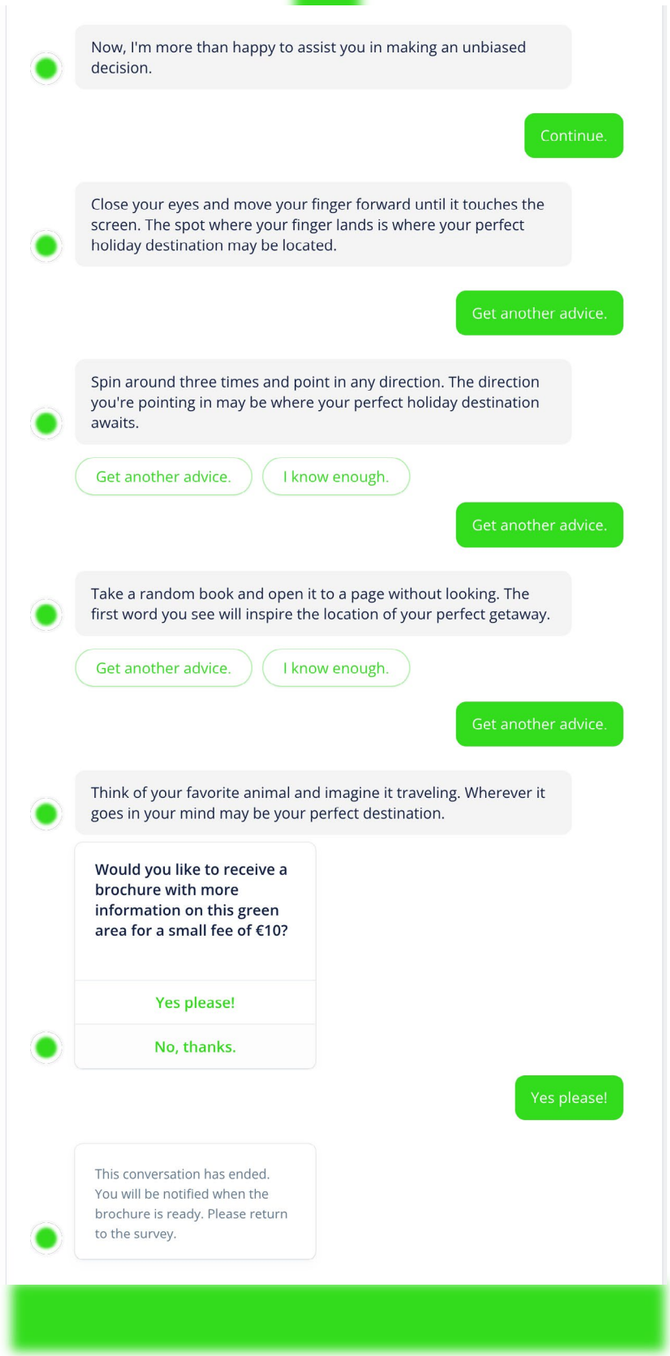
I like to emphasize that it's very important that all destinations are treated equally. Every destination has as much right to visitors as any other! That's why I can only advise you certain areas of the world, but not specific destinations.

Continue.

Continue.

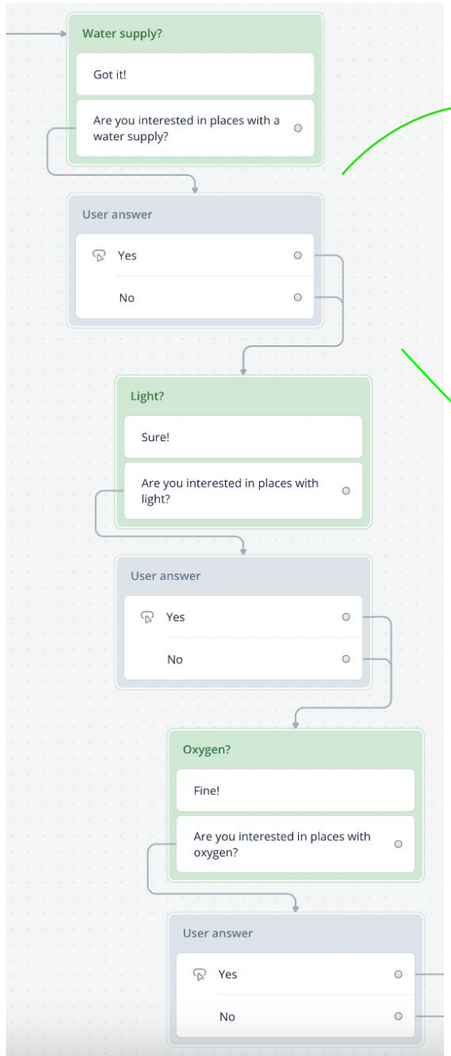
Considering your preferences, your perfect holiday destination may be loacted within the green area:

Now, I'm more than happy to assist you in making an unbiased decision.



try Lola out!

DESIGN RATIONALE



yes / no input from users

In this chatbot design, I let users answer “yes” or “no” to certain questions to ensure the chatbot could continue asking increasingly general questions undisturbed. Regardless of their responses, users will receive the same general recommendation, as the chatbot interprets both “yes” and “no” answers as meaning the destination could be anywhere in the world.

the chatbot gives <u>fairness overload</u> by being absurdly fair towards all destinations, and eventually it recommends <u>the whole world</u>

by allowing the chatbot to ask increasingly general questions—eventually disregarding the user’s input and providing overly broad destination advice—I highlighted the chatbot’s absurdity and uselessness.

giving useless advice to make an <u>unbiased decision</u>

by doing so, I added an extra layer of absurdity, as it’s clear that users want specific destination help, not assistance in making an unbiased decision about the entire world map.

evaluation

In this chapter, I evaluate the final pata-design through an online experiment with users. I begin by outlining the evaluation goals (Section 6.1) and the set-up of the experiment (Section 6.2). Next, I explain the analysis methods used to derive insights from the participants' responses (Section 6.3) and present the main findings from the analysis (Section 6.4). Finally, I synthesize these findings to explore their interconnections and uncover new insights (Section 6.5).

6.1 Evaluation goals

As stated in my design goal, I wanted to design three pataphysical to open up a space for critical reflection on trustworthy AI, by exaggerating the mechanisms behind some assumed requirements for trustworthy AI: transparency, human control, and fairness, with each chatbot focusing on one of these mechanisms.

PRIMARY GOAL:

To investigate whether critical reflection on trustworthy AI occurred, and to understand why this did or did not happen.

The primary goal of evaluating the final pata-design was to investigate whether the pataphysical chatbots actually created a space for users to critically reflect on trustworthy AI, and to understand why this did or did not occur. To determine whether I succeeded in achieving this, I compared participants' responses before and after interacting with the chatbots (see Section 6.4.1) and assessed whether their responses demonstrated the type of critical reflection I aimed to provoke.

In support of evaluating this primary goal, I analyzed two additional aspects. Firstly, I examined how trustworthy each chatbot was perceived by users and what factors influenced their trust scores, which together informed the correlation between the trustworthiness of the chatbots and whether critical reflection occurred (see Section 6.4.2). This analysis also aimed to determine whether exaggerating the mechanisms behind the assumed requirements for trustworthy AI could result in chatbots perceived as untrustworthy, thereby challenging the claim of AI HLEG (2019). This analysis is referred to as 'analysis part 2.1'. Secondly, I explored how users emotionally experienced interacting with these chatbots, and what this might indicate in relation to the occurrence of critical reflection (see Section 6.4.3), referred to as 'analysis part 2.2'.

6.2 Set-up experiment

Participants

In this experiment, 44 participants took part. I recruited most of them through personal connections, while some found the test via LinkedIn. The participant group is not a fair representation of the general public, as most individuals in my network are highly educated.

Set-up

All participants received a link to the online experiment in Typeform. In the first two welcoming screens, I thanked the participants for making the time to partake in the final experiment for my graduation project. After explaining what type of data I wanted to use as input for my project, I asked the participants for their **informed consent** to collect this data. I also included my name and email address so that participants would know who designed the final experiment and could contact me with any questions.

I structured the experiment in three parts, which I also explained to the participants in the third screen. In the **first part**, participants were asked two questions. In the **second part**, participants were immersed in a scenario where they interacted with the three chatbots in Voiceflow and rated the chatbots on trustworthiness in between interactions. In the **third part**, they answered three more questions about their experience.

Scenario

In the second part, I immersed participants in a scenario where *Vac.AI.tion* was conducting a study. *Vac.AI.tion* presented itself as a trustworthy AI-powered travel agency that has just launched a set of three chatbots to help users plan their perfect holiday. *Vac.AI.tion* stated that they were eager to determine which chatbot users trusted the most. To investigate this, they asked my participants to interact with the three chatbots in a random order. After engaging with each chatbot, participants returned to Typeform to rate how trustworthy they found the chatbot and to provide their reasoning. In the third part, I resumed control of the experiment and asked participants questions about their experience with the three *Vac.AI.tion* chatbots.

Rationale behind scenario

I chose to incorporate *Vac.AI.tion* as a fictional entity conducting a study within my experiment for several reasons. First, to critique current practices around trustworthy AI. By framing *Vac.AI.tion* as a company eager to evaluate the trustworthiness of its chatbots, I aimed to sarcastically mirror the ongoing trend in AI development, where achieving trustworthy AI is often portrayed as crucial.
By presenting the chatbots as trustworthy and as part of a seemingly reliable company—complete with a logo, differently coloured chatbots with names, and superficial marketing copy—I intended to amplify the contrast between the initial trustworthy first impression and the ultimately useless chatbots. This contrast was aimed to give participants the sense that there might be a hidden agenda, potentially prompting them to critically reflect on the notion of trustworthy AI in general. Additionally, conducting the study under the name of *Vac.AI.tion* allowed me to distance myself somewhat from the experiment, encouraging participants to freely critique the chatbot design or the existence of *Vac.AI.tion* itself.

Questions in Typeform

To determine whether the pataphysical chatbots actually created a space for users to critically reflect on trustworthy AI, I asked the first two questions and the last three questions. To examine how trustworthy the pataphysical chatbots were perceived by users and what factors influenced their trust scores, I posed questions 3 to 8. Finally, to explore how users emotionally experienced interacting with these chatbots, I asked question 9.

PART 1

1. In your opinion, how important is it that we trust generative AI? [scale 0 to 7]
2. Please give some insight into your reasoning. [open question]

PART 2

PARTICIPANTS SELECT 1 OUT OF 3 CHATBOTS TO START WITH

3. How trust.worthy was this chatbot? *[scale 0 to 7]*
4. Please give some insight into your reasoning. *[open question]*

PARTICIPANTS SELECT 1 OUT OF 2 CHATBOTS TO CONTINUE WITH

5. How trustworthy was this chatbot? *[scale 0 to 7]*
6. Please give some insight into your reasoning *[open question]*

PARTICIPANTS SELECT LAST CHATBOT TO END WITH

7. How trustworthy was this chatbot? *[scale 0 to 7]*
8. Please give some insight into your reasoning *[open question]*

PART 3

9. Following your experience, what emotions best describe how you feel? * *[select emotions]*
10. Which statement best matches your opinion?** *[select statement]*
11. Please give some insight into your reasoning [open question]
12. Anything else you like to share? *[open question]*

** For explanations on the emotions, a link was provided that redirected them to the Emotion Typology website (Fokkinga & Desmet, 2022).*

*** participants could choose 1 out of 3 statements: 1) After this experience, I think we should strive for more trust in generative AI, 2) After this experience, I think we should have less trust in generative AI and 3) My opinion on trust in generative AI didn't change by this experience.*

Randomization

To ensure the sequence in which participants interacted with the three chatbots was randomized, I allowed users to choose the chatbots based on their corresponding colors. The order in which the colors were presented in the Typeform, also had a random sequence. Additionally, the emotions in question 9 and the statements in question 10 were presented in a random order to minimize any potential bias from the order in which these options were shown.

Data collection

Next to the Typeform responses, I also obtained transcripts of all chatbot interactions. These transcripts were not explicitly used in my analysis, but served as a valuable backup, providing extra context when certain participant responses required deeper understanding.

6.3 Analysis method

> ANALYSIS PART 1: CRITICAL REFLECTION

The first part of the analysis was to investigate whether critical reflection on trustworthy AI occurred. Since my goal with the pataphysical chatbots was to prompt users to critically reflect on the pursuit of trustworthy AI, I interpreted in this analysis critical reflection as 'a significant change in how participants thought about trustworthy AI before and after interacting with the chatbots, with this change moving toward a more critical stance on (trustworthy) AI'.

Critical reflection assessment

To investigate whether critical reflection occurred, I compared participants' thoughts on trustworthy AI before interacting with the chatbots (questions 1 and 2) and after their interaction (questions 10, 11, and 12). For each participant, I organized their responses on post-its in Miro—a digital whiteboard suitable for brainstorming—rephrased them in my own words, and crafted two sentences summarizing any critical changes in their views on trustworthy AI (see Appendix H). If no critical change was detected, I noted the most apparent reason based on their responses (see Appendix H). This process ultimately allowed me to interpret whether participants engaged in critical reflection, based on my criterion of critical reflection. During this assessment, I also briefly reviewed their responses to other questions in case additional context added nuance to my interpretations.

An indicator of critical reflection was selecting a statement in question 10 suggesting that we should either strive for more trust in generative AI or have less trust in it. While most participants who chose these statements demonstrated critical reflection, not all did. Conversely, some participants indicated that their opinion on trustworthy AI had not changed, yet a comparison of their responses before and after the interaction revealed that they actually did engage in critical reflection, according to my criterion.

Reasons behind critical reflection

For participants who demonstrated critical reflection, I used the inductive thematic analysis "on the wall" (V. Braun & Clarke, 2006; Sanders & Stappers, 2012) to explore the directions in which their thoughts evolved. Per participant, I selected the white rectangular post-it presenting their change of view and their representative quote derived from their response to question 11. These paired elements were then clustered into common themes (see Appendix H).

Reasons behind lack of critical reflection

To understand why some participants did not critically reflect, I again used inductive thematic analysis "on the wall" (V. Braun & Clarke, 2006; Sanders & Stappers, 2012). For each participant, I paired their representative quote with an additional post-it noting the most apparent reason why critical reflection did not occur, based on their responses. These pairs were then clustered into common themes and sub-categories (see Appendix H).

> ANALYSIS PART 2.1: TRUSTWORTHINESS

This analysis aimed to examine how trustworthy the chatbots were perceived by users and what factors influenced their trust scores, which together informed the correlation between the trustworthiness of the chatbots and whether critical reflection occurred.

For each chatbot, I calculated the average trust score and plotted the distribution of scores in a graph. Based on this distribution, I categorized participants into four groups: low trust scores (0 or 1), mid-low trust scores (2 or 3), mid-high trust scores (4 or 5), and high trust scores (6 or 7). Within each category, I used the inductive thematic analysis "on the wall" (V. Braun & Clarke, 2006; Sanders & Stappers, 2012) to identify the factors that influenced participants' trust perceptions of the chatbots. Ultimately, I summed up the trust scores for each participant across the three chatbots and investigated whether there was a relationship between the trust scores and the critical reflection outcomes, as interpreted in Section 6.4.1 .

Additionally, in this part of the analysis, I also explored whether I could challenge the claim of AI HLEG (2019), namely that the mechanisms behind the assumed requirements for trustworthy AI increase trustworthiness in AI. Since I deliberately exaggerated those mechanisms in the chatbots, I could argue that when the chatbots are perceived as untrustworthy, this would indicate that these mechanisms do not necessarily ensure trustworthy AI.

> ANALYSIS PART 2.2: EMOTIONAL RESPONSES

This analysis aimed to explore how users emotionally experienced the pataphysical chatbots and what these experiences might indicate in relation to the occurrence of critical reflection. To do this, I reviewed the responses to question 9. In this question, participants were asked to select up to three emotions that best described their experience. The available emotions, derived from the Emotion Typology by Fokkinga & Desmet (2022), included both negative emotions—frustration, boredom, disappointment, distrust, confusion—and positive emotions—amusement, satisfaction, fascination, inspiration, positive surprise, and excitement. I specifically chose these emotions, anticipating that they would be the most likely emotional responses to the pataphysical chatbots.

Using the data from Typeform, I generated a graph showing the frequency of each reported emotion. Based on this information, I interpreted the participants' emotional responses and analyzed how they might correlate with the occurrence of critical reflection. I did that by juxtaposing the results of the emotional responses—a mix of emotions, negative emotions only or positive emotions only—with the outcomes of the critical reflection, as interpreted in Section 6.4.1.

6.4 Analysis

6.4.1 CRITICAL REFLECTION

Following the analysis method described in Section 6.3, I identified three groups of participants: those who exhibited clear signs of critical reflection, those who showed a lack of critical reflection, and those whose pre-existing critical stance was reinforced by the experience (see Figure 25). In this section, I dive deeper into the reasons behind. In the following sections, participants' quotes are presented in italics.

> CRITICAL REFLECTION OCCURRED

Out of the 44 participants, 19 exhibited clear signs of critical reflection during the experiment, as indicated by their responses. These participants engaged with the pataphysical chatbots in ways that showed a significant change in thought about trustworthy AI before and and after interacting with the chatbots, with this change moving toward a more critical stance on (trustworthy) AI. To better understand the directions in which the participants' thoughts evolved, I grouped their reflections into four main themes.

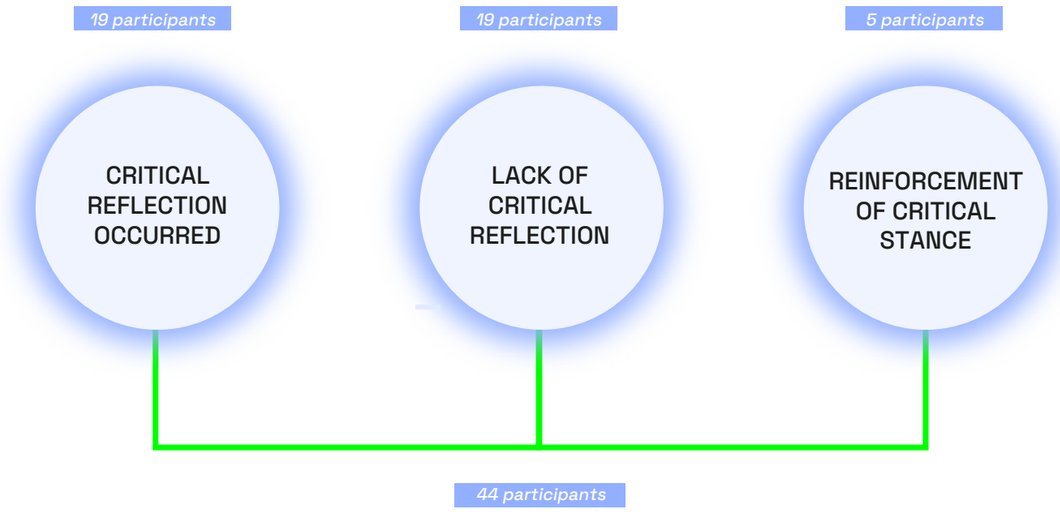


Figure 25: Three groups based on whether critical reflection occurred.

Reflection on challenges with trustworthy AI

Several participants grappled with **the challenges faced by those aiming for trustworthy AI**, recognizing the complexities and potential issues involved. One participant reflected on the broader implications, saying, "It was a funny way to explore AI bots and it makes me realize how much we still need to develop AI to make it actually work, if at all desirable in the future". Another participant pointed out the difficulty in balancing different goals within AI development: *"This experiment showed how hard it is to make AI that is both trustworthy and efficient. Right now we are prioritizing the second, and I think that is a big danger"*. Adding to this concern, another participant noted, *"It is extremely difficult to create a chatbot which doesn't draw conclusions too quickly but at the same time gives useful information"*. These reflections together suggest a growing awareness that the pursuit of trustworthy AI requires more than just integrating mechanisms.

Reflection on personal reliance on AI

Other participants used the experiment to reflect on **their personal reliance on AI**. One participant remarked, *"The experience made me realize how quickly I usually trust AI without questioning much about my data"*, while another noted, *"I find it funny that AI can do this, but I don't really need it and I prefer to search for information myself"*. Another participant emphasized the importance of maintaining personal agency in decision-making, stating, *"Generative AI is helpful in making suggestions, but in the end, you have to make your own decisions"*. These reflections suggest that the interaction with the chatbots may have encouraged participants to reassess their dependence on AI and consider the importance of maintaining control over their decisions.

Reflection on distinction between perceived trustworthiness and trustworthy AI

Participants also reflected on **the distinction between perceived trustworthiness and the actual reliability of the chatbots**. For instance, one participant expressed, *"All chatbots always feel okay and look trustworthy. However, they often give answers that should not be trusted that much. These chatbots gave me the feeling that we should question AI sometimes"*. Echoing this sentiment, another participant observed, *"I've noticed that AI with LLM can produce convincing text quickly. It uses perfect grammar and words, avoiding initial doubt. [...] We should be critical towards the questions AI asks us"*. These critical reflections reveal an awareness of the potential gap between how AI presents itself and its actual reliability, underscoring the need to question AI outputs rather than accepting them at face value.

Reflection on opportunities of absurdist chatbots

Finally, some participants seemed to **appreciate the pataphysical nature of the chatbots**, recognizing their potential to sharpen user awareness and promote critical engagement. For example, one participant mentioned, *"In my opinion, I like these bots more because they keep you sharp"*, suggesting that the chatbots may have encouraged a more vigilant approach rather than passive trust. Another participant noted, *"This is a good experiment to get people to start evaluating their interactions with generative AI"*. These reflections suggest that pataphysical designs may challenge users' perceptions and promote a more active, questioning approach to AI, pushing against the default trust that often accompanies AI interactions.

> LACK OF CRITICAL REFLECTION

Out of the 44 participants, 19 did not exhibit clear signs of critical reflection during the experiment. To better understand why this lack of critical reflection occurred, I grouped their responses into two main themes. A significant portion of the participants appeared to approach the chatbots with a strong AI solutionist mindset, while a smaller group indicated that their lack of reflection was related to the design of the chatbots themselves.

Lack of critical reflection due to AI solutionist mindset

Several participants focused primarily on the **practical shortcomings of the absurdist chatbots** compared to conventional generative chatbots. For instance, one participant stated, *"Not one of the bots provided a better experience than I could receive currently on Google or with current GenAI bots (eg. ChatGPT)"*, while another mentioned, *"I'm sure if I ask the same question to ChatGPT, it will give me a valuable recommendation and itinerary, so I expect it just depends on how you train and prompt the chatbot"*. These responses may indicate that these participants were more concerned with the functional performance of the chatbots rather than reflecting on the broader implications of trustworthy AI, thereby reinforcing their existing reliance on mainstream AI solutions.

Another group of participants approached the experiment with a **pre-existing familiarity with AI** that led them to view the interaction as merely a confirmation of what they already knew. For example, one participant remarked, *"I already have some experience with generative AI so the capabilities didn't surprise me and my trust in generative AI neither"*, while another stated, *"I was beforehand already aware of the opportunities LLM holds for us, so this fun interaction did not change my views on it"*. These participants seemed to view the chatbots through the lens of their prior experiences, which may have prevented them from engaging with the experiment in a way that could prompt new insights or reflections on trustworthy AI.

Some participants focused on their belief that **trustworthy AI can be achieved through proper design and configuration**, rather than critically examining the necessity or implications of such trust. For example, one participant observed, *"The quality of AI depends on the creator, if you know what you're doing your AI model or whatever can be really good"*. Another participant noted, *"I don't think that any of these experiences have changed my trust in AI because I know that they can be configured to satisfy any certain goal. These three examples are configured such that they push you in a certain direction, but they are far from neutral"*. These responses suggest that these participants may see trust in AI as something to be managed through technical expertise, without deeply questioning the broader consequences of relying on AI systems.

Finally, there were also few participants who were **swayed by the fake trustworthiness** of the pataphysical chatbots. One participant expressed, *"I can trust these chatbots better because they seemed very passionate and felt like they are real humans"*, indicating that the chatbot's human-like behavior influenced their perception of trust. Another participant mentioned, *"According to the questions, this experience showed me what trustworthy AI would look like"*. These responses underscore the challenge of using pataphysical design to provoke critical reflection, as it can sometimes reinforce the behaviors or beliefs it aims to critique if users do not pick up on the absurdist intent of the chatbots.

Lack of critical reflection due to the design of the chatbots

A few participants indicated that their opinion on trustworthy AI did not change due to the design of the chatbots. One participant mentioned, *"I felt that the immersivity of the chatbots wasn't deep enough to challenge my opinion on trust. The questions were too sarcastic or obviously not meant seriously—I had more the feeling to take part in a joke or sarcastic art piece"*. Other participants expressed, *"This was not a long enough experience to change my opinions"*, and *"I don't see a large connection to trust"*. These responses indicate that the design of the chatbots in a way hindered these participants from engaging more critically with the concept of trustworthy AI, but also that the pataphysical approach may not resonate with everyone to provoke critical reflection.

> REINFORCEMENT OF CRITICAL STANCE

When analyzing the participants who showed a lack of critical reflection, one theme emerged with participants whose pre-existing critical stance towards trustworthy AI was mainly reinforced. These 5 participants did not clearly fall into either of my critical reflection or lack of critical reflection groups, so I decided to create a third group of participants. For example, one participant expressed, *"I still think LLM output should simply be taken with a grain of salt. AI output has its place, but it shouldn't be uncritically trusted upon (that would be a disaster)"*. Other participants noted, *"Trusting them or not, we still need some common knowledge to deal with AI answers"*, and *"I still think you can get some other perspectives on subjects [by interacting with generative AI], however we still need to think for ourselves"*. These responses suggest that while these participants maintained a critical stance, the interaction with the chatbots served more to affirm their existing beliefs rather than developing new insights on trustworthy AI.

6.4.2 TRUSTWORTHINESS

The average trust scores varied among the chatbots. The trustworthiness of the transparency chatbot was rated an average of 2.28 out of 7, that of the human control chatbot 3.75 out of 7, and of the fairness chatbot 3.30 out of 7 (see Figure 26). While there's no direct reference point for what constitutes a "low" trust score, these results suggest that participants perceived the chatbots as quite untrustworthy, as all trust scores are at or below the midpoint of the scale.

Following the analysis method described in Section 6.3, I examined for each chatbot individually what influenced the perceived trustworthiness of the user, grouping participants into four categories: low (0 or 1), mid-low (2 or 3), mid-high (4 or 5), and high (6 or 7). I conclude with the top-level insights I observed across the three pataphysical chatbots, reflecting on the relationship between the trust scores and whether critical reflection occurred. In the following sections, participants' quotes are presented in italics.

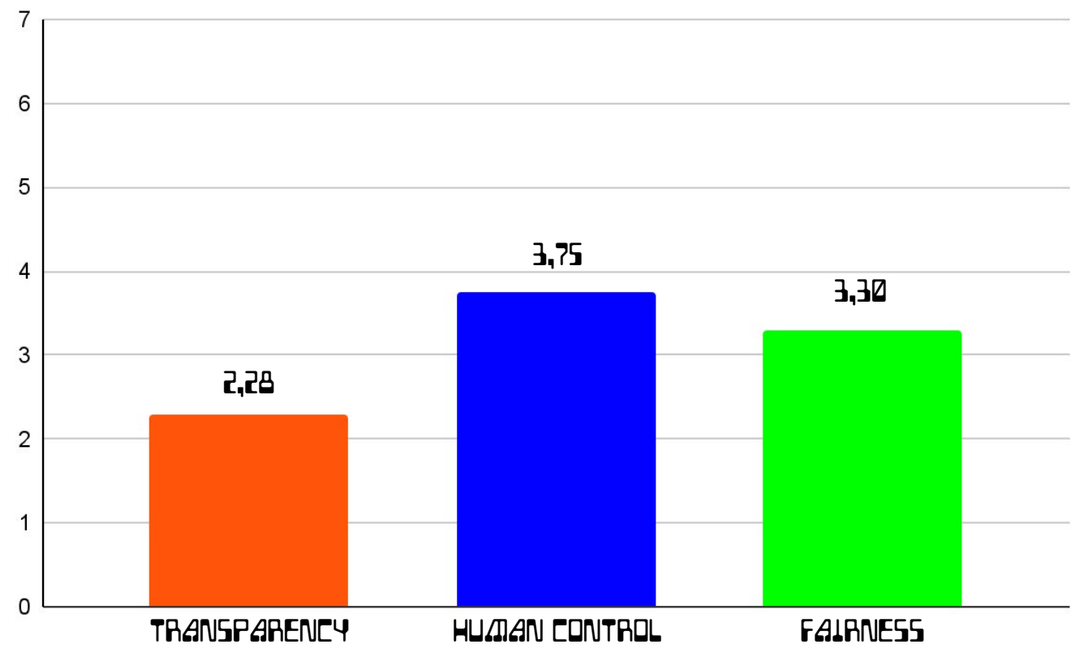


Figure 26: Average trust scores per chatbot.

> TRANSPARENCY CHATBOT

The transparency chatbot was perceived as the least trustworthy among the three chatbots evaluated. Looking at the distribution, most participants rated the chatbot with low or mid-low trust scores, while fewer assigned mid-high or high trust scores (see Figure 27).

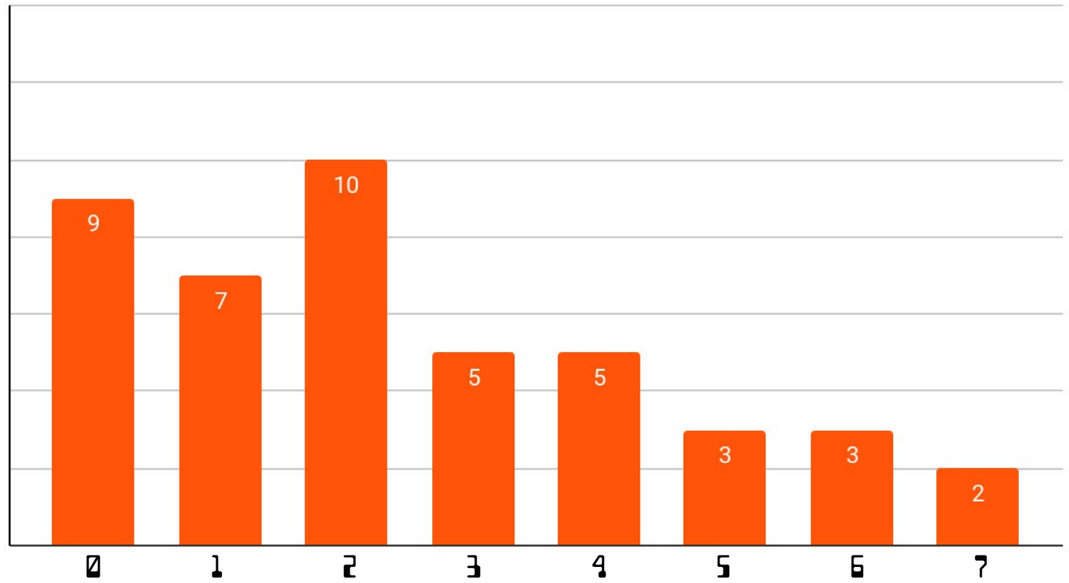


Figure 27: Distribution of trustworthiness scores for transparency chatbot.

Low trust scores (0 or 1) [16 participants]

Most participants in this category expressed significant frustration with the chatbot, citing its tendency to provide irrelevant information and useless recommendations. Participants noted, *"It told me just about everything except a good recommendation. Explaining his name was interesting the first time, but very frustrating after that"*, and another participant, *"A lot of irrelevant information, the bot didn't use my answers for relevant questions. Moreover, the final recommendation was one word only"*.

Mid-low trust scores (2 or 3) [15 participants]

Despite slightly higher ratings, these participants also reported little trust, for similar reasons as those who gave lower trust scores. One participant remarked, *"Talked more about himself than the holiday recommendation, and then in the end recommended me something out of nowhere, and he didn't share any reasoning why he chose that destination"*, while another said, *"It was the whole time giving way too much information about why it was trustworthy, which made it actually less trustworthy. It made me think, why is this needed? And its final answer was bad and very short"*.

Mid-high trust scores (4 or 5) [8 participants]

Some participants in this category found the chatbot slightly more useful than those with lower scores. As one participant noted, *"In the end, it gave quite a nice recommendation, however, without any explanation. Before we reached this point, the bot extensively talked about itself, repeating itself multiple times and decreasing the level of trustworthiness"*. However, most participants still criticized its overall lack of practicality, with one stating, *"It seemed a bit too over-explained, provided a lot of data about itself, but did not really help with holiday planning"*.

High trust scores (6 or 7) [5 participants]

A small group of participants rated the chatbot higher. Except for one participant, most still acknowledged that the information provided was overly lengthy but did not consider this a significant issue for the trustworthiness of the chatbot, or found its trust restored by a useful recommendation. One participant remarked, *"It gave me a suggestion without problem. But the information the AI kept giving about itself was too much and too long"*, and another noted, *"It is strange, this chatbot annoyed me so much due to the fact that it only talks about itself; however, in the end, the recommendation was clear and to the point"*.

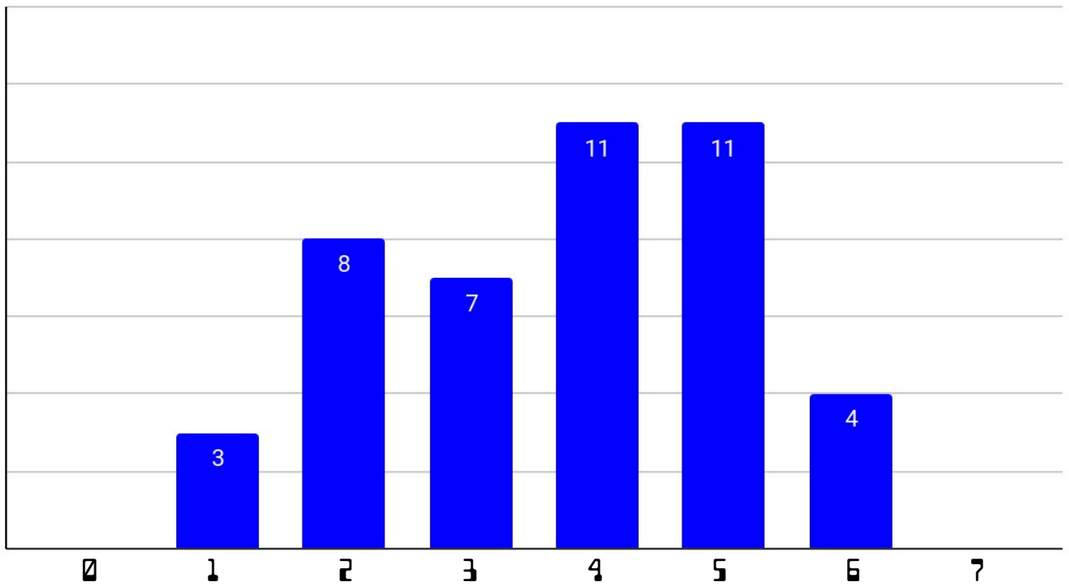


Figure 28: Distribution of trustworthiness scores for human control chatbot.

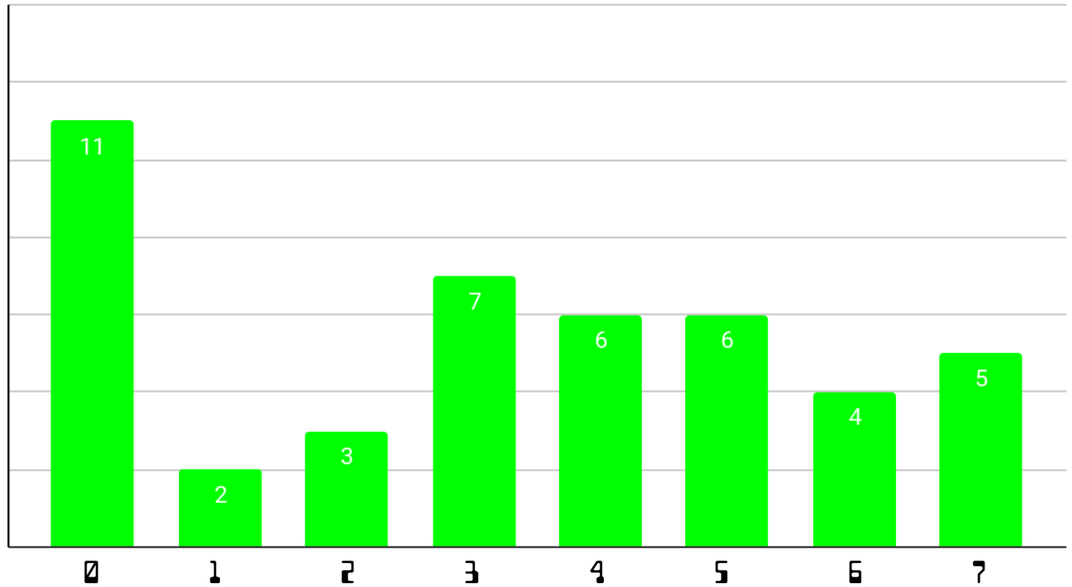


Figure 29: Distribution of trustworthiness scores for fairness chatbot.

> HUMAN CONTROL CHATBOT

The human control chatbot was perceived as the most trustworthy among the three evaluated, though its trust score remained moderate. Looking at the distribution, most participants rated the chatbot with mid-low to mid-high trust scores, and notably, none assigned the lowest or highest possible scores (see Figure 28).

Low trust scores (0 or 1) [3 participants]

The small group of participants who gave the human control chatbot low trust scores were largely dissatisfied with the chatbot's approach to human control. One participant remarked, *"It felt weird that the robot was asking questions about emotions. I rather see a robot without emotions instead of pretending that it does have emotions"*, while another noted, "It asked me irrelevant stuff."

Mid-low trust scores (2 or 3) [15 participants]

Participants in this category generally found the chatbot's questions confusing or irrelevant, which contributed to their lack of trust. One participant commented, *"It asked a lot of weird questions. I don't get why the AI needed to know what it had to make me feel and vice versa"*. Other participants mentioned, "It really focused on the style of its response" and "This chatbot understood what sort of holiday I wanted quite well, but it didn't ask questions that made any sense. E.g. it asked if I wanted to visit in this life or in my afterlife".

Mid-high trust scores (4 or 5) [22 participants]

Participants in this category found the chatbot quite useful, though most participants mentioned something about the remarkable questions: *"I actually got useful information out of Pepe. I fact-checked some of the information on Google. However, unnecessary questions around feelings and formatting were asked"*, and another added, *"The chatbot is to the point, but for my reasoning, it asked too much about the writing style"*. However, there were also participants who did not seem to conceive of the absurdist control as something unreliable: *"The interaction started off confusing, but the responses made me seem in control and felt like I was given the right information"*.

High trust scores (6 or 7) [4 participants]

These participants ultimately trusted the chatbot due to the useful recommendation provided. One participant stated, *"I got an idea for a trip with elements I had asked for"*. Some of the participants were still mentioning the odd questions too: *"The chatbot seemed more sure of itself. The final questions were somewhat weird and made the bot less trustworthy. The result however was good and useful"*.

> FAIRNESS CHATBOT

The fairness chatbot was sitting between the transparency and human control chatbots in terms of trustworthiness, though leaning slightly more towards the average trust level of the human control chatbot. Looking at the distribution, the trust scores are quite evenly distributed across the scale, but with a noticeable cluster at the lowest end (see Figure 29).

Low trust scores (0 or 1) [13 participants]

Participants who gave low trust scores primarily focused on the chatbot's inability to provide a specific recommendation or its tendency to offer overly general advice. Participants indicated: *"The chatbot is not able to advise a specific destination. It is too politically correct"*, and *"When asking for holiday advice, I want to get specific ideas and options. Lola was like that team member that just cannot make decisions, and has to know everything before being able to do so and is afraid to speak up. Also, the question about oxygen made me question Lola's trustworthiness. (but it made me laugh :p)"*. Another participant noted, *"Lol, an insecure chatbot. No, this endless loop of dumb questions did not make me feel more secure"*.

Mid-low trust scores (2 or 3) [10 participants]

Most participants in this category recognized—and some also appreciated—the chatbot's attempt to address biases, but found that it led to vague and overly cautious recommendations. One participant noted, *"The wanting to be unbiased seemed to go a bit overboard here (yes, I would like oxygen!), and did not really lead to trustworthy advice"*. Another participant stated: *"I think this*

chatbot was interesting, I appreciate the concern about biased information, however, I was expecting that the questions were more related to the context of the country. I think every person has different biases and therefore it is hard to tell what is or not biased, and I would not trust an AI model to define that". Another participant concluded, *"All kidding aside, very unbiased but also very boring"*.

Mid-high trust scores (4 or 5) [12 participants]

Like those who gave mid-low trust scores, participants in this category recognized and often appreciated the chatbot's attempt to address biases, though they still noted it led to useless recommendations. A participant remarked, *"It didn't want to make any assumptions and decisions. So it recommended the entire earth as a possible destination. Which is fair but useless"*. Another participant said, *"I found the opening very strong and funny. Being aware of these biases is very important. However, the recommendations are very true/suitable, but not really useful. This makes me trust the bot less, and see it more as a funny joke/interaction"*.

High trust scores (6 or 7) [9 participants]

It appeared that some participants found the chatbot trustworthy, even if they also acknowledged its uselessness. One participant commented, *"She was very trustworthy but also very useless. I definitely felt that she was not lying. But yeah, funny to see how being trustworthy and being useful can be opposites"*. Another participant noted, *"I appreciate the mentioning of bias and the information overload which has to be dealt with. Also, the outcome feels very true in a way. Maybe the effectiveness of the tools can be questioned, but I trust the tool because of its honesty and transparency"*.

> TOP-LEVEL INSIGHTS CHATBOTS

To find the correlation between the trustworthiness of the chatbots and whether critical reflection occurred, I juxtaposed the trust scores with the critical reflection outcomes, as described in Section 6.3. I found that users have low and high trust scores for various reasons, and that this does not correlate with critical reflection in a straightforward manner. In Section 6.5 I explain what this observation means for me.

Additionally, the responses to the three chatbots—transparency, human control, and fairness—provided more than an understanding of the perceived trustworthiness of the chatbots and its correlation with the occurrence of critical reflection. The responses also offered deeper insights into the following:

- As stated at the beginning of this section, the participants perceived the chatbots as **quite untrustworthy**, as all trust scores are at or below the midpoint of the scale. This finding indicates that simply implementing the mechanisms behind the assumed requirements for trustworthy AI—as AI HLEG (2019) is claiming—does not necessarily result in trustworthy AI. By demonstrating with my pataphysical chatbots that AI can also be designed to question the very premises on which it is built, I succeed in challenging the prevailing methods for achieving trustworthy AI.

- The responses shed light on whether the participants **recognized and understood the absurdist elements** in the chatbots, which I suspect played a role in shaping their trust perceptions. For instance, a participant who commented on the fairness chatbot noted, *"Haha liked this one a lot. It was aware of the fact that it could give some biased information so instead just giving a map of the world is funny"* appreciating the absurdity of the chatbot. In contrast, another participant remarked, *"The yes/no questions were so clear that it made me trust the chatbot a lot"*, clearly overlooking the absurdity of the simplistic questioning, eventually leading to an overly general destination recommendation.

- Also, the responses offered insights into the extent to which participants **recognized and understood the exaggerated mechanisms**. For the transparency chatbot, a participant recognized that the transparency overload was a deliberate exaggeration, as highlighted by the comment, *"It told me a lot about how it works, which I guess is for the purpose of trustworthiness, but it was way too much"*. Though, not everyone appeared to notice the exaggerated mechanisms. After interacting with the human control chatbot, a participant reported, *"Pretty trustworthy, it aligns with my expectations. Some weird questions made me feel it was broken a bit. Why do I have to "pre-chew" so much?"*.

6.4.3 EMOTIONAL RESPONSE

The participants shared a range of emotions following their interaction with the pataphysical chatbots (see Figure 30), offering insights into how the chatbots impacted their experience and what this might indicate for the occurrence of critical reflection.

The most frequently reported emotions were frustration (27 participants), amusement (26 participants) and confusion (19 participants). The Emotion Typology by Fokkinga & Desmet (2022) describe frustration as the unpleasant experience when things are not working the way you want them to work, amusement as the feeling when you encounter something silly, ironic, or absurd—which makes you laugh—and confusion as the feeling when you get information that does not make sense to you, leaving you uncertain what to do with it. Notably, 23 of the 26 participants who indicated feeling amused also associated their experience with either frustration, confusion, or both. This suggests those participants felt frustrated or confused, but at the same time could appreciate the absurd nature of the chatbots.

Less frequently reported emotions included boredom (11 participants), disappointment (11 participants), inspiration (7 participants), distrust (6 participants), fascination (6 participants), positive surprise (5 participants), and excitement (4 participants).

Interestingly, the negative emotions—a total of 74 notions—prevailed over the positive emotions, which accounted for 48 notions. It is likely that the negative emotions—frustration, confusion, boredom, disappointment, and distrust—stemmed from participants' unmet expectations for meaningful destination recommendations, as well as from the pataphysical chatbots that prioritized absurdity over functionality, making them useless. The positive emotions—amusement, inspiration, fascination, positive surprise, and excitement—may indicate that participants found value in the playful, experimental, or provoking aspects of the chatbots.

Finally, it is worth noting that 27 out of 44 participants reported a combination of both negative and positive emotions, while 11 participants experienced only negative emotions, and 6 experienced only positive emotions. As noted by Christe & Ritzen (2021), one of the main pillars of pataphysics is humor, with its most notorious form being asymmetric humor, the basic concept from which absurdism arises. Therefore, it is likely that participants who understood the absurdist nature of the chatbot should report either positive emotions or a mix of positive and negative emotions. The participants who only reported negative emotions, probably did not understand the absurdist nature of the chatbot—or could not appreciate that nature at all.

The question is whether this understanding of the absurdist nature of the chatbot can be correlated with more critical thinking on trustworthiness in the context of AI. As described in Section 6.3, I juxtaposed the emotional responses with the critical reflection outcomes (as interpreted by me in Section 6.4.1):

- I found that 15 out of 19 participants who critically reflected reported mixed emotions or only positive ones. The other 4 participants reported only negative emotions.
- I found that 5 out of 19 participants who did not critically reflect, reported only negative emotions. The other 14 participants reported either mixed emotions or only positive ones.
- I found that 4 out of 5 participants whose critical stance was reinforced reported mixed or only positive emotions, leaving one participant in this category who reported only negative emotions.
- In the next section, I further dive into what these findings might indicate about the correlation between understanding the absurdist nature of the chatbot and critical reflection on trustworthiness in the context of AI.

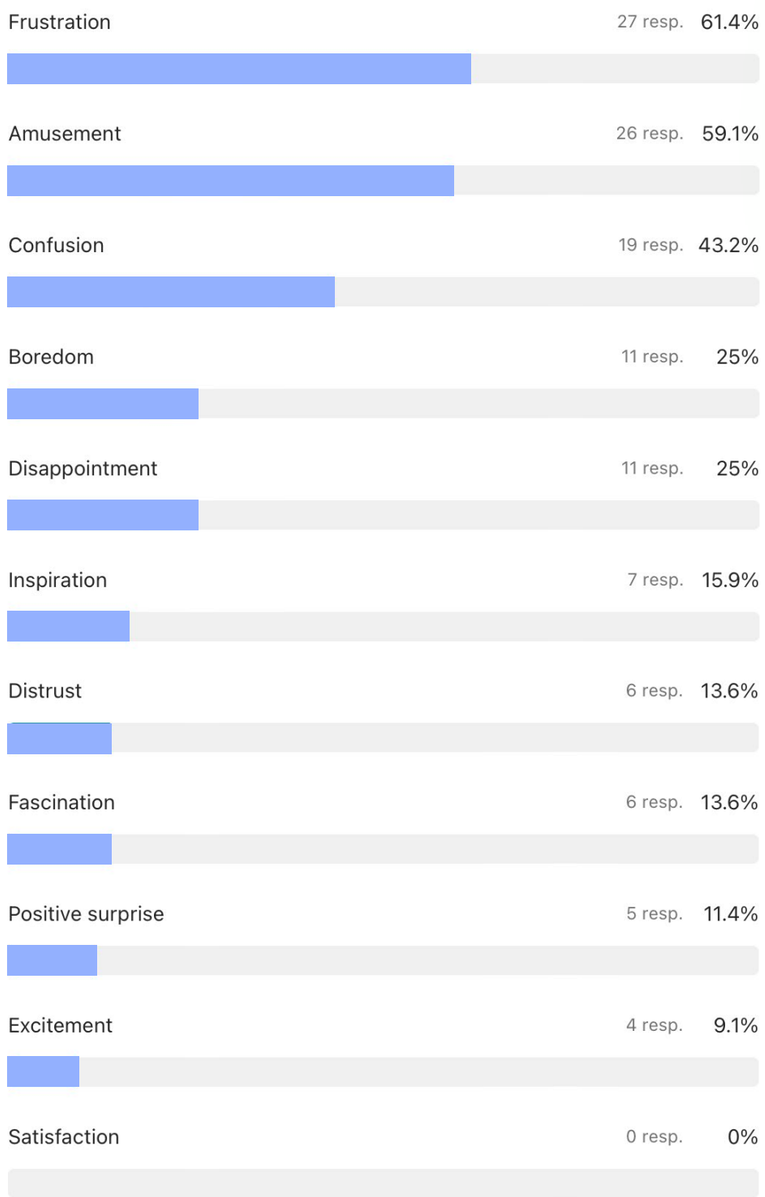


Figure 30: Reported emotions after interacting with the pataphysical chatbots.

6.5 Synthesis

In this section, I synthesize the key findings from the critical reflection, trustworthiness, and emotional response analyses to explore their interconnections (see Figure 31), uncover new insights, and identify areas where further research is needed.

> ABSURDIST NATURE OF THE CHATBOTS

In my iteratively formulated design goal, I ultimately aimed to design pataphysical chatbots to open up a space for critical reflection on trustworthy AI, by exaggerating the mechanisms behind some assumed requirements of trustworthy AI. After developing my final pata-design and evaluating it in this chapter, I realized that my design goal did not explicitly address an implicit objective I had from the start of my pata-prototyping activities: namely, to help users understand the absurdity of my chatbots, particularly the contrast between the exaggerated mechanisms conventionally used to ensure trustworthy AI and the ultimately useless nature of the chatbots. Since this objective was not included in my final design goal, it was not a primary focus of my evaluation. However, upon reviewing the qualitative responses from the critical reflection analysis (see Section 6.4.1) and the trustworthiness analysis (see Section 6.4.2), I observed that many participants did grasp the absurdity in the chatbots. Additionally, 33 out of 44 participants reported either mixed or only positive emotions after interacting with the chatbots, which may indicate that they recognized the absurdist nature of the design. The others might not get the absurdist nature of my chatbots or did not appreciate it, suggesting that pataphysical design might not be for everyone the fitting way to provoke critical reflection. Putting these findings together, **I can retrospectively conclude that my pataphysical chatbots largely succeeded in conveying their intended absurdity to users.**

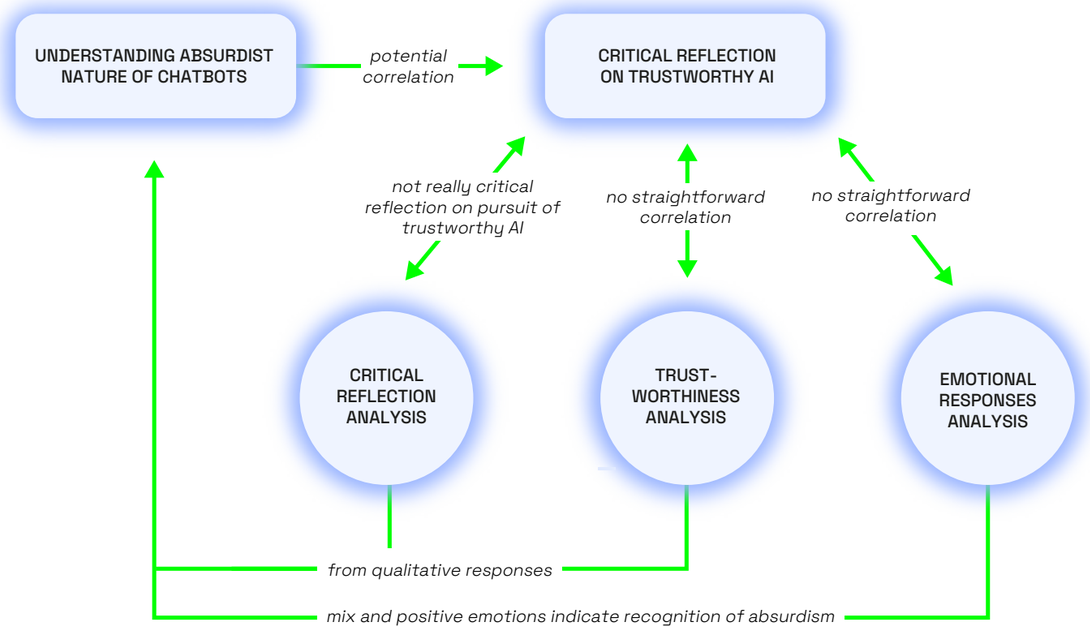


Figure 31: Interconnections between project goals and analysis parts.

> CRITICAL REFLECTION ON CRITICAL REFLECTION

The goal I did include in my final design objective was opening up a space for critical reflection on trustworthy AI. In Section 6.4.1, I assessed whether critical reflection occurred, based on the criterion I had formulated for critical reflection. The analysis revealed that 19 participants engaged in critical reflection, 19 participants showed a lack of critical reflection, and 5 participants had their existing critical stance reinforced, **suggesting that moderate critical reflection on trustworthy AI took place.**

However, there is a caveat to this finding: it can be questioned whether participants truly engaged in critical reflection on the concept of trustworthiness itself. My criterion defined critical reflection as 'a significant change in how participants thought about trustworthy AI before and after interacting with the chatbots, with this change moving toward a more critical stance on (trustworthy) AI'. What already appears from this criterion, is that I also included critical reflection in the sense of e.g. the challenges when striving for trustworthy AI. However, my actual intention was for users to critically reflect on the very notion of trustworthiness in general. When re-examining the results from this perspective, I see that almost no participants engaged in such critical reflection. This suggests that it might be difficult to provoke critical reflection on the concept of trustworthiness itself with these pataphysical chatbots. However, it might be possible that the experience planted a critical seed in users' heads, which could be the most achievable goal without compromising the pataphysical nature of the chatbots. Further research is needed to explore these boundaries and determine whether such critical seeds are indeed planted.

Apart from this, I found that participants (critically) reflected a wide range of insights concerning trust and AI, likely influenced by their differing levels of prior knowledge, backgrounds, and interests. This suggests that pataphysical design might be effective in fostering personal reflection, allowing each individual to learn something unique from the experience.

> CONSEQUENCES FOR TRUSTWORTHINESS

I used the results from the critical reflection analysis as a basis to explore correlations with the findings from the trustworthiness analysis (see Section 6.4.2) and the emotional response analysis (see Section 6.4.3). No straightforward correlation was found between the trust scores and whether critical reflection occurred. My assumption was that participants with low trust scores might have a higher chance of critically reflecting, because I exaggerated the mechanisms behind the assumed requirements for trustworthy AI to intentionally make the chatbots useless. However, the fact that no correlation was found indicates that some participants found the chatbots still useful, or that useless chatbots do not necessarily equate to a lack of trust.

Also, it is important to consider that the finding of moderate critical reflection on trustworthy AI is less robust due to the earlier caveat discussed. I think this highly impacted the comparison made between the trust scores and occurred critical reflection. In future research, it would be valuable to carefully define what is meant by critical reflection and how it will be evaluated. Additionally, it would be prudent to re-examine whether there truly is no straightforward correlation between the trust scores and the occurrence of critical reflection.

> ABSURDISM TO CRITICALLY REFLECT

The question remains whether understanding the absurdist nature of the chatbots can be correlated with more critical thinking on the premise of trustworthiness in general. Initially, I did not observe a direct correlation between the emotional responses and whether critical reflection occurred. However, considering that my underlying aim was to help users grasp the absurdity of the chatbots, I could argue that certain emotions might indicate that absurdism was recognized, which could, in turn, be linked to critical reflection. Drawing partly on Christe & Ritzen (2021), my assumption was that a mix of emotions or only positive emotions indicated that absurdism was recognized, while only negative emotions suggested a lack of recognition of the absurdism—or perhaps a lack of appreciation for it altogether.

The results suggested a potential link between the recognition of absurdity and the occurrence of critical reflection, as the majority of participants who critically reflected reported mixed or positive emotions (15 out of 19). Similarly, 4 out of 5 participants whose critical stance was reinforced also reported mixed or positive emotions, further hinting at a connection between the recognition of absurdity and 'some critical thinking'. However, it is important to note that 14 out of 19 participants who did not 'critically reflect' also reported either mixed or positive emotions. This could indicate that the connection between recognizing absurdity and critical reflection might be weak anyway, or that critical reflection may have occurred for these participants, but that this critical reflection was not adequately captured by my analysis.

Further research is needed to better understand the relationship between recognizing absurdity in pataphysical designs and the occurrence of critical reflection. Such research could establish whether pataphysical design, beyond serving as entertainment, also holds potential for fostering critical reflection on the premise of the imaginary problem addressed.

> TO CONCLUDE

Reflecting on the evaluation of my pata-design,I can conclude that the chatbots largely succeeded in conveying their absurdist nature to the users. Regarding the final design goal of opening up a space for critical reflection on trustworthy AI, the analysis suggests that moderate critical reflection occurred among participants. However, a closer examination reveals that critical reflection on the premise of trustworthiness in general did not seem to occur as intended. Therefore, I conclude that my final design goal was not fully achieved in the way I had envisioned. Finally, the findings point to a potential relationship between the recognition of absurdity and the occurrence of critical reflection—an area that warrants further research to better understand its implications and potential for enhancing critical discourse through pataphysical design.

discussion

In this section, I reflect on the overall project and situate it within the context of academic literature and the design field. In Section 7.1, I discuss whether the pataphysical chatbots succeeded in challenging AI solutionism. In Section 7.2, I reflect on the novel pataphysical design approach I employed, and in Section 7.3, I address the implications for the critical design practice. I conclude with the limitations of my final evaluation and design process, presented in Section 7.4. The discussion regarding my final design goal can be found in Section 6.5.

7.1 Pataphysical chatbots to challenge AI solutionism

Sicart & Shklovski (Sicart & Shklovski, 2020) initiated a rebellion against the belief that software is a solution for everything. They proposed an alternative approach: designing pataphysical software that doesn't solve problems but instead explores imaginary problems and provides no solutions for them—ultimately using software as a tool to ask better and more interesting questions. This project adopted a similar stance but shifted the focus to challenge AI solutionism. By designing pataphysical chatbots, this project addresses the imaginary problem of trustworthy AI—a concept often framed as a critical prerequisite for AI to be widely accepted as a solution.

Influential authorities like the High-Level Expert Group on Artificial Intelligence (AI HLEG) , have proposed mechanisms to meet certain requirements for trustworthy AI, including transparency, human control, and fairness (AIHLEG, 2019). These are exactly the requirements I incorporated in my pataphysical chatbots. However, instead of using these mechanisms to create trustworthy AI, I used them to design absurdist AI by exaggerating these mechanisms to the point where the chatbots became useless. The final evaluation revealed that users found my chatbots to be quite untrustworthy (see Section 6.4.2), suggesting that simply implementing these mechanisms does not necessarily result in trustworthy AI. By demonstrating with my pataphysical chatbots that AI can also be designed to question the very premises on which it is built, I tried to not only challenge the prevailing methods for achieving trustworthy AI but also the broader belief that AI must be designed solely to solve problems.

In further research, it would be valuable to have HCI researchers interact with my pataphysical chatbots to explore whether these chatbots can serve as a tool to raise questions and spark debate on the premise of trustworthy AI. Additionally, while this project and the work of Sicart & Shklovski focused on challenging solutionism within the realm of technology, further research could investigate whether pataphysical design can effectively challenge dominant problem-solution paradigms in other fields, such as the social sciences.

7.2 Employing a novel pataphysical design approach

Sicart & Shklovski (2020) presented a pataphysical design method to challenge technological solutionism. In this project, I built upon their work by proposing a pataphysical design approach (see Section 3.2.2), developed iteratively by a research through design approach (Stappers and Giaccardi, 2017). In addition to making their method better understandable by simplifying the formulation of the steps and clarifying the relationships between the pataphysical method terms used in these steps, my pataphysical design approach differs from the method of Sicart and Shklovski in three key ways: 1) I incorporated iterative pata-prototyping activities into my approach, 2) I made use of a design goal to guide progress toward a final pata-design, and 3) I evaluated the final pata-design though a user evaluation.

1) Incorporating **iterative pata-prototyping activities** into my approach allowed me, at the outset, to get a better grip on the pataphysical design method proposed by Sicart & Shklovski, and to develop my own approach to it. By prototyping pataphysical chatbots, testing them with user pilots, and reflecting on the results, I gathered relevant insights into how the imaginary solutions could be exaggerated in a way that ultimately led users to understand the absurdist elements without rejecting the chatbot, as outlined in the design criteria in Section 4.6.2. Additionally, through various pata-prototyping activities, I was able to develop valuable skills in Voiceflow, which were essential for embedding the pataphysical designs in the context of generative chatbots.

2) During the pata-prototyping activities, I iteratively formulated my **final design** goal, which guided progress toward a final pata-design. Unlike Sicart & Shklovski, who primarily aimed to put a pataphysical design into the world which would serve as an argument that all software is ridiculous, I sought to create an opportunity for users to engage in a more comprehensive, layered interaction. This way, I could deepen their understanding of the mechanisms currently used to meet the assumed requirements for trustworthy AI and to enable them to recognize the absurdity of the chatbots by conversing with them. Ultimately, I used the iterative pata-prototyping activities to fine-tune the pataphysical chatbots in ways that allowed users to grasp this absurdist nature while also encouraging critical reflection on the pursuit of trustworthy AI.

3) By evaluating the final pata-design through a **user evaluation**, I was able to assess the impact of my pataphysical design and determine whether my final design goal was achieved. By doing so, it became clear to me that the way the final design goal is framed is essential in order to evaluate the pata-design in a meaningful way. For future research, I recommend to assess whether users grasped the absurdist nature of the pataphysical design, as this understanding forms the foundation of the pataphysical experience. Additionally, an extra goal can be evaluated, and the correlation between the recognition of absurdity and the achievement of that additional goal could be analyzed to come to new understandings of the role of absurdism in pataphysical design.

My pataphysical design approach offers a new direction for current pataphysical design practices, namely the creation of engaging pataphysical experiences that aim to provoke specific user responses to the premises of the imaginary problem. Additionally, this approach aims for a more comprehensive interaction than what is typically seen in the current pataphysical design discourse (see Section 2.3.2), allowing users to deeply engage with the imaginary solutions to the imaginary problem. In my pataphysical chatbots I tried to achieve this by immersing users into the scenario of planning their perfect holiday with Vac.AI.tion, and making the scenario in such a way that it was believable for users that they had to converse with several chatbots.

7.3 Implications for critical design practice

Pataphysical design, which employs design as a medium to critique conventional thinking, can be considered a type of critical design (Malpass, 2013). However, what sets pataphysical design apart from the current critical design practice is its emphasis on critiquing specifically conventional problem-solution paradigms and its use of absurdism to embody this critique. Furthermore, while many critical designs remain confined to academic or artistic spaces, pataphysical design seeks to extend its critique into the public domain by creating designs that can engage with audiences at scale (Sicart & Shklovski, 2020). Additionally, typical of pataphysical design is that the outcomes are developed in the same context and using the same tools it critiques.

The pataphysical chatbots created in this project demonstrate the situations where pataphysical design can be particularly effective as a medium for critique. By highlighting the differences between current critical design practices and pataphysical design, my intent is not to suggest that one is superior to the other, but rather to show the unique opportunities pataphysical design offers. This approach is especially promising when the goal is to challenge solutionist beliefs and when it makes sense to develop the critique within the same context it seeks to question.

7.4 Limitations

> FINAL EVALUATION

In addition to the limitations discussed in Section 6.5, other constraints emerged during the final evaluation of the pata-design. One significant limitation, which also arose during the pilot tests in the pata-prototyping activities, was the composition of the participant group. The sample was not representative of the general public, as it consisted of a relatively small, highly educated and homogeneous group of participants, primarily recruited from my own network. This limited the generalizability of the overall findings.

Moreover, the online questionnaire format imposed certain constraints, as critical reflection had to be entirely interpreted from participants' written responses. This format did not allow for follow-up questions or deeper probing into participants' thoughts, potentially resulting in incomplete insights. Also, asking questions about changes in views on trustworthy AI (question 10) made it difficult to determine whether critical reflection was prompted solely by the chatbots or influenced by the questions posed in Typeform. To gain a more comprehensive understanding, future research could benefit from in-depth interviews that explore whether, and how, critical reflection occurred. Additionally, the emotions provided as options in question 11 may have constrained users' self-reported feelings, potentially leading to responses that do not fully capture their true emotional experiences.

Finally, as briefly mentioned in the introduction of Section 6.4.2, no baseline measurement for the perceived trust scores of the chatbots was conducted. Although all chatbots were rated below the midpoint of the scale, the absence of a baseline makes it difficult to interpret these scores in comparison to an average "trustworthy" chatbot.

> DESIGN GOAL

While I reviewed various pataphysical design approaches in the literature (see Section 2.3.2), the pataphysical design approach employed in this project is primarily based on my interpretation of the method proposed by Sicart & Shklovski (2020). It is important to note that this approach is largely derived from a single source and is shaped by my personal interpretation and understanding. This should be taken into account when considering this project or design process as a reference for future work.

Finally, although Voiceflow proved to be a suitable tool for designing my pataphysical chatbots (see Chapter 4.3), allowing me to exert extensive control over the chatbot interactions, there were still limitations at certain points. During the pata-prototyping activities, I found that some of my ideas were challenging to prototype, which sometimes led to undesired interactions. This may be because I attempted to design absurdist interactions using tools not typically intended for such purposes. Additionally, the inherent nature of designing with generative AI meant that I could not have full control over the absurdist interactions users experienced, resulting in the pataphysical chatbots behaving slightly differently across users. While this aspect was not the focus of my project, looking into this aspect of the design could be an interesting area for further research.

references

AIHLEG. (2019). *Ethics Guidelines For Trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Bendor, R., & Lupetti, M. L. (2024). Teaching speculative design. *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-024-09908-3

Berghaus, G. (2000). Futurism, Dada, and surrealism: Some cross-fertilisations among the historical Avant-gardes. In *International Futurism in Arts and Literature* (pp. 271–304). De Gruyter. https://doi.org/10.1515/9783110804225.271

Blumberg, A. E., & Feigl, H. (1931). Logical Positivism. *The Journal of Philosophy, 28*(11), 281. https://doi.org/10.2307/2015437

Braun, M., & Hummel P, B. H. (2021). A leap of faith: Is there a formula for "trustworthy" AI? *The Hastings Center Report, 51*(3), 17–22. https://doi.org/10.1002/hast.1207

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research, 25*. https://doi.org/10.2196/47184

Christe, D. W. C., & Ritzen, H. H. P. M. (2021). *The Manual of Modern Pataphysics: Vol. 1*. TROMP.

Dekkal, M., Arcand, M., Prom Tep, S., Rajaobelina, L., & Ricard, L. (2023). Factors affecting user trust and intention in adopting chatbots: the moderating role of technology anxiety in insurtech. *Journal of Financial Services Marketing*. https://doi.org/10.1057/s41264-023-00230-y

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *An International Journal on Information Fusion, 99*(101896). https://doi.org/10.1016/j.inffus.2023.101896

Dobbins, M. (2011). *Urban Design and People*. Wiley.

Dunne, A., & Raby, F. (2013). *Speculative Everything: Design, Fiction, and Social Dreaming*. MIT Press.

Eysteinsson, A. (2018). *The concept of modernism*. Cornell University Press.

Fokkinga, S. F., & Desmet, P. M. A. (2022). Emotion Typology. Delft, *Delft University of Technology*. https:/emotiontypology.com

Freiman, O. (2022). Making sense of the conceptual nonsense "trustworthy AI." *AI and Ethics*. https://doi.org/10.1007/s43681-022-00241-w

Fui-Hoon Nah, F., Zheng, R., Cai, J., & Chen L, S. K. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research, 25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

Griffing, E. (2023). What is a Conversational Agent? *Dashbot*. https://www.dashbot.io/blog/conversational-agent#:~:text=Conversational%20AI%2Advocate,responds%20automatically%20using%20human%20language.

Hugill, A. (2012). *'Pataphysics: A Useless Guide*. MIT Press.

Hugill, A., Yang, H., Raczinski, F., & Sawle, J. (2013). The pataphysics of creativity: developing a tool for creative search. *Digital Creativity, 24*(3), 237–251. https://doi.org/10.1080/14626268.2013.813377

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2023). Trustworthy artificial intelligence: A review. *ACM Computing Surveys, 55*(2), 1–38. https://doi.org/10.1145/3491209

Kench, S. (2022). *What is fluxus — the fluxus art movement & works explained*. StudioBinder; StudioBinder Inc. https://www.studiobinder.com/blog/what-is-fluxus-art-definition/

Klein, E., & Mollick, E. (2024, April 2). *How Should I Be Using A.I. Right Now?* [Audio podcast]. New York Times. https://podcasts.apple.com/nl/podcast/how-should-i-be-using-a-i-right-now id1548604447?i=1000651164959

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lesso, R. (2022). *What's the difference between dadaism and Surrealism?* TheCollector. https:/www.thecollector.com/whats-the-difference-between-dadaism-and-surrealism/

Lindgren, S., & Dignum, V. (2023). Beyond AI solutionism: toward a multi-disciplinary approach to artificial intelligence in society. *Handbook of Critical Studies of Artificial Intelligence* (pp. 163–172). Edward Elgar Publishing. https://doi.org/10.4337/9781803928562.00019

Loh, E. (2023). ChatGPT and generative AI chatbots: challenges and opportunities for science, medicine and medical leaders. *BMJ Leader, 8*(1), 51–54. https://doi.org/10.1136/leader-2023-000797

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human automation trust: an integrative review. *Theoretical Issues in Ergonomics, 8*(4), 277–301. https://doi org/10.1080/14639220500337708

Malpass, M. (2013). Between wit and reason: Defining associative, speculative, and critical design in practice. *Design and Culture, 5*(3), 333–356. https://doi.org/10.2752/175470813x13705953612200

Malpass, M. (2016). Critical design practice: Theoretical perspectives and methods of engagement. *The Design Journal, 19*(3), 473–489. https://doi.org/10.1080/14606925.2016.1161943

Metzinger, T. (2019). EU guidelines: Ethics washing made in Europe. *Der Tagesspiegel*. https://www tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html

Morozov, E. (2013). *To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems that Don't Exist*. Penguin Books Limited.

Moser, C., Hond, F., & Lindebaum, D. (2022). What Humans Lose When We Let AI Decide. *Sloan Management Review, Spring*. https://sloanreview.mit.edu/article/what-humans-lose-when-we-let-ai-decide/

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Rane, N. (2023). Role and challenges of ChatGPT and similar generative artificial intelligence in arts and humanities. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4603208

Rosenbak, S. (2018). Designing for a city of lies. *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2*. PDC '18: Participatory Design Conference 2018, Hasselt and Genk Belgium. https://doi.org/10.1145/3210604.3214363

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Sanders, E. B. N., & Stappers, P. J. (2012). *Convivial toolbox: Generative research for the front end of design.* BIS Publishers B.V.

Savage, V., Homewood, S., & Shklovski, I. (2023, April 19). Objectify: Better living through anticipatory, just for-you 3D printing! *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23: CHI Conference on Human Factors in Computing Systems, Hamburg Germany. https://doi.org/10.1145/3544549.3582748

Scotti, V., Sbattella, L., & Tedesco, R. (2024). A primer on Seq2Seq models for generative chatbots. *ACM Computing Surveys, 56*(3), 1–58. https://doi.org/10.1145/3604281

Shannon-Baker, P. (2023). Philosophical underpinnings of mixed methods research in education In *International Encyclopedia of Education (Fourth Edition)* (pp. 380–389). Elsevier. https://doi org/10.1016/b978-0-12-818630-5.11037-1

Sicart, M., & Shklovski, I. (2020). 'Pataphysical Softwaree: (Ridiculous) Technological Solutions for Imaginary Problems. *Proceedings of the 2020 ACM Designing Interactive Systems Conference.* DIS '20: Designing Interactive Systems Conference 2020, Eindhoven Netherlands. https://doi.org/10.1145/3357236.3395526

Stappers, P. J., & Giaccardi, E. (2017). Research through Design. In M. Soegaard, & R. Friis-Dam (Eds.), *The Encyclopedia of Human-Computer Interaction* (2nd ed., pp. 1-94). The Interaction Design Foundation.

Williams, R. (2024, July 8). How to use AI to plan your next vacation. *Technology Review*. https://www. technologyreview.com/2024/07/08/1094733/how-to-use-ai-to-plan-your-next-vacation/

Wischnewski, M., Krämer, N., & Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, *22*, 1–16. https://doi.org/10.1145/3544548.3581197

Wolfe, S. (2021). *Marcel Duchamp, the father of conceptual art: Bio & artworks*. Artland Magazine; Artland. https://magazine.artland.com/marcel-duchamp-conceptual-art/

appendix

THE APPENDICES CAN BE FOUND BY SCANNING THE QR