Pontifícia Universidade Católica
do Rio de Janeiro

Gabriel Diniz Junqueira Barbosa

Evaluating the Extended Metacommunication Template as an epistemic tool for the sociotechnical design of machine learning systems

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor : Profa. Simone Diniz Junqueira Barbosa
Co-advisor: Profa. Clarisse Sieckenius de Souza

Rio de Janeiro
August 2022

Pontifícia Universidade Católica
do Rio de Janeiro

Gabriel Diniz Junqueira Barbosa

Evaluating the Extended Metacommunication Template as an epistemic tool for the sociotechnical design of machine learning systems

Dissertation presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the Examination Committee:

Profa. Simone Diniz Junqueira Barbosa
Advisor
Departamento de Informática – PUC-Rio

Profa. Clarisse Sieckenius de Souza
Co-advisor
Departamento de Informática – PUC-Rio

Prof. Edgar de Brito Lyra Netto
Departamento de Filosofia – PUC-Rio

Prof. Bruno Feijó
Departamento de Informática – PUC-Rio

Rio de Janeiro, August 25th, 2022

Gabriel Diniz Junqueira Barbosa

Bachelor in Computer Engineering in the Informatics Department of the Pontifical Catholic University of Rio de Janeiro. Has a developer, in industry, and as a graduate researcher, collaborating in research on the topics of human-computer interaction, information visualization, and data science alongside other graduate students.

Bibliographic Data

Diniz Junqueira Barbosa, Gabriel

Evaluating the Extended Metacommunication Template as an epistemic tool for the sociotechnical design of machine learning systems / Gabriel Diniz Junqueira Barbosa; advisor: Simone Diniz Junqueira Barbosa; co-advisor: Clarisse Sieckenius de Souza. – 2022.

132 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2022.

Inclui bibliografia

1. Semiotic Engineering – Teses. 2. Sociotechnical Design Tools – Teses. 3. Responsible Design – Teses. 4. Ethical Reflection – Teses. 5. Machine Learning – Teses. 6. Engenharia Semiótica. 7. Ferramentas de Design Sociotécnico. 8. Design Responsável. 9. Reflexão Ética. 10. Aprendizagem de Máquina. I. Diniz Junqueira Barbosa, Simone. II. Sieckenius de Souza, Clarisse. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

I dedicate this work to my family, friends, colleagues, mentors, and all those who have supported me through this research journey.

Acknowledgments

I would first like to thank my advisors, past and present, Clarisse, Hélio, and Simone for their continued support during this research project. Second, I would like to thank the various professors I have had the pleasure to learn from at the Informatics Department at PUC-Rio. I would also like to thank those involved in the EMAPS project, who afforded me the opportunity to participate in interdisciplinary discussions that have taught me so much. My peers and colleagues at SERG, IDEIAS, and DasLab were also essential on this journey. I am thankful for their collaboration and solidarity throughout. Finally, I would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the partial financing of this work under grant number 001. I would also like to thank the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) for the partial financing of this work under grant number E-26/201.719/2021 (266916).

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Diniz Junqueira Barbosa, Gabriel; Diniz Junqueira Barbosa, Simone (Advisor); Sieckenius de Souza, Clarisse (Co-Advisor). **Evaluating the Extended Metacommunication Template as an epistemic tool for the sociotechnical design of machine learning systems**. Rio de Janeiro, 2022. 132p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

This dissertation presents the Extended Metacommunication Template, a tool based on a set of guiding questions derived from the theory of Semiotic Engineering. We report the results of a study we conducted to evaluate the tool's impacts on the design process of machine learning systems. By having designers and developers answer a set of questions, the tool aims to help them reflect on their interpretations of the design solution, while allowing them to revisit the presuppositions behind it. We then describe a speculative design study and analyze its results, identifying emergent themes that help us understand how the proposed tool may be used. Among the relevant themes identified are: the reflective practice of design, the designer's focus on their use of language, the process of attributing responsibility to the people involved, the use of the ethical framework provided to them, the bioethical principles, and the ways in which the extension of the template may be used.

Keywords

Semiotic Engineering; Sociotechnical Design Tools; Responsible Design; Ethical Reflection; Machine Learning.

Resumo

Diniz Junqueira Barbosa, Gabriel; Diniz Junqueira Barbosa, Simone; Sieckenius de Souza, Clarisse. **Avaliando o Template de Metacomunicação Estendido como uma ferramenta epistêmica para o design sociotécnico de sistemas de aprendizagem de máquina**. Rio de Janeiro, 2022. 132p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação apresenta o Template de Metacomunicação Estendido, uma ferramenta, denominada Template Estendido de Metacomunicação, baseada em um conjunto de perguntas-guia derivadas da teoria da Engenharia Semiótica. Relatamos os resultados de um estudo para avaliar seus impactos no processo de design de sistemas de aprendizagem de máquina. Ao solicitar que designers e desenvolvedores respondam um conjunto de perguntas, a ferramenta busca auxiliá-los a organizar mentalmente suas interpretações da solução de design, ao mesmo tempo que revisitam as pressuposições por trás dela. Descrevemos, então, um estudo de design especulativo e analisamos seus resultados, identificando temas emergentes que nos ajudam a entender como a ferramenta proposta pode ser utilizada. Dentre os aspectos identificados mais relevantes estão a prática reflexiva do design, a atenção ao uso de linguagem, a atribuição de responsabilidade às pessoas envolvidas, o uso do arcabouço ético de apoio fornecido, os princípios da bioética, e as possíveis formas de se usar o template estendido.

Palavras-chave

Engenharia Semiótica; Ferramentas de Design Sociotécnico; Design Responsável; Reflexão Ética; Aprendizagem de Máquina.

Table of Contents

List of Figures

List of Abbreviations

AI – Artificial Intelligence

EMT – Extended Metacommunication Template

HCI – Human-Computer Interaction

MC – Model Card

ML – Machine Learning

MM – Metacommunication Message

MT – Metacommunication Template

SemEng – Semiotic Engineering

1
Introduction

Developers cannot imagine all of the possible situations in which their systems will be used. There are just too many factors are involved, with the future seldom being what we envision. This is one of the essential characteristics of design problems. However, computer scientists tend to treat the artifacts that they build as solely engineering problems, rather than design projects. As such, they can often ignore the social context of the programs that they build, leading to unforeseen consequences.

As computational artifacts have become ever more present, we can clearly see their impacts in our daily lives. Social media, for example, has seemingly led to cases of polarization (Tucker et al., 2018), filter dysmorphia (Ramphul and Mejias, 2018), and attention problems (Firth et al., 2020). The developers who constructed these systems may not have even considered the possibility that these consequences might occur, but they occurred nonetheless.

Even if our ability to conceive of the consequences of the artifacts that we build is limited, it is not insignificant. Other issues, such as the presence of abuse in these social platforms could have probably been foreseen and acted upon before their introduction and widespread use. Reflection about the possible consequences of the artifacts we build, given their capability to scale and generate significant social impacts, is a crucial part of their responsible design and development.

Computing artifacts are very diverse, with certain types bringing their own challenges. This adds another level of complexity to the matter, since different types of technologies have their own properties that need to be considered. Having an understanding of how certain aspects of a given technology interact with the contextual influences of the circumstances in which they are deployed is an essential part of gaining expertise in solving problems with them.

Let us take Machine Learning as an example. One way to look at this type of technology is through the lens of design materials. A notable example of this type of effort is the work of Qian Yang (Yang et al., 2018), which analyzes how developers consider Machine Learning models as part of their overall systems. In doing so, they often look at the more general properties

that these technologies have, rather than more specific technical issues. They are considered instruments to be used for the sake of the broader system. In the case of Machine Learning models, there are various dynamics that need to be considered, like their dependencies on trends in data, a lack of transparency about their internal logic, among various others.

When developers are building systems around Machine Learning models, these properties can have significant impacts on the constructed artifact as a whole. Their lack of transparency, for example, creates new requirements for systems that need to enable user decisions based on model outcomes. As research has shown, not understanding what these outcomes mean and how they were decided upon may lead to mistakes Gunning et al. (2019). Given this possibility, the design of the system itself may need to be adapted, in order to allow for greater explainability.

The issue of transparency is just one example of how the general properties of different types of technologies can significantly impact how we design and develop our systems. Other types of technologies, such as the variety of devices involved in the Internet of Things, have their own implications for the systems' designs. Understanding the implications of their use usually requires some degree of experimentation accompanied by explicit reflection. Neglecting these dynamics may result in short-sighted designs that fail to account for the contextualized impacts of these technologies. It may result in irresponsible designs.

Given the social harms that we can already observe and the different types of computational artifacts, each with their own dynamics, there has been an increased interest in studying the socio-technical impacts within Computer Science. Since the social dynamics involved are usually not the topic of study for most Computer Science research, interdisciplinary inquiry has been deemed to be key. New conferences, such as the FAccT conference[1] (focused on Fairness, Accountability, and Transparency) have served as a catalyst for these new kinds of scientific research, accepting contributions from fields as diverse as philosophy, computing, and law. By having works from a variety of fields, discussing the same topic of Fairness, Accountability, and Transparency of computing systems in the same venues, these conferences encourage a more holistic approach to the topic, rather than each specific discipline only tackling the implications for their own area of study.

These conferences are also the venues where relevant design tools geared at promoting responsible design are proposed and discussed. Notable examples include the Datasheets for Datasets Gebru et al. (2018), the Model Cards

[1] https://facctconference.org/

Mitchell et al. (2019), among various others. They may focus on different aspects of design and development problems, but all of them are seen as helping developers create fairer, more transparent, and more accountable systems. Working with these can also help individuals consider the wider social context to which their solution will be introduced.

From within the field of Human-Computer Interaction, the theory of Semiotic Engineering can be useful for these kinds of inquiries. By framing the production of computational artifacts as a metacommunicative process between the system's designer and its users, this theory is focused on the meanings that are inscribed into these computing systems. To do so, it employs concepts from Peircean Semiotics, looking at these artifacts as a system of signs to be analyzed. Given this communicative focus, abstracting more specific technical choices in a system's implementation, it may provide a common ground for discussion with people from other disciplines that may lack a technical understanding of Computer Science. They could then be invited to engage with the design of these artifacts. Of course, having computer scientists involved will still be essential, but having these discussions around metacommunication may make the topic more approachable and encourage interdisciplinary discussions.

Seeking to contribute to the topic of responsible design of sociotechnical systems, in this dissertation we present an extension of the Metacommunication Template of Semiotic Engineering, propose a qualitative study based on a set of speculative design sessions to investigate the use of the Extended Metacommunication Template (EMT), and conduct a qualitative analysis of the data collected, identifying relevant, emerging themes. The EMT can be seen as a form of representation of the designers' intentions as they are inscribed in the computational systems that they build. Their instantiation via the development of a design then creates the *metacommunication message*. Our extension is based on a set of guiding questions that seek to help those involved in the creation of computational systems to explicitly structure their own metacommunication message, while having them reflect on a variety of aspects of their design and their potential sociotechnical consequences.

This dissertation is structured as follows. In chapter 2 we discuss the theoretical background behind our proposed contributions. Then, in chapter 3 we describe related work, pointing out what we can learn from them and where the differences between our contribution and theirs lie. Afterwards, in chapter 4, we present the Extended Metacommunication Template, relating it to existing work in Semiotic Engineering as well as describing the rationale behind each of its parts. Moving on, in chapter 5 we describe the study

design we conducted, based on speculative design sessions, to analyze how our proposed extension may impact developers' reflections during design. In chapters 6 and 7, we report and discuss the study results. Finally, we conclude this dissertation in chapter 8, highlighting the contributions of our work and pointing to future, promising directions.

2
Theoretical Background

In this chapter, we discuss some of the work that serves as theoretical background for our proposal. The understandings contained in each of these works are what led us to our conclusions and inspired our proposed contribution. Among the topics discussed are the sociotechnical nature of machine learning systems, the issue of responsible design, and the field of Semiotic Engineering, which constitutes the major theory supporting this proposal.

2.1
Sociotechnical Aspects of Machine Learning Systems

As Cooper and Foster (1971) explain, sociotechnical analysis consists of observing the interplay between social and technical systems. In addition to the dynamics for each of these systems there are others that are specific to the interactions between these two. In a larger scope, societal practices can gain new meaning when being mediated by technical artifacts. A clear example of this is social media, where various social dynamics are conducted through digital platforms. In these cases, the technical affordances can significantly change these social interactions given the new possibilities and limitations that are created.

Various types of technologies can be analyzed through such a lens. All technologies exist within a social context, but the way in which this contextualization takes place can differ between these different types. For example, an artifact based on ubiquitous computing can impact, and be impacted by, a social context in a different way than an artifact based on artificial intelligence. The former may bring with it issues of privacy and consent, while the latter may bring issues of fairness and transparency. Their internal, technical dynamics change the interactions with social systems.

The same types of variations can also occur in between different social systems. Technical artifacts exist within cultures, each with their own customs, habits, and practices. A single artifact, with its own dynamics, can interact in different ways with different cultures. For example, a system based on machine learning can have different meanings when used for entertainment and when used for financial decisions. In the first case, its decisions may not have great

impact, with occasional errors being acceptable. In the second case, specific decisions may end up having disastrous effects for the people involved. A person in dire need having their loan application rejected can make all the difference.

This dual relationship between the social and technical dynamics is a part of what makes each situation unique. Pragmatically, it would be impossible to study and evaluate each pair of social and technical systems. That is why it is often necessary to take a broader look at these phenomena, studying clusters of systems instead of specific ones. For software development, our main interest is in developing a specific technical system. Having a fixed technical artifact to focus on, what we then need to do is study how it interacts with different groups of social systems, according to our interests as creators of the artifact. How it impacts professionals in different fields. How it impacts people from different cultures. All of these can be worthwhile topics to study and gain greater insight on.

In this proposal, we chose to focus on systems based on machine learning (ML) models as the group of technical systems we seek to work with. Even this grouping is quite broad, with varying types of algorithms and tasks existing. However, it is a group of technologies that is being widely adopted and has already had significant social impact (Zhang et al., 2021).

Systems based on Artificial Intelligence (AI), and more specifically ML, are becoming an ever greater part of our lives. Given the quality of their performance in a variety of well-defined decision tasks and the facilities of being a digital artifact, such as their speed and constant execution, they have become attractive alternatives to human decision-making (Mitchell, 1997). As they start to mediate a variety of activities, we are starting to see effects in a societal scale. Recommendation engines behind social media feeds play an enormous role in our social interactions, for example.

As most of these social impacts tie back to a set of technical decisions made by developers when building the system, understanding both sides (social and technical) together becomes essential. Seeking to address this new form of inquiry, going beyond merely looking at isolated technical challenges, researchers have started to consider AI systems as sociotechnical systems (Makarius et al., 2020). Bridging this gap between technical decisions and their social impacts is not a trivial endeavor, and has been a topic attracting increasing interest.

Among the various types of AI systems currently in use, we can highlight Machine Learning models as being especially popular nowadays (Zhang et al., 2021). Their ease of implementation and deployment, alongside the quality of their performance have made it so. It is not the only type of Artificially

Intelligent model, but given its widespread use, we will be focusing on it. Even within this category of models, there is great variety. However, they all share a common characteristic: they learn from the data.

Among the various definitions of what Machine Learning is, Mitchell (1997) defines it as a set of algorithms that improve their performance through repeated interactions with data. Throughout the learning process, the algorithm makes a prediction, learns from the data, and tries to improve its internal mechanisms, seeking to improve performance. The way in which they "learn" from the data is entirely based on probability and statistics, making it hard for a human being, whose understanding is greatly ontological, to accompany the learning process and translate it into ontological understandings (Pearl, 2019).

Machine Learning algorithms are geared towards specific decision-making tasks (Carbonell et al., 1983). In order to fulfill these tasks, the models take into consideration the instances' features, which provide them with information about the instance's nature. Classification, for example, requires that models make decisions about which class an instance belongs to, while regression algorithms try to estimate numeric values for instances' target variables (Maglogiannis, 2007). There are numerous other types of tasks beyond these two, each with their own requirements and dynamics (Carbonell et al., 1983), such as clustering (Rokach and Maimon, 2005) and generative (Goodfellow et al., 2014) tasks, to name a few. Understanding what type of decision-making task is involved in our systems is essential to understanding their sociotechnical role, as they impose conditions to our design that may end up having significant social impacts.

In addition to being based on several types of tasks, these models may also be supervised, unsupervised, or even reinforced (Carbonell et al., 1983). Supervised models learn from the data by trying to make decisions about the instances being fed to them and comparing the end results with what was labeled in the data. Unsupervised models do not have these labels available to them, being then forced to try and identify trends in the data and using other metrics to assess their success, such as the entropy of the clusters formed. Reinforced models learn from feedback as they behave in the environment they are situated in (Sutton and Barto, 2018). They somewhat escape the traditional training, testing, and deployment cycle that most Machine Learning algorithms are subjected to.

As with all other technical factors we have discussed, the type of learning involved in any of these algorithms can also have social impacts. Let us take the case of Reinforcement Learning models (Sutton and Barto, 2018), illustrated

by the example of Tay, a chatbot developed by Microsoft that was meant to learn how to respond to inquiries based on its interactions with other people on social media (Neff and Nagy, 2016). Since its learning process was continuous, with the model learning new behaviors from its environment, developers had few means of identifying whether the patterns they learned violated social norms. As such, once the model began engaging with people on the internet its communicative patterns began expressing various hateful concepts, such as racist and anti-Semitic expressions (Neff and Nagy, 2016). Understanding how these models relate to the data from which they learn their behaviors therefore appears to be a relevant point of reflection for designers seeking to implement systems around them.

There are also algorithm-specific issues that should be considered. There are various general types of algorithms that are usually associated with Machine Learning, such as decision trees, clustering algorithms, and neural networks. Each has their own specific properties, which afford us different options. Decision trees, for example, allow us to explicit understand the criteria behind any given decision, while a neural network would not. In a real design situation where transparency is key, this would be a crucial difference that could lead a designer to opt for the former type of algorithm rather than the latter, even if it meant losing some degree of performance. This is just one example of the many aspects that we need to take into account when discussing a specific family of algorithms and deliberating on which to choose for a system that we may be building.

When analyzing the models we are dealing with, considering the intersection between these various factors (decision-making task, type of learning process, type of algorithm, etc.) seems to be appropriate. Specific dynamics may arise in each configuration. A classification model based on reinforcement learning through a neural network can have its own implications for design that are more than just the sum of its parts.

Despite all of this variance, we can also observe the commonalities that exist in Machine Learning algorithms and some of their implications. The dependence on data, for example, is universal within that group. Since all Machine Learning algorithms depend on data in order to learn desired behaviors, there is always a matter of ensuring that this source of information is representative of the reality that we want the model to replicate. In addition to the issue of data, another common element that is worth mentioning are the methods for statistical learning, which make it more difficult for people to understand the underlying logic, given their mathematical complexity.

Going beyond just analyzing the models themselves, understanding them

as sociotechnical artifacts also requires that we look at the contexts in which they are inserted, since these will be greatly influential to their social impacts. If we go back to the example of Tay, beyond the specifics of the implementation of the model, a crucial part of what occurred was the context in which Microsoft opted to insert it: social media. Having a chatbot based on reinforcement learning present in an environment where hateful speech often appears means that we risk the model learning these behaviors. Were it inserted in another, more constrained context it may have been able to mostly learn appropriate behaviors.

In the end, understanding Machine Learning models as sociotechnical systems may require us to focus on the relationship between internal technical decisions and external social dynamics. It is through these relationships that the social consequences we are trying to understand will take place. Ignoring them means only understanding a part of the whole picture of our system's design. Systems do not exist in a vaccuum, after all.

2.2 Responsible and Reflective Design

Since critiques of recent social impacts of Machine Learning systems have become mainstream (Levy, 2021), there have been increased discussions about the responsibility that developers have for the systems that they build (Matthias, 2004; Tigard, 2021). Differently from other research topics, it appears to be a discussion being held both within academia, industry, and society in general. Given that our work tries to help designers reflect on the possible sociotechnical impacts of the systems that they build, the responsibility that they hold creates an incentive for them to engage in such a reflective process.

Developers are the ones who build the systems that users use. Their roles may differ, with some being managers, designers, programmers, etc., but all of them have a role in shaping the final creation. As such, they hold some level of responsibility towards those who are affected by the system that they created. In Aristotelian tradition, for someone to be held responsible they would have to satisfy two conditions: first, they would have to be have some agency over the situation, which developers clearly do, since they are the ones doing the implementation; second, they would have to have some level of understanding about the consequences of their actions (Coeckelbergh, 2020a). In most software development that does not involve the use of an artificially intelligent component, both of these conditions are reasonably met, even if they are unable to conceive of all of the potential consequences of their actions.

Matthias (2004) outlines some of the ways in which different levels of intelligent automata can make it difficult for us to ascertain the responsibility of their creators. He argues that as developers distance themselves from explicitly determining the agent's internal logic it becomes more difficult to determine whether developers understand the potential consequences of what they are doing. He points to autonomous agents, those that learn from feedback from the environment in which they are inserted, as the epitome of this situation. The developer is unable to predict the behaviors that the model will learn and, as such, would be less responsible for its consequences. Of course, other thinkers disagree.

Tigard (2021) argues that the concept of moral responsibility is flexible enough to account for the role that developers have in their creations' consequences. He argues that by knowingly abdicating the possibility of monitoring the learned behaviors of the models that they build, they are accepting the risks that come with such a decision and can thus be held liable for whatever outcomes occur. Differing from Matthias, his argument states that any limitations to the developer's understanding are accepted by the developers themselves. If someone consciously chooses not to know the potential consequences of their actions, then they cannot claim innocence when others try to hold them accountable. Unfortunately, simply knowing that they can be held responsible does not make it easy for us to identify those responsible.

Coeckelbergh (2020b) outlines two practical challenges to ascertaining who the responsible parties are for the impacts of an artificially intelligent system. The first is the issue of "many hands," which results from the presence of multiple developers in the construction of computational artifacts and whose functions are not fully defined. This can then make it difficult to identify whose actions led to specific outcomes. The second issue is that of "many things," which discusses the interactions that exist between multiple agents and how they may alter each others' behavior. When a developer builds a system, how can they predict what it may learn from other systems, which they have not built? These two are just some of the practical challenges that people may face when trying to determine who the responsible parties are in real-world development scenarios.

The limitations that developers face in understanding situations when designing and developing their systems can also be understood along the lines of Horst Rittel's notion of "wicked problems" (Rittel and Webber, 1973; Buchanan, 1992; Skaburskis, 2008). His theory states that every design situation is based on an incomplete definition of the problem, given the infinite variables that may be relevant to the end product. Being unable to conceive of

all possible contextual factors, designers must then try and identify the most relevant ones and frame their solutions around them. Despite what developers usually face in computer programming, where the problems they are trying to solve by coding are more well-defined, acknowledging the sociotechnical role of the systems that they build requires recognizing the "wickedness" of the problems they face. There are usually too many social variables for them to analyze thoroughly.

During their process of designing the system, developers will make various decisions based on their current understanding of the problem at hand. The reasoning behind these decisions can be a valuable asset for ascribing responsibility, since it can allow us to trace the presuppositions that led to any design choices that led to poor outcomes. Moran and Carroll (1996) have done valuable work along these lines, discussing how to capture developers' design rationale. They mostly argue that understanding the rationale behind design decisions requires us to know the alternatives that had been considered, which was chosen, and the justification behind that choice. This fits nicely with Rittel's theory of design, where designers navigate through various possible solutions, trying to find the most appropriate one. As developers design their systems, they will probably discuss various options until they reach their final conclusions. Capturing these discussions and allowing us to trace back from technical decisions to their original design rationale may be very relevant to ensuring responsible design, as we could then possibly identify whether relevant alternatives were overlooked or abandoned, and, if so, why.

Multiple representations have been proposed by researchers to capture developers' design rationales. Noble (1988) originally proposed IBIS (Issue-Based Information Systems) as one possible alternative. In it, rationale would be captured as sets of alternatives to a given decision, each with arguments supporting or objecting to it, taking into consideration both the evidence for or against an option's adoption. Another relevant form of representation is QOC (Questions, Options, Criteria), proposed by MacLean et al. (1991). Decisions that need to be made are seen as Questions, with alternative solutions being considered as Options, which in turn are evaluated against a set of criteria. Based on this evaluation, some would be selected according to the set of Criteria that they fit to. With this setup, we would have the underlying argumentation in the form of fitting Options with Criteria, resulting in the final decision. Various other types of representations have been proposed, but these are among the most popular, especially in software development.

The representation we propose in this dissertation does not explicitly require designers to consider multiple alternatives in such a structured way.

However, it does greatly focus on the justifications behind their final design decisions, thereby giving us a glimpse into their reasoning when conceiving of the artifact. The two aforementioned representations tend to have limited room for argumentation, having a more structural focus, rather than an interpretative one. We made the choice, for the representation we propose, to have a greater focus on the arguments made rather than the alternatives considered.

The ability to successfully justify a design decision, taking into consideration most relevant factors identified and using them in arguments for said decision, usually requires some level of education on the practice of Design. As Lawson and Dorst (2013) discuss, as designers practice their trade, they tend to gain a greater level of expertise in the variety of tasks involved, as well as a better understanding of the design process as a whole. Through repeated experiences, either ending up in success or failure, they are usually able to start identifying which strategies are appropriate for a given situation and which are not. In doing so, they can also develop their own "style" as designers, having their own preferences that are expressed into the objects that they build. This process can also be helped along through certain tools or methods.

As Dorst and Lawson argue, for designers to successfully gain greater expertise usually requires that they do a great deal of reflection on what went right or wrong in each design project. Donald Schön was one of the pivotal thinkers about the role of reflection in our various practices, eventually developing the notion of a "reflective practitioner" (Schön, 1979). He argued that great designers, architects, engineers, and various other professions, were able to reflect in depth about the actions that they engaged in in their professional activities. The moments in which these reflections took place were also deemed essential, with more proficient individuals reflecting on their actions at almost the same time as they engaged in them. On this topic, he differentiated between "reflection-in-action," where practitioners would reflect on their actions as they did them, and "reflection-on-action," where they engage in reflection after the fact. Indeed, he recognizes that all reflections in actuality occur after actions, but argues that making them closer can lead to a more dynamic practice, allowing for professionals to quickly consider different alternatives, reflect on their possible consequences, and choose the one that seems the most appropriate. Having seen the outcomes of a decision, they could then go back and improve their own process of reflection, such as learning to consider new, previously overlooked factors or paths.

A similar process takes place during ethical deliberations, when a person considers what the appropriate action would be in relation to its patients.

Mark Johnson, in his book Moral Imagination (Johnson, 1993), outlines how people, when making ethical decisions, use their imagination to consider what the consequences of those decisions might be. This imaginative process, he argues, is usually based on peoples' cognitive conceptual systems, which change according to their reflections on their lived experiences. As individuals make certain moral choices, deeming them to be appropriate, and end up finding that they were wrong, they tend to learn from their mistakes after understanding what they missed or misunderstood. As such, reflection plays a role in shaping our imagination, which, in turn, ends up affecting our moral deliberations. This would open up the possibility that individuals could improve their ability to consider the consequences of their actions, thereby making them more effective moral deliberators.

This is one of the main points of the design approach that we propose in this dissertation. We attempt to tie a process of reflecting on the justifications behind design decisions to the process of moral deliberation that occurs alongside it. After all, most social impacts of developers' technical decisions have ethical connotations. By having them examine their underlying rationale, having to answer our proposed tool's guiding questions, and connecting it to the possible ethical issues that may come of it, we try to bridge the relational gap, as Coeckelbergh (2020a) puts it, making developers more aware of the relationship that their decisions have with the consequences to all stakeholders involved. This association, we argue, takes place through reflection.

In order to support the component of ethical deliberation, the representation we propose allows for the use of preexisting ethical frameworks, such as the Bioethical principles (Beauchamp et al., 2001). These can serve as a source of inspiration, allowing developers to consider the situations that they face in different ways, according to the frameworks that they use. In doing so, they may also engage in analogical and metaphorical thinking, as explained by Holyoak and Thagard (1996) and Lakoff and Johnson (2008), to determine what the appropriate course of action should be.

However, for these efforts to work, developers need to be able to connect more abstract ethical principles with more practical technical decisions, as Morley et al. (2021) explain. Without this ability to relate design decisions with ethical decisions, ethical deliberations become unable to guide the actions that developers make when developing their systems, given that few concrete changes would be identified.

In the representation we propose in this dissertation, by having ethical deliberations occur alongside design decisions, aided by the use of ethical frameworks, we seek to connect abstract decisions with their potential, concrete

implications. As such, developers would be more liable to learn the full impact that their technical decisions could have, adapting according to the situation's needs.

Of course, responsible design is not only a matter of individual ability, but also requires an environment in which it can thrive. Efforts to promote ethical development that do not result in any material changes are often called "ethics-washing," where an entity uses the appearance of ethical behavior without actually following through and changing what is necessary (Wagner, 2018). Winograd proposes the concept of "ethicking" as what should be strived for, where ethical deliberation becomes one of the main components of everyday practice, instead of some separate process that occurs after the fact.

The possibility of "ethics washing" seems to be present in all proposals for responsible design. Researchers and practitioners may propose a variety of tools and methods that, when used well, can have significant impacts in ethical practice, but if the different entities that use them only do so poorly and in a shallow manner, then their material contributions will not be had. A company can say that they use a variety of processes, methods, and tools that ensure that their development is ethical, but if they do so without the necessary dedication and care, then are actually performing "ethics washing" rather than responsible practice.

2.3 Semiotic Engineering

In this proposal, we present a new design tool based on the theory of Semiotic Engineering. In doing so, we seek to aid in developers' deliberation about the sociotechnical nature of the artifacts they build.

Since we are trying to promote reflection about the developers' decisions, Semiotic Engineering presents itself as an appropriate theory to explore, given its focus on the meanings involved in the system's creation.

Semiotic Engineering is a semiotic theory of human-computer interaction that frames user interactions with computing systems as a computer-mediated metacommunication (de Souza, 2005). According to it, the designer, when creating the system, imbues it with signs that communicate their design vision for the system as a whole. At interaction time, the user tries to interpret these signs and understand how the system actually functions. Of course, different individuals may have different interpretations based on the same representations. It is this process of producing and interpreting these signs that is the focus of Semiotic Engineering.

At the core of the theory is the Metacommunication Message, which

is generalized as the Metacommunication Template. As its name suggests, it provides developers with some guidance as to what information they should consider when building their systems, instantiating the final message. Semiotic Engineering's Metacommunication Template is formulated as follows:

> *"Here is my understanding of who you are, what I've learned you want or need to do, in which preferred ways, and why. This is the system that I have therefore designed for you, and this is the way you can or should use it in order to fulfill a range of purposes that fall within this vision."* (de Souza, 2005)

By analyzing the meanings of the signs imbued into the system by its creators, we can gain some insights into how they envision the artifact they are creating. Of course, real-world systems are developed by a collective of individuals, so the final Message being observed could be a compilation of their multiple interpretations.

There are two main evaluation methods in Semiotic Engineering (de Souza and Leitão, 2009): the Semiotic Inspection Method (SIM) (de Souza et al., 2006), and the Communicability Evaluation Method (CEM) (Prates et al., 2000). SIM, as its name suggests, is an inspection method geared at analyzing the three types of signs of a given system: static signs, whose meaning is related to the system state and is independent of interaction; dynamic signs, whose meaning is related to the system's behavior and can only be grasped through interaction; and metalinguistic signs, whose meaning refers to the signs themselves. CEM, instead on focusing on trying to discern some of the meanings as inscribed by the original designer (the emission of the metacommunication message) focuses on analyzing how these meanings are perceived by the systems end users (the reception of the metacommunication message). It is through this user observation that researchers may be able to identify communicability breakdowns that are of interest to them. Both have a firmly semiotic focus, trying to analyze different interpretations of the signs present in these computational artifacts.

As we take this interpretative focus, we can also consider the presuppositions behind the system's rationale, granting us a glimpse into the conceptual systems of those involved. By observing certain signs and what they communicate, we try to infer some of the designers' beliefs that justified the various decisions that make up the final artifact. This process of reflection can be done from the outside, by observers of the finalized system, or from the inside, by the developers as they start to conceive of the system and the signs that make it up.

If we consider the Metacommunication Template itself, some of its features may lend themselves to more reflection on the part of developers. An example of this is the relational focus that the Template takes, by framing the message as being communicated from the first person, assumed by the developers, to the second person, assumed by the system's stakeholders, *i.e.*, those that will be affected by the system's use. This focus on the relationship between the two parties can be critical to reflections about responsible design, as outlined in section 2.2. The template also tries to bring to light the justifications for the design decisions taken by the developers, in stating "here is what I have developed for you, and why."

In addition to the Metacommunication Template, another relevant form of representation proposed within Semiotic Engineering is MoLIC (Modeling Interactions as Converstations) (Barbosa and de Paula, 2003). As a modeling language, it takes a more structuralist approach segmenting the metacommunication message into several scenes and utterances, illustrating the system's interactive structure. Like the Metacommunication Template, it also has components that can lend themselves to reflective practice, such as the notion of explicitly outlining the presuppositions behind certain interactive decisions thorugh the tag "presup." This allows a designer to connect their underlying rationale, especially its justifications, into the interactive model itself. By having the more technical and structural decisions alongside the presuppositions, it can also be a good candidate to model the sociotechnical nature of the system being constructed.

Despite having a different focus, the Metacommunication Template's and MoLIC's approach can actually complement one another. The former has a more holistic focus, mainly interested in the system as a whole, while the other focuses on the structure that makes up that whole. Certain types of reflection could be better suited to one focus over another. Therefore, having a connection between the structured Metacommunication Message, as an instantiation of the Template, and the MoLIC model that describes it, could help developers take their more general reflections and tie them to more specific technical decisions. For example, a developer might identify a general communicative issue when structuring their Metacommunication Message through the Metacommunication Template, but be unable to identify what part of the system is responsible for it. They could then analyze the underlying MoLIC model and its connections with the Metacommunication Message to identify which design decisions might be at fault.

Going beyond Interaction Design, Semiotic Engineering has also been expanded to Software Development as a whole, with the proposal of the

SIGNIFYI suite (de Souza et al., 2016). This set of tools focuses on different aspects of traditional software engineering and try to bring a more semiotic perspective to the metacommunicative aspects involved. The three main tools proposed were the SIGNIFYIng Message, the SIGNIFYIng Model, and the SIGNFYIng API. As their names suggest, they focus on messages exchanged within the development team, the use of various models throughout the development process, and the construction and dissemination of APIs that make up some of the system's functionalities. In each of these fields, myriad communicative fails can emerge that can lead to significant misunderstandings.

Within the suite, we can specifically focus on the SIGNIFYIng Message as the closest to what we put forward in this proposal. It allows for brief Metacommunication Messages to be exchanged within the development team, trying to ensure that the meanings involved are well-understood. Instead of focusing on the broader Metacommunication Message that represents the system as a whole, these specifically mention parts of it. It is similar to what MoLIC proposes, in this respect.

One of the main differences between the SIGNIFYIng message and our proposed extension of the metacommunication template is that ours seeks to assist the developer in formulating their metacommunication message, while theirs is solely focusing on communication it. Ours is also focused on the development process as a whole rather than partial communications that may occur within it. Eventually, the two may even be used together, with the proposed set of guiding questions helping to structure the SIGNIFYIng messages that are formulated during the software development process. All the while still bringing attention to the sociotechnical impacts of the parts of the system being discussed.

Inspired by all of the theoretical background work we have just discussed, we propose to extend the metacommunication template with a set of guiding questions that focus on the sociotechnical aspects of the systems being built. In doing so, we also try to promote reflective practice by having developers critically analyze their own interpretations of the artifacts under construction. There are other related work in the literature that seek a similar goal, each with their own strengths and limitations. Our proposal may also work in tandem with some of these, with them complementing each other. We discuss these issues and opportunities further in chapter 3.

3
Related Work

In this chapter, we focus on relevant work that is related to ours, discussing similarities and differences between our contributions and theirs. This is of paramount importance in the effort of locating our work within the wider literature. There are two main kinds of works that will be discussed: works that propose design or documentation tools for machine learning systems from a sociotechnical perspective, and frameworks that may utilize these tools and may eventually work with our proposed extension of the Metacommunication Template as well.

3.1
Design and Documentation Tools

Trying to support better design processes while also allowing for more accurate documentation about what is being conceived, practitioners and researchers have proposed various tools. Many complement one another, creating a more comprehensive design process. Traditionally, design and development processes utilize many such tools throughout, so the conception of said tools tends to already take into account the types of processes it may be inserted in. In this section, we take a look at some of the most closely related tools to ours, highlighting any similarities, differences, and what we were able to learn from them.

The first tool we can discuss is the Model Card (Mitchell et al., 2019). It was proposed a documentation tool for Machine Learning models containing details beyond the most technical, such as any ethical issues involved. It combines a set of more technical criteria, such as details about the training data used, to more social aspects, such as acceptable and unacceptable uses of the model. Our proposed contribution more closely resembles these this social focus, more so than the technical one. Their work does not seek to guide developers in filling the tool, only outlining what information they deem relevant. Ours, on the other hand, seeks to assist in the Template's instantiation via the guiding questions we provide.

Datasheets for Datasets (Gebru et al., 2018) is another contribution that is closely related to the Model Cards. Instead of focusing on documenting Ma-

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 3.1: Model Card sections and topics. (Mitchell et al., 2019)

chine Learning models, however, they focus on documenting the sociotechnical aspects of the datasets that are used to train them. Unlike the Model Cards, that use affirmative criteria for developers to fill with the related information, the Datasheets use questions that should be answered in order to generate the resulting document. This is a similar approach to the one we take. However, our focus is on the sociotechnical aspects of the system as a whole, rather than specifically focusing on datasets.

Another interesting type of representation proposed is that of Nutrition Labels, originally put forward for the context of privacy by Kelley et al. (2009). This simplified representation was found to be more easily processed by users (Kelley et al., 2010), making it an attractive alternative to extensive, and notoriously hard to interpret, privacy terms. Nutrition labels were also proposed for other computing artifacts such as Machine Learning datasets (Holland et al., 2018), tackling a similar problem to the Datasheets for Datasets. Our proposed extension sacrifices brevity for the sake of assisting developers in their conception of the artifact, covering key questions in different stages of software development.

Given that, as we have seen, there are various tools available that touch on sociotechnical aspects of computing systems, choosing which to use and how to fit them into the design process is key. In this sense, Shen et al. (2021)'s work can be of interest to us. They conduct a speculative design study with multiple combinations of design tools to understand the benefits and perils of using them together. In their study, they focused on the Model Cards, which we have discussed previously, the Persona Cards, which are basic design personas, and Checklist Cards, which provide developers with a set of steps that they should follow in their design process. What they found was that using certain tools without the others could generate a lopsided design process. For example, participants could end up focusing excessively on the models themselves without paying much attention to the stakeholders involved. This study serves as a key reminder that the extension of the Metacommunication Template that we propose would be adopted alongside other design tools. We would then have to study how it could cooperate with other prototypical design tools, which we do not do in this work and leave as future work.

Going beyond proposing individual design tools, some researchers and practitioners have also proposed entire toolkits to be used in the design of computing systems. A notable example of this is the HAX Toolkit, based on the Guidelines for Human-AI Interaction (Amershi et al., 2019), proposed by researchers from Microsoft Research. This toolkit includes a set of guidelines and design patterns that can help developers in designing systems that involve

some degree of interaction between humans and artificially-intelligent models. Having these predetermined recommendations can be valuable to professionals, especially those without a significant understanding of design processes. For those with greater design expertise, choosing their own tools according to their situation, may lead to better results.

There are still other types of frameworks that are not focused on developers, but rather on stakeholders. The Action-Oriented AI Policy Toolkit for Technology Audits is one such framework (Krafft et al., 2021). In their work, they propose a set of guidelines and methods that seek to assist people with little technological understanding in conducting audits of governmental policy that involved the use of Artificial Intelligence. Their study was focused the case of Seattle, Washington, where the use of facial recognition in the city was abolished after public pressure (Elamroussi, 2021). Given that the topic of Artificial Intelligence can be quite technical, their proposal allowed individuals to focus on salient, sociotechnical aspects of the policy being proposed, allowing individuals to make their own value assessments, even without a completely technical understanding of the technology. Our work focuses on the sociotechnical component of these systems, while also capturing the developers' perceived meanings on the system's construction, which could eventually allow for outside analysis and scrutiny, such as occurs in the AI Policy Toolkit.

Analyzing some of the existing tools and toolkits that allow for sociotechnical considerations is essential to locate our proposal within the broader literature. Unlike some of the other examples observed, we opt for a more abstract focus on the system's meanings, rather than focusing on specific technical details. In doing so, we seek to allow for the participation of even non-experts in the examination of the underlying rationale, such as occurs in the AI Policy Toolkit. As we have seen with the Value Cards paper (Shen et al., 2021), understanding how our proposal works with other design tools is greatly important, given that it would probably benefit from the use of other tools, such as personas and Model Cards. By understanding some of these aspects, developers may have a better grasp on how to used our proposed extension, adapting it into their existing design and development processes.

3.2 Frameworks for Responsible Design

In addition to design tools and methods, several frameworks have been proposed that can help developers and stakeholders consider the sociotechnical aspects of computing systems. These can make use of various design tools and

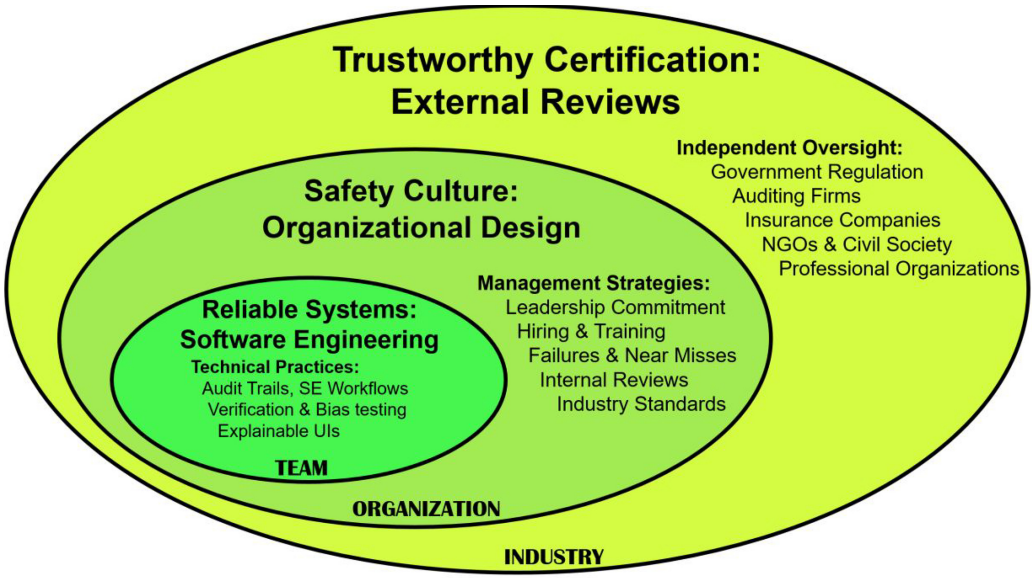

Figure 3.2: Shneiderman's audit framework for human-AI systems. (Shneiderman, 2020)

methods in order to fulfill the tasks they set forward. It is worthwhile to discuss these frameworks as these can also help us locate some of the roles that our extension of the Metacommunication Template can fill within the design and development process.

One of the first frameworks we can discuss is Shneiderman's Audit Framework for Human-AI systems (Shneiderman, 2020), shown in figure 3.2. It defines a series of levels in which we can analyze the impacts of these types of computing artifacts: the team level, the organization level, and the industry level. For each of these stages, he proposes a set of interventions that can allow for better auditing of these systems' performance. For example, at the team level, he proposes the use of Audit Trails, that enable tracing back the sources of issues that may have been identified. In his proposal, he mostly defines it as the use of system logs in conjunction with conceptual models, however, at a design level, our proposed extension of the Metacommunication Template might serve a similar purpose, allowing developers to trace outcomes back to the design decisions that led to them.

Another framework that is relevant to our work is that of organizational transparency, both about the products themselves and the processes that create them, of course, being cautious not to expose intellectual property. Schnackenberg and Tomlinson (2016) discuss how organizations can use transparency to engender trust in their stakeholders along three lines: information disclosure, clarity, and accuracy. It is through these three variables that institutions can choose how much relevant information to divulge and how. After all, as they have concluded, transparency only has a positive return if it matches

the demands of stakeholders. Felzmann et al. (2019) take a legal approach to these transparency requirements by analyzing the demands laid out in the General Data Protection Regulation (GDPR) of the European Union. Similarly to the work of Schnackenberg and Tomlinson, they propose that transparency should be understood as relational in nature, between the companies and their stakeholders, also taking into consideration the various contextual factors that might impact how transparency communications should take place.

Within the context of artificially-intelligent agents, transparency can also be incredibly relevant, as noted by Wortham and Theodorou (2017). When stakeholders interact and depend on these agents, a certain degree of trust is required. For it, some modicum of transparency is usually required. If people do not understand the underlying logic behind these agents' activities, they tend to be less likely to trust it (Schmidt et al., 2020). As such, companies that seek to create agents that are useful to their consumers may need to strive to provide them with sufficient information to engender trust in these intelligent systems.

Despite being more focused on assisting design conception, our proposed extension of the Metacommunication Template may also provide greater transparency to stakeholders. Since designers and developers answer the guiding questions to structure their Metacommunication Message, the resulting document would be a somewhat accurate representation of their intentions in building the system. This sort of information could be relevant to users and other affected members of society, since they would allow them to critique the reasoning behind the software being developed and compare their expectations with reality. By having some understanding of what developers intended to build, users might also trust it more, since they would have an idea of what the system sought to achieve, and how. This hypothesis is not tested in this work, being left as future work for now.

A final framework we can consider is that proposed by Cobbe et al. (2021), that lays out a set of requirements to allow for an automated decision-making process that is more accountable through the notion of *reviewability*. The basic notion is that of allowing recourse for any given automated decision by challenging the process that went into building the underlying autonomous agent. By focusing on the process itself, rather than the underlying logic of the model, they try to ensure that these reviews are not hindered by the opacity of the models themselves. The processes can be transparent, even if the trained models themselves cannot. This proposal fits interestingly with our proposed extension of the Metacommunication Template, given that our guiding questions ask developers about their decisions during the design

process, prompting them to document them for future analysis. This type of conceptual information might make for an interesting topic for outside reviews of automated decision-making processes, since they directly relate to the developers' mindsets at the moment of development.

We find it useful to understand how our proposed extension of the Metacommunication Template fits into some of these existing frameworks since it allows us to consider its different dimensions and possible uses. As with our analysis of other related design tools, it also helps us locate our proposal within the wider literature that touches on the sociotechnical aspects of Machine Learning systems. Having described how we perceive our extension's fit with related work, we can now move on to describing the guiding questions that comprise it.

4
Extended Metacommunication Template

In this chapter, we discuss our proposed contribution to the theory of Semiotic Engineering, which is the extension of the Metacommunication Template with a set of guiding questions. This work has already been published on the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) in 2021 (Barbosa et al., 2021). We first discuss the extension's questions and the sections in which they are organized (section 4.1). Then, we will discuss the rationale behind this proposal, as well as the questions' framing (section 4.2). Finally, we discuss some of the ethical aspects involved with the Extension, such as the possibility of using multiple pre-existing ethical frameworks to help with the developers' ethical deliberations about their design choices (section 4.3). Through these steps, we seek to provide a comprehensive understanding of our proposed Extension of the Metacommunication Template.

Before we start these discussions, however, it is important for us to remember the original formulation of the Metacommunication Template, which is:

> *"Here is my understanding of who you are, what I've learned you want or need to do, in which preferred ways, and why. This is the system that I have therefore designed for you, and this is the way you can or should use it in order to fulfill a range of purposes that fall within this vision."* (de Souza, 2005)

The set of guiding questions that we have proposed as an extension of this original template were:

1. **Analysis**
   - 1.1. What do I know or don't know about (all of) you and how?
   - 1.2. What do I know or don't know about affected others and how?
   - 1.3. What do I know or don't know about the intended (and other anticipated) contexts of use?

1.4. What ethical questions can be raised by what I have learned? Why?

2. **Design**

2.1. What have I designed for you?

2.2. Which of your goals have I designed the system to support?

2.3. In what situations/contexts do I intend/accept you will use the system to achieve each goal? Why?

2.4. How should you use the system to achieve each goal, according to my design?

2.5. For what purposes do I **not** wnat you to use the system?

2.6. What ethical principles influenced my design decisions?

2.7. How is the system I designed for you aligned with those ethical considerations?

3. **Prototyping, implementation, and formative evaluation**

3.1. How have I built the system to support my design vision?

3.2. What have I built into the system to prevent undesirable uses and consequences?

3.3. What have I built into the system to help identify and remedy unanticipated negative effects?

3.4. What ethical scenarios have I used to evaluate my design?

4. **Continuous, post-deployment evaluation and monitoring**

4.1. How much of my vision is reflected in the system's actual use?

4.2. What unanticipated **uses** have been made? By whom? Why?

4.3. What anticipated and unanticipated **effects** have resulted from its use? Whom do they affect? Why?

4.4. What ethical issues need to be handled through system redesign, redevelopment, policy, or even decommissioning?

Let us now discuss each of these in greater detail, relating them to the aspects of the design process on which they seek to promote greater reflection.

4.1 Template's Sections and Questions

The guiding questions in our proposed extension of the metacommunication template are split into four sections: analysis; design; prototyping, implementation, and formative evaluation; and continuous, post-deployment evaluation and monitoring. Each of these is related to a specific stage of the design process, starting with the gathering of information about the problem being tackled and ending with the monitoring of the final solution in its environment. By covering each of these steps, we try to promote a more comprehensive understanding of the proposed solution, starting with its presuppositions and continuing until after its actual deployment.

4.1.1 Analysis

The first stage in our extension of the Metacommunication Template focuses on Analysis. In it, developers and designers are asked about their understanding of the situation at hand, including those involved. As with all design projects, presuppositions about the context for which a solution is being proposed can greatly impact which options are deemed most appropriate, resulting in different designs. For a single situation, multiple designers may end up reaching very different conclusions based on the evidence they found and the understandings they drew from it. It is also natural for analysis to be the initial stage of a design process since it allows for the gathering of information prior to making essential design decisions, so as to make these better informed. The same will probably happen with our proposed extension, with answers to these earlier questions serving as a basis for answers to those further on ahead.

4.1.1.1 *What do I know or don't know about (all of) you and how?*

The first question in our proposed extension starts off by asking the developer to consider what they know about the stakeholders involved, who would be the second person involved in the Metacommunication Message. Given the conversational focus of the Metacommunication Template as a whole, making them aware of the receiver of the message being produced sets up the framing that will be assumed throughout. Another notable point in this question is the fact that it asks developers about what they know and about what they know that they do not know. By being prompted to face the gaps in their knowledge about those involved, they may then be more cautious in making their decisions, acknowledging the uncertainty involved.

4.1.1.2
What do I know or don't know about affected others and how?

Going beyond those directly involved with the system, the second analysis question asks developers to consider those whom their system's use may impact, even if they are not directly related to it. Especially with larger systems, such as in the case of social media, their use may end up creating new dynamics that involve even those that do not directly engage with them. In this question, we try to have these systems' creators reflect on this possibility and how these people may be indirectly affected. Of course, this is often more difficult since most of our data-gathering efforts, such as surveys and interviews, tend to be more focused on our system's users, and not on the wider community that may be affected by its use.

4.1.1.3
What do I know or don't know about the intended (and other anticipated) contexts of use?

In addition to knowing who the people involved, and affected, are, developers must also understand the contexts in which their systems will be used. Various interactions can vary significantly according to these contextual factors, so they are worth considering. Different cultural contexts, for example, can lead to very different interpretations of the signs involved, possibly leading to communication breakdowns. Contexts can also afford different possibilities, even if the developers have not considered them previously. There are various examples of technological appropriation where a system is used in ways unforeseen by its creators due to possibilities afforded by their context, wherein some interactions can take on a new meaning and fulfill a different task altogether (Riemer and Johnston, 2012).

4.1.1.4
What ethical questions can be raised by what I have learned? Why?

At the end of this section, we present them with their first ethical question. Here they are prompted to reflect on the ethical implications of their current understanding of the design situation they are facing. This could encompass issues such as knowing private information about their end users, not knowing some crucial information that can lead to a poor experience for a given type of stakeholder, and various others. It is important to note that the information presented in this Analysis section will serve as the basis for the decisions that will be made later on, so having them reflect on their ethical implications early on may also prime them for ethical reflections when making

the design decisions themselves.

4.1.2 Design

Having discussed their current understanding of the design situation, our questions then move on to the design decisions themselves. In this section, developers start to be asked about their actions in the design process, moving to a more active role. These questions seek to provide a comprehensive look at the solution they are conceiving, while not yet worrying with some of the more technical details of its implementation, which will come at the following section. By having them formulate answers to these questions, we are also making them structure their understanding of their solution and look for the right words and expressions to communicate it to the stakeholders involved. Remember, these questions still fall within the conversational framing set up by the Metacommunication Template, so thinking about the receivers of the message is key.

4.1.2.1 *What have I designed for you?*

The first question in the design section is very straightforward. It asks the developer to synthesize an explanation of the solution they are presenting to the stakeholders involved. The logic behind this solution and the decisions that make it up will be discussed in the following questions. In this way, developers are tasked with presenting their proposed solution as a whole and then breaking it down and justifying it piece by piece.

4.1.2.2 *Which of your goals have I designed the system to support?*

This question discusses the main reason that users will end up engaging with the system, which is to fulfill their own goals. However, no system is able to help with everything, so the system's designer needs to select which of the users' objectives will be supported by the system, as the question states. How successful the system is can then be measured in relation to how successful the users are in reaching these goals.

4.1.2.3
In what situations/contexts do I intend/accept you will use the system to achieve each goal? Why?

Going back to the issue of context, this question asks about those contexts in which the developers intend the system should be used, and those that they accept but do not regard as ideal. Given how stakeholders will end up appropriating the system, developers are unable to limit how, where, and when it will be used. They can, however, declare the kinds of situations that were considered during the design process. Those that fall beyond this scope would not have been explicitly considered and the system's functioning may not be guaranteed. Warning stakeholders of this can lead them to engage with the system more cautiously, especially when deviating from the designer's intentions.

4.1.2.4
How should you use the system to achieve each goal, according to my design?

Moving on to "how" stakeholders will be able to achieve their goals, this question asks the developer to state how they envisioned the process. We could think of this as the preferred path for stakeholders' interactions with the system. However, this does not mean that the ways described in the answer to this question are the only ones for them to achieve their goals. They may end up appropriating the system and using it in ways not envisioned by its creators, possibly leading to undesired consequences.

4.1.2.5
For what purposes do I not want you to use the system?

Related to the issue of goals and means, this question then starts to restrict the user's appropriation of the system from the point of view of its creators. In it, developers are asked to state which purposes should not be pursued through the system's use. These would be goals that they may find immoral or that they have not prepared the system to be able to handle. Both of these situations can fall within the scope of this question. However, it is also clear that just declaring what goals they deem undesirable is insufficient for stopping these types of uses. Restrictions need to be made, as we will see in other questions further on.

4.1.2.6
What ethical principles influenced my design decisions?

The first ethical question for the Design stage touches on the topic of the developers' guiding principles behind the system's design. These do not have to be a part of an existing ethical framework. They can also be personal principles that they abide by in their practice. Protecting the user's privacy, ensuring equity of outcome, all of these can serve as examples of the types of principles that can be mentioned in this question. Having the developers explicitly declare them can allow them to understand and reflect on some of their own values, possibly also making them transparent to any stakeholders involved.

4.1.2.7
How is the system I designed for you aligned with those ethical considerations?

The second ethical question is strongly connected to the first, focusing specifically on the fit between the developers' stated ideals and their actual practice. By having them justify how their design decisions fit with their declared principles, we may lead them to find contradictions between the two. This can, in turn, create a process of reflection, not only upon the decisions themselves, but also upon whether they truly hold these principles to heart. Eventually, the conclusions drawn from this process of reflection may even lead them to change their own design practice, sticking more closely to certain principles or abandoning others.

4.1.3
Prototyping, implementation, and formative evaluation

Having discussed the abstract conceived solution, we now move on to the Implementation section. Here, developers are asked about more concrete choices made when instantiating their design. Despite being closer to the actual programming of the computing artifact, this section is not exclusive to those who know how to code. Individuals lacking in technical know-how might still be able to answer these questions, albeit in less detail. Policy makers, for example, could look at them as a policy making challenge surrounding a system, implementing rules rather than actual computer code. The essence of this section is trying to turn the abstract ideas conceived in the previous section into a concrete artifact that can actually result in the outcomes their creators desire.

4.1.3.1
How have I built the system to support my design vision?

The first question asks developers to relate the development process behind the system's implementation. Similar to what occurred in the Design section, they are first asked to relate it as a whole and are then prompted to discuss specific aspects of it, justifying some of their decisions. Here they are also prompted to reflect on the fit between how they went about developing the system and their original design vision. During actual development various constraints may be identified that can deviate from the original design. As such, they are prompted to reflect on any of these adaptations and how they relate to the original vision.

4.1.3.2
What have I built into the system to prevent undesirable uses and consequences?

Going back to the design questions, one of them asked for which purposes the stakeholder should not use the system. Here developers are asked about how they went about implementing restrictions into the system to avoid these sorts of uses. Especially when discussing sociotechnical issues, constraints are essential to ensure that stakeholders are not abusing the system in significant ways. An example of such a constraint would be content moderation processes which prevent users from uploading harmful content to content-sharing websites. Once the system is developed and in the hands of end users, developers will not be present to make their will known so implementing restrictions into the system is one of the few recourses they have at their disposal.

4.1.3.3
What have I built into the system to help identify and remedy unanticipated negative effects?

Despite all of our efforts in considering multiple scenarios and factors, there is always the possibility that something may be overlooked. That is why this question asks developers about any mechanisms imbued into the system that seek to identify negative effects that have been overlooked. This usually involves some form of measurement of relevant indicators of negative trends, even if their causes are unknown. Uncertainty is ubiquitous in sociotechnical systems, so it is often worth considering the possibility that many things may have been overlooked.

4.1.3.4
What ethical scenarios have I used to evaluate the system?

This section's ethical question focuses on the scenarios that developers may have considered when during the development process. As Johnson explains in Moral Imagination (Johnson, 1993), the ability to envision scenarios and simulate what occurs in them is essential to much of ethical deliberation. It allows us better reflect on the potential, situated consequences of our system's use. In addition to their relevance for ethical deliberation, scenarios are also usually involved in most software development processes. They allow developers to consider some prototypical situations in which their systems may be used. As such, these serve as an ideal setting for them to test whether their system actually works as intended. Therefore, considering these scenarios has the dual benefit of not only serving to ensure the system's robustness through testing, but also of determining the main ethical situations envisioned and reflected upon during development.

4.1.4
Continuous, post-deployment evaluation and monitoring

Once the system is built, it must then be tested to ensure that it works as expected, as best practices dictate. Given how the system's use can change over time, especially given the possibility that users appropriate the system and figure out new ways of interacting with it, monitoring its continuous use is also necessary. This is especially true for sociotechnical analysis, given how social dynamics are ever-changing, along with their relations with technical artifacts. This is the focus of this section's questions, which serve as the end point for our proposed extension.

4.1.4.1
How much of my vision is reflected in the system's actual use?

The first question asks about the fit between expectations and reality. Even during the Implementation stage, where some iterative evaluation is conducted, the system is not actually deployed in the real world. Once it is, the way in which the various stakeholders interact with it can significantly deviate from what was expected and identified during the earlier stages of the development process. As such, analyzing the differences between what was expected to happen and what actually did can serve to inform developers about how and how much the consequences of the system's use may shift as well.

4.1.4.2
What unanticipated uses have been made? By whom? Why?

As was frequently mentioned, appropriation by stakeholders can result in various types of unanticipated uses. Developers are asked to identify not only what these were, but also who was involved, and why they took place. Unanticipated uses are not necessarily a bad thing, but they are worth paying attention to since they may lead to unexpected outcomes. These can also vary according to context, with some uses being made in certain situations and not in others. Mapping out these possibilities is essential for developers to ensure that no significant negative consequences arise.

4.1.4.3
What anticipated and unanticipated effects have resulted from its use? Whom do they affect? Why?

Since, at this stage, the system is already in use, developers can start to observe some of its effects. Some may have been anticipated, while others were unexpected. In any case, being aware of the consequences of the system's use is necessary to ensure that no harm is being done. In addition to the effects themselves, developers are also asked to map out the individuals being affected and the causes behind it. Articulating all of this information can allow them to have a more holistic understanding of the impacts that their system is having on stakeholders. It is also a part of their ethical responsibility to be aware of them, since they were the agents who built the system that is now affecting the stakeholders, who are the patients in the situation.

4.1.4.4
What ethical issues need to be handled through system redesign, redevelopment, policy changes, or even decommissioning?

Finally, developers are asked about any ethical issues identified during this process of evaluation and monitoring and how they would go about fixing them. Just being aware of harmful impacts is insufficient, ethically speaking. They also have a responsibility to fix what they can. The question itself already offers some possible types of interference that can be made. Developers could try and conceptually redesign the system, redevelop it, possibly fixing any eventual design or implementation errors, create policies surrounding its use that change how stakeholders interact with it, or even, in the most drastic cases, decommission the system as a whole. For every issue they were made aware of, developers are tasked with deciding on which approach to rely on. Doing so, they may adapt the system to the circumstances identified, ensuring

that they are being responsible towards their stakeholders who, after all, are the ones for whom the system was created.

4.2 Rationale and Framing

When conceiving our extension of the metacommunication template (MT) we first analyzed some of the original template's key characteristics. Our analysis was focused specifically on the reflexive processes involved in instantiating a metacommunication message based on the template, since it is this process that our proposed extension is meant to assist.

One of the main aspects of the original MT is its linguistic framing, representing the developer's intentions in natural language. This can constitute a significant deviation from standard practice in Computing, where most documentation employs more technical jargon and specific notations. This use of technical language can make it difficult for individuals without significant technical expertise to be able to understand what is being said in these documents, limiting their participation, reflection, and critique (Luck, 2003). The MT, by employing more accessible language, can make it possible for a wider variety of stakeholders to be involved in the development process, making it easier for them to negotiate the meanings behind the artifact in question. If we consider a participatory design paradigm (Schuler and Namioka, 1993), where the system's conception should involve as diverse a set of stakeholders as possible, going beyond just the development team, this level of transparency and ease of understanding can be valuable. From a standpoint of simple creativity it can also be worthwhile to have multiple perspectives on the artifact's creation, possibly expanding the amount of scenarios considered, which might then lead to a more robust design.

In addition to employing more accessible language, the original MT also imposes a relational setting wherein a first person (the system's developer/designer) communicates with a second person (the system's user). Since design can be seen as relational in nature, with designers constructing artifacts that act upon stakeholders, this framing can help keep designers constantly aware of those who their actions will affect. This is essential for the promotion of responsible design, since forgetting about those affected by the potential impacts of the things being built may lead designers to overlook important issues, especially if they depend on how these stakeholders will behave. Within the context of software development, we also posit that this sort of framing can also be helpful in dismantling the notion that developers are simply solving technical problems rather than designing systems with social implications. If

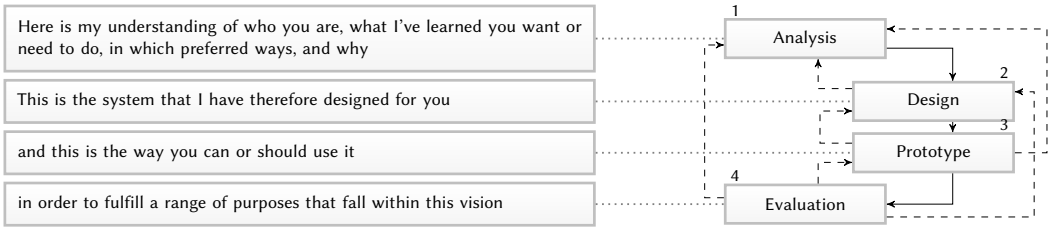

Figure 4.1: Correspondence between the sections of our extension and the original metacommunication template. (Barbosa et al., 2021)

they are always forced to face the individuals involved with their system, they may start to accept that their behavior cannot be entirely foreseen and therefore requires the acceptance of a significant level of uncertainty, which is often overlooked.

The MT itself is also based on traditional lifecycles of the design and development process, which we initially outlined in our FAccT paper (Barbosa et al., 2021). It can be broken down into four stages: analysis, design, implementation, and evaluation and monitoring. These also relate to specific snippets of the original metacommunication template, as seen in figure 4.1. In each of these stages, developers have to consider a multitude of factors and make various decisions. The goal of our proposed intervention is then to call their attention to some of the most crucial issues related to the sociotechnical dimension of the artifact under construction.

In order to try and achieve these goals, we devised a set of guiding questions that ask developers about their understanding of the situation they are designing for and the decisions they made based on these assumptions. All of the questions were formulated from the point of view of the designer, as the first person, in relation to a stakeholder, as the second person. Hence, when asked about their knowledge about the stakeholder, for example, the designer would have to ask themselves "What do I, or don't I, know about you?". By presenting the questions in such a way we try and take full advantage of the existing first person, second person framing that already exists in the original Template. Another potential added benefit is epistemic. By asking themselves about their own understandings, designers could then engage in a process of self-reflection, critically analyzing their presuppositions and possibly recognizing gaps or misunderstandings in their knowledge about the design situation.

Since the MT's representation of the design process can be separated in four stages (analysis, design, implementation, and evaluation), so can the guiding questions. Indeed, as was mentioned previously, we have directly conceived of the questions in relation to the stage in which they would be

most relevant. This leads us to having most questions about the designer's understanding of the situation being present in the "Analysis" stage, for example, while most questions about conceptual decisions would be a part of the "Design" section. Not only were the design questions tailored to the stage of the design process they were related to, but also the ethical questions. Going back to the example of the "Analysis" stage, since it deals with the designer's knowledge, or perceived knowledge, about the context they were designing for, the related ethical question directly touches on the ethical implications involved in what they know, or don't know, about the situation and those involved. By structuring these questions according to the stages of the development process, we seek to provide a comprehensive and holistic view of developers' intentions in building the artifacts.

However, in traditional development projects there are usually various individuals involved, with some participating in all of its stages and others only interfering in specific ones. This creates the problem of "many hands," as Coeckelbergh (2020b) put it, where it becomes unclear where the responsibility lies for the consequences of certain decisions. Since there are multiple individuals involved, each with their own conception of what the artifact they are building is, it is to be expected that there will be some level of disagreement about the artifact's nature. Of course, developers may not even be aware of such disagreements, believing that they view the artifact under construction in the same way. By having the individuals involved in the design process express their vision of the artifact according to our extension of the Metacommunication Template, it is our hope that some of these differences of interpretation can come to light and be dealt with appropriately.

Having acknowledged that individuals may interpret the system in different ways, developers then have the opportunity to discuss these differences in order to try and find a more consolidated and unified view. This is where the natural language framing can come in handy. Since the Template allows for the representation of the designer's intentions in plain language, it opens the door for discussions that are less hindered by differences in technical know-how, allowing for more stakeholders to be involved in the negotiation of these meanings. The same would not be possible if these discussions were to occur in terms of a technical notation of a modeling language, for example, of which most stakeholders know little about. As we have discussed previously, there can be multiple advantages to having as broad a set of individuals involved in these negotiations as possible, such as taking advantage of the multiplicity of perspective and the collective creative potential, both of which might lead to a more robust design.

Even though we have not yet sought to propose a method for these negotiations to take place, existing work on the dynamics of negotiation may already serve to fill this gap, somewhat. Regardless of how, if those involved with the computing system's development were able to go from a distributed view of what was being built to a more unified one, responsibility for eventual outcomes could be shared between those involved. Of course, this is a bit idealistic, since real-world development situations are filled with political dynamics to be considered which may result in an unbalanced process of negotiation where some voices matter more than others. There is probably nothing that a design tool can do with this respect. However, even in ideal circumstances where earnest negotiation is possible, without the appropriate tools to represent the developers' interpretations of the situation and the artifact, these types of discussions on the meanings involved might not be successful or even occur. That is why we propose this extension to the MT, even as we recognize that certain conditions, such as the freedom to engage in earnest reflection, need to be met for its successful use.

4.3 Support for Ethical Deliberation

Another key characteristic of our proposed extension of the metacommunication template is its potential for ethical deliberation on design decisions. Especially when discussing systems as sociotechnical in nature, there are numerous ethical issues that are often overlooked in traditional design processes. Hence the creation of more targeted efforts to research how to build software more responsibly, as is the case with the creation of the FAccT conference. The template's conversational, and therefore relational, framing, connecting the first person and the second person, is essential in this regard.

Most ethical issues can be understood as relational in nature (Coeckelbergh, 2020a). Agents act upon patients and are thereby liable to them. The same occurs within design in general, and software development more specifically. Developers analyze a design situation and its requirements and then go about building solutions that will impact not only those directly interacting with the system but also other around them. As such, it can be important to reflect on the possible ethical connotations of the design decisions that go into a piece of software, connecting some of their social implications with the technical decisions that create them. Our extension of the Metacommunication Template maintains the original Template's relational framing, which positions the Metacommunication Message as one written in the first person, by the designer, for a second person, the user. By maintaining this framing, the

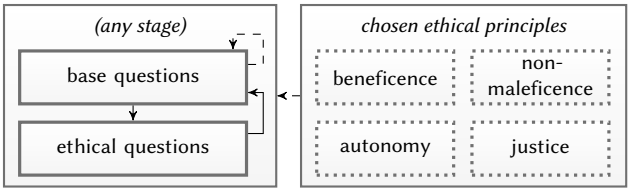

Figure 4.2: Relationship between the EMT's base and ethical questions. (Barbosa et al., 2021)

Metacommunication Template and our extension of it try to keep developers aware of those that their systems will impact, going beyond just the problem-solving mindset that often takes place within software development processes. Certain design personas, for example, can serve a similar purpose, even though we specifically try and maintain this awareness at the moment where those creating the system are trying to represent their intentions and understandings in natural language. Of course, there are multiple relationships to be considered, with the situation almost never involving only a single designer or developer and a single stakeholder.

Trying to make this process of ethical deliberation even more explicit, we also added questions specifically geared towards the topic at the end of each section. These would always relate to the previous design questions presented earlier in the same section, leading those formulating their Metacommunication Message to go back and reflect on what they stated earlier, as seen in figure 4.2. We thought this to be crucial not only to promote ethical reflections during the process of trying to answer them, but also to bring awareness that all of the questions have their own ethical implications that can be considered even before reaching the final ethical questions. After all, as we have posited, all of the design process has some ethical dimension to be considered.

In terms of where we placed these questions about ethical implications, we chose to put them at the end of each section to push developers to go back and look at their previous statements about their design intentions and reflect on them. This has an added benefit of forcing them to read what they wrote earlier and thereby play the role of consumers of the sentences they produced. In doing so, they can reflect on the meaning behind what was written, possibly leading them to identify that what was written was not exactly what was meant, which already serves as a sign that some level of semantic analysis is being conducted, even if they are unaware of it. This directly serves the purpose of this proposed extension by connecting the analyses of the meanings involved with the ethical implications behind them. Having read what was previously stated and reflected on its meaning, designers would then direct their attention to any ethical issues they could find, deliberating on whether their design

choices were appropriate or whether they required some adjustment.

As we have already mentioned when discussing the design-focused questions, ethically-focused questions are also tailored to the stage of the design process they reside in. Of course, in answering these questions, developers may end up reflecting on questions of other sections as well, due to the non-linear nature of most design processes. By having these ethical questions closer to the design questions they most resemble, instead of at the end of the set of guiding questions as a whole, we sought to also promote ethical reflection throughout the process, rather than it being a process that only starts after all of the relevant design decisions were made.

Instead of relying solely on the individuals' preexisting capabilities for ethical reflection, we also phrased these questions in such a way as to allow for the use of ethical frameworks to aid with their considerations. Especially for those that are not used to reflecting on such issues, using well-established theories, such as ethical principles, virtues, etc., can help them with framing the situation they are reflecting on in different ways, as Schön (1979) discussed in The Reflective Practitioner. More frameworks might then mean more frames and lenses to be applied to the specific situation, allowing one to have a more comprehensive view of the situation and the artifact being proposed.

They can also serve a creative purpose, since they may provide options that developers might never have considered in terms of the possible ethical conditions, as is argued by Johnson (1993). This can be further expanded with the use of multiple ethical frameworks, which can create a wider set of source analogs for the developers' imaginative processes involved in their ethical reflection (Holyoak and Thagard, 1996). However, it is worth noting that these existing frameworks ought only work as starting points, with the developers looking to them when unable to conceive of more ethical scenarios but never being constrained to only considering those that fit neatly with the frameworks themselves.

In addition to having multiple frameworks involved, it may also be essential to have a diverse set of individuals involved. During reflective processes, there is a significant imaginative component, as argued by Johnson. These creative processes often require individuals to draw from past experiences to serve as source analogs for the new situations they conceive of, as is explained by Holyoak and Thagard (1996). Individuals from different cultural backgrounds can bring their different sets of values to the analysis, which can significantly impact the conclusions reached in the deliberation process. As such, multiple individuals, preferably from diverse backgrounds, involved in these discussions can help ensure that the set of scenarios considered is wider than it would be

if generated by a single individual.

Of course, having multiple individuals involved also presents new challenges. As discussed in the previous section, there tends to be a significant political component to most software development processes, with some peoples' voices having a greater impact than others. As such, the benefits brought about by having a diverse set of individuals, for example, may end up not occurring if only individuals of a certain group get to make decisions. Interpretative differences may also come up, even if the Extended Metacommunication Template provides a form of common ground for these discussions to take place between individuals with differing levels of technical expertise. Sometimes people just see the world differently, which then can then require a significant level of negotiation to reach a consolidated position. This especially true with ethical issues, where individuals' specific moral values can come into play in a significant way.

To sum up our desired ethical contributions with our extension of the Metacommunication Template, designers must first be able to conceive of some of the most salient ethical implications of their design. For this, our extension provides them with a set of guiding questions that can lead them to revisit their answers to previous questions, reflect on their meanings, and identify possible ethical issues. Beyond this aspect of individual deliberation, we also have collective dynamics to consider. Most software development processes are not conducted by singular individuals, after all. In trying to reach collective consensus as to the ethical issues involved, developers will probably need to engage in some level of discussion and negotiation on the meanings involved and what their final decisions should be. Given that there may be individuals with differing levels of technical expertise within the group, the Metacommunication Template, with its natural language representation, provides them with a common ground for these discussions, even if it cannot ensure that consensus will be reached. With these interventions, we seek to stimulate individual reflection and collective discussion on some of the ethical connotations of the design decisions that make up these sociotechnical computing systems.

5
Evaluating the Metacommunication Template

Now that we have presented and discussed our proposed extension of the Metacommunication Template, we can move on to how we propose to evaluate its use in design processes. To do so, we have devised a study based on a speculative design process where the Metacommunication Template is used to represent the designer's intentions. In order to contrast our questions' impacts with other design tools, we also utilize the Model Cards as alternative design tools, comparing the effects of both.

In this chapter, we present our proposed speculative design study, analyzing its design, the materials used in it, and the procedure we intend to follow in examining the resulting data.

5.1
Speculative Design Study

Our study seeks to analyze how the use of the Extended Metacommunication Template can impact the considerations made during design and development process of a Machine Learning (ML) system. To do so, we opted for a speculative design scenario so as to free the participants from pragmatic constraints, allowing for a wider range of considerations (Auger, 2013). Real development scenarios can present constraints for such an analysis, such as disincentivizing reflections on activities that are not ongoing. In speculative design studies there is a risk, however, that the resulting design be too imaginative, based on assumptions that are unreasonable and would not hold in reality. It then becomes imperative for the researcher to ensure that participants stay within the realm of possibility during their speculations (Malpass, 2013).

Critical methods, such as speculative design sessions, could be well-suited for the Extended Metacommunication Template, since they engage participants in explicit reflection on the validity of the information in front of their eyes. The Extended Metacommunication Template seeks to do something similar in regard to the coherence of the design being represented. If the underlying assumptions for a given decision are deemed invalid, then so might the decision itself be. This constant process of critically analyzing

the presuppositions and preconditions expressed in the Metacommunication Template is an essential part of the guiding questions, seeing as they, especially the ethical ones, constantly direct designers to revisit their previous statements and reflect on their meanings (Morrow and Brown, 1994; Strydom, 2011).

5.1.1 Study Design

In terms of study setup, we have chosen to adopt a within-subject design, wherein each participant is exposed to both the Extended Metacommunication Template and the Model Card, which we have chosen as a base of comparison given its role in assisting developers in the creation of ML systems. The order in which these tools were presented is controlled so as to limit ordering effects (Day et al., 2012). Each of the study's groups can be seen in figure 5.1.

Participants take part in two speculative design sessions, one with each of the tools proposed. In them, they are presented with a design brief explaining what they are tasked with designing, as well as relevant contextual information. These scenarios are further explored in sub-section 5.1.2.

The use of multiple design scenarios is necessary due to the possible learning effects that may be associated with the proposed process (Lazar, 2017). A participant that engaged in a speculative design session about a certain context with the Extended Metacommunication Template might just repeat their previous findings when asked to follow the same procedure, with the same context, but with the Model Card. Since we want to minimize this learning effect, we provide them with different scenarios for each tool so as to avoid the situation described above.

In terms of how we recruit participants for the study, we use a convenience sample (Clark, 2017), given our institutional proximity with individuals

Study Group	Tool A	Scenario A	Tool B	Scenario B
1	Model Card	Education scenario	Extended Metacommunication Template	Financial scenario
2	Model Card	Financial scenario	Extended Metacommunication Template	Education scenario
3	Extended Metacommunication Template	Education scenario	Model Card	Financial scenario
4	Extended Metacommunication Template	Financial scenario	Model Card	Education scenario

Figure 5.1: Study groups according to design tool and scenario used.

experienced in the design and development of Machine Learning systems. We do not plan to filter participants according to any demographic variable, only collecting this information to inform our later analysis. All that we ask is that participants have at least an undergraduate-level understanding of Machine Learning systems, being aware of their dependence on data, the general dynamics of model training and testing, as well as the types of tasks for which they could be used.

Given the currently ongoing COVID-19 pandemic, and the social distancing restrictions that come along with it, the study will be conducted remotely. This approach presents new opportunities, as well as new risks (Bolt and Tulathimutte, 2010). On the positive side, conducting these types of remote user studies can facilitate the scheduling process, by reducing transportation times; participant recruitment, given the possibility of participating in the study from the comfort of their own homes; and data collection, reducing the amount of extra equipment involved in the study, such as multiple cameras and audio recorders. On the downside, it creates a dependence on the participant's own equipment and circumstance, in the form of a functioning computer with a stable internet connection and a microphone; as well as the researcher's. If internet connections cut off, for example, maintaining a study session going can be difficult, having to rely on back-up plans, like the use of mobile internet connections. However, given the aforementioned restrictions, we find that this is the only available approach for this type of synchronous user study at the moment.

The design sessions are held via Google Meet. Pending the participant's authorization, they are also recorded, capturing both screen footage and the meeting's audio. In order to register what is taking place on the participant's side, we ask that they share their screen in the video conferencing website so that the researcher can see it and record it. This way, we can also relate what is being said to what is actually taking place on the screen.

Before the first speculative design session, researchers explain the study's terms and conditions to the participant via an informed consent form, asking for their approval and permission to start recording. This is standard practice when conducting user studies (Lazar, 2017). These terms contain the main risks and benefits to the study's participant, as well as any mitigating strategies adopted by the researchers involved. In our case, the main concern pertains to the participants' identities, which should be kept private from anyone that is not directly involved with the study. As such, video and audio are transcribed so as to protect the participants' anonymity. This study's informed consent form can be found on Appendix A.1.

Once approval is given, researchers then ask participants a few questions about their background information. These range from topics such as how experienced they are with Machine Learning technology to whether they were familiar with the tools they would be using in the study. This information is important so that we better understand their profile and are able to relate it to the resulting study data. For example, a more experienced Machine Learning practitioner might be able to conceive of certain things that a novice would not. This sort of finding would only be possible if we knew their level of experience beforehand and is why we have devised the profile questions, which are further explored in subsection 5.1.2.

During the speculative design sessions, researchers ask participants to follow the think-aloud technique, voicing their thoughts so as to be captured by the audio recording (Lazar, 2017). This information can be valuable as it can express some of their efforts in interpreting the guiding questions, relating it to scenarios, and various other processes.

After each of the speculative design sessions, researchers interview the participants on their experiences in them. This is another crucial source of information, as it allows us to probe into their retrospective perceptions of the tool. After the second, and final, session, the interviews also contain comparative questions that ask participants to contrast their experiences with each of the tools involved in the study. In addition, we also introduce leading questions that induce certain topics of conversation that may not have naturally arisen in the more contained questions in the interview script. We only do so after the final session so as to not interfere with the results of the sessions themselves.

The resulting study procedure can then be represented in figure 5.2, which represents each of the steps involved. The tools used in each of the sessions depend on which study group the participant belongs to, as outlined in figure 5.1.

Having discussed the procedure we will follow, we can now move on to discussing the materials used in the study in greater detail.

5.1.2 Study Materials

As mentioned previously, this study depends on various materials to allow for the participant's successful engagement with the tools presented to them. Namely, it requires:

- Design tools (Extended Metacommunication Template and Model Card, one in each session)

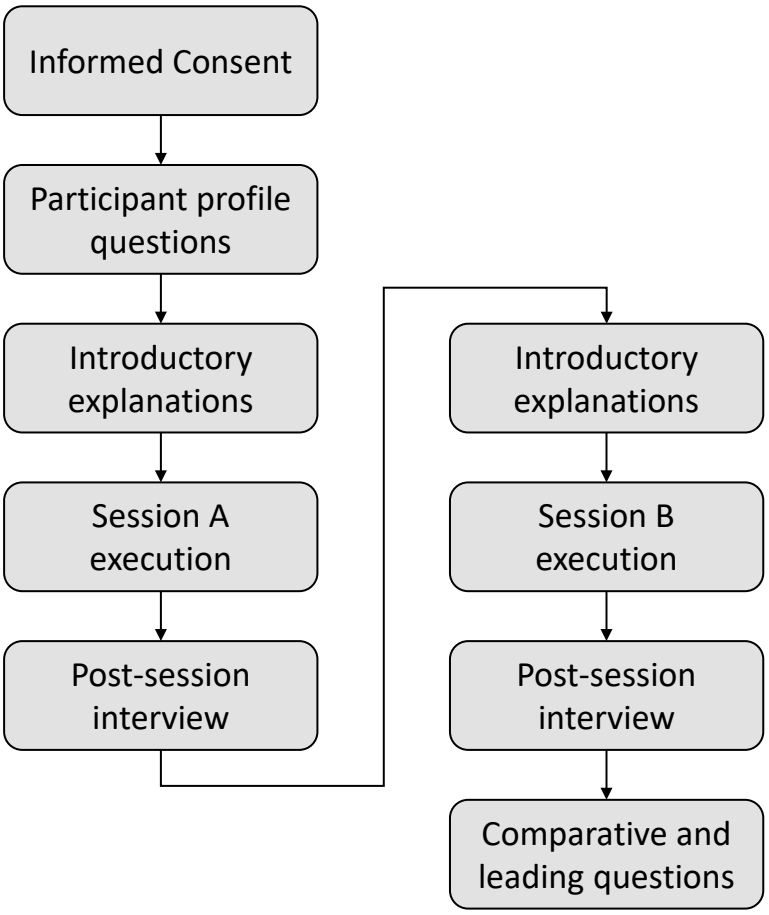

Figure 5.2: Flowchart representing study procedure.

- Design briefs for both scenarios
- Summary of the Bioethical Principles
- Interview Scripts (before and after each speculative session and at the end of the study)

It is important that we discuss each of these materials in depth since they may influence the study results. Interview questions, for example, should not induce answers in the participants, thereby making them biased in one way or another. Our goal in creating these documents is to allow for the speculative design sessions to occur without reducing the validity of the data obtained. For this discussion, we analyze each of them individually.

5.1.2.1 Design Tools

Since this study seeks to compare the effects that our proposed extension of the Metacommunication Template and the Model Cards have, then they obviously make up the main materials involved in the procedure itself. They

are the documents that participants will fill out, expressing their design vision. We first need to discuss how this should take place. Then, given that we have already discussed our proposed extension's guiding questions in depth, here we mostly discuss the Model Cards structure and how participants go about expression their visions in it.

As we have discussed, our study is to be conducted remotely due to the currently ongoing COVID-19 pandemic. As such, participants need to be able to edit these documents online. We create multiple versions (one for each participant) of these documents as Google Docs and give participants permission to edit them during the session. This way, the study can be conducted safely, and possibly even be more comfortably. Of course, this is not without risks, as the participants' connections may be unstable, making it difficult for them to interact with the online document. However, restricted internet connections are a risk that all online studies face to some degree. Such is the nature of remote studies.

Moving on, we can now have a more in-depth discussion on the Model Cards' structure and how participants may engage with it during the speculative design sessions.

Model Cards are not structured in the form of questions, like our proposed extension of the metacommunication template is. Instead, they outline a set of dimensions to consider and document about various aspects of the model's creation and use. They also do not follow any specific process for documentation, with the possibility of filling the fields in any given order. This is different from what should occur with the Extended Metacommunication Template, where answers should be provided sequentially.

In terms of fields that need to be filled, the Model Card contains the following:

- *Model Details*: Basic information about the model, including date of creation, type of model utilized, licenses, and other related information.
- *Intended Use*: Use cases envisioned during model development pertaining to the users involved, the uses themselves, and uses that are out-of-scope.
- *Factors*: Contextual factors to be considered about the model, such as demographic information, technical attributes, and various others.
- *Metrics*: Metrics chosen to represent the real-word impacts of the model.
- *Evaluation Data*: Details on datasets utilized for quantitative analysis regarding the information contained on the card.

- *Training Data*: Information about the dataset utilized in model training. May not be possible in various cases due to intellectual property restrictions.
- *Quantitative Analysis*: Methodology utilized for the quantitative analysis of model performance, according to the chosen metrics.
- *Ethical Considerations*: Ethical considerations that went into the model's development, including any mitigation strategies utilized.
- *Caveats and Recommendations*: Additional concerns that were not covered in previous sections.

As we can see, the Model Card is shorter than the Extended Metacommunication Template, containing 9 fields in contrast to the Template's 19 questions. This, however, does not ensure that working with one of the tools will take less time than with the other, since the Model Card's fields are expansive in their scope, possibly keeping the designer engaged for a long duration. This is one of the details that we plan to look into in our study, relating the kind and depth of reflection with the time taken to execute it.

The Model Card does not require that developers fill its fields in any specific format. They may choose to write their considerations about a specific criterion in the form of a paragraph or they may choose to separate it into various points. Both are acceptable in regard to the Model Card. In contrast, the Extended Metacommunication Template requires that developers maintain the linguistic framing it presents them with, as it is essential to the epistemic process that it seeks to induce in them.

In terms of the subject matter involved in the Model Card, we can notice that it also maintains a sociotechnical perspective on the Machine Learning models' development. It goes beyond purely technical details, asking developers to consider the contexts in which these models will be inserted, how they will be used and who will be using them, among other type of information that tie in the technical decisions made about the model with their potential social impacts.

Despite also possessing a field that mentions ethical considerations, the Model Card does not seek to assist developers in their reflective process. This is a key difference between it and the Extended Metacommunication Template. Our guiding questions explicitly draw the developers' focus to aspects where ethical issues may exist, possibly leading them to identify some that would not otherwise come to them with a less guided model.

Despite their differences, we argue that there are sufficient similarities between the Model Card and the Extended Metacommunication Template for

them to serve as bases for mutual comparison. They occupy similar roles in the design and development process, consider similar information, and may be used by people in similar roles.

Having understood how people may use the Model Card and why we have chosen it as a basis for comparison with the Extended Metacommunication Template, we can now proceed to discuss the design briefs provided to participants in conjunction with both these tools.

5.1.2.2 Design Briefs

In order for the speculative design sessions to occur, participants must first understand what they are being tasked with designing. That is why we present them with design briefs that outline some of the key information that they will need to take into consideration during the session. These contain information such as what the system's goals are, who the stakeholders are, what are some of the resources available, in which context their proposed solution will be inserted, and various others.

Both scenarios directly refer to situations with salient ethical implications. We chose to focus on these to facilitate the ethical reflection involved in the speculative design session. Domains that require more nuance to identify these types of issues might be unsuited for this sort of study, since it would take a longer amount of time to try and find them. Despite the ethical connotations being quite salient, there is no explicit mention of any issues in the briefing itself. Participants must still be able to conceive of them solely based on contextual information.

Our first brief asks participants to design and develop a Machine Learning system tasked with evaluating loan applications for a bank. It must consider the applicant's profile and financial history to establish the level of risk they represent to the financial institution. In it, the clients have defined that the overall goal of the system should be to maximize profits. This scenario was described as follows:

> *You are the leader of a development team hired by a financial institution. Your team must develop an AI system to make decisions about loan requests made to the institution that hired you. Your algorithm should analyze the risk of each potential client based on their profile and financial history, deciding whether or not to approve the loan, with the objective of maximizing the company's expected profit.*

Our second brief asks participants to conceive a system, based on a Machine Learning model, which rates college applicants. It should consider the applicant's past achievements and educational history to try and establish how qualified they would be in relation to other applicants. In this case, the final decision will not be the system's, but rather a selection committee that will make it based on the model's recommendations. This scenario's specific formulation can be seen in the quote below:

> *You are the leader of a development team hired by a university. Your team must develop an AI system to attribute grades to candidates to the university's programs. Your algorithm must analyze their previous accomplishments and attribute to each a grade that will be used by the selection committee to decide which students will be admitted or not. You algorithm will not be used to give them grades after they have been accepted into the university's programs.*

Despite their different domains, both scenarios can be seen as very similar. Both of them ask for some form of rating algorithm. In both cases, applicants would have to submit information that might then be fed to the algorithm, in order for it to reach its final decision. However, despite these similarities, there are still many ways to go about it. Participants may opt for regression algorithms, resulting in a quantitative score, or build classification models that rate participants into discrete categories, for example. This is the balance we try to maintain: sufficient similarity for both scenarios to be comparable while still allowing for various design strategies and ethical reflection.

5.1.2.3 Summary of the Bioethical Principles

In addition to the design tools and briefs, participants are also offered a summary of the Bioethical Principles. As we have mentioned in chapter 4, the Extended Metacommunication Template allows for the joint use of preexisting ethical frameworks in order to support developers' ethical deliberations.

In terms of possible ethical frameworks, we have opted to present participants with the Bioethical Principles based on the recommendation of Floridi and Cowls (2021). During their analysis of various proposals for the ethical development of Artificial Intelligence systems, they found that most could be decomposed into the original four Bioethical Principles (Beauchamp et al., 2001) of Autonomy, Beneficence, Non-Maleficence, and Justice, with the addition of a final principle of Explicability. In our work, we re-interpret

this final principle as the demand that stakeholders are able to understand how a system functions and how it was conceived. We then see this as an issue of Communicability, which already underlies our proposed extension of the Metacommunication Template. Given this, we only present the four core Bioethical Principles to participants.

In our orientation at the end of the first session, we provide them with short explanations of the meaning of the four Bioethical Principles. Of course, it is up to them to interpret this framework within the context of their design scenario, appropriating and contextualizing it. The definitions we present them are as follows:

1. *Autonomy*: It is the duty to respect individuals' capacity to think, decide, and act based on their own thoughts and decisions, with freedom and independence.

2. *Beneficence*: It is the duty to actively do what is best for the patient, based on the actor's understanding.

3. *Non-Maleficence*: It is the duty not to do harm to the patient (stakeholder), based on the actor's understanding.

4. *Justice*: It is the duty to ensure that all patients are treated equitably, receiving the same level of quality in their treatment.

Participants will not be obligated to use this framework or follow these rules. They serve only as a source of inspiration for their own ethical reflections throughout. Indeed, there may even be certain requirements that clash with some of these principles, such as the first scenario's stated objective of maximizing profit. By considering this ethical framework, they may become aware of the multiple trade-offs that usually exist in a design situation and their ethical impacts.

Having explained how we planned to present the Bioethical Principles to the study participants, we have now finished analyzing the tools that will be necessary for the speculative design sessions. Moving on, we now discuss the interview scripts, which serve as our main data collection instruments.

5.1.2.4 Pre-session Interview Script

Before the speculative design sessions begin, researchers ask participants questions about their profile. This includes their previous experience with Machine Learning models, Semiotic Engineering, among other information. We

need to know these things in order to better contextualize the data resulting from their speculative design sessions.

In order to ensure that the same information is gathered from all participants, we have therefore built a script that will be followed by all researchers involved in the study. The questions in this script are:

1. *What is the area you are currently working on?*
2. *What is your level of education?*
3. *For how long have you worked in the development of Machine Learning models?*
4. *What is the process that you usually follow when developing Machine Learning models?*
5. *What is you previous level of knowledge about Semiotic Engineering?*
6. *Do you possess any previous experience with Model Cards or the Extended Metacommunication Template?*
7. *Are you interested in ethical problems in software design and development?*
8. *Do you possess any previous experience in considering ethical problems in software design and development?*
9. *When you think about ethical problems in software development, what comes to mind?*
10. *Have you had any previous experience with software that you consider unethical? What was that experience like?*

Overall, with this set of questions, we will be able to better understand the participants' backgrounds, how experienced they are with some of the concepts involved in the speculative design session, and how invested they have been in ethical deliberations in software development as a whole. These kinds of information are essential for our framing of the study's end results. For example, it would be relevant to figure out whether those that were involved in more in-depth ethical reflections were also those who previously claimed interest in the topic. With these questions, we believe that we will be able to draw these types of connections.

5.1.2.5 Post-session Interview Script

After each speculative design session is concluded, researchers then conduct another structured interview to get their perceptions on the experience as a whole. Some of them go beyond the experiences themselves and look towards real-world use. This information will complement their utterances during the session, which they are prompted by the researcher to provide, following the "think-aloud" technique (Lazar, 2017).

The same interview script will be followed regardless of which tool is involved in the session. It has been built so that it is applicable to both, focusing on the aspects that we believe will matter the most to our comparison. Having the same script for both sessions also facilitates our comparisons between the answers given for either tool.

The questions that will be asked after every speculative design session are:

1. *What did you think about the design tool?*
2. *How did the design tool influence your ethical reflection about the scenario presented? How did it aid you reflection? How did it limit it?*
3. *In a scale from 1 to 7, how would rate the design tool in relation to how it aided your reflection?*
4. *Was there anything that you included in the design tool that did not fit well in any of its sections? If so, what?*
5. *Was there anything that you would have liked to have included in the document but did not fit well in any of the design tool's sections? If so, what?*
6. *What did the design tool help you reflect on?*
7. *How did the design tool help you consider the consequences of what you were developing for the scenario proposed?*
8. *Which sections helped you reflect on the potential ethical issues?*
9. *Who did you consider would read what you were writing?*
10. *What other individuals affected by the model do you believe would benefit from having access to the resulting document?*
11. *What other individuals affected by the model do you believe would benefit from working with this design tool?*

As we can see, this set of questions is quite varied, inquiring about several aspects of the participant's experience with the design tool in addition to their thoughts on how they could be used in real development situations. Given how our study is focused on the ethical deliberations involved in the development of sociotechnical systems, these questions reflect that focus.

5.1.2.6 Final session Interview Script

After the second, and final, speculative design session, researchers ask additional questions beyond those presented in script above. This is due to the fact that it is then possible to ask them to explicitly compare their experiences with both tools. As such, we have defined the following additional questions:

1. *How would explain the differences between the design tools in a simple way to a new member of your development team?*
2. *How would you compare the tools in respect to the considerations you made in regards to the design scenarios? Why?*
3. *Considering that you have used the other design tool before using this one, how do you think that this session was influenced by that previous experience?*

Finally, we also devised a set of more leading questions that should only be asked when their topics were not brought up by the participants themselves. These would also only be applicable after the final design session, so as to not influence the answers to the previously asked questions. These final questions are:

1. *What did you think about the "I-You" formulation presented in the Metacommunication Template?*
2. *How do you think having to write down your design decisions impact your reflection?*
3. *How do you think that the structure of the questions, following the different stages and aspects of the development process, impacted your reflection?*
4. *How important were the Bioethical Principles to your reflection?*
5. *How did you feel in relation to the design scenario after your reflections?*
6. *How did the field about Ethical Considerations in the Model Card impact your reflection in the session?*

This final set of inductive questions is useful since it allows us to bring up certain topics that may not have been brought up in previous questions. Asking these questions after the fact means that such an influence is no longer possible and we can now discuss certain concepts with no fear of creating bias in our qualitative data. However, the analysis of all of the answers obtained in these questions should have the caveat of only having been raised after explicit probing. Were these topics brought up by participants themselves in previous questions, they would be considered even more relevant, as it would have occurred without them being explicitly mentioned. We just have to be careful not to equate the two phenomena.

Now that we have discussed our study's interview scripts, which serve as the main way for us to collect qualitative data, we can now move on to discussing how we are going to evaluate it.

5.1.3 Evaluation Methodology

Having already discussed the materials used in the study, we can have an idea of the data that will be available to us for analysis. Almost all of it will be in the form of subjective answers to our questions and their utterances during the design sessions. Only a single question takes on the form of a quantitative rating. Given this focus on qualitative evidence, we will opt to use qualitative methods and tools to evaluate the design tools that we are studying.

Since the speculative design sessions will be recorded in video and audio, our first step will be to transcribe them. This is essential for enabling us to better analyze what is said, but also to ensure the participants' anonymity. This transcription process will occur manually, without the use of any automated transcription software. In terms of transcription strategy, researchers will focus on fidelity, trying to keep the transcript as close to what was being said as possible, even if it means including long pauses, broken sentences, and grammatical errors. Once the transcription process is concluded, we can move on to analyzing the data.

Two researchers will be involved in the qualitative analysis process so as to try and ensure a less biased analysis of the results. Of course, qualitative methods are always subjective, as is the data itself (Marshall and Rossman, 2014). Having multiple individuals allows for better triangulation of codes and themes, possibly leading to a more robust analysis. Various steps will have them develop their analysis independently, while others may call for joint action.

To start our qualitative analysis of the data, both researchers will conduct an independent round of open coding (Saldaña, 2021). Looking at the data,

they will develop their own codes and attribute them to the sections of the data they believe they relate to. This will not be based on any pre-existing theory, solely reflecting their perceptions on the data. Of course, this does not mean that the researcher's perceptions themselves are not influenced by their previous experiences, appearing through the abductive reasoning behind their code creation (Lipscomb, 2012). In a general sense, both researchers will have an understanding of Machine Learning technology as sociotechnical artifacts, as well as of theories of design, like those presented in chapter 2.

In terms of how they will go about coding the data, both researchers will adopt the practice of using splitting codes. This practice consists of attributing codes to smaller sections of data, possibly having multiple codes for a given paragraph. It is the opposite of using lumping codes, which would attribute a single code to large swaths of data. They will also be allowed to use simultaneous coding, wherein a single part of the data may be attributed to multiple codes (Saldaña, 2021).

After the researchers have concluded coding all of the qualitative data, including both the Extended Metacommunication Template and the Model Cards, they will then get together and consolidate their codes into a single codebook. First they will try to match similar codes and decide on how to name the unified code. Any codes that do not have equivalent counterparts in the other researcher's codebook will be adopted into the consolidated codebook directly. That way, we will opt to have a more expansive set of codes, allowing for greater coding.

Once this consolidated set of codes is at hand, researchers will re-code all the data with it. For every code, they will register their justification why that section of the data should be coded that way. Doing so will be important for the next step.

Once they have finished going through all the data and re-coding it, the researchers will once again unify their codings. For sections that have no overlap, *i.e.*, only a single researcher coded them, they are included as they are. For sections that both researchers have coded, but with different codes, these disagreements will then be settled by both parties, looking at the rationales behind them. The justifications behind the final decisions should also be registered for future analysis. In the end, this process should result in a unified coding, with no major disagreements. Of course, the codings prior to this unification can also be subject to analysis further down the road. The entire coding process is illustrated in figure 5.3.

Once we possess a single unified coding, we can start to look for patterns in the data that may be indicative of significant themes. Sequences and co-

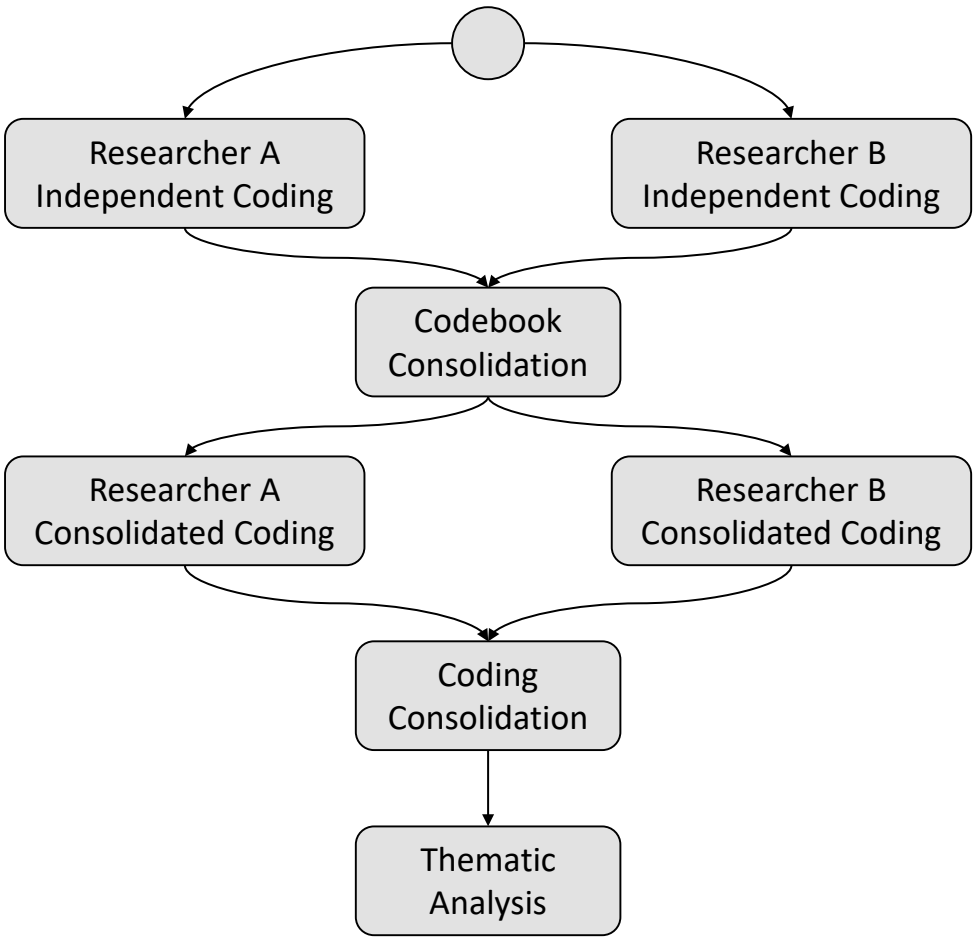

Figure 5.3: Qualitative coding procedure.

occurrences are just some of the patterns we will be looking for. We will also be able to separate the data according to the tools' structures, section by section, question by question. This separation may create new opportunities for analysis, such as co-occurrences within the same section, and other metrics.

Having found patterns in the data, we will go back and take a qualitative look at them. Patterns themselves still need to be interpreted and turned into conceptual themes. Both researchers will engage in this process together, rather than doing it independently and then getting together to consolidate.

Themes come in different levels of abstraction as well. We can identify lower-level themes, such as those represented by a single pattern, or higher-level themes, such as those represented by multiple trends in the data. Having a clear understanding of the level of abstraction we are dealing with is essential throughout the process of thematic analysis (Braun and Clarke, 2012).

In addition to finding these themes, we will also seek to describe the coding through various statistics and visualizations. This can help us get an overview of these patterns, possibly helping with the identification of the themes, as well. Code frequencies, code locations, and various other infor-

mation, can be used to calculate these metrics and create these visualizations. Adequately representing these patterns is paramount to ensuring that our qualitative study is transparent, so as to lend credibility to the themes we identify, as well.

Statistics and visualizations can also help us compare what happens in regard to each of the design tools. Going beyond merely contrasting which themes occurred with each tool, we can produce a more detailed analysis by looking at the lower-level codes. Another interesting possibility is to identify themes based on the contrasts between code occurrences for each of the tools. After all, one of the main interests of this study is in comparing the effects of the Model Card and the Extended Metacommunication Template.

Through this comprehensive analysis of the data, we seek to provide some insight into the potential impacts of the Extended Metacommunication Template's use, while also comparing it with the Model Card. Multiple researchers are needed to ensure the robustness of such a conceptual analysis, as well as multiple rounds of independent and joint coding. The subjectivity inherent to these qualitative studies is not something to be feared, so long as a reasonable level of transparency is involved, allowing readers to adequately frame the study's results.

Having executed the study, we describe the results obtained in the following chapter, including discussions on our participant sample, our researchers' positionality, and the consolidated codebook and semantic network we identified in the data.

6
Results

Having executed our study, we can now discuss the results we have collected. First, we must better understand the participants recruited for our study, characterizing them through the variables we have probed about during our interviews. Then, we ought to take a more inward look and analyze the positionality of the researchers who were involved in coding the data. Last, we describe the consolidated codebook that resulted from our open coding sections and the negotiated agreement of each code.

6.1
Participants

In our study, we sought to study how Machine Learning (ML) developers might use the Extended Metacommunication Template (EMT) to reflect upon their design. To this end, we sought to recruit people with some experience on the topic so that our findings are closer to what might happen with the tool's real-world use. During our study's run, we recruited 8 developers to participate in our speculative design sessions. Given that there were two sessions, one for the EMT and another for the Model Card (MC), with each lasting around an hour and a half, we were able to collect a considerable amount of data, totaling approximately 20 hours of recording. In order to better contextualize our findings, it is important that we first understand the profile of these participants and how their backgrounds may have led to certain reflections when using the EMT.

In terms of their ML expertise, all participants reported having at least an undergraduate-level understanding of how these models work and having participated in at least one project developing such models. There were still disparities within the group, with 5 of the participants (P1, P2, P4, P6, P5) having a post-graduate understanding of the topic. As such, we should keep in mind that some of our participants may have a greater technical understanding of these ML models that can impact the kinds of reflections they engage in.

Another key factor we asked our participants about was their expertise in Semiotic Engineering. This is especially important since the EMT was derived from this theory and might, therefore, be used differently by those

with greater familiarity with the theory. For example, it is possible that those who have participated in classes on Semiotic Engineering are more attentive to the communicative aspects of their design. As such, they might be able to engage in deeper reflections on their language use when recording their design in the EMT. Within our sample, 4 participants (P1, P3, P4, P5) reported having had classes on Semiotic Engineering, with one having heard of the EMT but never having engaged with it.

Finally, in terms of their interest in dealing with ethical issues, all but one (P8) of the participants related being interested in the subject. However, none of them reported having had to deal with them in any of their development projects. Despite being able to relate experiences that they may have had or cases they deemed unethical, none could claim to have professional experience in dealing with ethically sensitive systems, in their view. Given this, we ought to keep this lack of expertise in mind when analyzing the data collected via the EMT's use.

6.2 Coder Positionality

When reflecting on our findings, it is essential for us to take our positionality into account. How we perceive the data is influenced by our past experiences and the worldview they create. In our study, two researchers were responsible for coding the data; it is their background we must understand to accurately frame our analysis.

In terms of their identity, both of them are white men from a high socioeconomic class. This can indicate a possible blind spot towards issues related to the experiences of other genders and socioeconomic classes.

They were both educated in private schools and universities, with one of them having a bachelor's in Computer Engineering and the other one in Law. Both of them were graduate students in Computer Science at the time of the analysis. This difference in undergraduate backgrounds increased the diversity of considerations made while looking at the data. In addition, although the computer engineer had greater experience with qualitative methods, the lawyer was also experienced in the qualitative consideration of data, albeit not in a formal research setting.

The topic of responsible design of AI systems is at the core of both researchers' work. Besides, both of them took part in several academic discussions on the topic in multiple graduate courses.

Another factor that may have influenced our analysis was the coders' familiarity with the study's participants. All of them were graduate students

in the same Computer Science department. In some cases, they took the same courses or even worked together on research projects. As such, the researchers may have been influenced by their preconceived judgments about the participants when coding their transcripts. This effect may have been reduced by the anonymization of the transcripts, but since the coders were also the interviewers, it is entirely possible that they were able to recognize who the transcripts were from during the coding process.

Finally, it is also important that we consider the coders' relationship to the research context itself. This study was a core part of their graduate research and, as such, they may have felt pressured to produce interesting findings. Therefore, it is possible that some of the trends they identified were overstated. Of course, they were aware that they should resist these urges as best they could, but it is still worth noting that these pressures did exist.

6.3 Consolidated Codebook

As discussed in chapter 5, at the start of our analysis we conducted an open coding session of the data, with each of the two researchers being free to categorize utterances that they deemed relevant based on their understanding of the data in front of them. Through this, each would be able to develop their own framework of analysis by which to investigate any trends present in the transcripts of the speculative design sessions. This phase resulted in two independent codebooks, one with 77 codes and another with 55 codes.

Having our perspectives represented in two distinct codebooks, we then engaged in a session of codebook consolidation via negotiated agreement, wherein we would negotiate on which codes should be included in our finalized, consolidated set of codes. In doing so, we took advantage of our different understandings of the data we analyzed and tried to develop a set of instance categories that were useful to the aspects of the data we sought to investigate. For example, during this process we identified that the codes "*Consequence of Language*", "*Language Aspect in Communication*", and "*Focus on Semantics*" were very similar and could be consolidated into a generalized category of "*Designer's Focus on Language Use*". These were the kinds of decisions we had to make in order to arrive at our finalized codebook, which allowed us to better tag the instances in the data that we wanted to take a closer look at.

As a result of this process, we were able to arrive at a final set of codes that we used in a final round of data coding, *i.e.*, categorizing utterances of interest in order to study trends in the data. This consolidated codebook contained 62 different codes, organized around seven entities of interest in the

data, as follows:

- **Data**: Data Appropriation; Data Bias; Data Limitations.
- **Designer**: Designer Revisiting Previous Statement; Designer's Culture and Values; Designer's Ethics and Design Decisions; Designer's Focus on Language; Designer's Focus on Process; Designer's Focus on Technical Aspects; Designer's Goals; Designer's Limitations; Designer's Propositions about Scenario; Designer's Responsibility; Designer's Restrictions; Designer's Understanding; Designer's View on Reality Matching with Design.
- **EMT**: EMT Collaboration with Developers; EMT Collaboration with Stakeholders; EMT Non-Linearity; EMT Read by Developers; EMT Read by Stakeholders; EMT's Comprehensive Questions; EMT's Confusing Questions; EMT's Difficult Questions; EMT's I-You Framing; EMT's Question Impacts; EMT's Redundant Questions; EMT's Relation to MC; EMT's Space for Reflection;
- **Bioethical Principles**: Framing around Principle of Autonomy; Framing around Principle of Beneficence; Framing around Principle of Justice; Framing around Principle of Non-Maleficence; Mention of Principle of Autonomy; Mention of Principle of Beneficence; Mention of Principle of Justice; Mention of Principle of Non-Maleficence; Bioethical Principles' Impacts.
- **Stakeholder**: Stakeholder's Limitations; Stakeholder's Motivations; Stakeholders' Multiple Interests; Stakeholder's Responsibility; Stakeholder's Understanding.
- **User**: User's Bias; User's Interactions; User's Limitations; User's Motivations; User's Responsibility; User's Restrictions; User's Understanding.
- **System**: System's Appropriation; System's Bias; System's Ethical Impact Testing; System's Explanations; System's Fixes; System's Impacted Stakeholders; System's Impacts; System's Limitations; System's Outputs; System's Responsibility; System's Restrictions; System's Safeguards.

An in-depth description of each of these codes can be found in appendix B.

We chose to structure our set of codes around these entities since we found that they would be related to certain phenomena. Therefore, by having these codes be structured by both the entity they are associated with and the topic they are meant to represent, we are able to more easily compare these entities.

For example, participants often discussed the issue of bias, associating it with the data, the system, and/or the users themselves. By having three separate codes for each of these associations, we would be able to more easily contrast how the participants considered the issue of bias for each of these entities.

With this codebook at hand, we proceeded to separately re-code the utterances of interest we identified in our initial open coding session. In total, we classified 987 utterances according to our consolidated set of categories. Once both researchers had finished, another session of consolidation through negotiated agreement was conducted, with each instance where there was disagreement being considered and a final code being decided. This way, we were able to proceed with our analysis with a unified coding of the data, reflecting both researchers' perspectives.

From our experiences in this coding process, we were able to arrive at a semantic network that represents our interpretation of the study's data, as seen in figure 6.1. By structuring the different semantic entities and relationships we created a structure through which we can frame our discussions on the themes that emerged during the speculative design sessions.

In the following subsections, we discuss some of the relevant themes we identified during our analysis, each being associated with its own sub-structure within the semantic network. By looking at each of these themes and the part of the network in which they are represented, we seek to provide a better understanding of the whole of our interpretation by breaking it down into its essential parts. By understanding its fragments, we arrive at the meaning of the whole.

6.3.1 Reflective Design

As we have previously described in chapter 4, the main goal of extending the metacommunication template with a set of guiding questions was to promote greater reflection during the design process. It is then appropriate that we first discuss the issue of how reflective the participants' design activities were, and where their focus lied. Figure 6.2 showcases the main concepts we identified pertaining to this theme.

Essentially, the guiding questions in the EMT ask designers to express their understandings and intentions towards the design situation. In the data, participants would often state the facts they knew about the design situation and their own design goals, *i.e.*, what they wanted to achieve through the solution they were conceiving. This could be seen whenever they made claims such as: "*I know what your, candidate's, past achievements (school transcripts,*

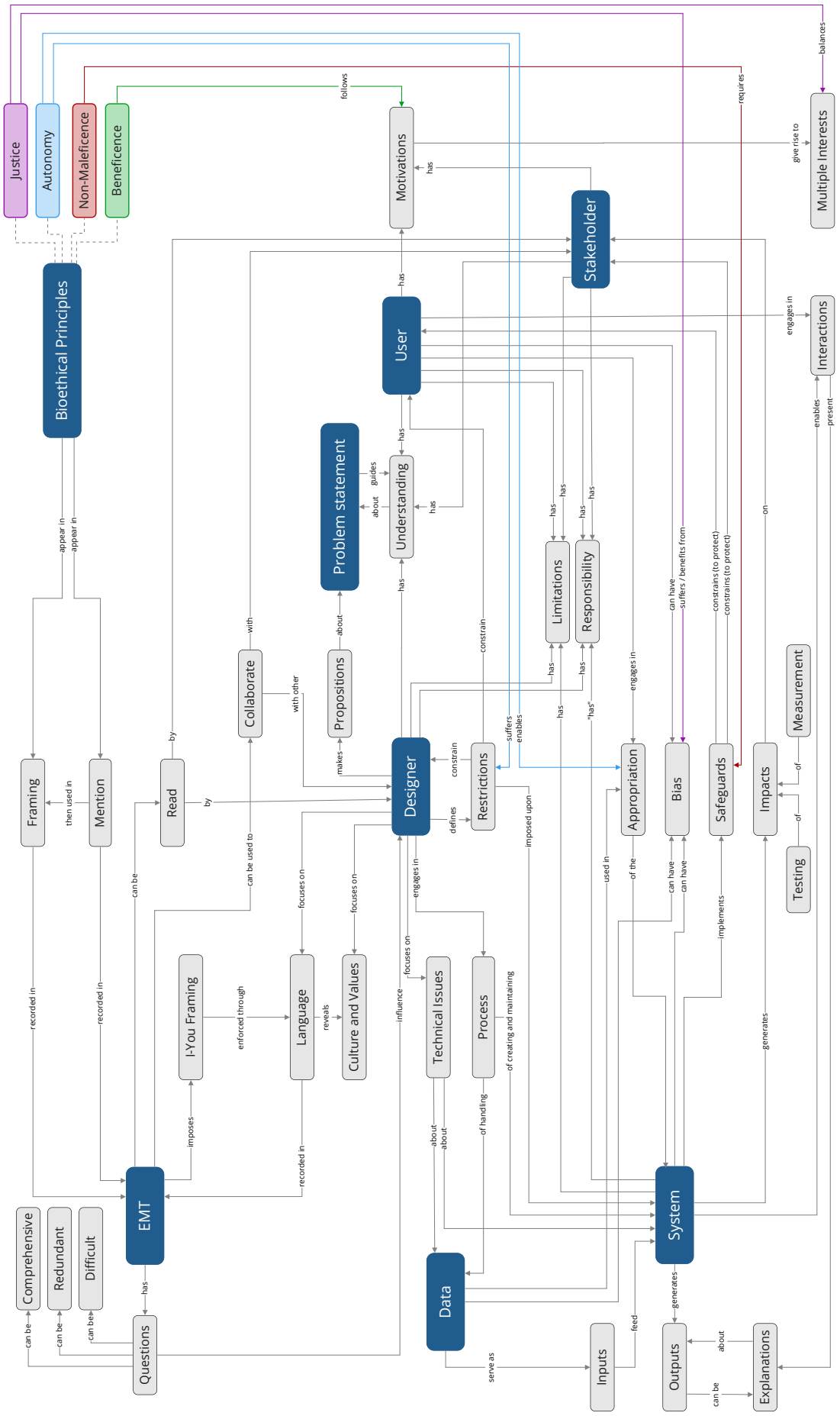

Figure 6.1: Semantic Network from EMT's Use in Speculative Design Sessions

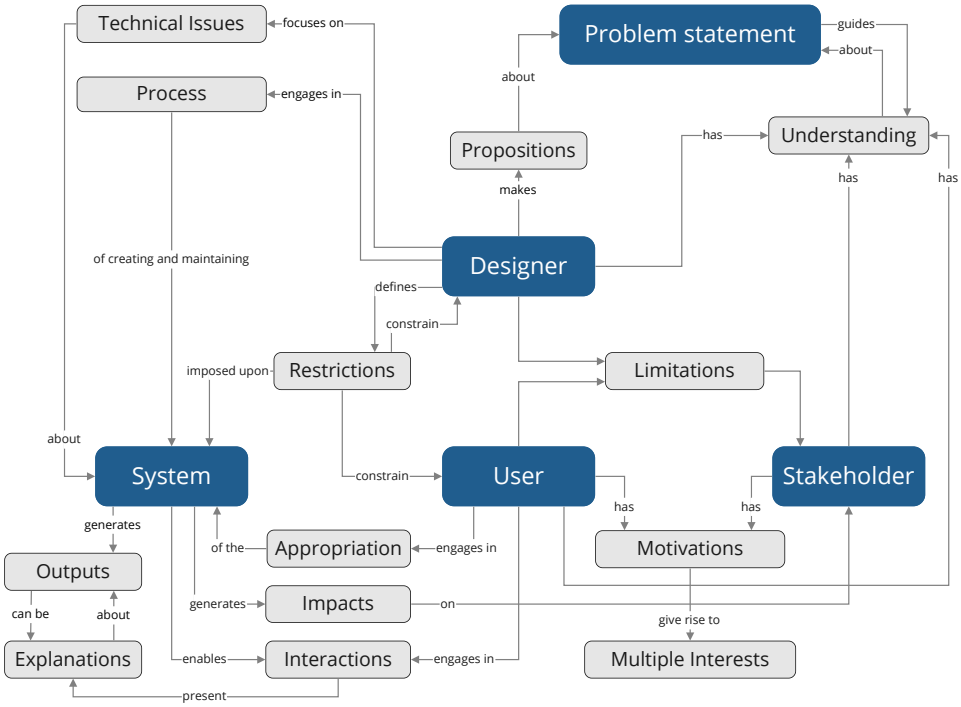

Figure 6.2: Semantic Network pertaining to the theme of *Reflective Design*

CV, recommendation letters, candidate's presentation letter) are." (P3, EMT Document) and "*I have built a system that takes a candidate's school transcript as input and generates a score as output*" (P2, EMT Document).

However, the EMT's questions also asked participants about what they did not know or might not be able to achieve. This led participants to reflect on their own limitations within the design situation, trying to find ways to overcome them and, when unable, making this clear in the design document. In our data, there were two main kinds of limitations that participants often reflected on: their limited knowledge and their limited control. Examples of this include: "*I do not know which factors you, the committee, deems more important for attributing scores to candidates*" (P3, EMT Document) and "*So, I do not think there is anything I could include there to prevent undesirable consequences*" (P8, EMT Transcript). Interestingly, we can draw comparisons between these two areas of focus and the conditions that are often associated with an agent's responsibility, namely their agency over the situation and their knowledge of the potential consequences of their actions. We explore the topic of responsibility further in section 6.3.3.

Related to their own limitations, we also identified a very similar phenomenon surrounding the participants restricting themselves in a few ways so as to ensure that they were behaving ethically. For example, they would often talk about how they might be missing some information but should not try

to obtain it through unethical means, as seen in the quote: "*So I do not have this information and should not go to the deep web to try to search for it*" (P4, EMT Transcript). Differently from cases where they noted their own limitations, these were cases where they recognized their agency and chose to forsake it, usually in order to avoid some harm befalling the stakeholders involved.

Participants recognized that there were various aspects of the system's use that they could not and should not control. The way in which the users might appropriate the system, *i.e.*, use it in ways that it was not initially planned for is an essential part of building a system that is used in the real world. These kinds of considerations can be seen in instances such as: "*... only at the admission stage; however, members of other universities' admission committees might want to use this system to evaluate students after being admitted*" (P2, EMT Document). This is not to say that they did not recognize the level of agency they had as the system's designers, only that they recognized the limits of said agency.

That being said, they did find a few ways to influence how the system would be appropriated in the real world. One of the ways in which they proposed they could do so was by providing explanations that would help users understand the meanings of the system's outputs and interpret them accurately, as seen in: "*I designed a system that makes it clear that your decision, as member of the committee, is the most important*" (P7, EMT Document). Another significant approach taken was to implement safeguards into the system that would impede certain forms of use. These were usually associated with trying to avoid harm, as in the case: "*It would be a simple question. You would have to confirm that you are not using the system in contexts X, Y, or Z. They might lie, but...*" (P1, EMT Transcript). In the end, participants recognized that even these measures would still be limited. For example, safeguards might be avoided, as seen in: "*When asked for the context of use confirmation, the user lied*" (P1, EMT Document).

In regard to their own actions in building the system, they mostly focused on two themes: technical issues and the development process itself. In the former, they mostly discussed technical minutiae that needed to be handled in order to successfully build the system. The issue of model explainability was one such issue, as seen in the quote "*I am seeing it as a black box algorithm which I think has, as inputs, the person's profile and financial history with the final decision as the output*" (P8, EMT Transcript). In the latter, participants would focus on the design and development process itself, critically analyzing the activities involved.

Given the scenario's ambiguity, designers would often have to make

propositions about it in order to be able to determine certain relevant factors for their design. For example, the educational design scenario instructs them to build a system that takes into account a potential student's "past achievements" when deciding whether or not to approve them for college admissions. It would then be up to them to propose what said achievements might be, as in the case: "*It's because, like, I can consider that I would have access to the person's history, their publications...*" (P3, EMT Transcript). An important thing to note is that the study was based on speculative design sessions, wherein the participants imagine certain aspects of the scenario in order to conceive their design. However, this also occurs in real-world situations. An essential part of design is manipulating the problem statement, creating conditions and redefining key features in order to enable the design's implementation, as discussed by Schön (1979).

Another key aspect of the EMT that lends itself to ethical reflection is the fact that participants were asked, through the ethical questions at the end of each section, to revisit their previous statements and find potential ethical issues. However, we were pleasantly surprised to find that this also occurred with other statements unrelated to the ethical side of things. Participants would, for example, look back at previous design decisions while considering the system's impacts post-deployment, as in the case of P2, who said: "*What I introduced in the system... yeah, then we go back to this previous issue of paternalism*", referencing back to a previous consideration of theirs where they concluded that they should not implement more safeguards into the system so as to not be too overbearing, even at the risk of allowing for greater abuse. In our view, the presence of these references back to previous statements indicates that the EMT was able to support the non-linear nature of most design processes, where decisions made are constantly revisited and re-evaluated.

As was intended with the template's extension, participants appeared to be able to reflect on their design decisions and their own situation in the ways we have just discussed. They themselves also claimed so, arguing that the EMT afforded them space for reflection through its guiding questions, as stated in: "*Look, this EMT explores more details about what the tool's objectives are. So it ensures that I, who am answering, will think about more dimensions than I would at a lower level, with greater technical detail. [...] Ensuring that more dimensions are being explored here*" (P5, EMT Transcript). They also compared this depth of reflection with what was achieved with the Model Card, stating that they were able to achieve a deeper level of reflection with the EMT, exemplified in the following snippet: "*Not here, here I not only had to explain but also had to think about potential errors and such. Here, I think*

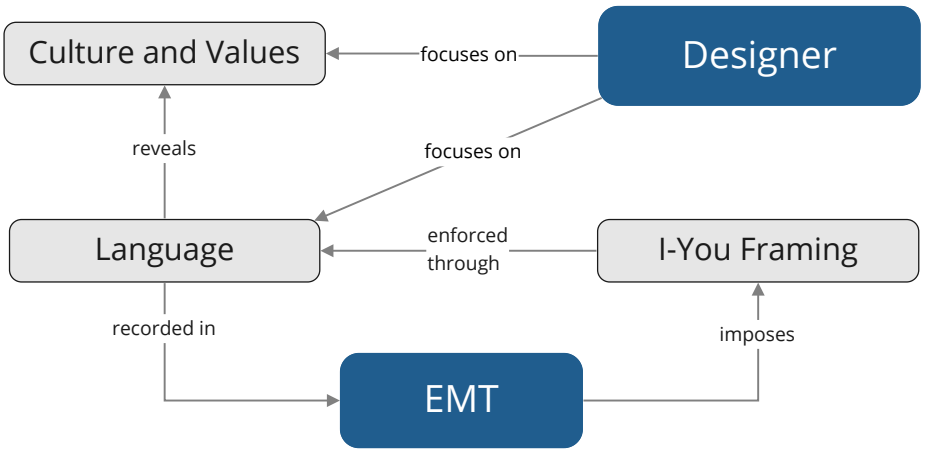

Figure 6.3: Semantic Network pertaining to the theme of the *Designer's Language Use*

I had a greater provocation to think ethically than in the other" (P2, EMT Transcript). However, not all participants agreed that the tool helped, with P8 claiming that: "*The questions promoted discussion, but did not really change my mind*" (P8, EMT Transcript). This may be due to their belief that the system and themselves were not responsible for the system's impacts, which we will discuss in greater length in section 6.3.3.

6.3.2 Designer's Language Use

The EMT is framed around the communication between the designer and stakeholders. As such, participants were forced to represent their design intentions through natural language. In our data, we were able to identify a few interesting phenomena related to how participants tried to accurately convey their meanings while also trying to conform to the EMT's framing.Figure 6.3 showcases the main concepts we identified pertaining to this theme.

One of the essential parts of the guiding questions is the directional framing that they impose on the designer, what we also call the "*I-You Framing*." All questions prompt participants to reflect on their own understandings and intentions and then try to communicate them directly to the stakeholder in question. In our data, we found that this was often not a trivial task for them, with some participants stating that they found it difficult to conform to it, mostly due to their past experiences with other development documentation, as exemplified in: "*To start with, I think that I am not very accustomed to saying 'you and I', you know? I mostly write in the third person*" (P5, EMT Transcript).

Despite these difficulties, participants mostly expressed that this framing helped them be aware of the stakeholders and be more empathetic towards them. In fact, there were also a few comments about how they often forget that there are people that will be impacted by the system and see their design and development process as simple problem solving, as shown in: "*Because, when developing, you end up being focused on the project, you want to optimize some value, maximize some value, and you don't worry about it. [...] By connecting these two, it gives a more personal factor to this structure, to this description, allowing for more space for discussing ethical, human questions*" (P2, EMT Transcript). In terms of their empathy towards stakeholders, they made statements similar to: "*I found this valuable because it makes me think directly about the people that are going to use our software, and this goes a bit beyond just mentioning the user. It makes the situation more human. You can better place yourself in the user's situation. There would be a bit more empathy*" (P1, EMT Transcript).

In addition to maintaining this framing, participants were also very aware of their choice of words and whether they were able to communicate the meanings they sought to convey. This is directly related to the semiotic aspect of the EMT, which tries to make them reflect upon the signs that they chose to represent their message and, by considering their meanings, analyze their own interpretations. This focus on their language use can be seen in the following quote: "*I was going to put that it may 'lead to problems', but this would be a bit generic. What am I thinking about here? It can generate... false... I'm gonna say 'doubtful scores to students'*" (P2, EMT Transcript).

By considering their own language use, they also came face to face with their own culture and values. For example, P1, after having expressed the guiding principles of their design stated: "*Yes... because when you are an engineer, your ethical principles... it's kind of complicated, you know? You want to maximize gain*" (P1, EMT Transcript). In this case, the participant was not only forced to reflect on their own intentions, but also on those that are usually involved in their own professional culture, considering the tendency of engineers to adopt an optimization mindset. This was an exceptional case, however, with most participants reflecting on their own values and not necessarily making generalizations about the cultures they belong to. It is possible that this potential for cultural reflection can be realized through the collaborative use of the EMT, but this falls outside the scope of this study, where participants engaged with the document separately.

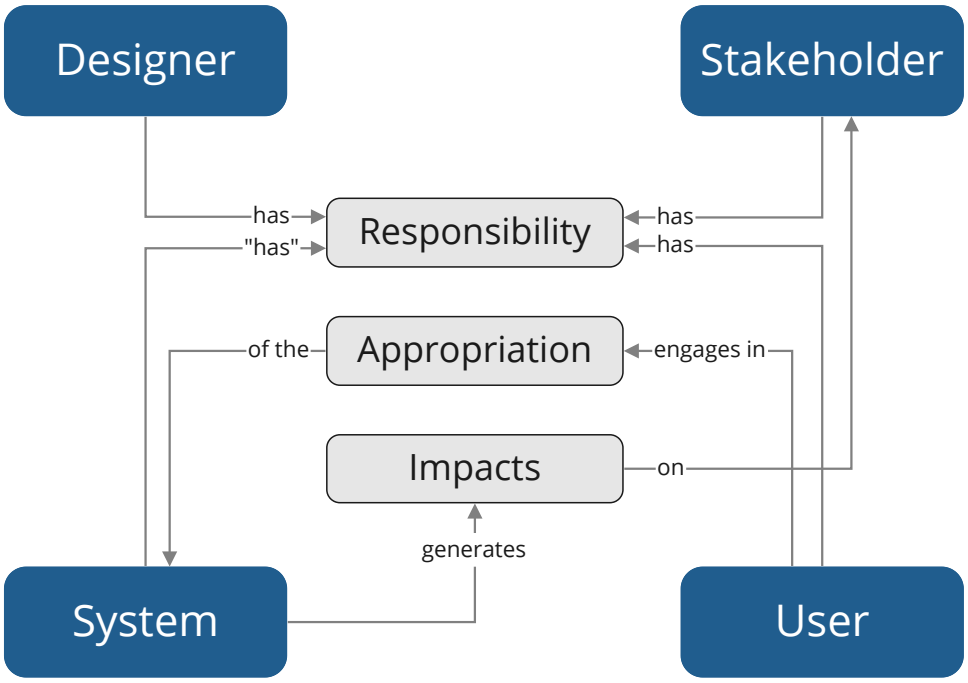

Figure 6.4: Semantic Network pertaining to the theme of *Responsibility Attribution*

6.3.3 Responsibility Attribution

Given the EMT's ethical focus, the topic of who was responsible for the consequences of the system's use emerged as a key theme in the data. Figure 6.4 contains some of the key concepts related to how participants determine who were responsible for the system's impacts. In addition to considering what these impacts might be, participants also spent a considerable amount of effort in identifying who caused them to occur. They mostly adopted the perspective that those deemed responsible would be the ones that would have to adapt their behavior in order for these outcomes to change. There were no cases where they sought to identify those responsible in order to punish them.

Through this reflective process, everyone was considered to be potentially responsible, from the designer to the users and stakeholders. In some instances, even the system itself was framed as responsible even though they are incapable of moral deliberation and thought.

For their part, participants were mostly able to recognize their own responsibility when it came to the actions that they took during the development process. This can be exemplified in: "*So, I think that this information is sensitive and that it should not be available in any way. I could end up doing harm to my user in some way, right?*" (P4, Transcript). However, in another notable

case, P8 also shared this responsibility with the stakeholders that would have hired them to implement the system, as stated in: "*I hire you to implement an algorithm that will provide you with the percentage of similar people who did or did not pay. Then comes the question of who is the owner of the system*" (P8, EMT Transcript). In essence, P8 saw themselves as a problem solver, tasked by their client with solving a problem, and it would be up to the people involved to be responsible for the solution's impacts.

In the end, all of the participants recognized that they were not in full control of the situation and that the system would be used by and impact other people. Given the EMT's questions about anticipated and unanticipated uses and consequences, they were then forced to try and determine who would be responsible when the system would have already been deployed in its actual use context. These could be either the users, who interact directly with the system, or other stakeholders, as shown in the following quotes: "*It is this owner of the system, they are the ones who decide when it is used... when they accept that it might be used*" (P8, EMT Transcript) and "*When I thought about it, the people in the bank I am working with... if there is a problem, it is their problem, that is what I think about the responsibility for these consequences, between these 'I' and 'You', in this way*" (P4, EMT Transcript).

Going beyond the actual people involved, there were also a few cases where responsibility was attributed to the system itself. An example of this phenomenon was: "*I do not want this scenario to happen, so I want to have access... that it gives this to us, and this list is the system's responsibility*" (P4, EMT Transcript). In contrast, there were also those who repeatedly stated that the system could not be responsible for its consequences, since it was only a tool to be used rather than an agent, as expressed in: "*The algorithm is the same thing, it is a tool and the person that uses the tool is responsible for the consequences*" (P8, EMT Transcript). We interpret this to mean that, even though the EMT leads to discussions on who is responsible, it still allows designers to deviate responsibility from themselves in some cases.

6.3.4 Effects of Bioethical Principles

One of the key features of our study was the use of the bioethical principles as a means of supporting participants' ethical reflections. It is important to remember that they were not obligated to engage with the principles during the speculative design session. In the end, we found that the presence of a supportive ethical framework did, in fact, assist them in priming and framing their reflections throughout the process. Figure 6.5 displays the

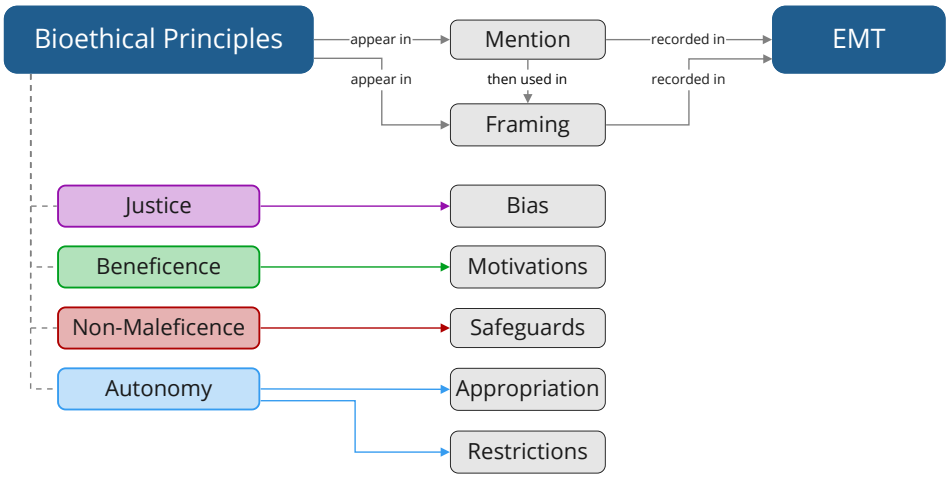

Figure 6.5: Semantic Network pertaining to the theme of the *Effects of Bioethical Principles*

main topics that were impacted by the use of the bioethical principles.

Most participants themselves stated that the presence of the bioethical principles served as a productive starting point for their ethical considerations, even if they eventually started relying on their own moral values, as stated by P1: "*They served to provide a starting point. Because I think that everybody has ethical principles, it is a part of peoples' personality but when the time comes you can end up forgetting something, it takes a while to get in a rhythm. So I think that these principles helped me start this process, even if I still have my own principles*" (P1, EMT Transcript). However, not all participants found them as significant, as exemplified in the following statement: "*To be honest, the impact was low. Low because in this round, as with the Model Card, I rarely saw myself doing this (switches browser tab to the document containing the bioethical principles), you know? [...] It probably had an impact, sure, but I did not find it essential for my discussions*" (P2, EMT Transcript).

Regardless of their opinions on the matter, all of the participants seem to have framed their reflections through some of the bioethical principles. This framing process was usually implicit in their statements, as they adopted different perspectives on the design situation based on a certain principle. For example, one of the most common topics of reflection was related to how just the system was, as in: "*I do not want the candidate to hide certain information when applying, since I want a just model*" (P4, EMT Document). This quote also showcases another interesting phenomenon since this consideration, framed around the concept of Justice, occurred in one of the EMT's regular design questions. When it came to framing, participants engaged in it when answering various questions and not only the ethical ones.

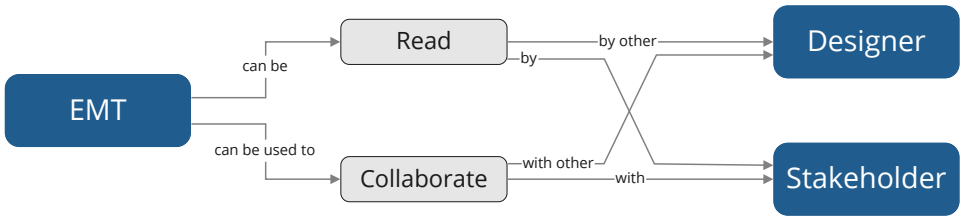

Figure 6.6: Semantic Network pertaining to the theme of the *Extended Metacommunication Template's Use*

In addition to using the bioethical principles as conceptual frames, participants also mentioned them explicitly. Usually, these explicit mentions would occur while the participant was answering the EMT's ethical questions, like in the quote: "*... with each of their risks, we would be able to work on this ethical principle of Justice*" (P5, EMT Transcript). We found that these mentions could occur prior to an ethical consideration, serving as a guide for reflection, or after the consideration had already been made, as a form of classification of the issue identified.

When explicitly mentioned before an ethical consideration was made, the following statements tended to be framed around the principle in question. This served as an indication that the participants were using said principles to scaffold their reflection, *i.e.*, choose an ethical principle, and then try to find an issue that would match it. For example, a participant could mention the principle of Autonomy and then start discussing how the user should be able to challenge or correct some of the models' decisions, as stated in the quote: "*My system should be able to accept suggestions from a human reviewer so that it correct certain mistakes*" (P4, EMT Document).

Although our findings pertain to the use of bioethical principles, it is possible that these observations can generalize to other ethical frameworks as well. However, this process of framing and scaffolding may be different with other types of frameworks. A utilitarian framework, for example, may not be as easy to convert into conceptual frames for the participants' ethical reflection. Further studies will be necessary to study the use of the various types of ethical frameworks.

6.3.5 Extended Metacommunication Template's Use

Participants saw a few ways in which the EMT could be used in real-world development situations. Figure 6.6 showcases how we modeled these potential uses in our semantic network. Given that we asked them about who might

benefit from reading or using the template, most of our findings surround the creation and consumption of these documents. Something to consider is that certain scenarios may afford certain uses that others will not. For example, a situation where an understanding of the system's internal functioning should be kept secret may not allow for the developers to release the filled-out EMT to the stakeholders involved. In these cases, it is possible that some level of secrecy may be required, as mentioned in: "*So I am afraid of giving users these nefarious ideas. Not in all of the questions, of course, but maybe in some of them. I think they could read it if some details remain confidential*" (P4, EMT Transcript).

Keeping with the topic of the EMT's potential readers, most participants saw value in providing others access to the filled-out document. The two main groups of interested parties they identified were the development team itself and the stakeholders that might be affected by the system's use. Participants found that the development team might benefit from being aware of the potential ethical implications of the system's use and by being able to compare the reality of their development with their original design intentions, as stated in: "*... at least in my view, this is more of a document for those involved in the development team, so that they can think about these ethical issues. What was the initial purpose of the design that was not fulfilled in the final solution, that kind of stuff*" (P2, EMT Transcript). On the other side of things, they also found that stakeholders might benefit from being able to read the EMT written by the developers, since they would gain greater insight into their intentions while building the system and be aware of its potential for harm, as stated in: "*Yeah, I think the university's administration, I don't know, could be interested in reading the document to understand at a high level how the system was designed and how it can help the committee without doing them harm, considering the risks involved*" (P7, EMT Transcript).

Another key topic we discussed was the potential for others to also answer the EMT's questions. As with the previous issue of who might read the EMT, participants mostly saw the stakeholders and the developers as interested parties. There were also two ways in which they might collaborate with the system's designer: either by separately filling out their own versions of the EMT and then consolidating them together, or by collectively answering the questions in real time. The former can be observed in P4's statement: "*Oh yeah, when I am talking about my team, for sure. Then I would imagine a scenario where everybody answers the questions and then we go and exchange ideas and compare. I think that would be good*" (P4, EMT Transcript). The latter can be observed in P3's statement: "*I am imagining users participating*

so that they could talk about which ethical issues they see and that they would want to be handled" (P3, EMT Transcript).

These findings show that the participants mostly agreed with our original idea that the EMT might be used for collaborative design, with multiple parties being involved. By having them express their perspectives on the design situation, we posit that they are able to identify a wider set of issues, resulting in a more robust design. Future work on how the EMT might be used collaboratively would be required for us to understand how this might actually take place, but our results so far seem to indicate that developers may be interested in this kind of use.

Having described each of these themes, we can now move on to a more cross-cutting analysis of the observed phenomena. Certain trends, such as the recognition of one's limitations and the necessity of recognizing others' responsibilities encompass multiple of our themes.

7
Discussion

Having analyzed our study's results and emerging themes, we can now discuss some of the relevant phenomena that we have observed. These are not necessarily contained in a single, specific theme. Instead, they serve as a more generalized description of what we have observed in the data.

7.1
Limited Knowledge and Design Propositions

One of the key aspects of the Extended Metacommunication Template's (EMT) questions is that they provoke the author to not only consider what they do know, but also what they do not. From what we have seen, participants were able to acknowledge the limits of their understanding of the design situation. This, in turn, helped them recognize how the lack of some information may limit their design decisions, since these would be undermined by a great deal of uncertainty.

Having recognized gaps in their understanding and how they can limit their design, participants then resorted to making propositions about the design scenario. They assumed something was true or false so that they were able to make a decision. This choice would then be as solid as the propositions that allowed it to be made.

However, in addition to allowing them to make necessary decisions, these propositions would also create new requirements. If something is true, then it might imply a whole host of other issues that may need to be tackled through the solution.

This process of making such proposals while trying to narrow down the design problem to a manageable state was a continuous process throughout the participants' experiences with the EMT. This matches Lawson and Dorst's (Lawson and Dorst, 2013) observation that more experienced designers treat the design problem itself as design material, constantly manipulating it so as to allow for a better solution.

It was difficult, however, to discern between propositions that would occur with the regular use of the EMT and those that were brought about by the fact that they were engaging in a speculative design session and, therefore,

did not have all of the necessary information. As it stands, we are unable to ascertain the EMT's potential for promoting the designer's acknowledgment of their limited understanding and the manipulation of the design problem.

7.2 Recognizing Agency and Responsibility

Given their roles as the solution's designers, participants had a significant amount of agency over the situation, which they did recognize. In addition, they were also able to identify that their control was limited when it came to how the system would be used at use time.

Users and stakeholders would be able to appropriate the system and use it in ways that were not necessarily intended by the designer. As they deviate from the designer's vision, the consequences of the system's use also differ from their expectations.

Given that they are the ones in control during use time, our study's participants ended up attributing some of the responsibility for the system's impacts to the users and stakeholders themselves. For example, if one was to abuse a system, they would be at least partially responsible for the consequences.

This is not to say that participants did not recognize that they still had some level of asynchronous agency at use time since they built the system's functionalities. Having recognized the possibility of users deviating from their design vision and the potential for them to influence their actions during use time through the system's functionalities, our study's participants focused on two strategies to try to avoid undesirable consequences.

The first strategy focused on the use of explanations. The key idea they had was that through the use of explanations they might guide the user's interpretations of the system's functionalities, ensuring that there were no misunderstandings and that their intentions, as designers, were clear to the users. A common example of this touched on the issue of how users might interpret the machine learning models' scores for both the educational and financial scenarios. Some participants were adamant that they should not be the only information used for a decision as to whether a student should be accepted into the university or a loan application should be approved. In order to reduce the chances that the scores were used in this undesirable way, in their view, they stated that they would add explanations that made it clear to users that they should not be used as the sole criteria for the decision. Through the use of additional information, they guide the way that the system's mechanics are understood correctly.

The second strategy they employed was to develop safeguards. This would be a more forceful approach than the use of explanations, since it would add interactions that make it more difficult for an undesirable use to be made, instead of just guiding the user's understanding. Having recognized potential for abuse, intentional or not, participants came up with different mechanisms to try and stop it, ranging from a simple confirmation screen that asked you to assert that you were not using a piece of information in a forbidden context to the use of strict access control and anonymization of sensitive information.

In the end, having recognized that their control would be limited at use time, participants noticed the importance of trying to foresee potential undesirable uses and implement mechanisms that try to stop them. Responsibility would be shared with users and stakeholders, but since the designer might be able to identify these potentials for abuse and come up with ways to stop them, they would still be somewhat responsible.

7.3 Language, Communication, and Introspection

Another key aspect of the participants' experiences with the EMT was how they dealt with the language they used when trying to communicate something to the solution's users and stakeholders. Interestingly, they were often critical of their own choice of words, trying to find the correct ones to convey the meanings that they seek to communicate. This matches the EMT's semiotic grounding, focusing on framing the development process of metacommunication and, therefore, drawing attention to how the representations they include in the system may be interpreted and signified by users and stakeholders.

By reflecting on the representations, they were also able to analyze the meanings themselves. In some cases, while trying to find the right words to communicate something, they ended up considering the underlying values and presuppositions behind what they were trying to tell the users and stakeholders. In an even more interesting case, one participant not only looked at a value they held but also considered how they learned it from the culture they were inserted in; in this case, the culture of engineering, which incentivizes practitioners to try and optimize any given situation.

Given the communicative focus of the EMT, participants also focused on the "you" of the situation, *i.e.*, the users and stakeholders to whom they are trying to convey the metacommunication message. In alignment with the tool's semiotic grounding, participants not only had to consider what the words they chose meant to them but also how they might be interpreted by the receivers of the message. Via this process, they often placed themselves in the shoes of

the system's users and stakeholders in order to try to imagine how they might signify the chosen representations.

By placing themselves in the shoes of those affected by the system's use, participants noted that they felt more empathetic towards them. The task of communicating with stakeholders led them to constantly think about the stakeholder's existence, place themselves in their shoes, consider how they might interpret the system, and also how they might feel being affected by it. For the participants, the results of this process were a constant reminder that there will be people affected by the system's use and how they might feel these impacts. These are important factors for designers, since they may lead them to be more invested in their design and also be aware of the user's perspective throughout.

7.4 Bioethical Principles as Instruments

An important part of our study was the use of the bioethical principles. From what we observed during the participants' experiences with the EMT, they were used as a means to start the reflective process. In other words, they served as primers for ethical reflection. Whenever they had difficulty finding a relevant issue to reflect on, they would rely on the bioethical principles to either provide them with a different perspective on the situation or even serve as a structural pillar around which they would identify ethical problems.

Participants may have been mindful of these effects or not. Even if they did not explicitly turn to the set of principles, being aware of their existence may have influenced their interpretations of the situation at hand. For example, being aware of the existence of a principle of Justice may have made them more prone to balancing the benefits and harm done to stakeholders. This possibility was even explicitly mentioned by one of our participants, who stated that, although they did not consciously turn to the set of principles, they may have been affected by them.

As mentioned in section 6.3.4, we noticed that these principles were used either as conceptual frames or as scaffolds to new ethical findings. The latter implies a more conscious use of the principles, while the former may be indicative of a potentially unconscious influence.

An interesting thing to then consider is how different types of ethical frameworks could be used in a similar role. For example, how would a utilitarian ethic work with the EMT? Participants would probably be able to use conceptual frames that seek maximal utility, but what about the scaffolding process? Future studies could try to investigate this process, relating how

certain structural aspects of an ethical framework work with the EMT's use of guiding questions and focus on metacommunication.

If we achieve a better understanding of how these different types of ethical frameworks could be used in conjunction with the EMT, it would allow us to tailor our choices to the design context in question. Later, how we integrate the EMT into the design and development is also a matter of design. More specifically, of metadesign.

7.5 Extended Metacommunication Template's Appropriation

As our participants often noted, the way in which the EMT might be used varies from situation to situation. It is context-dependent. Certain situations may afford new possibilities in terms of when, how, with who, etc. we can use the extended template. In others, they may add restrictions that impede certain forms of use. It is therefore important, the participants noted, to be attentive to the design and development scenario at hand and try to find an effective way to use the EMT.

One of the most frequently mentioned topics was that of the restrictions imposed on how the EMT may be appropriated by users and stakeholders. Participants would often mention how the artifact being developed may be proprietary information and may, therefore, not be accessible to stakeholders outside of the company that owns it. Usually, these participants would then state that perhaps some of the information discussed may be provided to users, while others may be hidden. This selective disclosure that may be required by the need to protect intellectual property and avoid exposing vulnerabilities in the system is an interesting process that should be relevant to study in the future.

When asked to consider how the EMT might be used in their usual design and development process, participants would also discuss the connection between the EMT's structure and the phases in the development process. In the study, participants were asked to answer all of the extended template's questions in one sitting. However, in real development situations, it may be more interesting for designers and developers to answer them as the project advances through its different stages. Another possibility would be to try and answer all of them at the start while making it clear which answers are predictions and which are based on facts that they have already ascertained. Then, at each stage in the process, these answers could be revised, changing as they gained a better understanding of the design situation. In essence, what most of the participants noted was that the tool might be integrated

into the process in different ways. Future studies would be required to better understand the advantages and disadvantages of these different settings.

From the stakeholders' point of view, participants stated that it might be valuable for them to be able to access at least some of the instantiated EMT, since it might provide them with a better understanding of the system's internal functioning. They mostly attributed this ease of understanding to the language employed in answering the guiding questions, which is directional and less dependent on jargon. However, as stated previously, restrictions may require designers to limit the information provided to the stakeholders. This then creates a process of trying to balance out the need to protect sensitive details and the desire to have users and stakeholders better understand how the system functions and what its potential impacts might be. In the end, this is just one more of the decisions that need to be made when deciding how to integrate the EMT into an actual design and development process.

7.6 Limitations of our Analysis

In the end, our qualitative analysis was not without its limitations. From the people in our participant sample to the potential for bias in our observations of the data, there are a few issues we should discuss to more accurately frame our findings.

In terms of our sample, in addition to having a limited number of participants (8), it was also quite uniform, with all those involved having studied in the same academic institution. Also, there were no individuals currently working in industry, potentially creating a blind spot for the dynamics involved in these companies. Taking both of these factors together, we end up with a limitation regarding how representative these findings might be. Future work involving other participants and including practitioners would be welcome to ensure that our insights are indeed applicable to a wider population of developers of machine learning systems.

Another limiting factor in our study was our reliance on speculative design sessions. Despite allowing participants to have a wider range of considerations, not being constrained by limitations that would be present in actual development scenarios, it ended up relying greatly on their imaginative capabilities. This created some discrepancies in the depth of reflection in each session, since some people may have been more creative than others. Additionally, this also meant that, in the last section of the EMT, participants not only had to think of their considerations but also determine what would have happened with the system's use. This double requirement may have made an-

swering this section's questions more difficult than it would have been in a real development scenario.

Finally, we, as researchers analyzing the data, may have suffered from bias. First, as mentioned in section 6.2, both of the researchers involved in this study were familiar with the study's participants. This may have led them to take their previous experiences with these individuals into account when analyzing the data they generated. Of course, this effect may have been reduced because, during the coding process, all instances were associated with their anonymous codes and not their names. Second, one of the researchers was also involved in the proposal of the EMT. Because of this, it is possible that they suffered from confirmation bias, *i.e.*, they were more sensitive to phenomena they expected to find in the data. This may have meant that phenomena that were just as significant were overlooked and that the significance of those they did expect to find was overestimated. However, the other researcher who analyzed the data was not one of the proponents of the EMT, serving as a counter-balance and reducing this potential for bias.

Despite these limitations, our findings still seem to show some of the EMT's potential for promoting greater reflection during the design process. Of course, it is important to note our insights are limited to our own perspectives and participant sample. Further work will be required to ensure that these findings not only generalize to the EMT's potential users but also explore different aspects of the EMT that fell outside the scope of this work, such as its potential for collective reflection. Through these additional studies, we may better understand the impacts of the extended template's use, with this work only serving as an initial step on this journey.

8
Conclusion

In this dissertation we proposed an extension to Semiotic Engineering's metacommunication template (MT) through a set of guiding questions and investigated some of the impacts of its use through a speculative design study. Through the analysis of the data collected from our 8 participants' interactions with the design document, we identified a few relevant themes that emerged from its use. In addition to giving us a greater understanding of how the Extended Metacommunication Template (EMT) may be used to promote reflection, this analysis also helped us identify promising opportunities for future work.

As discussed in chapter 4, we sought to extend the metacommunication template through a set of guiding questions that ask developers to reflect on their own understanding and intentions and try to communicate them to their solution's stakeholders. To do so, they would employ what we called an "I-You Framing," which they inherit from the original MT. The idea behind it is that designers would be forced to recognize their role as creators of the system and directly face the people that may be impacted by its use. This would, in turn, lead to greater reflection due to their feeling of responsibility.

We then proposed and executed a study based on speculative design sessions, as described in chapter 5. Our goal was to collect empirical data that could provide us some insights into the themes that could emerge through the EMT's use. Given that we wanted to allow participants to engage in a wide range of design and ethical considerations, we opted to use a speculative method that was free from pragmatic constraints. However, this also came with its downsides, as it limited how representative our findings were of real-world situations. From this limitation, we recognize that future work involving real development situations is necessary to ensure that the findings described in this dissertation generalize beyond purely speculative activities.

For our study, we recruited 8 participants with at least an undergraduate-level understanding of the development of Machine Learning systems, as described in chapter 6. We conducted multiple rounds of coding the qualitative data we obtained from their study sessions, starting with an open coding process, where each researcher is free to identify categories that they think

might be useful for the later analysis and then proceeding to consolidate them into a finalized codebook. By following this process, we sought to take advantage of the difference in perspective that the two researchers had, potentially making their final interpretation of the data more robust and less biased. This process resulted in a set of 62 codes, which we then used to tag the data once again, laying the groundwork for our qualitative analysis.

From the coded data, we could identify five relevant themes that emerged during these speculative design sessions, discussed in chapter 7. The first pertained to their reflections throughout the design session, wherein they continuously re-evaluated not only their own design decisions, but also their understanding and position within the design situation. The second mostly focused on their use of language and how attentive they were to their choice of words, trying to find the right ones to communicate the meanings they sought to convey. The third was related to their attempts to attribute responsibility for the system's actions to those involved, including themselves. The fourth mostly focused on the use of the bioethical principles and how different ethical frameworks may be used to support ethical reflection with the EMT. The fifth and final theme then pertained to how the EMT might be appropriated and used in real development situations.

Our analysis was not without its limitations. However, from these limitations we were able to identify a few opportunities for future work. Since our sample comprised mostly people that hailed from an academic setting, it would be necessary to replicate this study with professionals currently acting in industry so that we can understand how these companies' dynamics may change how the EMT is used. Given the limitations of speculative design methods, future work should also include investigations based on actual development projects. Finally, despite falling outside the scope of this study, we were able to observe an interest in the collaborative use of the EMT, which would require further studies involving teams of developers and stakeholders.

In the end, the goal of our proposal of the Extended Metacommunication Template was to encourage a more reflective design of machine learning systems, taking into account not only technical issues but also social ones and how they relate to one another. Further work will be necessary for us to better understand what the actual impacts of the EMT's use can be, and we hope that the findings described in this dissertation serve as a motivator for these future investigations. All for the sake of creating more robust, useful machine learning systems.

Bibliography

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., and Horvitz, E. (2019). Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, New York, NY, USA.

Auger, J. (2013). Speculative design: crafting the speculation. *Digital Creativity*, 24(1):11–35. Publisher: Routledge _eprint: https://doi.org/10.1080/14626268.2013.767276.

Barbosa, S. D. J., Barbosa, G. D. J., Souza, C. S. d., and Leitão, C. F. (2021). A Semiotics-based epistemic tool to reason about ethical issues in digital technology design and development. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 363–374, New York, NY, USA. Association for Computing Machinery. tex.ids= BarbosaEtAl2021SemioticsbasedEpistemicToola.

Barbosa, S. D. J. and de Paula, M. G. (2003). Designing and Evaluating Interaction as Conversation: A Modeling Language Based on Semiotic Engineering. In Jorge, J. A., Jardim Nunes, N., and Falcão e Cunha, J., editors, *Interactive Systems. Design, Specification, and Verification*, Lecture Notes in Computer Science, pages 16–33, Berlin, Heidelberg. Springer.

Beauchamp, T. L., Beauchamp, P. o. P. a. S. R. S. a. t. K. I. o. E. T. L., Childress, J. F., and Childress, U. P. a. H. P. o. E. J. F. (2001). *Principles of Biomedical Ethics*. Oxford University Press. Google-Books-ID: _14H7MOw1o4C.

Bolt, N. and Tulathimutte, T. (2010). *Remote Research: Real Users, Real Time, Real Research*. Rosenfeld Media. Google-Books-ID: FHo3DwAAQBAJ.

Braun, V. and Clarke, V. (2012). Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, quali-*

tative, neuropsychological, and biological, APA handbooks in psychology®, pages 57–71. American Psychological Association, Washington, DC, US.

Buchanan, R. (1992). Wicked Problems in Design Thinking. *Design Issues*, 8(2):5–21. Publisher: The MIT Press.

Carbonell, J. G., Michalski, R. S., and Mitchell, T. M. (1983). An Overview of Machine Learning. In Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors, *Machine Learning*, pages 3–23. Morgan Kaufmann, San Francisco (CA).

Clark, R. (2017). Convenience Sample. In *The Blackwell Encyclopedia of Sociology*, pages 1–2. John Wiley & Sons, Ltd. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781405165518.wbeosc131.pub2.

Cobbe, J., Lee, M. S. A., and Singh, J. (2021). Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 598–609, New York, NY, USA. Association for Computing Machinery.

Coeckelbergh, M. (2020a). *AI Ethics*. MIT Press. Google-Books-ID: Gs_XDwAAQBAJ.

Coeckelbergh, M. (2020b). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4):2051–2068.

Cooper, R. and Foster, M. (1971). Sociotechnical systems. *American Psychologist*, 26(5):467–474. Place: US Publisher: American Psychological Association.

Day, B., Bateman, I. J., Carson, R. T., Dupont, D., Louviere, J. J., Morimoto, S., Scarpa, R., and Wang, P. (2012). Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of Environmental Economics and Management*, 63(1):73–91.

de Souza, C. S. (2005). *The Semiotic Engineering of Human-computer Interaction*. MIT Press. Google-Books-ID: 0yjnotmvtGkC.

de Souza, C. S., Fontoura de Gusmão Cerqueira, R., Marques Afonso, L., Rossi de Mello Brandão, R., and Soares Jansen Ferreira, J. (2016). The SigniFYI Suite. In Sieckenius de Souza, C., Fontoura de Gusmão Cerqueira, R., Marques Afonso, L., Rossi de Mello Brandão, R.,

and Soares Jansen Ferreira, J., editors, *Software Developers as Users : Semiotic Investigations in Human-Centered Software Development*, pages 49–125. Springer International Publishing, Cham.

de Souza, C. S. and Leitão, C. F. (2009). Semiotic Engineering Methods for Scientific Research in HCI. *Synthesis Lectures on Human-Centered Informatics*, 2(1):1–122. Publisher: Morgan & Claypool Publishers.

de Souza, C. S., Leitão, C. F., Prates, R. O., and da Silva, E. J. (2006). The semiotic inspection method. In *Proceedings of VII Brazilian symposium on Human factors in computing systems*, IHC '06, pages 148–157, New York, NY, USA. Association for Computing Machinery.

Elamroussi, A. (2021). This Washington county is the first to ban facial recognition technology, official says. *CNN*.

Felzmann, H., Villaronga, E. F., Lutz, C., and Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1):2053951719860542. Publisher: SAGE Publications Ltd.

Firth, J. A., Torous, J., and Firth, J. (2020). Exploring the Impact of Internet Use on Memory and Attention Processes. *International Journal of Environmental Research and Public Health*, 17(24):9481.

Floridi, L. and Cowls, J. (2021). A Unified Framework of Five Principles for AI in Society. In Floridi, L., editor, *Ethics, Governance, and Policies in Artificial Intelligence*, Philosophical Studies Series, pages 5–17. Springer International Publishing, Cham.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., and Crawford, K. (2018). Datasheets for Datasets. *arXiv:1803.09010 [cs]*. tex.ids: GebruEtAl2018DatasheetsDatasetsa, GebruEtAl2019DatasheetsDatasets arXiv: 1803.09010 tex.xnote: arXiv: 1803.09010.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*,

4(37):eaay7120. Publisher: American Association for the Advancement of Science.

Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv:1805.03677 [cs]*. arXiv: 1805.03677.

Holyoak, K. J. and Thagard, P. (1996). *Mental Leaps: Analogy in Creative Thought*. MIT Press. Google-Books-ID: 8ZRHYv59154C.

Johnson, M. (1993). *Moral Imagination: Implications of Cognitive Science for Ethics*. University of Chicago Press.

Kelley, P. G., Bresee, J., Cranor, L. F., and Reeder, R. W. (2009). A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, pages 1–12, New York, NY, USA. Association for Computing Machinery.

Kelley, P. G., Cesca, L., Bresee, J., and Cranor, L. F. (2010). Standardizing privacy notices: an online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1573–1582, New York, NY, USA. Association for Computing Machinery.

Krafft, P. M., Young, M., Katell, M., Lee, J. E., Narayan, S., Epstein, M., Dailey, D., Herman, B., Tam, A., Guetler, V., Bintz, C., Raz, D., Jobe, P. O., Putz, F., Robick, B., and Barghouti, B. (2021). An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 772–781, New York, NY, USA. Association for Computing Machinery.

Lakoff, G. and Johnson, M. (2008). *Metaphors We Live By*. University of Chicago Press. Google-Books-ID: r6nOYYtxzUoC.

Lawson, B. and Dorst, K. (2013). *Design Expertise*. Routledge. Google-Books-ID: fXBTAQAAQBAJ.

Lazar, J. (2017). *Research methods in human computer interaction*. Elsevier, Cambridge, MA, 2nd edition edition.

Levy, M. G. (2021). RE:WIRED 2021: Timnit Gebru Says Artificial Intelligence Needs to Slow Down. *Wired*. Section: tags.

Lipscomb, M. (2012). Abductive reasoning and qualitative research. *Nursing Philosophy*, 13(4):244–256. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1466-769X.2011.00532.x.

Luck, R. (2003). Dialogue in participatory design. *Design Studies*, 24(6):523–535.

MacLean, A., Young, R. M., Bellotti, V. M., and Moran, T. P. (1991). Questions, Options, and Criteria: Elements of Design Space Analysis. *Human–Computer Interaction*, 6(3-4):201–250. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/07370024.1991.9667168.

Maglogiannis, I. G. (2007). *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press. Google-Books-ID: vLiTXDHr_sYC.

Makarius, E. E., Mukherjee, D., Fox, J. D., and Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120:262–273.

Malpass, M. (2013). Between Wit and Reason: Defining Associative, Speculative, and Critical Design in Practice. *Design and Culture*, 5(3):333–356. Publisher: Routledge _eprint: https://doi.org/10.2752/175470813X13705953612200.

Marshall, C. and Rossman, G. B. (2014). *Designing Qualitative Research*. SAGE Publications.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. ACM. tex.ids: MitchellEtAl2019ModelCardsModela event-place: Atlanta, GA, USA.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill. Google-Books-ID: EoYBngEACAAJ.

Moran, T. P. and Carroll, J. M. (1996). *Design Rationale: Concepts, Techniques, and Use*. CRC Press. Google-Books-ID: LAoHEAAAQBAJ.

Morley, J., Floridi, L., Kinsey, L., and Elhalal, A. (2021). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. In Floridi, L., editor, *Ethics, Governance, and Policies in Artificial Intelligence*, Philosophical Studies Series, pages 153–183. Springer International Publishing, Cham.

Morrow, R. A. and Brown, D. D. (1994). *Critical Theory and Methodology*. SAGE. Google-Books-ID: wnY5DQAAQBAJ.

Neff, G. and Nagy, P. (2016). Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, 10(0):17. Number: 0.

Noble, D. a. R. (1988). Issue-Based Information Systems for Design. In *Computing in Design Education [ACADIA Conference Proceedings] Ann Arbor (Michigan / USA) 28-30 October 1988, pp. 275-286*. CUMINCAD.

Pearl, J. (2019). The Limitations of Opque Learning Machines. In Brockman, J., editor, *Possible Minds: 25 Ways of Looking at AI*, page 3. Penguin Press.

Prates, R. O., de Souza, C. S., and Barbosa, S. D. J. (2000). Methods and tools: a method for evaluating the communicability of user interfaces. *Interactions*, 7(1):31–38.

Ramphul, K. and Mejias, S. G. (2018). Is "Snapchat Dysmorphia" a Real Issue? *Cureus*, 10(3):e2263.

Riemer, K. and Johnston, R. B. (2012). Place-making: A Phenomenological Theory of Technology Appropriation. page 19.

Rittel, H. W. J. and Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, 4(2):155–169.

Rokach, L. and Maimon, O. (2005). Clustering Methods. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 321–352. Springer US, Boston, MA.

Saldaña, J. (2021). *The Coding Manual for Qualitative Researchers*. SAGE. Google-Books-ID: X7T5DwAAQBAJ.

Schmidt, P., Biessmann, F., and Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4):260–278. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/12460125.2020.1819094.

Schnackenberg, A. K. and Tomlinson, E. C. (2016). Organizational Transparency: A New Perspective on Managing Trust in Organization-Stakeholder Relationships. *Journal of Management*, 42(7):1784–1810. Publisher: SAGE Publications Inc.

Schuler, D. and Namioka, A. (1993). *Participatory Design: Principles and Practices*. CRC Press. Google-Books-ID: pWOEk6Sk4YkC.

Schön, D. A. (1979). *The Reflective Practitioner: How Professionals Think in Action*. Routledge, London.

Shen, H., Deng, W. H., Chattopadhyay, A., Wu, Z. S., Wang, X., and Zhu, H. (2021). Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 850–861, New York, NY, USA. Association for Computing Machinery.

Shneiderman, B. (2020). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4):26:1–26:31.

Skaburskis, A. (2008). The Origin of "Wicked Problems". *Planning Theory & Practice*, 9(2):277–280. Publisher: Routledge _eprint: https://doi.org/10.1080/14649350802041654.

Strydom, P. (2011). *Contemporary Critical Theory and Methodology*. Taylor & Francis. Google-Books-ID: N2yXpOD508YC.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press. Google-Books-ID: uWV0DwAAQBAJ.

Tigard, D. W. (2021). There Is No Techno-Responsibility Gap. *Philosophy & Technology*, 34(3):589–607.

Tucker, J. A., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., and Nyhan, B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. SSRN Scholarly Paper ID 3144139, Social Science Research Network, Rochester, NY.

Wagner, B. (2018). Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping? In *Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping?*, pages 84–89. Amsterdam University Press.

Wortham, R. H. and Theodorou, A. (2017). Robot transparency, trust and utility. *Connection Science*, 29(3):242–248. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09540091.2017.1313816.

Yang, Q., Scuito, A., Zimmerman, J., Forlizzi, J., and Steinfeld, A. (2018). Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, pages 585–596, New York, NY, USA. Association for Computing Machinery.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchmendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J., and Perrault, R. (2021). The AI Index 2021 Annual Report. Technical report, Human-Centered AI Institute, Stanford University, Stanford, CA.

A
Study Materials

In this appendix we present the main study materials as they were presented to the study participants. All of the documents are in Portuguese, the language in which the study was conducted.

The list of materials is as follows:

1. Informed Consent Form (section A.1)
2. Design Tools (section A.2)
3. Design Briefs (section A.3)
4. Summary of Bioethical Principles (section A.4)
5. Interview Script (section A.5)

A.1
Informed Consent Form

Termo de Consentimento Livre e Esclarecido

Natureza da Pesquisa

Nós, pesquisadores responsáveis pelo projeto de pesquisa “Avaliação de Métodos para Reflexão e Comunicação sobre Sistemas de Aprendizado de Máquina,” sob coordenação dos professores Simone Diniz Junqueira Barbosa, Clarisse Sieckenius de Souza e Hélio Côrtes Vieira Lopes, do Departamento de Informática da PUC-Rio, lhe convidamos a participar como voluntário neste estudo.

Nossa pesquisa visa investigar como *stakeholders* avaliam e refletem sobre modelos de aprendizagem de máquina. Entre outros pontos, estamos interessados em entender como desenvolvedores abordam potenciais questões éticas levantadas pelo sistema que constroem auxiliados por documentos propostos na literatura técnica. O objetivo deste estudo não é avaliar pessoas, mas sim quais problemas e reflexões são estimuladas pelo artefato a ser apresentado ao participante. Através desta pesquisa espera-se identificar problemas e oportunidades de melhoria destes modelos e métodos.

Esta pesquisa envolve entrevistas que serão gravadas em forma de vídeo e áudio. Posteriormente, esse conteúdo será transcrito de forma a garantir o anonimato do participante. Vale ressaltar que toda participação neste estudo é inteiramente voluntária.

Benefícios

Os benefícios esperados incluem a avaliação e possível aperfeiçoamento de métodos para reflexão sobre modelos de aprendizado de máquina e seus potenciais impactos em partes interessadas. Essas contribuições se darão em forma de eventuais publicações científicas. No entanto, não prevemos benefícios a curto prazo para participantes deste estudo.

Riscos e Desconfortos

Participação neste estudo pode implicar a ocorrência de alguns desconfortos, que buscamos minimizar da seguinte forma:

1. **Desconforto físico:** Cansaço ou aborrecimento caso a sessão seja longa (acima de 2 horas). A fim de minimizá-lo, buscamos minimizar a duração do experimento, focando apenas nas questões mais relevantes para o objetivo do estudo.
2. **Constrangimento por causa de gravação de áudio ou vídeo:** Gravação de áudio ou vídeo ao longo do experimento só será feito mediante o consentimento do participante. Além disso, uma vez que o material coletado seja processado, ocorrendo uma transcrição anonimizada de seu conteúdo, ele será descartado.

3. **Quebra da segurança digital de dados armazenados:** A fim de evitar quaisquer quebras de segurança, dados coletados serão armazenados em ambiente seguro (mídia ou máquina sem acesso à internet ou em área protegida por senha). Além disso, o material coletado será desassociado de sua identidade para garantir seu anonimato e privacidade.
4. **Qualquer tipo de incômodo ou constrangimento:** Você pode interromper a pesquisa a qualquer momento e sem qualquer prejuízo, penalização ou constrangimento. Não ficará registrado que você iniciou sua participação no estudo e optou por interrompê-la.

Garantia de anonimato, privacidade e sigilo de dados

Esta pesquisa se pauta no respeito à privacidade, ao sigilo e ao anonimato dos participantes. Todos os dados brutos serão acessados somente por pesquisadores envolvidos nesta pesquisa e anonimizados para análise ou divulgação. O uso que faremos dos dados coletados durante o teste é estritamente limitado a atividades e publicações científicas. Qualquer imagem, vídeo, ou áudio divulgado será disfarçado para impedir a identificação dos participantes que nela aparecem.

Qualquer dado resultante do estudo só será apresentado para membros do quadro de professores e orientadores de forma anônima, sem conteúdo de áudio eventualmente gravado durante a entrevista.

Divulgação dos resultados

Os dados agregados e análises realizadas poderão ser publicados em publicações científicas e didáticas. Ao divulgarmos os resultados da pesquisa, nos comprometemos em preservar seu anonimato e privacidade, ocultando ou disfarçando toda informação (seja em texto, imagem, áudio, ou vídeo) que possa revelar sua identidade. As informações brutas coletadas não serão divulgadas.

Acompanhamento, assistência, e esclarecimentos

Todo material coletado será arquivado por no mínimo cinco anos. A Professora Simone Diniz Junqueira Barbosa, do Departamento de Informática da PUC-Rio, será a responsável por arquivar o material de pesquisa ao longo deste período e até o seu descarte. A qualquer momento, durante a pesquisa e até cinco anos após o seu término, você poderá solicitar mais informações sobre o estudo ou cópias dos materiais divulgados. Caso você observe algum comportamento que julgue antiético ou prejudicial a você, você pode entrar em contato para que sejam tomadas as medidas necessárias. Ao final deste termo você pode encontrar formas de contato conosco ou com a Câmara de Ética em Pesquisa da PUC-Rio.

Liberdade de recusa, interrupção, desistência e retirada de consentimento

Sua participação nesta pesquisa é inteiramente voluntária. Sua recusa não trará nenhum prejuízo a você, nem à sua relação com os pesquisadores ou com a universidade. A qualquer momento você pode interromper ou desistir da pesquisa, sem que incorra nenhuma penalização ou constrangimento. Você não precisará sequer justificar ou informar o motivo da interrupção ou desistência. Caso você mude de ideia sobre seu consentimento durante a sessão de estudo, basta comunicar sua decisão aos pesquisadores responsáveis, que então descartaram seus dados.

Consentimento

Eu, participante abaixo assinado(a), confirmo que:

1. Recebi informações detalhadas sobre a natureza e objetivos da pesquisa descrita neste documento e tive a oportunidade de esclarecer eventuais dúvidas;
2. Estou ciente de que minha participação é voluntária e posso abandonar o estudo a qualquer momento, sem fornecer uma razão e sem que haja quaisquer consequências negativas. Além disto, caso eu não queira responder a uma ou mais questÕes, tenho liberdade para isto;
3. Estou ciente de que minhas respostas serão confidenciais. Entendo que meu nome não será associado aos materiais de pesquisa e não será identificado nos materiais de divulgação que resultem da pesquisa. Professores orientadores só terão acesso ao material anonimizado;
4. Estou ciente de que a minha participação não acarretará qualquer ônus e que as atividades previstas na pesquisa não representam nenhum risco para mim ou qualquer outro participante;
5. Estou ciente de que sou livre para consentir ou não com a pesquisa, conforme as opções que marco abaixo:

Sobre a **coleta e uso de dados**:

[ ] Não autorizo o uso das informações coletadas descritas neste documento;

[ ] Autorizo o uso das informações coletadas conforme as condições descritas neste termo;

Sobre a **gravação de áudio**:

[ ] Não autorizo a gravação em áudio do que eu disser durante o estudo;

[ ] Autorizo a gravação em áudio do que eu disser durante o estudo;

Sobre a **gravação de vídeo**:

[ ] Não autorizo a gravação em vídeo das atividades que eu realizar;

[ ] Autorizo a gravação em vídeo das atividades que eu realizar;

A.2
Design Tools

Model Card

Instruções de Preenchimento

Preencha as seções do documento de acordo com o cenário apresentado. Pode haver informações que você terá que supor, ou precisasse tomar alguma ação para obter e seja difícil especular. Pedimos que deixe marcado caso isso ocorra, e também comente como poderia explorar o ponto ou obter a informação.

Caso tenha alguma dúvida, basta perguntar para o entrevistador.

Detalhes do Modelo

Informações básicas sobre o modelo.

- Consideração Exemplo

Usos Pretendidos

Casos de uso que foram pretendidos durante o desenvolvimento.

- Consideração Exemplo

Fatores

Considerações sobre outros fatores que possam impactar a performance do modelo.

- Consideração Exemplo

Métricas

Métricas deveriam ser escolhidas para refletir potenciais impactos do modelo no mundo real.

- Consideração Exemplo

Dados de Avaliação

Detalhes sobre o(s) dataset(s) usado(s) para avaliação quantitativa do modelo.

- Consideração Exemplo

Dados de Treino

Detalhes sobre o(s) dataset(s) usado(s) para treinar o modelo.

- Consideração Exemplo

Métodos de Análise Quantitativa

Fatores utilizados para análise quantitativa do modelo e resultados encontrados (fatores podem ser isolados ou para grupo na interseção).

- Consideração Exemplo

Considerações Éticas

Considerações éticas sobre o modelo criado.

- Consideração Exemplo

Cuidados e Recomendações

Cuidados e recomendações sobre os possíveis usos do modelo.

- Consideração Exemplo

Template de Metacomunicação Estendido

Instruções de Preenchimento

Preencha as seções do documento de acordo com o cenário apresentado. Pode haver informações que você terá que supor, ou precise tomar alguma ação para obter e seja difícil especular. Pedimos que, caso isso ocorra, deixe explícito e também comente como poderia explorar o ponto ou obter a informação.

Para cada pergunta guia, liste algumas respostas relacionadas à ferramenta que você está criando. Em todas as suas respostas, você deve manter um relacionamento “Eu-Você,” como se estivesse falando diretamente com o usuário final de sua ferramenta.

Ao final de cada seção há as perguntas éticas. Nesses casos, você deve revisitar suas afirmações anteriores buscando possíveis questões éticas, se apoiando nos Princípios da Bioética quando necessário. Questões éticas devem ser relacionadas às respostas que lhe levaram a identificá-las.

Caso tenha alguma dúvida, basta perguntar para o entrevistador.

Perguntas Guia

Análise

1. *O que eu sei, ou não sei, sobre você(s), e como eu sei isso?*

- Consideração Exemplo
    - Consideração Ética

2. *O que eu sei, ou não sei, sobre outras pessoas afetadas, e como eu sei isso?*

- Consideração Exemplo
    - Consideração Ética

3. *O que eu sei, ou não sei, sobre os contextos de uso que planejo, ou antecipo?*

- Consideração Exemplo
    - Consideração Ética

4. *Quais são as questões éticas que são levantadas pelo que eu aprendi? Por quê?*

-- Analise as considerações anteriores relacionando com esta pergunta --

Design

1. *O que eu projetei para você?*

- Consideração Exemplo
    - Consideração Ética

2. *Quais dos seus objetivos eu projetei o sistema para apoiar?*

- Consideração Exemplo
    - Consideração Ética

3. *Em que situações eu pretendo, ou aceito, que você use o sistema para atingir os seus objetivos? Por que?*

- Consideração Exemplo
    - Consideração Ética

4. *Como você deveria usar o sistema para atingir os seus objetivos, de acordo com o meu projeto?*

- Consideração Exemplo
    - Consideração Ética

5. *Para quais fins eu **não** quero que você use o sistema?*

- Consideração Exemplo
    - Consideração Ética

6. *Quais os princípios éticos que guiaram o meu projeto?*

-- Analise as considerações anteriores relacionando com esta pergunta --

7. *Como que o sistema que eu projetei para você está alinhado com esses princípios?*

-- Analise as considerações anteriores relacionando com esta pergunta --

Prototipação, Implementação e Avaliação Inicial

1. *Como eu construí o sistema para cumprir o meu projeto?*

- Consideração Exemplo
    - Consideração Ética

2. *O que eu introduzi no sistema para impedir usos e consequências não desejados?*

- Consideração Exemplo
    - Consideração Ética

3. *O que eu introduzi no sistema para ajudar a identificar e remediar efeitos negativos não antecipados?*

- Consideração Exemplo
    - Consideração Ética

4. *Quais são os cenários éticos que eu usei para avaliar o sistema?*

-- Analise as considerações anteriores relacionando com esta pergunta --

Avaliação contínua pós-implementação

1. *Quanto da minha visão foi refletida no uso real do sistema?*

- Consideração Exemplo
    - Consideração Ética

2. *Quais usos não antecipados foram feitos? Por quem? Como?*

- Consideração Exemplo
    - Consideração Ética

3. *Quais efeitos antecipados e não antecipados ocorreram com o uso do sistema? Quem está sendo afetado? Por quê?*

- Consideração Exemplo
    - Consideração Ética

4. *Quais questões éticas precisam ser manejadas através de re-projeto, re-desenvolvimento, ou até desconstrução do sistema?*

-- Analise as considerações anteriores relacionando com esta pergunta --

A.3
Design Briefs

Cenário Financeiro

Você é o líder de uma equipe de desenvolvimento contratada por uma instituição financeira. Sua equipe deve desenvolver um sistema de Inteligência Artificial para tomar decisões sobre empréstimos para a empresa. Seu algoritmo deve analisar o risco de cada cliente em potencial com base em seu perfil e histórico financeiro, e assim decidir se autorizará o empréstimo ou não, com o objetivo de maximizar o lucro esperado da companhia.

Você também deve (preencher o seguinte formulário/responder às seguintes perguntas) sobre o processo de desenvolvimento e potenciais problemas identificados no produto final. Estas questões são relacionadas a potenciais problemas éticos e considerações práticas sobre o sistema em desenvolvimento. Você pode não possuir todas as informações necessárias para responder adequadamente todas as perguntas/campos; contudo, nesses casos foi pedido que você inclua tópicos que você espera irão aparecer na versão final e que você pretende investigar.

Cenário Educacional

Você é o líder de uma equipe de desenvolvimento contratada por uma universidade. Sua equipe deve desenvolver um sistema de Inteligência Artificial para atribuir notas aos candidatos ao programa da universidade. Seu algoritmo deve analisar as realizações anteriores dos candidatos e atribuir uma nota a cada um deles que será utilizada pelo comitê de seleção para decidir quais estudantes serão admitidos ou não. Seu algoritmo não será utilizado para atribuir nota a novos trabalhos dos candidatos.

Você deve (preencher o seguinte formulário/responder às seguintes perguntas) sobre o processo de desenvolvimento e potenciais problemas identificados no produto final. Estas questões são relacionadas a potenciais problemas éticos e considerações práticas sobre o sistema em desenvolvimento. Você pode não possuir todas as informações necessárias para responder adequadamente todas as perguntas/campos; contudo, nesses casos foi pedido que você inclua tópicos que você espera irão aparecer na versão final e que você pretende investigar.

A.4
Summary of Bioethical Principles

Princípios da Bioética

Quatro princípios "prima facie" da bioética.

- Proporcionam uma forma simples e culturalmente neutra de abordar questões éticas em práticas clínicas.
- Auxiliam profissionais de saúde na tomada de decisões que refletem questões morais no ambiente de trabalho

Os 4 Princípios da Bioética

Autonomia

- Capacidade para indivíduos pensarem, decidirem e agirem com base em seus próprios pensamentos e decisões com liberdade e independência.
- Para respeitar a autonomia, deve-se possibilitar que indivíduos cheguem às suas próprias conclusões.
    - Essas conclusões devem ser respeitadas quer eles concordem ou não com elas.

Beneficência

- Ativamente fazer o que for melhor para o paciente.
- Baseado em um juízo objetivo do médico, e no que ele acredita ser melhor para o paciente.
- Decisões médicas podem entrar em conflito com visões do paciente, portanto podendo entrar em conflito com autonomia.
- Sobreposição de decisão do paciente sobre o médico é conhecido como paternalismo médico. Isso nunca ocorre na prática.

Não-maleficência

- Não causar danos ao paciente.
- Atuação profissional tem como objetivo o que é melhor para o paciente, beneficência, mas cirurgias acarretam em riscos.
- Prática deve ponderar risco e benefícios de potenciais tratamentos, ou beneficência ou não-maleficência.

Justiça

- Todos os pacientes em circunstâncias semelhantes devem receber de forma igual o melhor tratamento possível.
- Fator chave na alocação de serviços/recursos
    - Para aumentar fundos para serviços de atendimento a acidentes e emergências, é justo restringir recursos de saúde mental.
- Limitações de tempo e recurso significam que nem todos os pacientes recebem o melhor tratamento possível.

A.5
Interview Script

Roteiro de Entrevista

Termo de Consentimento

- Leitura do termo de consentimento e obtenção do consentimento
- https://www.overleaf.com/project/5fd97b7e6336f3305a0aecf1

Perguntas Preliminares

Antes de começarmos o estudo, gostaríamos de fazer algumas perguntas sobre você para conhecermos melhor o seu perfil.

1. Qual sua área de atuação no momento?
2. Em que área você teve sua educação formal?
3. Há quanto tempo você trabalha no desenvolvimento de modelos de AM?
4. Qual processo você costuma seguir para desenvolver modelos de AM?
5. Qual seu conhecimento prévio sobre Engenharia Semiótica?
6. Você possui algum conhecimento prévio sobre o Model Card ou o Template de Metacomunicação Estendido?
7. Você possui interesse em problemas éticos no design e desenvolvimento de software?
8. Você possui alguma experiência em considerar problemas éticos no design e desenvolvimento de software?
9. Quando você pensa sobre problemas éticos, o que lhe vem à mente?
10. Você já teve contato com algum programa que você considera antiético? Como foi essa experiência?

Estudo

“O estudo se baseia em um cenário sobre o qual vou pedir para você refletir. Primeiro, vamos ler o cenário. Depois, vou explicar os princípios da Bioética, que você poderá usar para ajudar a sua reflexão. Finalmente, vou pedir para você (preencher o formulário / responder as perguntas do template) levando em consideração o cenário e os princípios da Bioética.”

“Quanto ao cenário, pode haver pontos onde seja difícil especular sobre uma informação desejada a partir das informações presentes no cenário. Nesses casos, pedimos que você deixe claro como você exploraria esse ponto e obteria as informações necessárias.”

Leitura do Cenário

Cenário 1 (Financeiro)

Você é o líder de uma equipe de desenvolvimento contratada por uma instituição financeira. Sua equipe deve desenvolver um sistema de Inteligência Artificial para tomar decisões sobre empréstimos para a empresa. Seu algoritmo deve analisar o risco de cada cliente em potencial com base em seu perfil e histórico financeiro, e assim decidir se autorizará o empréstimo ou não, com o objetivo de maximizar o lucro esperado da companhia.

Você também deve (preencher o seguinte formulário/responder às seguintes perguntas) sobre o processo de desenvolvimento e potenciais problemas identificados no produto final. Estas questões são relacionadas a potenciais problemas éticos e considerações práticas sobre o sistema em desenvolvimento. Você pode não possuir todas as informações necessárias para responder adequadamente todas as perguntas/campos; contudo, nesses casos foi pedido que você inclua tópicos que você espera irão aparecer na versão final e que você pretende investigar.

Cenário 2 (Educação)

Você é o líder de uma equipe de desenvolvimento contratada por uma universidade.Sua equipe deve desenvolver um sistema de Inteligência Artificial para atribuir notas aos candidatos ao programa da universidade. Seu algoritmo deve analisar as realizações anteriores dos candidatos e atribuir uma nota a cada um deles, a qual será utilizada pelo comitê de seleção para decidir quais estudantes serão admitidos ou não. Seu algoritmo não será utilizado para atribuir nota a novos trabalhos dos candidatos.

Você também deve (preencher o seguinte formulário/responder às seguintes perguntas) sobre o processo de desenvolvimento e potenciais problemas identificados no produto final. Estas questões são relacionadas a potenciais problemas éticos e considerações práticas sobre o sistema em desenvolvimento. Você pode não possuir todas as informações necessárias para responder adequadamente todas as perguntas/campos; contudo, nesses casos foi pedido que você inclua tópicos que você espera irão aparecer na versão final e que você pretende investigar.

Leitura sobre Princípios da Bioética

“Podemos agora passar para os princípios da Bioética, que você poderá considerar para ajudar a sua reflexão sobre o cenário.”

Preenchimento da Ferramenta

“Antes de começar, vamos dar uma passada pela ferramenta para mostrar como ela funciona.”

== *Feita a Explicação da Ferramenta* ==

“Vou pedir então que você abra os links que mandei por mensagem no chat e compartilhe sua tela com essas abas. Você deve conseguir acessar o cenário, a explicação sobre os princípios da Bioética, e a ferramenta que você vai preencher. Peço também que fale o que você está pensando enquanto preenche a ferramenta.”

== *Feita o Preenchimento da Ferramenta* ==

Perguntas Pós-Estudo

10. O que você achou do Model Card / Template de Metacomunicação Estendido? (probe: por favor elabore)

11. Como o Model Card / Template de Metacomunicação Estendido influenciou sua reflexão ética no cenário apresentado? Como auxiliou em sua reflexão? Como limitou ou prejudicou sua reflexão?

12. Em uma escala de 1 a 7, como você avalia o formulário / perguntas em relação a auxiliar sua reflexão? (1 = prejudicou muito; 7 = ajudou muito)

13. Houve algo que você incluiu que não se encaixava bem em nenhuma das seções? Se sim, o quê?

14. Houve algo que você gostaria de incluir mas não se encaixava bem em nenhuma das seções? Se sim o quê?

15. Sobre o que o formulário / as perguntas o/a ajudou a pensar? (probes: sobre o sistema, o processo de desenvolvimento, consequências de uso do sistema a seus usuários e a sociedade como um todo)

16. Como (o Model Card / Template de Metacomunicação Estendido) lhe ajudou a considerar as consequências daquilo que você estava desenvolvendo como solução para o cenário dado? Por exemplo, considerando stakeholders, minorias, questões econômicas, etc.

17. Quais seções o/a ajudaram a refletir sobre potenciais problemas éticos? (probe: como?)

18. Quem você considerou que ia ler o que você estava escrevendo?

19. Quais outros indivíduos afetados pelo modelo você acredita que se beneficiariam do documento resultante? (probe: usuários, reguladores, pessoas impactadas)? Como?

20. Quais outros indivíduos afetados pelo modelo você acredita que se beneficiariam de **passar** pelo processo de preencher o formulário / responder às perguntas?

Perguntas após a ***Segunda Sessão***

22. Como você explicaria a diferença entre as ferramentas de forma simples para uma nova pessoa de sua equipe?

23. Como você compararia as ferramentas em respeito à variedade de considerações que você fez sobre o cenário? Por que?

24. Considerando que você usou anteriormente o (Model Card, Template Estendido), como você acha que esta sessão foi influenciada? O que mais?

Perguntas opcionais após segunda sessão (indutivas)

25. O que você achou da formulação “eu-você” no Template de Metacomunicação?

26. Como você acha que ter que escrever e escolher as palavras para representar o que você estava pensando impactou sua reflexão?

27. Quanto à separação das perguntas nas diferentes fases e aspectos do desenvolvimento, o que você acha que isso impactou na sua reflexão?

28. O quão importante você acha que os princípios de Bioética foram para sua reflexão?

29. Como você se sentiu em relação ao cenário após o seu processo de reflexão?

30. Como o campo aberto sobre reflexões éticas no Model Card influenciou sua reflexão?

B
Consolidated Codebook

The codebook we arrived at through an initial round of open coding followed by a consolidation process via negotiated agreement contains a total of 62 codes, organized around 7 categories. Each following section contains the codes pertaining to one of these categories.

B.1
Data

1. **Data Appropriation**: Participant discusses ways in which the data involved may be used in the real world.
2. **Data Bias**: Participant discusses the biases represented in the data being fed to the system.
3. **Data Limitations**: Participant discusses the data's inherent limitations.

B.2
Designer

1. **Designer Revisiting Previous Statement**: Participant mentions revisiting a previous statement.
2. **Designer's Culture and Values**: Participant discusses their own culture and values.
3. **Designer's Ethics and Design Decisions**: Participant discusses some of the ethical implications of their design decisions.
4. **Designer's Focus on Language**: Participant focuses on their language use, whether it communicates what they want to communicate, whether they fit with the question, etc.
5. **Designer's Focus on Process**: Participant provides a process-driven perspective on some aspect of their design.
6. **Designer's Focus on Technical Aspects**: Participant provides a technical perspective on some aspect of their design.

7. **Designer's Goals**: Participant discusses the goals of their design.
8. **Designer's Limitations**: Participant discusses their own limitations, as designers.
9. **Designer's Propositions about Scenario**: Participant relates some of their propositions about the design scenario.
10. **Designer's Responsibility**: Participant discusses their responsibility as designers.
11. **Designer's Restrictions**: Participant discusses restrictions imposed on themselves, as designers.
12. **Designer's Understanding**: Participant discusses what the designer knows about the situation.
13. **Designer's View on Reality Matching with Design**: Participant discusses how reality conforms or goes against what they have designed.

B.3 EMT

1. **EMT Collaboration with Developers**: Participant discusses how the EMT might be collaboratively used by designers and developers.
2. **EMT Collaboration with Stakeholders**: Participant discusses how the EMT might be collaboratively used by designers and stakeholders.
3. **EMT Non-Linearity**: Participant discusses how the EMT questions may be answered non-linearly.
4. **EMT Read by Developers**: Participant discusses how the EMT could be read by developers.
5. **EMT Read by Stakeholders**: Participant discusses how the EMT could be read by stakeholders.
6. **EMT's Comprehensive Questions**: Participant discusses how comprehensive the EMT questions are.
7. **EMT's Confusing Questions**: Participant discusses how some of the Template's questions are confusing.
8. **EMT's Difficult Questions**: Participant discusses some of their difficulties in answering EMT questions.

9. **EMT's I-You Framing**: Participant discusses the Template's I-You framing.
10. **EMT's Question Impacts**: Participant discusses some of the impacts of the Template's questions.
11. **EMT's Redundant Questions**: Participant discusses how some of the Template's questions are redundant.
12. **EMT's Relation to MC**: Participant relates aspects of the EMT to aspects of the MC.
13. **EMT's Space for Reflection**: Participant discusses how the Template provides space for reflection.

B.4 Bioethical Principles

1. **Framing around Principle of Autonomy**: Participant frames the situation through the lens of the principle of Autonomy.
2. **Framing around Principle of Beneficence**: Participant frames the situation through the lens of the principle of Beneficence.
3. **Framing around Principle of Justice**: Participant frames the situation through the lens of the principle of Justice.
4. **Framing around Principle of Non-Maleficence**: Participant frames the situation through the lens of the principle of Non-Maleficence.
5. **Mention of Principle of Autonomy**: Participant explicitly mentions the principle of Autonomy.
6. **Mention of Principle of Beneficence**: Participant explicitly mentions the principle of Beneficence.
7. **Mention of Principle of Justice**: Participant explicitly mentions the principle of Justice.
8. **Mention of Principle of Non-Maleficence**: Participant explicitly mentions the principle of Non-Maleficence.
9. **Bioethical Principles' Impacts**: Participant discusses the impact of having the Bioethical Principles at hand.

B.5 Stakeholder

1. **Stakeholder's Limitations**: Participant discusses the limitations of the stakeholders involved.
2. **Stakeholder's Motivations**: Participant discusses the motivations behind stakeholders' actions.
3. **Stakeholders' Multiple Interests**: Participant discusses the multiple interests that exist within the group of stakeholders.
4. **Stakeholder's Responsibility**: Participant discusses the stakeholder's responsibility for the system's impacts.
5. **Stakeholder's Understanding**: Participant discusses what the stakeholder knows about the situation.

B.6 User

1. **User's Bias**: Participant discusses ways in which users' decisions may be biased.
2. **User's Interactions**: Participant discusses user's interactions with the system.
3. **User's Limitations**: Participant discusses the user's inherent limitations.
4. **User's Motivations**: Participant discusses the motivations behind users' actions.
5. **User's Responsibility**: Participant discusses user's responsibility for the system's impacts.
6. **User's Restrictions**: Participant discusses restrictions imposed on the user, in terms of interactions, resources, etc.
7. **User's Understanding**: Participant discusses what the user knows about the situation.

B.7 System

1. **System's Appropriation**: Participant discusses ways in which the system involved may be used in the real world.
2. **System's Bias**: Participant discusses ways in which the system's decisions may be biased.
3. **System's Ethical Impact Testing**: Participant discusses how ethical impacts can be detected, via testing, and measured.
4. **System's Explanations**: Participant discusses the system's explainability.
5. **System's Fixes**: Participant discusses how the system's behavior may be fixed.
6. **System's Impacted Stakeholders**: Participant discusses which stakeholders may be impacted by the system's use.
7. **System's Impacts**: Participant discusses the impacts of the system's use.
8. **System's Limitations**: Participant discusses the system's inherent limitations.
9. **System's Outputs**: Participant discusses the outputs that the system provides.
10. **System's Responsibility**: Participant discusses the system's responsibility for its impacts.
11. **System's Restrictions**: Participant discusses restrictions imposed on the system in terms of interactions, resources, etc.
12. **System's Safeguards**: Participant discusses the system's safeguards put in place. About how they are created or about how they are avoided.