Pontifícia Universidade Católica
do Rio de Janeiro

José Luiz Nunes

Evaluating approaches for developers' ethical reasoning and communication about Machine Learning models

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
October 2021

Pontifícia Universidade Católica
do Rio de Janeiro

José Luiz Nunes

Evaluating approaches for developers' ethical reasoning and communication about Machine Learning models

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. Hélio Côrtes Vieira Lopes
Departamento de Informática – PUC-Rio

Rio de Janeiro, October 1st, 2021

José Luiz Nunes

Bachelor in Law at Escola de Direito do Rio de Janeiro of Fundação Getulio Vargas (FGV Direito Rio). Since 2019 he works as a researcher in the same Institution.

Bibliographic data

Nunes, José Luiz

Evaluating approaches for developers' ethical reasoning and communication about Machine Learning models / José Luiz Nunes; advisor: Simone Diniz Junqueira Barbosa; co-advisor: Clarisse Sieckenius de Souza. – Rio de Janeiro: PUC-Rio, Departamento de Informática, 2021.

v., 129 f: il. color. ; 30 cm

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

Inclui bibliografia

1. Aprendizado de máquinas;. 2. Considerações éticas;. 3. Avaliação de modelos;. 4. Transparência;. 5. Model Cards;. 6. Template Estendido de Metacomunicação;. 7. Engenharia Semiótica;. I. Barbosa, Simone Diniz Junqueira. II. de Souza, Clarisse Sieckenius. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

Acknowledgments

I would like to thank:

- my advisers, for their support and help with the research, as well as accepting a student from a different field, without them and everything they have taught me this dissertation would not have been possible.
- Gabriel Barbosa and Dalai Ribeiro, for their collaboration in the work carried for this dissertation, as well as all members of the IDEIAS-SERG research group for their input.
- I am grateful to Clara Almeida, Fernando Correa, and Guilherme Almeida, for their friendship and their time to carefully read and suggest changes to previous versions of this text. Also, to Kaline Santos, Alexandre Almeida, and Felipe Silva for their companionship, and all that I learned working with them.
- Thiago Bottino, Ivar Hartmann, and Moacyr Alvim, in name of everyone with whom I worked with during my undergrad. Thanks to them I learned what I know about research in Law and had the opportunity to work and study data science.
- not least, my family, especially my grandparents, parents, and sister for all their support throughout the years.

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

Abstract

Nunes, José Luiz; Barbosa, Simone Diniz Junqueira (Advisor); de Souza, Clarisse Sieckenius (Co-Advisor). **Evaluating approaches for developers' ethical reasoning and communication about Machine Learning models**. Rio de Janeiro, 2021. 118p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Machine learning algorithms have become widespread for a wide array of tasks. However, there is still no established way to deal with the ethical issues involved in their development and design. Some techniques have been proposed in the literature to support the reflection and/or documentation of the design and development of machine learning models, including ethical considerations, such as: (i) Model Cards and (ii) the Extended Metacommunication Template. We conducted a qualitative study to evaluate the use of these tools. We present our results concerning the use of the Model Card by participants, with the objective of understanding how these actors interacted with the relevant tool and the ethical dimension of their reflections during our interviews. Our goal is to improve and support techniques for developers to disclose information about their models and reflect ethically about the systems they design. Furthermore, we aim to contribute to the development of a more ethically informed and fairer use of machine learning.

Keywords

Machine Learning; Ethical Reasoning; Model Evaluation; Transparency; Model Cards; Extended Metacommunication Template; Semiotic Engineering;

Resumo

Nunes, José Luiz; Barbosa, Simone Diniz Junqueira; de Souza, Clarisse Sieckenius. **Avaliando técnicas para a reflexão ética e comunicação sobre modelos de aprendizado de máquinas para desenvolvedores**. Rio de Janeiro, 2021. 118p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O uso de modelos de aprendizado de máquina se tornou ubíquo para um leque diverso de tarefas. Contudo, ainda não há nenhuma forma estabelecida para refletir sobre questões éticas em seu processo de desenvolvimento. Neste trabalho, realizamos um estudo qualitativo para avaliar duas técnicas propostas pela literatura para auxiliar desenvolvedores a refletirem sobre questões éticas relacionadas à construção e uso de modelos de aprendizado de máquina: (i) Model Cards; e o (ii) Template Estendido de Metacomunicação. Apresentamos nossos resultados a respeito do uso do Model Card pelos participantes, com o propósito de entender como esses atores interagiram com a ferramenta, assim como a dimensão ética de sua reflexão durante nossas entrevistas. Nosso objetivo é melhorar técnicas para desenvolvedores disponibilizaram informações sobre seus modelos e que a reflexão ética sobre os sistemas que desenvolveram. Além disso, nosso trabalho tem como objetivo contribuir para o desenvolvimento de um uso mais justo e ético de sistemas de aprendizado de máquina.

Palavras-chave

Aprendizado de máquinas; Considerações éticas; Avaliação de modelos; Transparência; Model Cards; Template de Metacomunicação Estendido; Engenharia Semiótica;

Table of contents

List of figures

List of tables

1
Introduction

Machine Learning (ML) algorithms have become both ubiquitous in our society and a staple in many industries. These models have been rapidly deployed with expectations of huge efficiency and monetary gains, stemming from the automation of decision making. These high expectations were also coupled with promises of greater equality and the elimination of human error. Much research has gone into showing that these systems do indeed present bias, especially when using data captured from previous human behavior and decision making, and also from developers' own bias in deciding how to frame a problem, the variable to be predicted, and what to include in the model (Barocas and Selbst, 2016).

Fairness concerns, such as the issues just mentioned, have been accompanied by accountability and transparency. This has been the case at least since the inaugural 2014 Fairness, Accountability and Transparency Machine Learning (FAT-ML) conference, which provided an academic venue with focus on addressing such issues. According to the conference scope, this was partially in answer to a call to action from governmental organizations,[1] revealing that the topic's relevance was already identified outside of academia and the research context.

Substantive discussion has also arisen on the issue of explainability. Much research has been dedicated to this topic, which is not new. In the 1990's there was already research on making intelligent systems explainable, then based on knowledge representation and logical rules (*e.g.*, Swartout and Moore (1993)). This problem has shifted to deal with more complex algorithms that learn patterns from large amounts of data.[2] However, these issues have not yet been properly addressed in policy and practice. For example, the right to explanation was included in some early drafts of the European General Data Protection Regulation (GDPR), but ultimately excluded from the final text (Wachter et al., 2017).

The relevance of all these issues has also been recognized as part of a theoretical framework for the development of artificial intelligence (AI).

[1] https://www.fatml.org/schedule/2014/page/scope-2014

[2] Mittelstadt et al. (2019) provide an overview of current proposals.

They show up in various documents made by governmental and academic organizations. Floridi and Cowls (2019) include them by analyzing these documents, and conclude that the four principles of bioethics, with the addition of explicability, can be used as general principles for the adoption of AI in Society. We explore these topics in Chapter 2.

Summing up, there is a rising concern with how to adopt ML systems in society while avoiding social harm. This comes not only from Computer Science research, but also from the legal literature. For instance, Desai and Kroll (2017) discuss transparency and accountability, and argue that transparency understood as opening source code is not necessary for the desired compliance with legal rules and standards. They propose technical tools that would be able to verify desired characteristics in decision-making software. Therefore, these could be used to verify their compatibility with legal and policy standards.

This sets the stage for a general debate on ethical issues in algorithmic decision making. Our research investigated tools that aim to promote ethical reasoning of developers about models they contribute to. As will be shown, the tools under study can also be used as a way to share information about AI models with other stakeholders, allowing for an increase in transparency and accountability of the algorithms.

This research is relevant not only from an academic perspective, but it is a topic that has been acquiring policy relevance. For instance, we can cite the Brazilian bill "PL 21/2020", which creates a legal landmark for Artificial Intelligence, and includes a provision for public entities to request an impact report for entities that deploy these systems, which would be a documentation for the system including its potential risks and measures and safeguards taken to avoid them.

We conducted an interview study on tools that developers currently use to reflect on potential ethical issues about their software, which can also be revealed to other actors and stakeholders, without necessarily disclosing sensitive information, such as the source code or data used to train their software. To do this, we conducted an interview-based study with eight participants who had previous experience in developing ML models.

We asked participants to use two tools proposed in the Computer Science literature, which differ substantially in their representation and theory. The first one is the Model Cards (MC) representation, proposed by Mitchell et al. (2019). The second one is the Extended Metacommunication Template (EMT), a semiotic engineering tool proposed in Barbosa et al. (2021). One key design difference between these tools is that, while Model Cards are meant to bring transparency and provide information about the resulting model after being

deployed/implemented, the Template should be used since the design phase of the system, to guide the development, implementation, and even follow up work on the system in use.

Our main goal with the larger study is to understand the ethical considerations made by participants in our study about the system they were developing, and how these may have been influenced by and registered into these tools. A more general goal was to collect data to support a range of long-term studies that analyze the product of these artifacts and compare the ethical reasoning promoted by them. Furthermore, conducting more interviews in the future would also allow the comparison of other artifacts that may be proposed over time.

The results presented in this work are part of this larger ongoing research effort. In this dissertation, in particular, we focus on the data related to the Model Cards representation. We have chosen to narrow our scope in order to be able to make an in-depth analysis of the data collected. Due to time constraints, we have opted to leave an analysis of the data collected about the Extended Metacommunication Template for future work.

Our broader research question is "How do Model Cards and the Extended Metacommunication Template contribute to stakeholders' ethical reasoning and how do they differ?". The narrower question, which we intend to answer in this dissertation is "How do Model Cards contribute to developers' ethical reasoning and what ethical issues does it help identifying?". Our main contributions are a better understanding of how Model Cards may be used by developers, what relevant information is recorded into them, and insights into how to promote fairer AI systems, including through transparency and developers ethical reasoning.

The remainder of this document is organized as follows. Chapter 2 reviews the overarching literature and discussion around fair machine learning and transparency, as the background of what has been proposed for the ethical reflection upon ML systems. Chapter 3 presents related works, especially proposed tools, that share the goal of our research, and both proposals included in our study. Chapter 4 describes the methodology and design of the study conducted, while Chapter 5 details our methodology of analysis and coding of the data collected during our study. Finally, Chapter 7 goes over our main findings, and how they communicate to the literature we aim to contribute.

2
Background for this Research

This chapter describes relevant research that sets the background for fairness, transparency, and accountability discussion in AI. This overview depicts the overarching current debate we intend to contribute to, and how each of these topics connect with one another.

In recent years, ML algorithms have been deployed to support a wide array of tasks across different areas. This was followed by a number of reports of algorithms revealing failed predictions, sometimes related to specific population groups; opaque decision making; and human actors who could provide no further details on the issues, even if the result could be life-lasting negative consequences to the subject (*e.g.*, Brandenburg (2011)).[1]

A famous report on a wide range of such cases was written by O'neil (2016). She reported on a number of algorithms being used both by the public and private sector, whose applications were leading to wrong incentives, negative results to humans, or frustration due to the opacity of the decision process and inability of the organization to provide more information. She also suggested some causes for these problems, such as the use of biased data produced by human behavior.

Much academic research has been developed to address these issues and improve upon existing ML systems. One such instance of the increased importance and attention of this research was the creation of the yearly conference on Fairness, Accountability and Transparency. This Conference started as the FAT Machine Learning Conference, and was later included in the Association for Computing Machinery's (ACM) yearly calendar as ACM FAT (renamed, in 2021, to ACM FAccT).

Another ethical framework that has been suggested by some authors is the adoption of the four principles of bioethics. These were originally suggested by Beauchamp and Childress (2019), in their book first published in 1979.

We address here four areas of this literature to establish a theoretical background for our research: first the three concepts of (i) fairness;

[1]Much of this criticism could also be directed towards human decision-making processes conducted or even non machine learning algorithms, which are not the object of our work.

(ii) transparency; and (iii) accountability; followed by (iv) the four principles of bioethics.

The inclusion of these principles and theories does not imply we believe they are the only applicable ones, or the ones that should be adopted. These principles have found some consensus around their relevance and meaning in the literature, being used by authors even outside of the computer science community. For example, in his philosophical work, Coeckelbergh (2020) uses these principles while discussing bias and the impact of AI for the future of society. However, other ethical frameworks could be used to guide our use of AI and to apply to the analysis resulting from the tools shown in Chapter 3.

2.1 Fairness

The concept of fairness has been widely discussed in Philosophy, including what is fair both from a social and an individual perspective. However, the literature on fair machine learning has focused on definitions that can be stated in mathematical terms, thus distancing themselves from definitions discussed by moral and political philosophers (Binns, 2018). Most commonly used definitions for fairness in ML focus on formal mathematical definitions, and how to operate them within computational boundaries (Corbett-Davies and Goel, 2018). Heidari et al. (2019) show that such definitions can be reconciled with Equality of Opportunity concepts derived from the work of philosophers such as John Rawls, which require that "[t]hose who are at the same level of talent and ability, and have the same willingness to use them, should have the same prospects of success regardless of their initial place in the social system" (Rawls, 1971, p.63). This shows that, while they may not be directly related, or straightforwardly derived from them, they can share common assumptions.

Furthermore, definitions found in the literature have also been derived from legal definitions of discrimination used in American law, specifically those of disparate treatment and disparate impact (Corbett-Davies and Goel, 2018; Sunstein, 2019). Disparate treatment is the most direct case of discrimination, where a person of a specific group receives a different treatment just for being a part of it. One example of this would be for a public official to favor men over women (Kleinberg et al., 2018). In that case, the discrimination lies on the direct unequal treatment of individuals.

Disparate impact, by contrast, does not imply a direct difference in treatment simply because someone belongs to a group. It happens when a certain rule or policy has an adverse impact that is disproportionate for a

certain group. Thus, it does not require any discriminatory purpose to exist. Here, what characterizes discrimination is not the resulting unequal treatment for a group *per se*, but whether there is a valid justification for the policy. For example, if we require candidates for a job to be able to perform a certain task which a certain group of people might be less likely to accomplish, what will decide whether it is a case of disparate impact is whether that task is indeed relevant for performing the job in question.

Starting from these definitions carries some advantages. They are narrow and, although they will not cover every case of what could be considered unfair, they can be supplemented of other definitions of unfairness. Furthermore, they have been thoroughly used in decision making, thus being easier to operationalize. They can also be directly tested, as has been proposed by Galhotra et al. (2017).

With that in mind we can arrive at a few different definitions applicable to ML, as stated by Corbett-Davies and Goel (2018). A simple one would be to require that the algorithms do not make decisions based on protected characteristics, meaning that they would not be used as variables for training,[2] in what they call *anti-classification*. An alternative definition would be that of "classification parity", which requires that relevant performance measures of the algorithm remain the same across different groups. Thus, in order to test such metrics, the minimal requirement is to store these characteristics, even if they are not directly used in the model. Finally, what they call "calibration", which means that classification should have the same probability across different protected groups, conditional on risk score.

2.2 Transparency

Artificial Intelligent systems have commonly been regarded as black boxes (Rudin, 2019), especially outside of the Computer Science literature (Pasquale, 2015; Bathaee, 2017). The use of this term usually implies high complexity and lack of knowledge of how these systems work (*e.g.*, neural networks), of what causes them to return a given prediction, or of an explanation about their decisions.

Increasing transparency on the building process and specifications of these systems is also a way to support fairer systems, as it allows for easier finding and proving cases of discrimination, as long as it allows greater insight into the decision process and reasoning carried out by algorithms (Kleinberg

[2]This assumes there are no other correlated variables that could be used as a proxy by the algorithm.

et al., 2018). Moreover, transparency also supports higher accountability of intelligent systems, as discussed in the next section. This can be done even if such transparency does not result in an explanation of the inner working of algorithms.

Transparency can also be achieved by revealing information about the software development process. This information allows relevant stakeholders to understand decisions made during development that can impact the results, including trade-offs of different values or goals (Kleinberg et al., 2018). However, transparency does not imply that the source code must be made public or auditable by other entities.

One such example of information that could be disclosed is what type of data was used to train a model, which might imply it will reproduce some biased past human behavior, or that the data underrepresented a certain group of people, resulting in poor prediction for such cases. Another information would be the scenarios for which the software was developed and tested, which might imply different standards or expectations for the system, even if the output of the algorithm would be the same.

2.3 Accountability

Accountability is an issue closely related to the legal literature. In legal practice, it is concerned with making actors, especially those that create and make out the product or service, accountable for the results of the deployed ML software. This definition stems from political accountability. Furthermore, it is also concerned with how to align computer systems with existing legal and policy choices (Kroll et al., 2017).

Desai and Kroll (2017) argue that, while the given definition of accountability stands in legal literature, it is not understood as such in Computer Science. In Computer Science, the concept of accountability is "[a]bout making sure that software produces evidence allowing oversight and verification of whether it is operating within agreed-upon rules"(Desai and Kroll, 2017, p. 10).

It is important to note that these meanings are deeply intertwined. If our goal is to hold agents that employ AI systems accountable for the results of their decision making, then being able to generate evidence and oversight over those systems, and ensuring that they do (or that we can identify if they do not) comply with policy and legal rules is a necessary step.

Accountability is also directly related to fairness and transparency discussions. The stated technical side of accountability is a direct application of

transparency, of being able to inspect the developers' product, choices and decision making. Furthermore, one of the main agreed-upon rules intelligent systems should operate with are the few legally agreed-upon concepts of fairness and discrimination. Hence, accountability and transparency are what allow stakeholders and society to uphold software to our politically agreed fairness standards.

2.4 Four principles of Biomedical Ethics

The four principles of Biomedical Ethics were suggested as a moral framework to aid in the resolution of ethical issues in medicine. The four principles are (i) autonomy; (ii) non-maleficence; (iii) beneficence; and (iv) justice (Beauchamp and Childress, 2019). While these principles were established as a way to guide ethical decisions in the medical field, Floridi and Cowls (2019) argue that they provide a good general framework for the adoption of ML algorithms in society.

These principles adapt well to serve as a moral framework for ethical challenges posed by ML algorithms. However, Floridi and Cowls (2019) also warn that their exact meaning is contested, and that some similar terms have been used with different meanings. Furthermore, they also argue for the inclusion of explicability as an addition to this set of principles.

These principles were also used by Barbosa et al. (2021) as an ethical framework to exemplify the use of the Extended Metacommunication Template. We offer here a brief explanation of each of the four principles. These will focus on the definition given by Floridi and Cowls (2019), which considered multiple documents made by governmental, academic and multi-stakeholders organizations which tried to establish principles for the use of AI.

2.4.1 Autonomy

This principle is related to human agency. Even though ML algorithms replace human decision making in an array of tasks, we delegate human autonomy in these processes to technology. This should be done in a way that preserves human autonomy and power to decide, especially in more critical situations.

Some of the documents stated that the potential for humans to decide standards or principles should be protected; that humans should have their decision power preserved in some situations; be able to choose when and where decision making will be transferred to AI systems; or even be able to override

automated decisions. Floridi and Cowls (2019) identify that this principle imposes a twofold restriction (i) promote human autonomy; and (ii) restrict autonomy of machines, either by deciding where it will be used or by making it reversible.

2.4.2 Non-maleficence

The simplest way to state the non-maleficence principle is as "do no harm". This cautions us against negative consequences that may arise from the misuse of AI. While it is not clear exactly where these risks would arise from – the technologies themselves or people developing it –, it seems agreed upon that one well-established concern is for violations of privacy.

2.4.3 Beneficence

This principle states that AI should be used where it will promote benefits for humans. In other words, the development of ML systems should promote social benefits and human well being.

This principle may seem like an equivalent to non-maleficence, and that avoiding harm would mean promoting good. However, as simple as this principle may seem, it was stated differently in all six of the documents analyzed by Floridi and Cowls (2019).

2.4.4 Justice

The last of the four principles is the principle of justice. It recognizes that the capacity to delegate decisions is not equally distributed among people, nor are the benefits reaped from its adoption.

One common concern included in the justice principle is to avoid discrimination and bias towards certain groups. This is especially relevant considering that ML algorithms can reproduce human discriminatory behavior, creating a perverse feedback loop.

This principle is intrinsically related to that of fairness mentioned earlier. And while it is easy to agree that algorithms should be deployed in a just way, it may be very hard to reach a concept of justice that is widely agreed upon.

3
Related Work and Objects of Study

In this section we present the current state of the literature on our topic. We searched the proceedings of the ACM FAccT Conferences[1] and the FAT Machine Learning Conferences[2] –conferences focused on fairness, accountability, and transparency issues in computing-related areas, with special focus on machine learning–for proposals of tools to document AI algorithms. We present them, and other works they reference, in the following sections.

Then, we more extensively go through the work of Mitchell et al. (2019), which proposes the Model Card as a tool for reporting machine learning models, and Barbosa et al. (2021), which proposes the Extended Metacommunication Template, building upon previous Semiotic Engineering research. These tools not only document intelligent systems, but have the specific goal of aiding the ethical reflection of machine learning algorithms, and were used in our study, further described in Chapter 4.

3.1
Related Work

In this section we focus on other tools proposed for documenting the development process of AI algorithms and datasets, which serve as basis for their training or evaluation. We organize them in two different groups: the first focuses specifically on datasets and the second describes other tools to document AI systems. Finally, we briefly mention toolkits and guidelines, proposed mainly by industry actors to highlight the human agents that interact with AI systems.

3.1.1
Dataset Documentation

A group of works are dedicated to documenting datasets and making their information available to other stakeholders. Constructing or choosing, and analyzing a dataset are important steps in the development process of machine learning projects. Due to the work and cost associated to creating annotated

[1] https://facctconference.org/
[2] https://www.fatml.org/

and curated data, these are very important in machine learning projects, and are used for a vast amount of different models and distinct research projects.

These proposals aim to standardize this aspect of the development of AI systems, documenting their characteristics, how the data is analyzed, how it is distributed, and possible biases in the data. Among the works we found, these were the first to appear in the literature, and include both standardized and automated framework for the analysis of datasets, and documents with characteristics of their data that would be manually produced by developers.

Holland et al. (2018) published the first work of this group. Their proposal, named *Nutrition Label*, focuses on the generation of standardized labels to convey metadata and information about datasets, and to try to reflect a portion of standard exploratory analyses conducted by developers before utilizing data. They include seven different modules to make distinct aspects of information regarding the data available to other parties, each of which requiring different manual effort, and revealing varying elements of the data. Each of the present labels would compose the 'nutritional' ingredients of the dataset.

Holland et al. (2018) highlight the need for more empirical research regarding the use of the Labels and which information will be practical and desired to be included. They also predict that the adoption of Labels would foster the development of an environment that would also stimulate reflection about the use and construction of datasets.

Bender and Friedman (2018) target natural language datasets. They focused on dealing with issues identified in natural language processing tasks. They argue that new datasets in this area tend to be published with discussions about how they were annotated, but there is an informational gap regarding a characterization of the people who produced the data (speakers or writers), and those responsible for annotating it. Their proposal identifies the importance of context, and variations in language, and how its meaning is constructed through interpretation.

The Data Statements was designed to include not only information about the people who took part in producing it, but also the context regarding the language, and its use, such as the situation in which it was collected, or the type of language used – dialect and region, for instance. Furthermore, the authors also highlight that, while academic publications and model documentation should include this complete statement, other uses could include a briefer format, summarizing the information and also directing the reader to it. The proposed tool is described as being capable of mitigating different types of bias, better highlighting what the data represents and what it does not.

Gebru et al. (2020) present a similar proposal, but for all types of datasets. Their work on *Datasheet for Datasets* aim to address some specific goals for different stakeholders. For dataset creators and curators, they wish to promote reflective practices, including about underlying assumptions of the data and potential risks its use may carry. For dataset consumers, the increased transparency would inform the decision to use certain dataset for the task they wish to address. Finally, they also highlight the goal of increasing reproducibility of machine learning results, by enabling the creation of mirroring datasets.

They enumerate questions for seven different topics, and highlight their exemplifying nature and that the information should include and be personalized according to the specific use, domain, and other factors. The scope of the datasheet crosses the whole process of creating the dataset, from collection and processing to distribution and maintenance. It also includes use cases tested and envisioned by creators. In addition to the upfront work required to document the data, including finding factors that may cause potential fairness problems such as biases, they identify a limitation of their proposal when addressing datasets that are constantly updated, rendering part of the documentation out of date at a fast pace.

Miceli et al. (2021) conducted a study of both of these proposed documentation methods in the field of computer vision. They compare each of these general proposals, as well as ones made specifically for the publication of certain datasets (Choi et al., 2018; Seck et al., 2018). Furthermore, they conducted fieldwork by studying the process of two data collection companies and interviewing 30 experts that make use and request this data. Their work identifies difficulties for the adoption of the studied tools, and highlight the importance of collectively considering the social aspects that shape dataset for their effective documentation.

More recently, Hutchinson et al. (2021) propose a framework that identifies different stages in the development process of a dataset. Each of these should include an specific documentation which should include information considered necessary for appropriate accountability of actors for the use of the respective data.

Their work is based on identifying similarities between datasets in Machine Learning projects and computing infrastructure. In addition, they also shape their proposal on practices already adopted in software engineering. The result is a series of five documents, each targeted at keeping record of relevant practices and decisions made in each stage.

In addition to enabling better accountability for datasets, the authors

highlight the capability of their work to aid developers in the maintenance phase of their datasets, and other uses that may arise at this stage. This includes disseminating knowledge of failures identified, challenging decision making based on the data, and facilitating reuse of data, as well as improving the reproducibility of results using that dataset.

3.1.2 Cards

Another group of research has the objective of highlighting different values that influence the development of AI systems. This is especially relevant considering we have different social values at play, and decisions usually imply a trade-off between them.

One proposal in this direction is the Value Cards (Shen et al., 2021). The authors propose a methodology that include three cards to highlight different social values that may be at stake in decision making. Their goal is to "foreground the importance of social values and collective decision making via deliberation", in order to promote deliberation between stakeholders so they can understand each other's perspectives and values, as well as the inherent trade-offs that these will lead to, including those inherent to what may be seen as technical decisions and metrics (*e.g.*, maximizing accuracy).

They proposed and tested three different cards. Model Cards focus at different ML models and should capture trade-offs between choices in development of AI applications. Persona Cards depict perspectives and values of different stakeholders. Finally, Checklist Cards have the goal of enumerating social and technical considerations, which should be used to guider the deliberation and decision process. Shen et al. (2021) highlight the epistemic value of these model and checklist cards.

They investigated the effectiveness of their proposal in an educational setting, as part of a machine learning course. They found that students were able to actively engage with different perspectives and values, understand technical and social trade-offs, and even come up with different stakeholders not initially included in the material offered by researchers.

Raji et al. (2020) offer a distinct framework with a similar emphasis on values of the responsible organization to guide decision making. Aligned with the work of Hutchinson et al. (2021), they frame their proposal from an accountability standpoint, with the goal of creating a framework that allows internal auditing of AI systems, to be conducted before deployment in order to identify and avoid negative impacts. To structure their proposal, they used findings and proceedings used in other areas where internal audits already play

an important role in the accountability process, such as the aerospace, medical, and financial industry.

They leverage the proposals of Seck et al. (2018) and Mitchell et al. (2019), discussed in the following sections, to propose a framework that includes not only the production of datasheets and Model Cards by the development team, but a series of other artifacts produced through the auditing process. We include them in this category because they frame all the audit process starting by values and principles to be defined by the organization, which will guide and be verified by the use of the proposed framework.

The final stage of the process, labeled as Reflection Stage, should culminate by comparing their findings with the values put forward in the start of the auditing. By the end of the process, the organizations producing or deploying the system should be aware of design and product decisions that may clash with their ethical values, and take action accordingly, either by adapting and altering the system to mitigate identified risks, or by identifying use cases that should be excluded from the system.

The emphasis on values and ethical principles of stakeholders of the development process of this group of proposals is aligned with our research. However, they also have wider scope and rely on the interaction among members of development teams, and even other parties, as is the case of auditing teams. We direct the focus of our research to tools that can be directly used by developers, and may even be integrated into these framework as was done by Hutchinson et al. (2021) with Model Cards (Mitchell et al., 2019).

Finally, Arnold et al. (2019) propose the use of FactSheets to create AI documentation, focusing on transparency issues. The high-level sections in FactSheets are: statement of purpose, basic performance, safety, security, and lineage. Their proposal differs from others as it focus on disclosing information and shortcomings not of a specific ML model, but of the system as a whole.

The proposed questions and topics concern themselves with technical aspects and decisions taken during system design and development, including whether there were any tests for bias in the system. However, they do not highlight values and ethical principles that might have guided these choices, like other works cited in this section.

As follow-up work, Richards et al. (2020) describe a methodology for creating FactSheets that build upon this artifact. Moreover, Hind et al. (2020) present two interview studies to evaluate challenges in creating documentation for AI systems, and to present recommendations on how to collect and report relevant information to improve efforts in this area.

3.1.3 Toolkits and Guidelines

Many methods and tools have been proposed to identify issues in algorithms or datasets. A common topic has been identifying biases in datasets, creating automated tools to evaluate whether a trained model violates certain definitions of fairness, or proposing models that adhere to them by design (Bellamy et al., 2018; Wexler et al., 2019).

In addition to the research mentioned in the previous subsections, many proposals have been made by industry even outside of specific research and academic publications. These have been made through toolkits, guidelines, and a few services offered. It is worth mentioning that many of the cited related work include contribution from industry research centers.

These proposals are not necessarily in the same line as the other ones cited in previous subsections, but they share the broad goal of improving fairness and accountability in the use of AI through providing practices and methods to be adopted during the design stage by developers. One such example is the Human-AI Interaction HAX Toolkit, published by Microsoft.[3] According to the description provided, it comprises a set of tools to provide developers to take a human-centered approach, and include the Guidelines for Human-AI Interaction, which are declared as "best practices for designing human-interaction with AI-based products and features."

In the same vein, Google's project "People + AI Guidebook" intended to serve as a "Tactical guidance and best practices for designing human-centered AI products". Although it does not directly touch on issues of fairness, it includes issues such as building trust in AI systems, and highlights the issues users may have with them.

Loukides et al. (2018) defend the use of checklists as a practical and simple way for developers to evaluate their work and whether they have steered from ethical values. They offer a checklist to be used in data projects and place special emphasis on checking whether developers took action to test for certain issues in their system, and whether they ensured there were control mechanisms to address harmful results unaccounted for.

3.2 Model Cards

Mitchell et al. (2019) defined their Model Cards as "[s]hort documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions [...]". Their objective was to increase the transparency

[3]https://www.microsoft.com/en-us/haxtoolkit/

of the model by reporting its use cases, performance metrics, and known shortcomings, thus reducing the likelihood the model will be used in unintended scenarios. They also aimed to standardize the reporting of test results, benchmark metrics, and ethical considerations made by developers; and to create a means for stakeholders to benchmark and compare models not only along a performance axis, but also along an ethical axis.

The authors argue that different stakeholders (interested parties) will have different expectations and meanings assigned to the topics explored in Model Cards. Then, they present use cases relevant to these actors, and how their proposal is relevant to them, ranging from developers to policymakers and impacted individuals. For instance, they argue that the main interpretation for model developers will be of benchmarking and comparing the performance of their models to other existing models. By contrast, policymakers can understand cases where a machine learning system may succeed or fail, and how it will impact people.

According to the authors, their cards could aid developers since it would make them actively consider possible break points of their algorithms. For instance, they argue that slicing the evaluation of the algorithm across groups would highlight errors that are not equally distributed across groups, avoiding situations where algorithms are considered biased and unfair, as has happened a few times in recent years.[4]

The authors then detail each of the sections. This starts with "Model Details", which should include general information regarding the software and its developers, including organization, type, and academic reference. Next, the authors should write the "Intended Use" section, which should include the intended users of the software, the use cases envisioned by the developers and explicit ouf-of-scope uses, cases where the system could be applied but are out of its application range, or even for which there might exist better algorithms.

The next section is called "Factors". While this title is not very descriptive, it should contain details about the model's performance according to different factors. These should include the method for capturing the data; this can be relevant, for example, for details about the hardware that captured an image.

The most interesting and ethically relevant point of this section is the "groups" factor, which should include different characteristics that can

[4]A few such instances with Google include: (i) its tagging of individuals holding a thermometer "gun" differently based on skin color that came to light during the COVID 19 outbreak – `https://algorithmwatch.org/en/story/google-vision-racism/`; and (ii) Google Photos tagging black people as gorillas, an issue that came to light in 2015 – `https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai`

categorize the data. This becomes especially relevant when we think of people's attributes, including ethnicity, gender, sexual orientation, or health conditions, to name a few.

Next, they propose a "Metrics" section. This section should include information about the performance and some tuning of the algorithm. In addition to the metrics, they argue for a reasoning of why these metrics were used to measure performance instead of other available ones. Another suggested kind of information is the decision threshold. Finally, it should also include how uncertainty and variability are dealt with and estimated by the software.

The following two sections concern the data used in evaluation and training. This would include the dataset used for evaluation, processing steps, and motivation for use of that dataset. Ideally, the same information should be included for the training data and made available. However, the authors recognize this might not be feasible in many cases due to other interests, such as the data being proprietary.

This segues into the "Quantitative Analysis" section. Here the information should be disaggregated by the chosen relevant factors, and provide confidence intervals when possible. This section should include how the model performs with respect to each factor and their intersections.

In "Ethical Consideration", the developers should include what considerations went into the development, potential issues that were found or that could show up from the use of the model. This does not mean that all issues should have solutions, but that stakeholders and users should be informed about them. They also suggest the following questions should be explored on this section:

> **Data**: Does the model use any sensitive data (*e.g.*, protected classes)?
> **Human life**: Is the model intended to inform decisions about matters central to human life or flourishing (*e.g.*, health or safety)? Or could it be used in such a way?
> **Mitigations**: What risk mitigation strategies were used during model development?
> **Risks and harms**: What risks may be present in model usage? Try to identify the potential recipients, likelihood, and magnitude of harms. If these cannot be determined, note that they were considered but remain unknown.
> **Use cases**: Are there any known model use cases that are especially

> fraught? This may connect directly to the intended use section of the model card.

Finally, the card is closed off by "Caveats and Recommendations". This section is included in order to address concerns that were not considered to have been covered in the previous sections.

The authors then provide two examples of cards with information with hypothetical cases. The model structure, including the synthesis of the suggested content, can be found in Figure 3.1.

After summing up this proposal to increase algorithmic transparency, in the next section we describe the other proposal to be used in our larger study.

3.3 Extended Metacommunication Template

Barbosa et al. (2021) proposed the Extended Metacommunication Template, where they focus on the "1st-person [...] assessment of ethical dimensions of one's work. This proposal resembles the Model Card as developers are also expected to ethically reflect about their work; however, it is differentiated since this is not expected to be shared with other stakeholders."

The authors created a framework for developers to reflect on their algorithms. The Extended Metacommunication Template leverages the Semiotic Engineering theory of Human-Computer Interaction (HCI). Their epistemic tool aims to make developers reason in the 1st person about the ethical dimensions of their work, while explicitly highlighting the targeted 2nd person.

They acknowledge that development teams are composed of multiple professionals, with distinct backgrounds. Also, that the software development process can also be influenced by other stakeholders. To minimize potential issues arising from conflicting views, each interested party could write their own template and bring it to the discussion, so that they may arrive at a unified collective view of the system.

The authors highlight how their work can be aligned with different ethical frameworks, such as that of bioethics, which comprises the following principles: beneficence; non-maleficence; autonomy; and justice (Beauchamp and Childress, 2019) as mentioned in section 2.4.

3.3.1 Semiotic Engineering Concepts

In the authors' view, the original contribution of Semiotic Engineering is that it defines a specific object of investigation to the field: the communication

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.
  - 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 as shown in Mitchell et al. (2019)

process between developers and users, mediated through the system. With this object defined, it also developed tools and concepts to aid in its investigation.

The paper extends an existing tool in Semiotic Engineering, the Metacommunication Template. This tool was originally envisioned as a way to express the message designers were communicating to the users through their system interfaces (De Souza, 2005; de Souza and Leitão, 2009). The generic message was paraphrased by the author as follows: (De Souza, 2005, p.84)

> "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."

This template aims to support the design of the user–system interaction from the perspective that it constitutes a dialogue between the developer and the user. From this perspective, the system is a proxy for the developers, through which they communicate with the user, in a *computer-mediated communication* process.

Metacommunication can be understood as a two-level process. At a higher level, the designer is sending a complete message to the user about "how, where, when, what for, and why the user can communicate back with the system". The communication is also achieved at a lower level, by the user's direct interaction with the system, even though they do not receive the entirety of the message at once.

3.3.2 Extended Metacommunication Template Structure

The entirety of the Extended Metacommunication Template, with its sections and questions, is shown in Figure 3.2. The parts that comprise this document are: (i) Analysis (understanding needs and defining requirements); (ii) Design; (iii) Prototyping, implementation, and formative evaluation; (iv) Continuous, post-deployment evaluation and monitoring.

Each section is based on questions which highlight the designer's decision making and knowledge when creating the software. The sections also reflect the software development process. They also emphasize how the set of beliefs and expectations about the users and other stakeholders directly influenced by the resulting system.

These questions reflect the development team's knowledge, assumptions and expectations about the system they created and the people who may use

or be affected by its use. Furthermore, it is also backed by ethical questions related to the content of each section.

The first section, "Analysis", consists of the designer's knowledge about the system's users, other affected parties, and intended and anticipated uses. The ethical question should highlight potential issued raised by this knowledge.

The "Design" section reflects the contemplated goals supported by the software, and how it should be used to achieve these goals. It should also highlight uses that are not supported by the designers.

The "Prototyping, implementation, and formative evaluation" section is related to how what was previously stated influenced and was reflected into the system itself. The ethical question is concerned with what scenarios were used to evaluate the system.

Finally, the last section, "Continuous, post-deployment evaluation and monitoring", should reflect upon the actual use of the system, including unexpected uses, effects and ethical issues that have been raised. The ethical question should address what and how can ethical issues be addressed by the development team.

3.4 Discussion

Each of these artifacts adopt a different theory to support it and, as a result, their focus is drastically different, in spite of both aiming to promote ethical reflection. On a superficial level, the information that was envisioned as being included on Model cards by Mitchell et al. (2019) is much more technical, including information about the model, data used to train and test it, and performance metrics, for instance.

By contrast, the Extended Metacommunication Template focuses on how the development team, including designers, view their system and its user. Its focus is more on the subjective choices made by developers, the reasoning and reflection that motivated them. Its purpose as an epistemic tool is to support the development process, while bringing these issues to the forefront of it.

For instance, the ethical questions in the "Design" section explicitly aim to highlight ethical considerations and principles that were influential in the development process, and how the system is aligned with them. Furthermore, the "Prototyping, implementation, and formative evaluation" section includes a topic about what was built into the system to address undesired effects of its use. Finally, the section "Continuous, post-deployment evaluation and monitoring" inquires how the actual use of the system deviated from their vision, and whether any ethical issues have to be addressed at this stage.

Extended Metacommunication Template

1. **Analysis (understanding needs and defining requirements)**.
   (a) What do I know or don't know about (all of) you and how?
   (b) What do I know or don't know about affected others and how?
   (c) What do I know or don't know about the intended (and other anticipated) contexts of use?
   (d) *What ethical questions can be raised by what I have learned? Why?
2. **Design**.
   (a) What have I designed for you?
   (b) Which of your goals have I designed the system to support?
   (c) In what situations/contexts do I intend/accept you will use the system to achieve each goal? Why?
   (d) How should you use the system to achieve each goal, according to my design?
   (e) For what purposes do I **not** want you to use the system?
   (f) *What ethical principles influenced my design decisions?
   (g) *How is the system I designed for you aligned with those ethical considerations?
3. **Prototyping, implementation, and formative evaluation**.
   (a) How have I built the system to support my design vision?
   (b) What have I built into the system to prevent undesirable uses and consequences?
   (c) What have I built into the system to help identify and remedy unanticipated negative effects?
   (d) *What ethical scenarios have I used to evaluate the system?
4. **Continuous, post-deployment evaluation and monitoring**.
   (a) How much of my vision is reflected in the system's actual use?
   (b) What unanticipated **uses** have been made? By whom? Why?
   (c) What anticipated and unanticipated **effects** have resulted from its use? Whom do they affect? Why?
   (d) *What ethical issues need to be handled through system redesign, redevelopment, policy, or even decommissioning

Figure 3.2: Extended Metacommunication Template as proposed by Barbosa et al. (2021)

This difference is also reflected in which development stages are explicitly referred in the artifacts' sections. In the Model Card structure, most of the information concerns the final model and its data, and the data used to train and evaluate it. The exceptions are the intended use, which comprises the purpose envisioned for the software by its developers, and the open-ended ethical consideration section. On the other hand, each of the sections of the Extended Metacommunication Template goes through a different phase of the development process, starting from a conception and planning phase, reflected on the Analysis section, extending through monitoring the use of the system post-deployment, on the "Continuous, post-deployment evaluation and monitoring" section.

The Template thoroughly highlights potential scenarios where ethical issues could appear, even if these had not been previously considered by developers. These questions help to direct attention to the influence the system could have on its users and affected others.

Brandão et al. (2019) suggest that developers may not recognize the impact their AI system can have on its social context if not directly prompted by others, while focusing on technical aspects and measures. They identify this as a mediation problem of the social meaning, since developers had to be explicitly alerted by researchers about potential social issues that could be raised by its use in the context it was build for.

Furthermore, they also found that, even when alerted by potential problems, developers would often rely on someone else to help communicate the meaning of the artifact they produced and its traits, or to help with the task their system would perform in real-world scenarios (*e.g.*, having a person responsible for validating the result of their artifact). This suggests that, even when directly prompted to reflect on these issues, developers may not feel comfortable with communicating their thoughts, and may appeal to using the third person to mediate this meaning to relevant stakeholders.

These findings accentuate the importance of stimulating developers to reflect on the meaning and consequences of the artifacts they develop, as well as communicating it to other stakeholders. The tools described in this section have the goal of directing the focus in the development process to issues aligned with the ones raised by Brandão et al. (2019), through directing and creating means to communicate their ethical reflection.

Finally, it is possible that the mentioned differences between the Model Card and the Extended Metacommunication Template may affect the resulting developers' reflections. For example, the focus of the Model Card on the technical side, more commonly considered by developers, might not prompt the

same type of thought process. Furthermore, since the document produced by the use of each of these tools would constitute a means for communicating their choices and intention, shaped by each of the underlying views on development, verifying the similarity of developers' view of their use to the aforementioned findings is relevant to understanding the development of machine learning systems.

4
Study

We conducted an interview study to understand how potential stakeholders, specifically members of machine learning development teams, would use each of the presented tools. Each interviewee was presented a hypothetical development scenario and a brief outline of the bioethical principles,[1] and asked to fill either the Model Card or the Extended Communication Template with the relevant information. We also asked questions about each participant's background, and about their thoughts on the tool used and their own reasoning at the end of the interview.

Since the Extended Metacommunication Template and Model Cards require information about development and post-development stage issues, as well as performance information, such as tests used, performance metrics, and potential problems that have been identified when deploying the algorithm, the interviewees were instructed to imagine what would have happened in the scenario and fill in each section with topics about issues they believe would have been raised or expected to show up on that scenario, without detailing them.

This study focused on actors that take part in the development of machine learning systems. As stated by Mitchell et al. (2019), Model Cards can have different interpretations and be used differently by different stakeholders. We did not involve different stakeholders in our investigation and leave this topic to future inquiries. Likewise, we believe that the Extended Metacommunication Template can also have different interpretations by different stakeholders, which should also be investigated.

The following sections describe how the study was designed and conducted, and how we structured the collected data for analysis. A diagram of all stages of this process, from interview to analysis can be seen in Figure 4.1.

[1] We chose to use the bioethical principles to provide a basic common ground for ethical reasoning for all participants.

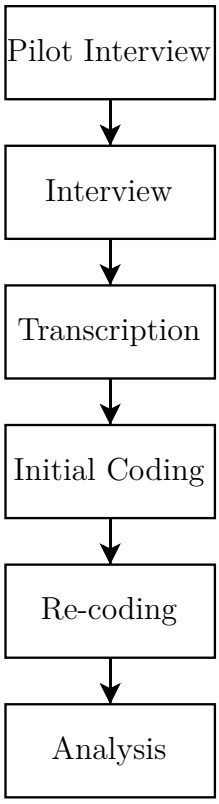


Figure 4.1: Stages of research from pilot interview to analysis.

4.1 Research Design

We used two different scenarios for the study, presenting different backgrounds and applications for the algorithm being developed. We chose to focus on algorithms that are used to take socially relevant decisions, *i.e.*, decisions that can have a big impact on someone's life. We have opted for this kind of scenario since there is a higher social concern for ethical issues in such cases, which might not happen otherwise. Hence, the scenarios chosen were: (i) an algorithm that decides whether to grant or deny a loan; and (ii) an algorithm that assigns a score to a college applicant, and this score decides who will or will not be accepted to the relevant program.

Although both scenarios present cases that can have a high impact on the subject of the decisions' life, it is still possible they might be seen differently by participants. The first scenario, where the system would predict default risk, is a relatively common machine learning application. On the other hand, college admissions is not a widespread AI application scenario.[2] In fact, our use of this scenario was inspired by a case at the beginning of the COVID-19

[2]A search at the ACM Digital Library for abstracts containing for terms "default" and "loan" machine learning yielded 105 results, while a search for college or school admission returned only 5 publications.

pandemic, where the exams for entry in universities in the UK was replaced by one generated automatically.[3]).

In order to gather more comprehensive information data, we opted to conduct two interview sessions with each participant so each one would explore both scenarios and both the Model Card and the Extended Metacommunication Template. As a result, we had four groups of participants, in which we varied the scenario, and the order in which each participant performed each task, since we believe the first session could have an influence over the second one. The division can be found in Table 4.1. Furthermore. an image of the interview process step by step can be found in Figure 4.2.

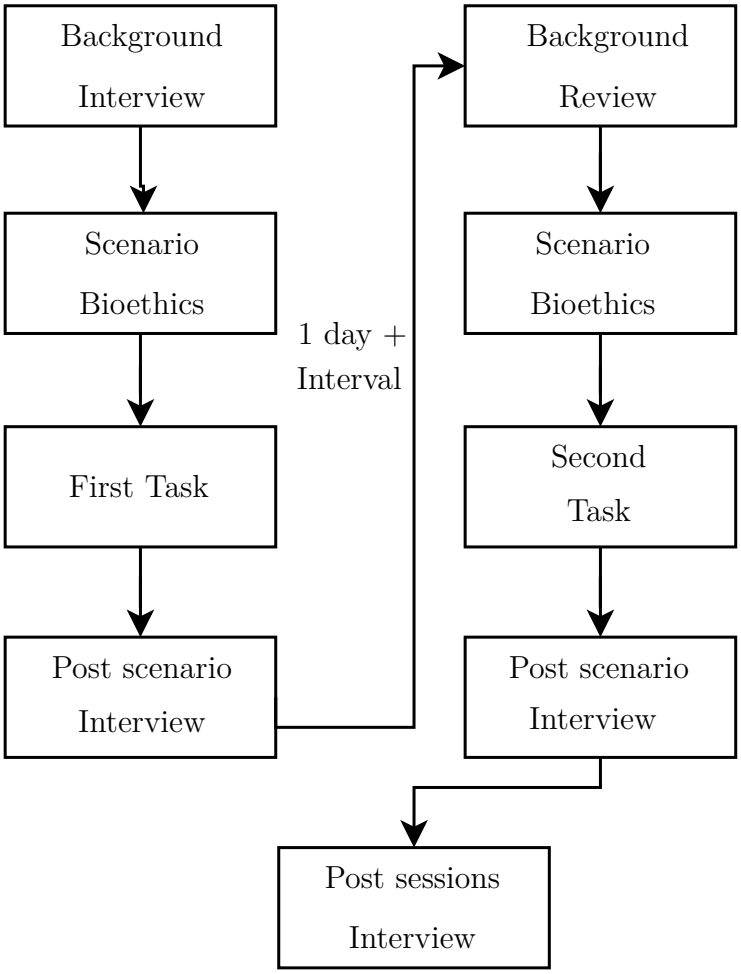


Figure 4.2: Interview process.

We now present the two scenarios. The underlined text varied: study participants working with the Model Card were tasked with filling out the corresponding form, and those working with the Extended Metacommunication Template were tasked with answering the corresponding questions.

[3]A-Level exams were initially substituted by a algorithms for the year of 2020. After much backlash, the use of this system was dropped (https://www.wired.co.uk/article/gcse-results-alevels-algorithm-explained. Visited on: 18/06/2021

Table 4.1: Allocation of representation and scenario in each session, for each group.

group	1st session	2nd session
A	MC - Financial Institution	EMT - University
B	EMT - Financial Institution	MC - University
C	MC - University	EMT - Financial Institution
D	EMT - University	MC - Financial Institution

Scenario 1

You are the leader of the development team of a *financial institution*. The role of your team is to develop an Artificial Intelligence algorithm to *make automated lending decisions for your company. Your algorithm must access the risk of each potential client based on their profile and financial history, and decide whether to grant the requested loan or not, with the goal of maximizing the expected profit of the company.*

You have also been tasked with filling out the following form | answering the following questions about the development process and potential issues with your final product. These forms are related to ethical and practical concerns of the system being developed, as well as the scenario it was envisioned and tested for. You may not possess all the information to properly fill out all the fields; however, you have been asked to include topics that you expect will show up in the final version and that you plan on further investigating.

Scenario 2

You are the leader of the development team *hired by a university*. The role of your team is to develop an Artificial Intelligence algorithm to *assign scores to students who have applied to the university program. Your algorithm must consider the student's previous accomplishments and assign a score to each student, which will be used by the selection committee to decide which students should be admitted or not. Your algorithm will not be used to grade any new material from the students.*

You have also been tasked with filling out the following form | answering the following questions about the development process and potential issues with your final product. These forms are related to

> ethical and practical concerns of the system being developed, as well as the scenario it was envisioned and tested for. You may not possess all the information to properly fill out all the fields; however, you have been asked to include topics that you expect will show up in the final version and that you plan on further investigating.

The documents participants used with each tool can be found in Figures 4.3 and 4.4.[4]

Model Card

- **Model Details**. Basic information about the model.
- **Intended Use**. Use cases that were envisioned during development.
- **Factors**. Other factors that could impact the model's performance.
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses of the model.
- **Training Data**. Details on the dataset(s) used to train the model.
- **Quantitative Analyses**. Factors used for quantitative analysis of the model and results found (factors may be isolated or combined on the intersection).
- **Ethical Considerations**. Ethical considerations about the model created.
- **Caveats and Recommendations**. Caveats and recommendations about possible uses of the model.

Figure 4.3: Sections to be filled out by participants.

To establish the participant's background, before going through the scenario, we asked some preliminary questions about participants' previous knowledge, including about the subject of our study. Upon completion of the scenario, we also interviewed participants to better understand whether and how the document aided and motivated their ethical reflection.

We asked the following set of questions. This list includes questions asked before working with the first scenario, to determine the participants background, as well as more speculative questions at the end, which the interviewers decided to ask or not in light of the participants' previous remarks.

[4]The code for generating this Model Card was obtained from Mitchell et al. (2019) The Extended Metacommunication Template was adapted from Barbosa et al. (2021). We used the template containing the briefer descriptions of each section provided by the authors for each of the tools, except for the Quantitative Analysis section of the Model Card, where a sentence commenting the section on the paper was adapted to provide the information given by authors in a shorter passage, since no explanation was provided in the template.

Before the first session (participant's profile questions)

1. What is your current line of work?
2. What was the area of your formal education?
3. How long have you worked in the development of ML?
4. What is your usual process for developing a ML Model (probe: gathering and cleaning data, selecting parameters, training, evaluation)?
5. What is your previous knowledge of Semiotic Engineering?
6. Do you have any previous knowledge of the Model Card or Extended Metacommunication Template?
7. How interested are you in ethical issues in design and development?
8. How experienced are you in considering ethical issues in design and development?
9. When you think about ethical issues, what comes to your mind?
10. Have you had any contact with software you consider unethical? How was the experience? (b)

At the end of each session (post-task questions)

11. What did you think about the Model Card/Extended Metacommunication? (probe: Please elaborate.)
12. How did the Model Card/Extended Metacommunication Template influence your ethical reasoning on the presented scenario? How did it aid in this reasoning? And how did it constrain or hinder it?
13. On a scale from 1 to 7, how do you rate how much the form / questions helped (1 = hindered a lot; 7 = helped a lot).
14. Was there anything you included that did not fit well in any of the sections? If so, what?
15. Was there anything you wanted to include but did not fit well in any of the sections? If so, what?
16. What did the form / questions help you to reflect upon? (probes: about the system, the development process, about the consequences of using the system to its users and to the society at large)

17. How did the (form/questions) help you think about the consequences of the model you were building? (probe: stakeholders, minorities, economical matters:?)

18. Which sections helped more in your reflection upon potential ethical issues? (probe: How?)

19. Who did you consider would read what you were writing?

20. What other stakeholders do you believe would benefit from the resulting document (*e.g.*, users, policymakers or impacted parties)? How?

21. What other stakeholders do you believe would benefit from going through this process of (filling out the form / answering the questions)?

After the second session (comparative questions)

22. In simple terms, how would you explain the difference between the two tools for a new member of your team?

23. How would you compare both tools concerning the variety of your reflection on the scenario?

24. Given that you have previously used /artifact/, how do you think it influenced you in this session? * What else?

Possible questions after the second session, depending on the previous answers

25. What did you think of the first person use in the Metacommunication Template?

26. Do you believe your word choice to represent your thoughts had any impact on your reflection?

27. Do you believe the division of question on different development steps impacted your reflection?

28. How important do you believe the Principles of Bioethics were to your reflection?

29. How did you feel about the Scenario after your reflection process?

30. Do you believe the open ethical reflection section of the Model Card influenced your reflection?

4.2 Execution and participants' profiles

Before recruiting participants, we ran a pilot study with one participant to evaluate the duration of each session, whether our questions could be comprehended by potential participants, and whether they were capable of stimulating reflections over the issues we wanted to investigate.

As a result, we made small changes in the wording of some of our questions. The last set of optional questions were added due to observations made in our pilot study. We also estimated each session would take around 1:30, which proved to be a good estimate.

Our study was then carried out with a total of eight subjects with graduate degrees in Computer Science. Seven of our participants had experience as members of teams developing systems that apply machine learning algorithms, while the remaining one only had experience with academic courses on the topic.

Our participants were allocated into each group in the order they agreed to take part in our study. Thus, the first participant was put into group A, the second in group B, and so forth. In order to avoid tiring participants, and to create some distance between each interview, we scheduled the two sessions with each interviewee with at least a day in between.

We opted to divide the two interviewers to allow for more time slots to conduct our study, and each one of them was responsible for conducting interviews with one of the tools. As a result, each one was responsible for half of the first and half of the second interview sessions.

All sessions were conducted through audio calls in Portuguese, to let participants more comfortable, speaking in their native language. At the beginning of our first session we would ask for the participant to give us consent to take part in the study, including recording of the session and sharing what they wrote or said anonymously. We asked for confirmation of the consent before starting the second session.

We chose to conduct semi-structured interviews. We prepared a well-defined script, including our sets of questions, presenting the scenario and the bioethical principles, and explaining each of the tools used. Throughout the session (*i.e.*, when participants read the scenarios and used the tool), interviewers provided clarifications when asked by participants[5] and asked *ad hoc* follow-up questions for participants to elaborate on certain points that were deemed particularly interesting or unclear. At the end of each session,

[5]Since most participants were using these tools for the first time, a common line of questioning by them was asking for explanation about certain sections.

the interviewer asked the post-task questions and, after the second session, the questions comparing the two models.

At the start of the first session, we asked the first set of questions, related to their background. Afterwards, we sent a link to a document containing the text of the relevant scenario, an outline of the bioethical principles, and the tool to be used during the session. Finally, we asked the participant the post-task questions, as described in the previous section.

The process was repeated for the second session, with two small changes: in the beginning of the session, we asked participants to confirm the background information from the previous session; and to conclude the second session, we asked the remaining comparative questions, as well as the more speculative ones, depending on their previous answers.

4.3 Transcription

After all the sessions, we created our text corpus to be able to conduct our analysis. For that we transcribed all the interviews.

Each audio recording was transcribed *verbatim* in full, including what was said both by the interviewee and interviewer. In addition to that, as we had video recording of the screen shared by participants, we also included visual cues when they were referenced during the speech, or when the context was needed to understand certain remarks.

In addition to that, we kept the artifact resulting from applying each tool, as written by each participant. As a result, we collected a total of 32 text documents of the interviews, two for each session. This set is the data we used in the next steps of our study.

While the documents produced by the use of the tool were relatively homogeneous in size, the sessions varied between 1:10 and 1:40 in duration, which is reflected in the length of the transcripts. Furthermore, participants would not necessarily talk as much as one another. To give a rough idea of the volume of text, the transcription of the absolute majority of the Model Card sessions (formatted using Times New Roman 12, A4 paper size, single spaced) took over 10 pages each.

Extended Metacommunication Template

1. **Analysis (understanding needs and defining requirements)**.
   (a) What do I know or don't know about (all of) you and how?
   (b) What do I know or don't know about affected others and how?
   (c) What do I know or don't know about the intended (and other anticipated) contexts of use?
   (d) *What ethical questions can be raised by what I have learned? Why?
2. **Design**.
   (a) What have I designed for you?
   (b) Which of your goals have I designed the system to support?
   (c) In what situations/contexts do I intend/accept you will use the system to achieve each goal? Why?
   (d) How should you use the system to achieve each goal, according to my design?
   (e) For what purposes do I **not** want you to use the system?
   (f) *What ethical principles influenced my design decisions?
   (g) *How is the system I designed for you aligned with those ethical considerations?
3. **Prototyping, implementation, and formative evaluation**.
   (a) How have I built the system to support my design vision?
   (b) What have I built into the system to prevent undesirable uses and consequences?
   (c) What have I built into the system to help identify and remedy unanticipated negative effects?
   (d) *What ethical scenarios have I used to evaluate the system?
4. **Continuous, post-deployment evaluation and monitoring**.
   (a) How much of my vision is reflected in the system's actual use?
   (b) What unanticipated **uses** have been made? By whom? Why?
   (c) What anticipated and unanticipated **effects** have resulted from its use? Whom do they affect? Why?
   (d) *What ethical issues need to be handled through system redesign, redevelopment, policy, or even decommissioning

Figure 4.4: Extended Metacommunication Template as presented to study participants.

5
Coding

This chapter describes the coding process. The process described here, and our experience during it, will also be leveraged to further study our topic, especially for the analysis of the data related to the Extended Metacommunication Template, which lied outside the scope of this dissertation.

5.1
Coding Characteristics

To analyze the data gathered through our interviews, we opted to code our data with the objective of conducting a Thematic Analysis (Braun and Clarke, 2012). Our corpus included what each participant spontaneously said during the session while using each tool, the answer to our interview questions, and the document produced by the use of the tool in each session.

One initial decision we made was to follow an inductive approach (Braun and Clarke, 2012), allowing our codes to emerge from exploring the data we collected, instead of pre-selecting a set of *a priori* categories. Our goal with this decision was to directly represent how participants interacted with the Model Card, and not describe it based on a pre-determined framework.

One factor that certainly influenced participants was our decision to provide bioethical principles to serve as a basis for ethical reasoning, which constitutes the background of our research. However, these principles are quite general, so participants' values may end up influencing how they apply them. Furthermore, we made sure to make clear to all participants that they did not need to use them, and that they could apply their own views about what is ethical or unethical.

Following the first consolidation step, described in the following section, we also considered some of the options for coding strategies listed by Saldaña (2009), especially simultaneous and descriptive coding, classified as first cycle methods. These were used as guidance on how to construct our codes, including naming and descriptions.

The presence of simultaneous coding is an important characteristic of our dataset. In other words, the same passage could receive multiple codes,

e.g., different codes for the same excerpt; or for a different code for the whole excerpt and for one or more segments in that passage.

We also opted to use a strategy of splitting codes when relevant, trying to tag only the relevant sentences. However, the size of the tagged excerpts varied widely, and some of the passages could correspond to whole paragraphs, when the participant kept exploring for a while the idea expressed in the code.

5.2 Coding process

This section describes the coding process we followed, especially the steps taken to code the data and arrive at our final codebook. It also contains the codebook reached at the end of the process.

To establish a preliminary codebook, two coders first coded the documents of two participants (P3 and P4) individually. Each coder arrived at an initial set of independent codes, which emerged from what these participants said or wrote that was related to the model they were developing, its potential impact, the tool they were using, or ethical issues they identified. We opted for independent initial codebooks to allow for more diversity in the codes created on this initial step, which could then be further refined and narrowed down to a more comprehensive and robust set of codes.

For this preliminary coding, we did not use any specific data analysis tool. Instead, we opted to first read each document on a text editor, highlighting passages that we believed were relevant and interesting for our study in order to obtain an initial overview of our data before committing to an initial set of codes. After finishing this process for P3 and P4, we reread their interviews and documents, making our initial coding.

Upon completion of this initial coding, we reviewed both sets of codes and their descriptions in each codebook. We merged some of the codes and opted to discard some of them, which we considered outside the scope of our intended analysis. One of the objectives during this process was to have a comprehensible list of a few dozen codes, which could be easily used during the coding process.

One of the coders had also coded a third participant, which allowed us to note that our initial codebook was considerably stable and was not significantly modified during the coding of that third participant. With this consolidation, we reached a new set of codes that would be used to annotate the remaining participants' data.

After this step, we maintained a shared codebook in case any changes arose from coding the remaining data. During this stage, if any of the coders

felt any change was necessary, they would contact each other to discuss the change. This procedure was used mainly to define the boundaries of some of the codes, including whether some cases should be included or not, and applied to around 10 excerpts.

Close to the end of the coding of the remaining data, we opted to make another review of our codes. Our goals with this revision were twofold: (i) to further consolidate the codes that were created after the previous step; (ii) to make some adjustments to the names and descriptions of the codes to better represent the uses that had been made during coding of the remaining participants.

Upon finishing the coding, we made a final revision of our codebook. This time, our focus was on splitting some of our codes to better represent the concept that was explored in the context they were used. The split codes were not entirely new, and were formed from the previous code in addition to an indication of the direction that was expressed by participant. For instance, the "ethics of development process" code, used to tag instances where participants reflected on the ethics of a certain aspect of the development process, was split to represent whether the passage reflected that the aspect was either ethical or unethical. After this process, each coder reviewed their own coding to incorporate the split codes into the data.

Table 5.1 shows the final version of the codebook, including general and split codes, and their description.

Table 5.1: Final codebook, including split codes.

Code name	**Description**
DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK	Diagnoses the existence of an ethical issue, and then (*a posteriori*) relates it to an element of an ethical framework.
DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK - AUTONOMY	Diagnoses the existence of an ethical issue, and then (*a posteriori*) relates it to the principle of autonomy.
DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK - BENEFICENCE	Diagnoses the existence of an ethical issue, and then (*a posteriori*) relates it to the principle of beneficence.
DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK - JUSTICE	Diagnoses the existence of an ethical issue, and then (*a posteriori*) relates it to the principle of justice.
DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK - NON MALEFICENCE	Diagnoses the existence of an ethical issue, and then (*a posteriori*) relates it to the principle of non maleficence.

ETHICS OF DEVELOPMENT PROCESS - ETHICAL DEVELOPMENT	Considers the ethical nature of certain design choices in themselves, not necessarily due to specific consequences. Related to the design and development process rather than to the actual use of the tool post-deployment.
ETHICS OF DEVELOPMENT PROCESS - UNETHICAL DEVELOPMENT	Considers the unethical nature of certain design choices in themselves, not necessarily due to specific consequences. Related to the design and development process rather than to the actual use of the system post-deployment.
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK	Implicitly considers an element of an ethical framework as a lens to frame their reflection on some subject (context).
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK - AUTONOMY	Implicitly considers the principle of autonomy as a lens to frame their reflection on some subject (context).
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK - BENEFICENCE	Implicitly considers the principle of beneficence as a lens to frame their reflection on some subject (context).
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK - JUSTICE	Implicitly considers the principle of justice as a lens to frame their reflection on some subject (context).
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK - NON MALEFICENCE	Implicitly considers the principle of non maleficence as a lens to frame their reflection on some subject (context).
GUIDING VALUES	Identifies or asserts a personal value as a guide for designing the artifact.
SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK	Explicitly starting from an element or principle of an ethical framework, tries to find an issue that is related to it.
SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK - AUTONOMY	Explicitly starting from the principle of autonomy, tries to find an issue that is related to it.
SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK - BENEFICENCE	Explicitly starting from the principle of beneficence, tries to find an issue that is related to it.
SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK - JUSTICE	Explicitly starting from the principle of justice, tries to find an issue that is related to it.

SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK - NON MALEFICENCE	Explicitly starting from the principle of non maleficence, tries to find an issue that is related to it.
BREADTH OF REFLECTION	Considers how much of the problem domain was considered during ethical reflection. Delimits the boundaries of their own reflection.
CONSIDERING CONSEQUENCES	Discusses how the tool helped them focus on the consequences of the artifact's actual use.
DIRECTED REFLECTION	Considers how the tool directed their reflection.
SHAPING REFLECTION	Asserts how the tool (and its sections or questions) influenced the reflection process.
SPACE FOR ETHICAL REFLECTION	Discusses how the tool aided them to consider ethically relevant issues through having a section/question with ethical focus.
AGENCY TO ARTIFACT	Attributes agency to the artifact they were developing.
CONSTRAINTS ON ARTIFACT	Considers design choices, implemented in the artifact or based on external rules, to prevent unethical uses of the artifact.
CONTEXTUAL INFLUENCES	Considers how the context in which the artifact would be developed or used would affect its results.
DESIRED TRAITS OF ARTIFACT	Defines a desired characteristic of the artifact.
ESTABLISHED ETHICAL FRAMEWORK	Asserts that bioethical principles are helpful for being a common-ground reference.
EXACERBATING PROBLEMS	Considers how design choices may make existing problems worse, based on a recognition that the *status quo* is already problematic.
EXPECTED READER	Discusses which stakeholders they considered would, or should, read the document resulting from the tool's use.
EXPOSURE	Mentions that they avoided stating something in the tool due to their responsibility for the artifact, and making the information available.
IMPACTED INDIVIDUALS	Considers who could be impacted by the artifacts they are developing. These would be the patients of the designer's actions, even if they may be agents in other moral relationships.
INSIDER UNDERSTANDING	Discusses how people not directly involved in the development process might not be able use the tool.
LIMITED KNOWLEDGE	Considers the limitation of the existing knowledge available for themselves, as designers, or for the artifact.
MEASURING IMPACT	Considers how they could measure the impact of the artifact's use.

PARTICIPATORY DESIGN	Considers how the participation of stakeholders in the tool-filling process would be beneficial, or how they could reap benefits from the resulting document communicating design choices.
REAL-WORLD USE	Defines use cases that could occur after the design process is finished. This can include use cases that were not accounted for.
RESPONSIBILITY FOR ARTIFACT	Remarks about their feeling of responsibility for the artifact being developed.
SUBJECTIVE PERSPECTIVE	Considers how their own subjective perspective on the process affects the artifact's development.
SYSTEM'S AUTONOMY - INCREASE	Defines that the artifact's autonomy should be increased, can be made upon defined conditions.
SYSTEM'S AUTONOMY - LIMIT	Defines that the artifact should have its autonomy limited.
UNDESIRED CONSEQUENCES	Defines possible consequences of the artifact's use they want to avoid.
USER GOALS	Considers what would be the stakeholders' goals with the artifact

After the final coding step, a superficial analysis of the coded segments led us to find out we had a high variance in the length of coded excerpts tagged by each coder. Upon this discovery, one of the coders re-tagged the dataset in order to bring both codings closer. As a result, we had a more similar distribution of the size of coded segments and the amount of coded segments for each coder.

To conduct the entire coding process, following the first step, all documents were imported into QDA Miner Lite,[1] a computer-assisted qualitative data analysis software (CAQDAS). The free version of the tool allowed us to conduct some basic analysis during and after the codification process. At the end of the process we were able to export all the coding data to CSV files and analyze it using Python libraries for data analysis and visualization.

[1] `https://provalisresearch.com/products/qualitative-data-analysis-software/freeware/`

6
Results

This chapter explores the data collected from our study. First, it details the preliminary steps we conducted to augment the amount of information we could relate to our data, then it describes the subset of our data used to conduct the following analysis and discussion. Next, we conduct an exploratory analysis of the coded data, including characteristics of the coded excerpts and where they were identified. Finally, we engage in more specific explorations of the codes, which are directly related to our research question of how Model Cards may contribute to designers ethical reasoning.

To properly explore the ethical reflection engaged by participants, using the research described in Chapters 2 and 3, and more easily visualize the relation between our codes, we restricted our analysis in this dissertation to the Model Card interview, and also to a subset of codes more closely related to ethical reflection, as described in Section 6.2.

6.1
Preliminary Steps

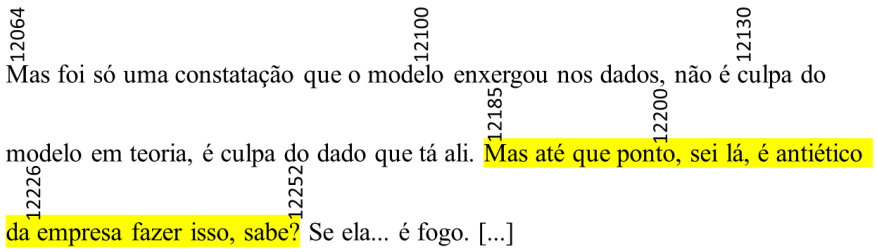
We opted to conduct a more thorough exploration of our data in Python, exporting each coder's dataset to a CSV file. The main objective of this decision was to better control our plots and to able to generate visualizations with information not directly available through QDA Miner, such as the section of the model card the text was written.

Our first obstacle was that, despite using the software export feature, the resulting file had considerable problems in its columns and separation, even after adjusting the character encoding.[1] To deal with this, we dropped columns that contained no information in any of the rows, and replaced a non-encoded whitespace character in the column names. In the end, we discarded no information from our dataset, but renamed the columns and discarded columns that were empty.

To further contextualize each of the excerpts and provide more information about the context in which it was used, we opted to add more information

[1]We found that the encoding that best loaded the QDA Miner data was"cp 1252" with a "," separator.

relating to the location where it was used within the document. To add this information, we took two steps. The first was simply to find each coded excerpt substring within the coded document and add columns to a dataframe containing the character locations of the document text where each substring started and ended. An example of this process is illustrated in Figure 6.1



Case	Text	Words	start_code	end_code
P2 - Model Card	Mas até que ponto, sei lá, é antiético da empresa fazer isso, sabe?	13	12185	12252

Figure 6.1: Example of identification of excerpt start and end character.

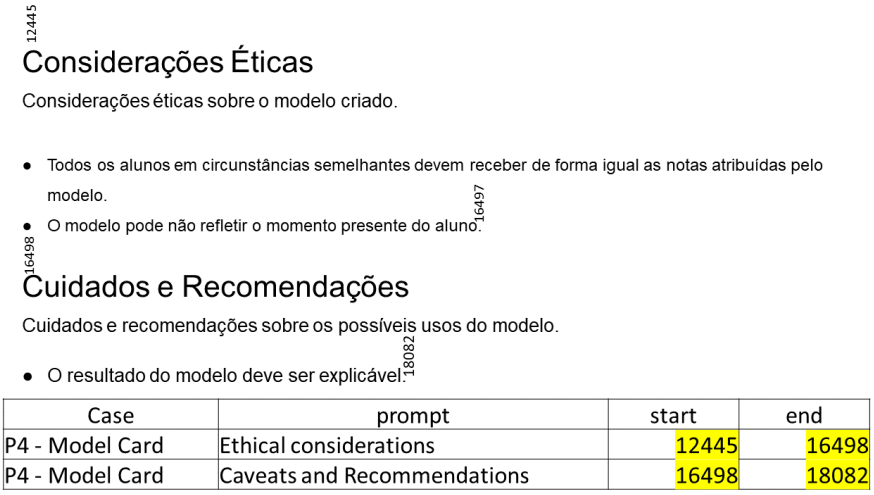
On a more important note, we also wished to obtain context about the coded segments regarding what section of the tool on which the interviewee was working. To address this, we manually compiled another file containing the character locations in the document text, where each of the tool's section began and ended. For the filled Model Cards, this was done simply by its structural division. However, for the interviews' transcripts, this was done either when participants vocalized something about starting a new section, reading its description, for instance, or by the context of what was being discussed or asked (Figure 6.2).

After doing this preliminary cleaning and enriching of our coded segments, we had the data upon which we conducted our analysis. All the code used for this process can be found in Appendix C.

6.2 Subset of codes and data used

As briefly mentioned, we opted to focus all of our analysis into a subset of the data we collected. This decision was made to address concerns of the time constraints, and to allow a more thorough analysis into our topic of ethical reflection using the research mentioned in Chapters 2 and 3.

Initially, this was done by limiting the first set of analysis to the documents and interviews that used the Model Card. This decision was made



Case	prompt	start	end
P4 - Model Card	Ethical considerations	12445	16498
P4 - Model Card	Caveats and Recommendations	16498	18082

Figure 6.2: Example start and end of each Model Card section.

based on the fact that the work on Model Cards was published earlier,[2] had already been adopted openly by certain initiatives,[3] and we also expected its questions and fields to be seen as closer to concepts with which developers interact routinely than those introduced by the Extended Metacommunication Template.

We also opted to focus all of our analysis into a subset of the codes we used, those more closely related to ethical reflection. Although we cannot expect to have been exhaustive about all that participants expressed with our coding, we covered a wide array of themes, including participants' perceptions about the tools in use and their own reflection.

In line with the initial objective of our research, we opted to further narrow down our scope of analysis only to codes directly related to the participants' ethical reflection while using the tool. This includes their reflection related to what was discussed in Section 2, and eventual use of the bioethical principles presented.

The selection of codes that were further explored is found in Table 6.1.

Table 6.1: Final codebook, including split codes.

Code name	**Description**
AGENCY TO ARTIFACT	Attributes agency to the artifact they were developing.
DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK	Diagnoses the existence of an ethical issue, and relates it to an element of an ethical framework (*a posteriori*).

[2]At the start of this research, Barbosa et al.'s 2021 article was not yet published, and was at an early draft stage.

[3]Google has adopted Model Cards as an initiative for its cloud models (https://modelcards.withgoogle.com/about).

DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK - AUTONOMY	Diagnoses the existence of an ethical issue, and then (*a posteriori*) relates it to the principle of autonomy.
DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK - BENEFICENCE	Diagnoses the existence of an ethical issue, and then (*a posteriori*) relates it to the principle of beneficence.
DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK - JUSTICE	Diagnoses the existence of an ethical issue, and then (*a posteriori*) relates it to the principle of justice.
DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK - NON MALEFICENCE	Diagnoses the existence of an ethical issue, and then (*a posteriori*) relates it to the principle of non maleficence.
ETHICS OF DEVELOPMENT PROCESS - ETHICAL DEVELOPMENT	Considers the ethical nature of certain design choices in themselves, not necessarily due to specific consequences. Related to the design and development process rather than to the actual use of the tool post-deployment.
ETHICS OF DEVELOPMENT PROCESS - UNETHICAL DEVELOPMENT	Considers the unethical nature of certain design choices in themselves, not necessarily due to specific consequences. Related to the design and development process rather than to the actual use of the tool post-deployment.
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK	Implicitly considers an element of an ethical framework as a lens to frame their reflection on some subject (context).
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK - AUTONOMY	Implicitly considers the principle of autonomy as a lens to frame their reflection on some subject (context).
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK - BENEFICENCE	Implicitly considers the principle of beneficence as a lens to frame their reflection on some subject (context).
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK - JUSTICE	Implicitly considers the principle of justice as a lens to frame their reflection on some subject (context).
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK - NON MALEFICENCE	Implicitly considers the principle of non maleficence as a lens to frame their reflection on some subject (context).
GUIDING VALUES	Identifies or asserts a personal value as a guide for designing the artifact.
IMPACTED INDIVIDUALS	Considers who could be impacted by the artifacts they are developing. These would be the patients of the designer's actions, even if they may be agents in other moral relationships.

RESPONSIBILITY FOR ARTIFACT	Remarks about their feeling of responsibility for the artifact being developed.
SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK	Explicitly starting from an element or principle of an ethical framework, tries to find an issue that is related to it.
SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK - AUTONOMY	Explicitly starting from the principle of autonomy, tries to find an issue that is related to it.
SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK - BENEFICENCE	Explicitly starting from the principle of beneficence, tries to find an issue that is related to it.
SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK - JUSTICE	Explicitly starting from the principle of justice, tries to find an issue that is related to it.
SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK - NON MALEFICENCE	Explicitly starting from the principle of non maleficence, tries to find an issue that is related to it.
SYSTEM'S AUTONOMY - INCREASE	Defines that the artifact's autonomy should be increased, can be made upon defined conditions.
SYSTEM'S AUTONOMY - LIMIT	Defines that the artifact should have its autonomy limited.
UNDESIRED CONSEQUENCES	Defines possible consequences of the artifact's use they want to avoid.

6.3 Coding consolidation

Upon starting to work on our analysis, and comparing both coders datasets, we noticed that each coder had tagged several segments which had not been tagged by the other. After filtering for the selected codes, we had one set of 81 and another one of 98 observations.

By applying a process to merge and compare segments of both codes based on lax criteria,[4] we were only able to find matches for 54 segments.

During this exploration we also made an initial evaluation of our agreement. Our initial evaluation lead to an Fleiss Kappa score of around 0.32. However, this was heavily influenced by the aforementioned fact that each coder had a high number of excerpts not codified by the other; ignoring such

[4]We matched pair of codes where the match were of at least 60% of the size of the longest coded excerpt, codes that contained or were contained for each of the documents. This process is prone to error, especially considering both coders used nested codes, which could lead to multiple matches for cases where both coders used nested codes in the same passage.

instances, our inter-coder agreement was at a substantial level of 0.62. We understood this as meaning that where both identified a relevant comment by participants, they tended to agree on the relevant code. However, in several cases, one of them would either not associate a passage with any of the codes, or associate it to a code not included in the final subset.

In light of these findings, we opted to create a combined set of all the coded excerpts and conduct a new coding step. To create a cohesive dataset, we joined all excerpts that received at least one of the tags included in the final codebook, by any of the coders.

After these steps, we had a set of 192 excerpts. All information related to the previous coding was removed so as not to unduly influence our consolidation. Then, those excerpts were coded again by both coders, using only the codes shown in Table 6.1. In addition, we included a code "Other", for when they believed another code (not in the final codebook) should have been used.

After each a round of independent coding, all cases that presented divergence between the coders were commented and discussed by them. In the end of the process, we reached our final coding by a process of discussing each case and reaching a consensus, denominated as negotiated agreement by Campbell et al. (2013). We reached a consensus in most coded excerpts, and in four of them we applied two distinct codes to the same excerpt. While recoding this set, we noticed that 31 excerpts were duplicated, referring to the same passage (give or take a couple of words), due to limitations of the algorithm used to combine the excerpts. As each pair of those excerpts was associated with the same code, we manually identified these cases and removed the duplicates from our dataset, opting to preserve the longest ones in all cases. Our final dataset is the result of this process. Each coded excerpt can be found as transcribed in Appendix A (in Portuguese).

6.4 Coding Exploratory Analysis

After the steps described above, our code dataset comprised 161 coded excerpts. Excluding the excerpts associated with the code "Other", this number was further reduced to 150. The following section provides an overview of the general characteristics of this dataset.

Figure 6.3 shows a distribution of the coded excerpts per section, in the order which they appear in the Model Card.

It is no surprise that the section with most excerpts was that of *Ethical Considerations*, as this was the focus of our analysis and of the subset of codes we narrowed our analysis to.

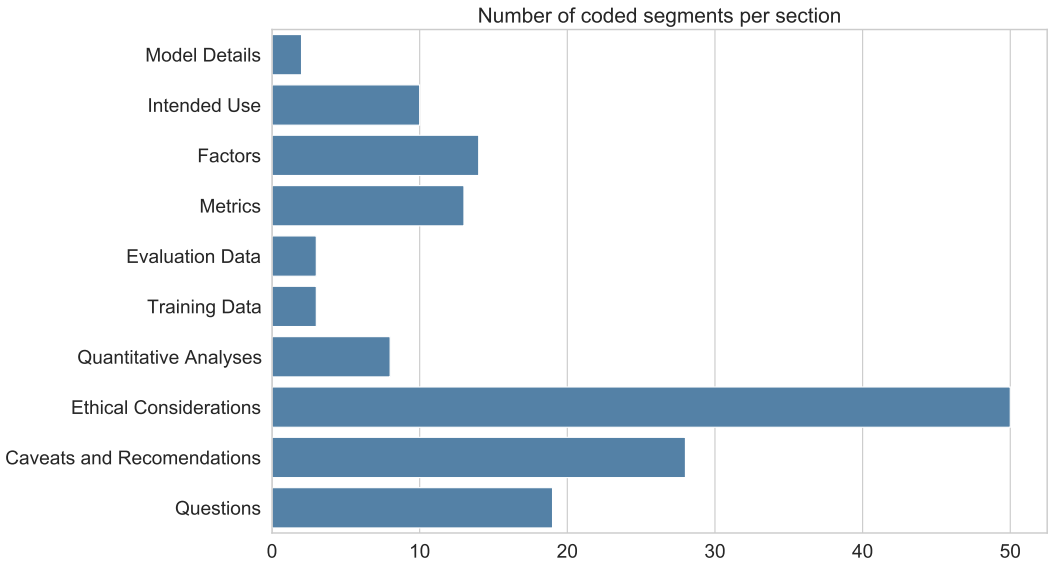


Figure 6.3: Distribution of coded excerpts per section

In addition to that, we can verify that the *Caveats and Recommendations* section was also a strong prompt for ethical considerations by the participants. This may be partly explained by the fact that it directly followed the *Ethical Considerations* section. Thus, as their reflection had just been directed towards that topic, it may have directed their focus and prompted them to include these considerations in the following section.

Figure 6.4 shows similar information, but divided by whether the code was found in the interview or the document containing the filled-out Model Card. The figure reveals that, as expected, in absolute numbers coded excerpts were much more frequent in the interviews, which had 119 codes assigned, when compared to the tools themselves, with a total of 31 assigned codes.

It is interesting to note that some of the sections only had coded segments in the interviews. This is a consequence of the fact that participants would discuss their ideas, going in depth into their thoughts and reasoning in the interview, while the resulting document was composed mostly of a summarized version of the result of that reasoning, not the reasoning process itself. Another consequence of this difference can be found on Figure 6.5. This figure shows that, as expected, excerpts in the interview tended to be longer than those extracted from the filled Model Card.

While both cases present a high variation, and the shorter passages were of similar length, the upper ends of the distribution were larger for the interviews. The same trend is found when we disaggregate the distribution for each section, as found in Figure 6.6.

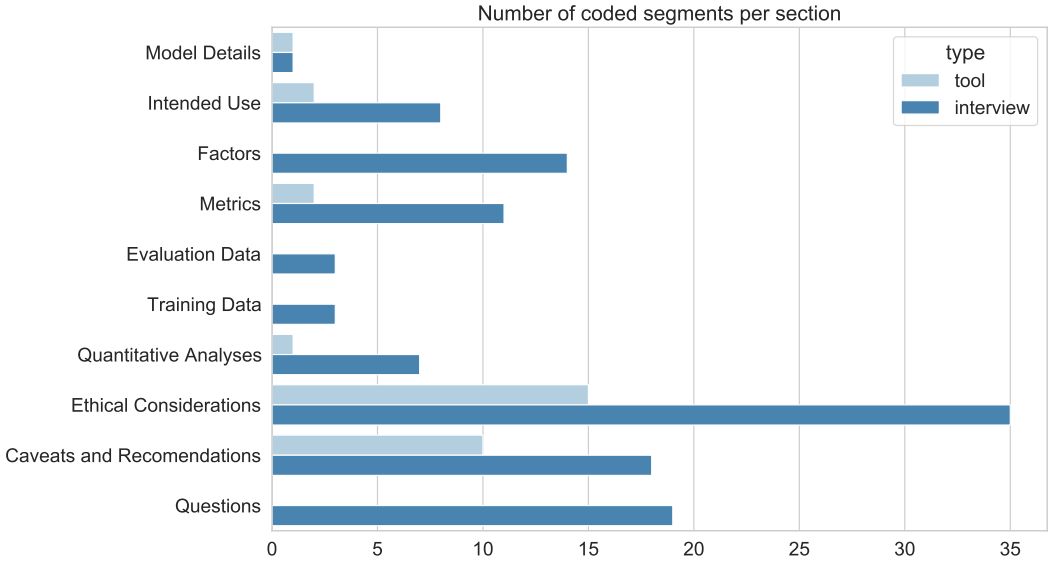


Figure 6.4: Codes distribution per section and type of document

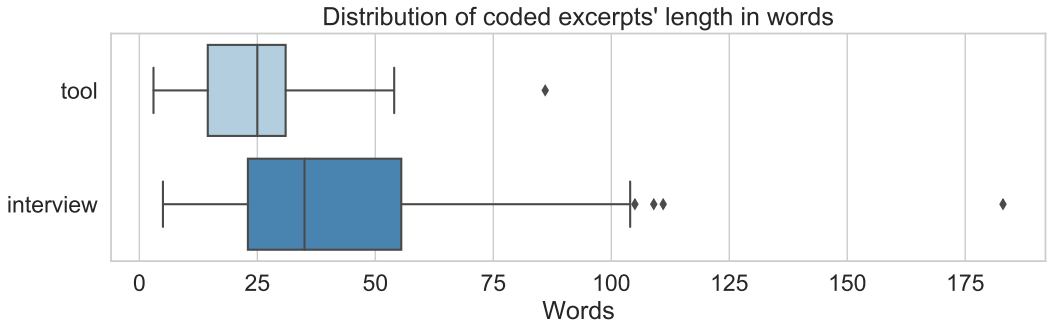


Figure 6.5: Distribution of coded segments length.

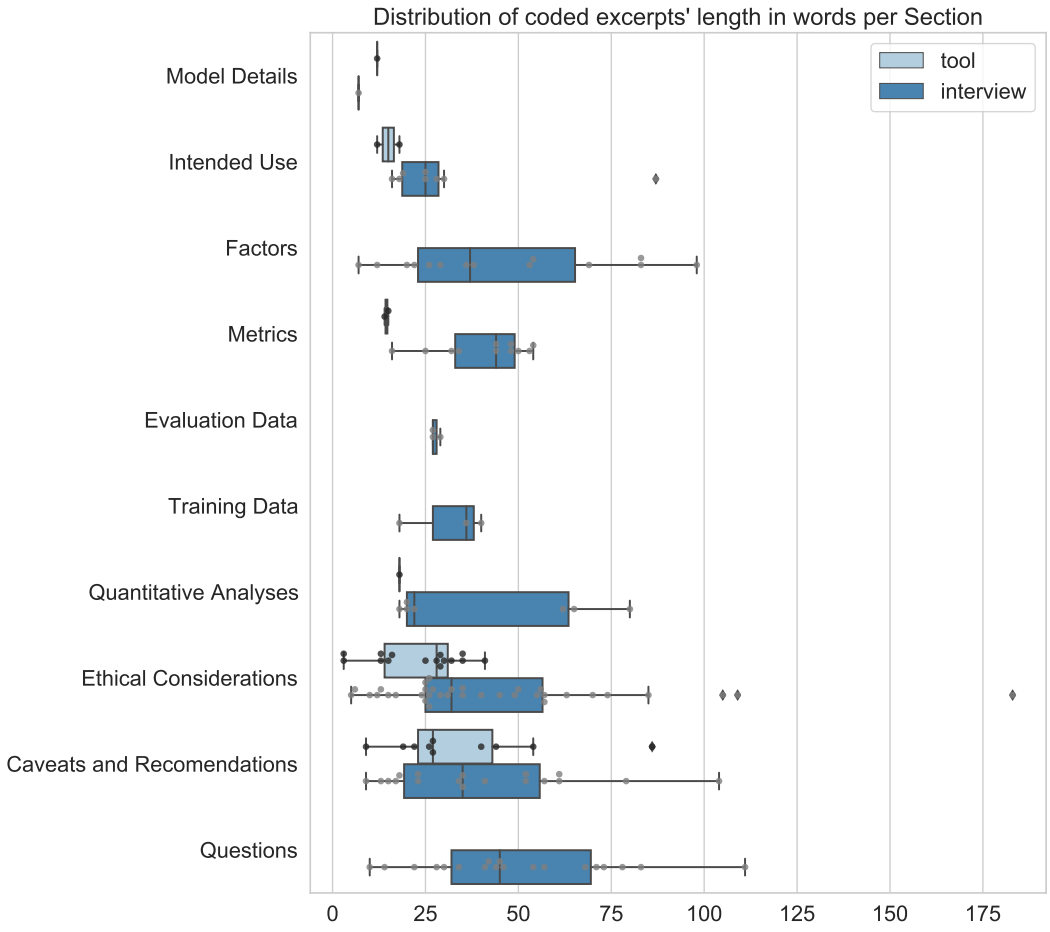


Figure 6.6: Distribution of coded segments length per section.

6.5 Code Occurrence and Frequency

In this section, we explore the coded excerpts themselves, the codes frequency, how they varied between each of our coded documents and among participants. Initially, we go through the frequency of codes, further constructing the analysis by breaking the frequency down by type of document, and among participants. Finally, we investigate the co-occurrence between codes, to try to understand how they were related in our data.

We simplified our codes for the plots in this section to better visualize them, while fitting them to the appropriate page size. All codes can be found in their original text in Appendix A.1. In our analysis, the numbers inside parentheses indicate the corresponding code fragment in Appendix A.

Fig 6.7 displays the frequency of each of the codes throughout our data.

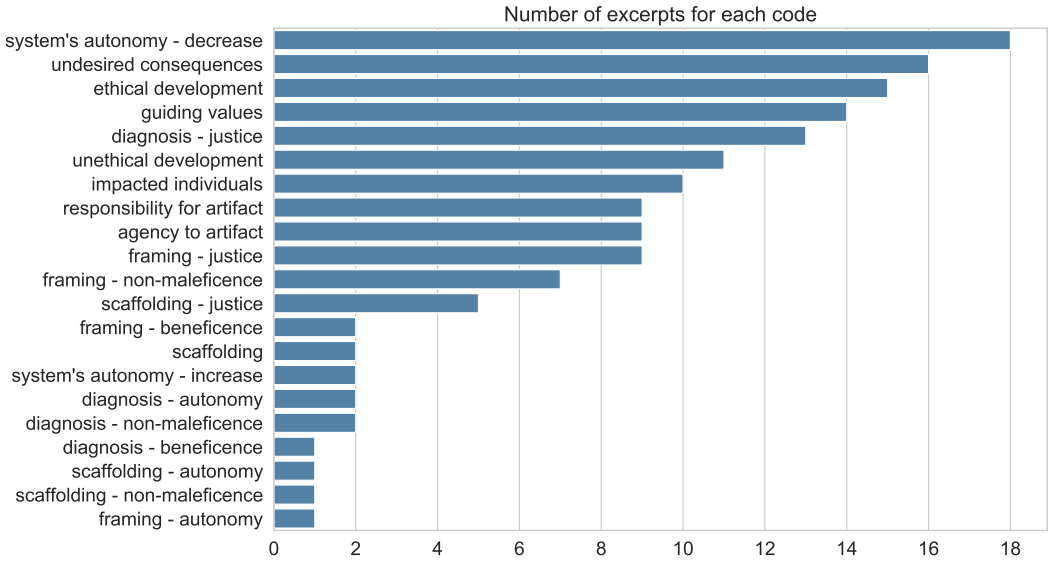


Figure 6.7: Number of excerpts associated to each code.

Our most frequent code was “System’s Autonomy - decrease”, which constituted instances of participants declared they believed the autonomy of their artifact should be limited over the relevant decision-making process. One example of these statements was made by Participant 1, and registered in his Model Card: he had previously stated that he envisioned the final grade attributed to candidates in the education scenario should be an aggregation of scores for different attributes: “At a second moment, the grades would be issued and a committee would ~~make a manual attribution~~[5] calculate the final grade”, the original text can be found in line 14 of Appendix A.1.

[5]In our analysis, we format in strikethrough text fragments that the participant inserted and later erased.

Other excerpts were even more explicit over this issue, stating that the output of the artifact should not be used without user supervision. For instance, Participant 3 stated "The model will aid the decision-making process of the [financial] institutions offering a value related to the risk of each client, the model should not decide whether the loan will happen or not." (49). On the same note, Participant 4 stated multiple times that "[t]he result of the selection should not depend solely on the grade given by the model" (65, 69, 71, and 72).

Another common theme of remark throughout our data were the "Ethics of Development Process" codes, related to the ethical nature of certain choices and actions made during the development process. On this issue, there was a large number of coded excerpts for both the ethicality and unethicality of aspects they identified.

Regarding ethical choices, Participant 8 declared "OK, I think that I should ensure that the data is well distributed to, for instance, characteristics like social class. In order to, for example, not benefit a group more than other" (155). On the other hand, he also recognized the negative effects of using some characteristics, affirming "These[social class, race and sex] are information I should be careful because I could be reinforcing existing prejudices that already exist."

Figure 6.8 shows the count of each code, differentiating between codes in the Tools and Interviews. Figure 6.9 complements this picture by displaying the difference in the number of excerpts for each code between the Interviews and the Model Cards documents.

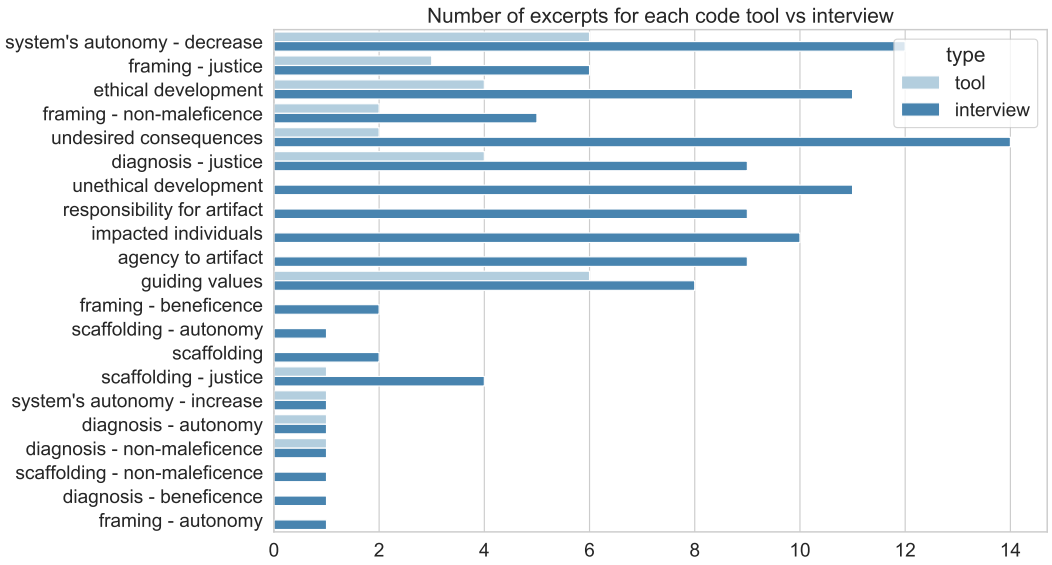


Figure 6.8: Number of excerpts for each code in tools and interviews.

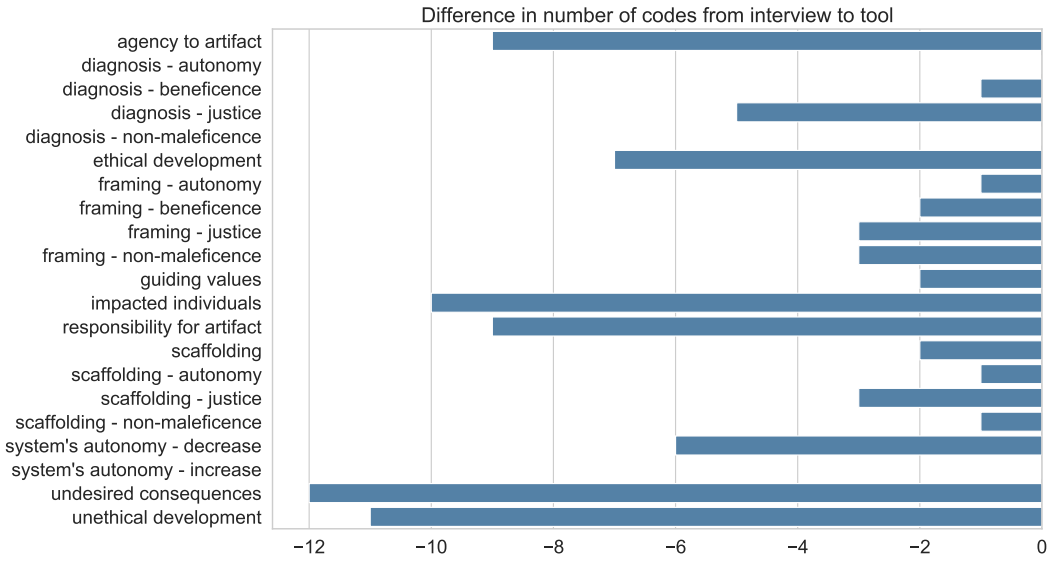


Figure 6.9: Difference in number of excerpts from interview to tool for each code.

By looking at both images, we can start to notice some interesting patterns regarding how the codes were identified in the data, and what participants said and wrote during our study. Specifically, the fact that while "Ethics of Development Process - Unethical Development" and "Impacted Individuals" were among the most frequent codes in our dataset, their occurrences were completely limited to the interviews, hence also presenting the highest difference in frequency from interviews to the tool.

This suggests that, while participants opted to delve into these issues in their own reflection upon the scenario, they seem to not have found the Model Card a suitable vehicle for documenting them, and did not include topics related to them on the final document. This is especially interesting considering all participants, when asked at the end of the interview, answered they did not exclude from the documents any information they believed was relevant for not fitting the Model Card structure, indicating the exclusion was a choice made by participants based on what they found relevant to record.

Another example of the "Unethical Development" code which was not transferred to the Model Card document was given by Participant 2: "Hold on, profession is important. Profession, education level. Then I can be a little unethical and ask for the address. I will already raise this ethical issue here. I am going to include it there: address" (28). Later in the Model Card, the participant raised the possibility of the artifact using an occasional relation between the address and income of the loan requester to deny future clients. However, we applied codes that focus on the participant identifying ethical problems to these segments, since they do not highlight the ethicality of the

development choice of opting to use this information, as was explicitly done in the interview.

This behavior also happened with Participant 3, which made the following statement during the interview regarding the inclusion of age in the artifact input: “When I was in this part here [Quantitative Analysis] I removed it, because I saw that it was not something... fair.” (60). Despite this realization during the interview, and the reflection guided by the Model Card, the participant did not explicitly mention any problem related to the use of this feature, which was actually left among the variables of the dataset in the Model Card.

The fact that no instance of “Impacted Individuals” was identified in the tools themselves is also interesting, since it may indicate that the resulting Model Cards did not highlight groups of individuals that could be impacted by the artifact being developed, although they were identified during the reflection process. Some of these comments were made during our questions, answering our inquiries about individuals and stakeholders that could benefit and be impacted by the Model Card. Participant 3 made a comment about the scenarios and how they felt similar to him due to the individuals that are directly affected by the artifact: “I would say that both scenarios are sensitive, you know, due to having to evaluate issues related to people. So... they are similar in that regard.” (64).

However, other excerpts were also expressed during the use of the Model Card, although not registered into them. Participant 6 made two remarks that included other actors that could be affected by the use of his artifact. These were: “maybe family members that have a bad financial history and this end up... influencing the answer of the system to that person” (108) and “[i]magine that a client has a, perhaps their parents have a bad financial history, but as he is just entering the market now, the thing is being used by his parents in order to receive the loan, do you understand?” (109).

Another code that presented the same pattern was “Undesired Consequences”, which only occurred twice in the Model Cards despite having an overall frequency of 16. This code was used to establish excerpts where participants identified possible consequences of their system that they considered undesirable, regardless of the reason.

One instance of this code can be found in passage 16 by Participant 1, which was also reflected in his tool: “Other information not mapped may be... may be inferred, it is necessary (typing into Model Card). I am not being able to explain this very well, but I mean that this dataset, imagining it had the candidate achievements, their publications, hence other information may be inferred. For instance, where he publishes, which are the main vehicles in

which he publishes, from there if I take some other information that may be used to profile this user, identify his interests and preferences. That may not be necessary for this selection, but it could be something... it is a possible product looking at this data, therefore [inaudible] certain care". The equivalent excerpt in his Model Card, found on Index 5 and 6 , received the code of "Framing based on the maleficence principle", since he framed that remark around not causing damage by misuse of the collected data, and "Undesired Consequences".

On the other hand, Participant 8 identified a possible consequence of including certain information into his model in the interview, which was not reflected into his final document. Reflecting over possible discriminatory results he stated: ´´I cannot use race, for example, as input, but... it may impact in a more indirect way which is that someone that comes from a lower social class had more issues during his education, because he may need to work for example, and had lower grades. It is not direct, but it is more indirect." (146). The fact this type of reflection was not transferred to the Model Card in most cases suggest our participants did not find it suitable to include it while using the tool.

We can see in Figure 6.10 the number of occurrences of each code per participant in our study.

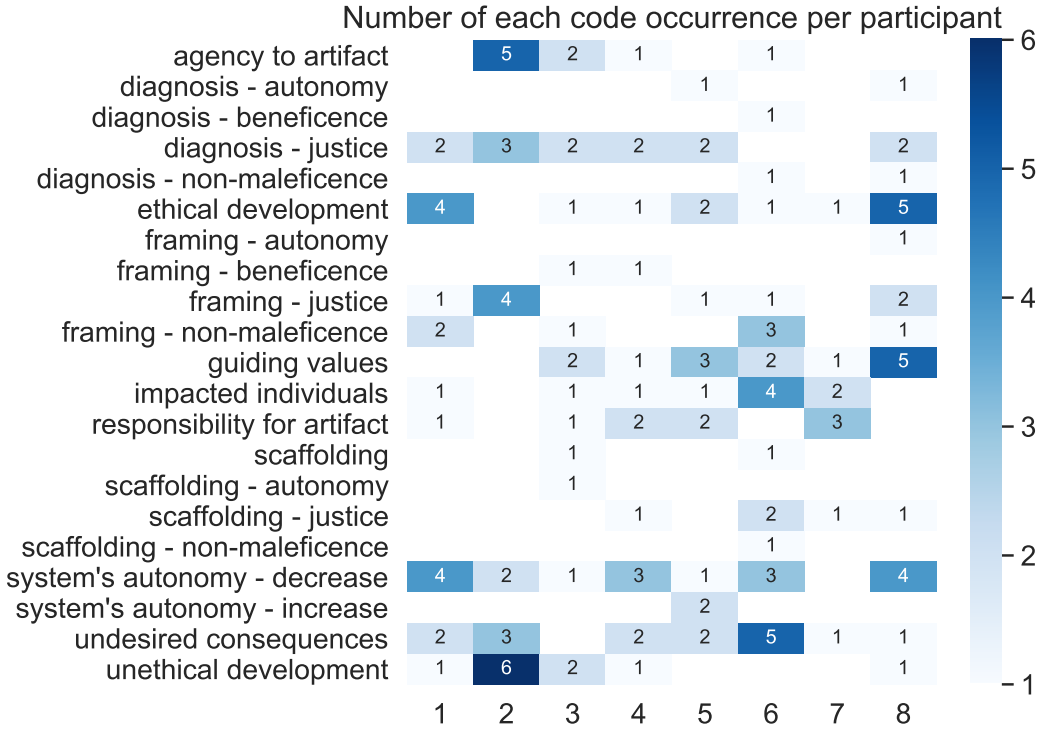


Figure 6.10: Number of occurrences of each code per participant.

The information regarding the code "System's Autonomy - decrease" shows that it was not only a frequent code throughout the study, but that it

was also a recurring theme across participants, being present in all of them except for Participant 7. The remaining ones made comments very similar to the one cited from Participant 4, stating that the artifact output should not have the final say over the decision making process. For instance, Participant 6 left the following recommendation at the relevant section: "Do not take the model as the only resource to approve or deny a loan. The model could be used to assist on the decision of an employee." (102).

On the flip side, the opposite code ("System's Autonomy - increase") was only used by Participant 5, who was the sole participant to establish conditions in order to increase the role taken by the artifact in the decision making process. This can be seen in excerpt 96: "Let's say that during the following two years we will make a mixed admission process. We will take the model's output and the opinion of the evaluators, and check whether it is OK. If it is OK, perhaps in the following year we can use only the model. It is something in that sense.". Despite this, he also expressed his general thoughts on limiting the autonomy of the artifact "I am always a bit uneasy to make something completely autonomous, completely automated" (94). It seems that this participant understood that it was desired of him, in the presented scenario, to create a fully automated systems, which he registered was against his own personal beliefs.

Another notable aspect is that, excluding Participant 2, all remaining participants had excerpts tagged with "Ethical Development", meaning we identified they were analyzing the ethical nature of an aspect of the development process. However, these codes were much more frequent for Participants 1 and 8, which might indicate they were especially aware of possible consequences of their choices. This is further supported by the fact that both of them also had high frequency of other codes, such as "System's Autonomy - Decrease".

Participant 1 also had a notable frequency of "Framing based on element of ethical framework - non maleficence", also related to reflecting on possible damages that arise from the use of the artifact under development. On the other hand, Participant 8 had a high count of excerpts identified as "Guiding Values", and putting these values as explicit may have also highlighted the option to incorporate ethical decisions into the development. One instance of this code was excerpt 149: "Here I believe that as I am considering it as a public university [the institution in the admission scenario], in my view I should give more opportunities for those, for example, do not have financial means to get higher education, to pay for a private university. Thus, I believe that this should be a metric. I'll think about how I can write this."

Connecting the codes to each of the Bioethical Principles, we can see that some of them were present in a wide array of participants, while others

were restricted to a few individuals. For instance, codes directly related to justice were present for all participants, while codes related to autonomy were used by all except for Participant 7.[6] Conversely, codes based on the non-maleficence principle were only present for Participants 1, 3, 6, and 8, and based on the beneficence principle in Participants 3, 4, and 6, while also having lower absolute frequency, as shown in Figure 6.7.

Figure 6.11 further breaks the information down between each participants' Model Card and Interview. One surprising fact is that Participant 7 had no coded texts in his Model Card, only in the Interview. This suggests that this session was exceptional in a way. One characteristic from this interview, which corroborates its peculiarity, was the choice made by the participant to define its scenario towards corporations, and not individuals, and might have deviated the reflection from the one had by other participants. This was shown in excerpt 127, coded as "Impacted Individuals": "It was... an option, I suppose, I flipped a coin. It could be a model for individual or corporations. Then I chose in the scenario to be focused on corporations.".

Another relevant fact we can notice from this figure, further discussed in Section 7.1, is that some codes occur in interviews but not in the Model Cards. This is especially noticeable in the contrast of ethical development, which appears in the Model Cards, and Unethical Development, which does not.

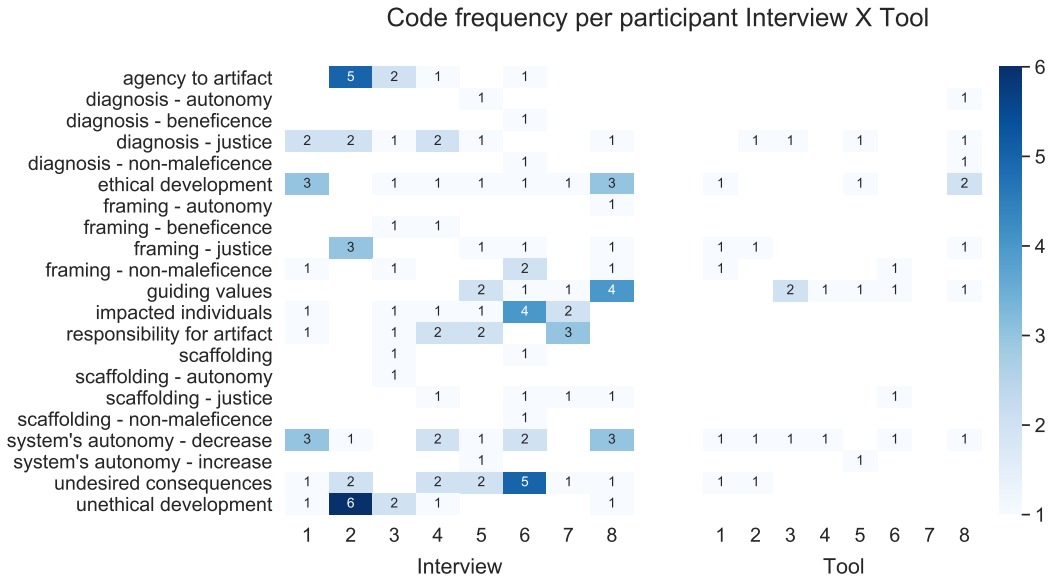


Figure 6.11: Number of occurrences of each code per coded document.

We can now shift our focus to the co-occurrence between codes. To look at

[6]This includes excerpts identified as related to the system's autonomy, since the more common option to restrict it implies the preservation of the autonomy of humans involved in the process.

this information, we generated the heatmap in Figure 6.12. The figure displays the number of documents in our study in which each pair of codes occurred within the same section. Using this information we can look for relations between codes that might have been shared between different participants. Similar plots for each individual participant can be found in Appendix B

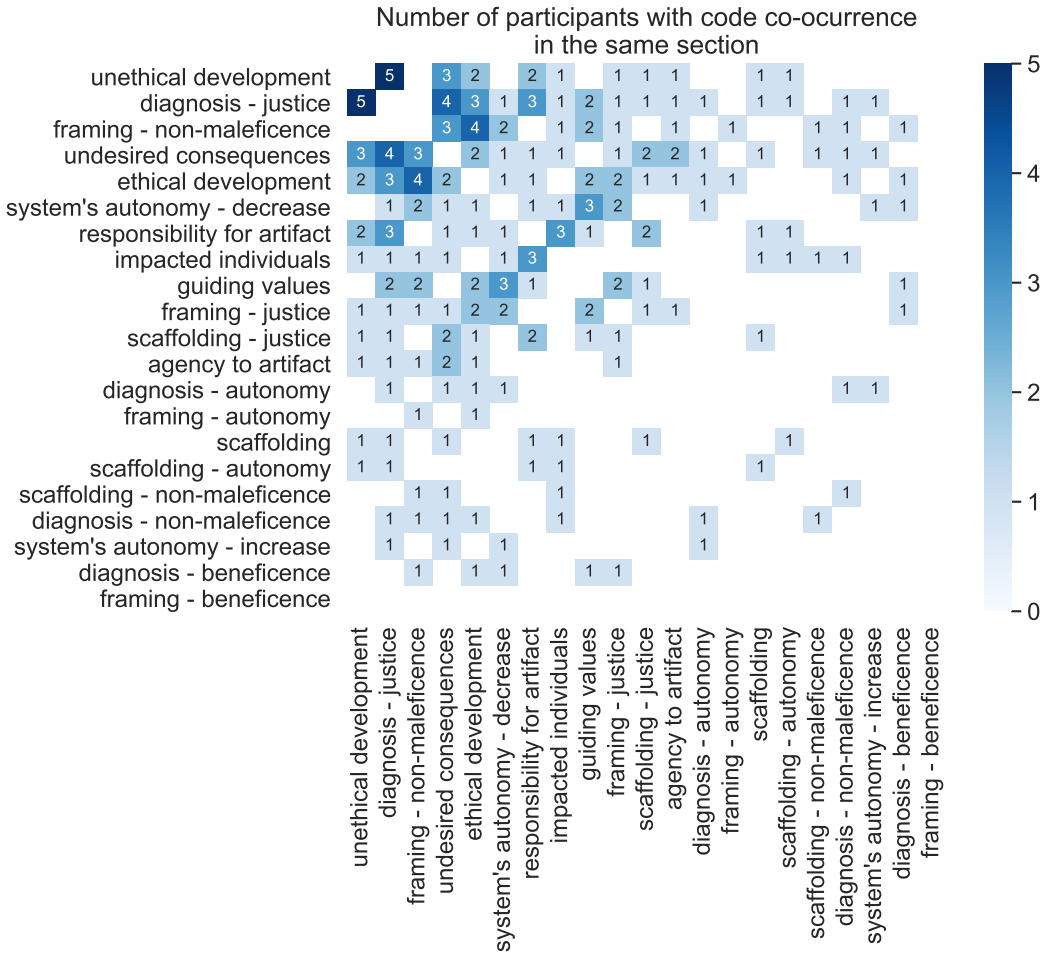


Figure 6.12: Number of sections for different documents with co-occurrences of each pair of codes in the same section.

Moreover, we conducted an analysis considering each code as a token. Afterwards, we generated bigrams of codes that occurred in the same document (both the Model Card and Interview for each participant). When we had two codes that started at the same place in the text, we opted to generate a bigram of both codes with the previous and following one.

To avoid counting as bigrams codes that were largely separated in the text, we analyzed the distribution of the distance between each pair of codes. We then calculated the median distance, and discarded all pair of codes that were not within that distance in tokens. Figure 6.13 shows the resulting data for all pairs that occurred more than once. It is important to note that the

order the codes were found is relevant, therefore we can see the same pair with different orders in the plot.

An interesting observation is the common presence of "Diagnosis of Element of Ethical Framework - Justice" among the pairs of the most frequent co-occurring codes. This may be explained in part by its high frequency –it was the fifth most common code in our dataset–, and its use by six participants. However, other codes that were even more frequent are not as dominant on this co-occurrence metric. We were unable to come up with an explanation for this behavior.

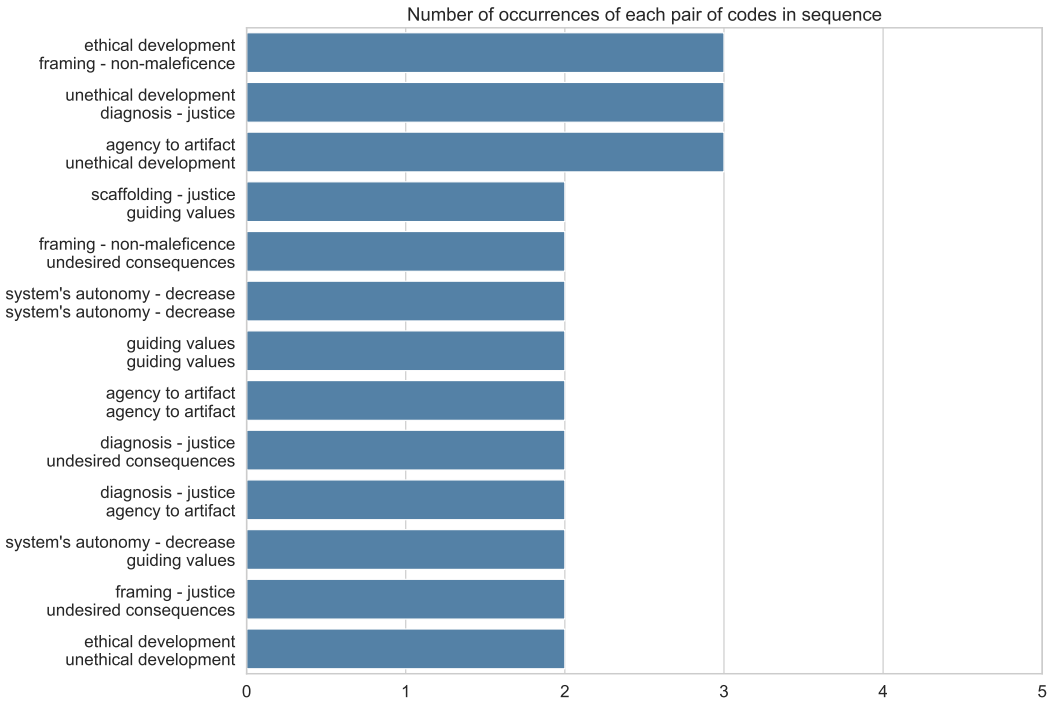


Figure 6.13: Frequency of each code bigram.

The most common pair of codes to co-occur for both analysis were "Diagnosis of Element of Ethical Framework - Justice" and "Ethics of Development Process - unethical development", which occurred with five participants, a very high rate considering these codes were present for six and five participants, respectively. This means that in all participants we identified an instance of the *Unethical Development* code, one of its occurrences was in the same section of an instance of the *Diagnosis - Justice* code.

The pairing of these codes suggested to us that participants were relating identified characteristics of development they considered unethical, and basing this analysis on the diagnostic of the presence an an issue they identified as a problem of justice.

Participant 2 offers a good example of this, where the relation was actually even stronger, and we coded each of two different excerpts from his

interview with both of these codes. These were the passages: “if [the company] is going to maximize its profit, it will leave people with lower education level excluded, because they know they would have more losses. And, overall, the people with lower level of education are more modest. Thus, it ends that these more modest group of people is gonna stay without credit, if compared with people with more money. It is an ethical discussion.” (37/38) and “Well... with basis on the address there is this issue... [writing] ‘there is a chance the model, when using the address as a training point, will have the tendency [to present discriminatory patterns]”’ (41/42).

Another instance of this relation was Participant 3’s already quoted comment regarding the removal of age as a feature, coded as “Unethical Development”. That comment was immediately followed by him diagnosing it as potentially causing discrimination, a problem of justice, in excerpt 61: “Exactly. Thus, i think it could lead to a certain... in way, like it or not, maybe for the companies it may mean something, but I am not certain it would be something that would not discriminate”.

The pair “Ethics of Development Process - ethical development” and “Framing Based on Element of Ethical Framework - non-maleficence” occurred in the same section for four participants. Here, we can interpret this as participants opting to take actions in the development process they view as ethical, with the goal of avoiding harmful results. This is further supported by the order they commonly occurred together, with the framing based on the non-maleficence principle following the development code, as seen in Figure 6.13.

We can see this in the Ethical Consideration section of Participant 1’s Model Card, excerpts 4 and 5. The participant first stated that “There is no data that allows the direct identification of characteristics such as gender, race, religion, income, address or any other information that is not related to the candidates’ achievements.”. However, despite the effort to avoid identification, he states that some things might still be inferred: “The origin of the candidate (state or city) may be inferred based on the information available. Other [information] not gathered may be inferred, beware the undue use of this dataset.”

Participant 8 had a similar train of thought, also reflecting about the dataset resulting from the development of his artifact. This can be seen in the comment “OK, I think I have to ensure the data is well distributed for characteristics such as social class. In order to, for example, not unduly benefit a group more than other.” (154), then followed by “I believe I also need to ensure an anonymization of the data, for issues like data leaks, or for example to keep information like academic history and family income from

individuals." (154). Although the cited excerpts form the pair of codes under discussion, we can also see within each of them this rationale, where they identify a development action they consider necessary, and frame this decision around avoiding negative results.

Another result we found is that six of the pairs seen in Figure 6.13 are repetitions of the same code. Both the codes "Diagnosis of Element of Ethical Framework - Justice" and "Ethics of Development Process - unethical development" also appear doubled, in addition to comprising one the most common pairs. We hypothesize two possible explanations for the presence of these repeated pairs: either that participants were further elaborating on the topics or they could just be repeating themselves, rephrasing previous statements.

To understand what happened, we looked for each of the instances where this occurred. In the case of the "System's Autonomy - decrease", it seems that these instances were indeed participants repeating and rephrasing their statements. For Participant 6, in excerpts 124 and 125, he first stated "It would be a caveat not to take the model as the only resource to approve or not the loan, right. In the scenario, hold on (changing written text). The model would not be used as the only resource to approve or not the loan, it would be used to facilitate the decision making of an employee." The following coded comment has the participant restating the role he envisioned for the model as a tool to be used by a company's representative, after being prompted by the interviewer.

However, Participant 8 had a sequence of remarks on the same topic where he was advancing his reflection further. He started this final comments by stating in excerpt 160, "I believe a possible recommendation is... I think I should have some step, some way to monitor this cases, or even, for instance, maybe go through a manual monitoring by someone.", which was then followed up by "I recommend a manual revision, monitoring, evaluation... to have a phase to calibrate and adjust of the parameters/model. [...]". Despite not being captured in this analysis due to the distance between the excerpts, the participant had another comment that elaborated on this issue and how he saw it, which also received the same code. He then stated "The idea here is for it [the model] to decide autonomously if the candidate should be accepted in the university or not. But perhaps there should be a step in which, for example, I can ask for a revision of my case, something in that sense, or to go through a manual evaluation, to have someone look at my data, check if there was no mistake, something like that."

This was also the case for the duplicated "Guiding Values" code. In the Model Card's Ethical Considerations made by Participant 3, he first stated in

excerpt 50: “The model will treat all clients equally, taking into account only economic factors.”. He then also included another point saying ´´The model should not be used in a discriminatory manner.”, which reinforces his previous point about maintaining the justice of the process.

A curious pair of codes that occurred twice, and in the same section, as seen in Figure 6.12, is that of “Ethical Development” followed by “Unethical Development”. The fact there were two instances where they followed each other indicates a stronger relation in how participants reflected upon the subject. By looking at the passages, we found that these instances actually indicated participants elaborating more on their views and the choices they made for the artifact, unlike for repeated codes.

In the case of Participant 1, this pair was followed by a third passage, coded as “Ethical Development”. In the first one (excerpt 12), he was reflecting on how he could conduct the development and evaluate the model without being discriminatory in his own view “Grades of groups of candidate by gender. How can you be sure that there is no segregation happening, without segregating. Because every time I divide a group of candidates by gender, in a way or another I am segregating. Because another factor here would be gender, color”. Then, he notices that using some of the information he just suggested was not correct: “Race is something that varies a lot, I just think... well, I am not sure that asking for this information about race is right.”. Finally, he goes back to building his process based on his previous reflection of what he considered correct “But... (writing) would verify. Well but race is not a variable that I would add, I think this is already a first step. Well, race, age perhaps... No, but in academia more experience, the person is older, thus it does not make much sense.”.

The other pair of these codes found in our data was further explaining Participant 4’s thought process, and what he decided to include in the “Ethical Considerations” section. First, he identified in excerpt 77 a rift between what he believed would be ethical to do, and what he interpreted from the scenario: “When I read the scenario, as I said, I considered this previous achievements, which the model do not take into consideration, at least the people that asked me, from the university, did not take into account this social history. Then, in the ethical considerations I am saying that the model that I was requested do not take these considerations, but I as a developer wanted to take these into account, and explained the reasoning.”

He then proceeded elaborating what his artifact does take into account: “Here, I included a kind of obvious consideration, that the model may not reflect the present state of the student. I can be a horrible student during

3 years of high school, make a fantastic admission exam e get into the best university there is. Then, my current state is of a good student, but in the past that was not the case. The model will always consider cases from the past, right and... it is bad to extrapolate that, slaying that if the candidate has a bad grade, to say he is a bad student at the moment. The best I can do is saying that he was a bad student in the past." (78).

This sets the picture of our interpretation of the data collected during our study regarding ethical reflection in our study. With this groundwork, in the next chapter we relate it to more general issues in the literature, and what we can use to improve the ethical reflection of developers during the design of AI systems.

7
Discussion

This chapter discusses possible implications of the findings of our study. Specifically, we consider how these relate to previous findings in the literature for communicating the meaning of AI systems, for the use of the Model Card as an epistemic tool for AI systems, and for future research that should be made to develop more ethical use of Machine Learning.

7.1
Ethical Development and codes registered into Model Card

One initial general finding regarding the codes in our study is that we identified a very high frequency of remarks concerned with the ethicality of the development process among participants, both from the perspective of ethical and unethical actions. However, these were not equally distributed across our data, with some participants concentrating most of the passages for each of these codes.

Participants 1 and 8 presented especially high frequency of excerpts related to "Ethical Development", in addition to also including high frequency of other codes, such as "System's Autonomy - Decrease" for both of them, and "Guiding Values" in the case of Participant 8. On the other hand, Participant 2 presented an unique case in other ways. His data contained half of our excerpts related to "Unethical Development". It also contained most of the passages we identified as attributing agency to the artifact.

Looking at their answers to our background questions, we found an interesting common trait between these participants. Participants 1, 2, 4 and 8 all had recently taken courses related to semiotic engineering in their graduate studies, over the the previous years. These courses have included in their topics issues like the ones discussed in this work, such as Mitchell et al. (2019) and explanations for AI systems, and even an early draft of Barbosa et al. (2021). It is expected these participants would have more reflections aligned with our background research, due to their previous exposure to them and engagement in discussing them.

Another fact we found that distinguishes these participants, was that while some of these codes tended to be recorded on the Model Card created

during the interviews, others were not. Especially when we consider that there were no passages regarding "Unethical Development" in the Model Cards, while 3 participants included its "Ethical Development" counterpart.

This indicates that, while participants recorded the development decisions they identified as being ethical, decisions they identified as being unethical were not recorded. Even if these decisions were not included in the system, after being identified by participants as leading to negative results, their acknowledgment is important, since communicating them to other developers, and potential users of the system, could allow them to avoid opting into these same decisions and their consequences.

Moreover, these are important when we consider our goal of improving the transparency and accountability of AI systems. For the first, disclosing options considered by the developers, and discarded for being considered unethical, reveals important information about the development and design process and allows other stakeholders to be aware of and also take these considerations into account in their own decisions.

On the accountability topic, we turn back to the definitions given in Section 2.3. Considering the Model Card can be a tool to increase the accountability of AI systems, either inside the organization using/developing it, or to outside stakeholders, it is relevant to document these decisions, so we can account not only for actual consequences of the system, but also for others that were identified and avoided by developers, or what they considered that ought to have been avoided but might be the consequence of other stakeholders' use of the system they developed.

Furthermore, disclosing and communicating this information is desirable and can be a way to expand knowledge about AI systems; it contributes to increasing transparency over the development process and designers' decisions. Recognizing what development options should actively be avoided, due to their unethical nature or negative results, is as important as acknowledging what scenarios were envisioned by designers. One such example, mentioned in Section 6.5, is the decision to not use certain variables as part of the model, or that under certain conditions the model violates certain definitions of fairness.

One explanation for this phenomenon may lie in how we teach and promote the design of software systems. One example of this can be found in the Human AI HAX Tool-Kit, made by Microsoft,[1] which includes the Human-AI Interaction guidelines. These are declared to be a set of tools to create systems involving human and AI "with people in mind from the beginning".

[1]https://www.microsoft.com/en-us/haxtoolkit/

Focusing on the guidelines, all steps include recommendations over what designers should do, and no focus on what should not be done or should be avoided. Starting by their proposed order, both initial steps advise on the communication of what the system can do, and how it performs at these tasks. However, there is no comment on what we know the system cannot do, or in what tasks we have found the system should not be used.

The same applies to the section dedicated to when the system is wrong. These include recommendations regarding facilitating the dismissal and correction of undesired behaviors of the system, but include none regarding the communication of cases where the AI system is known to produce undesirable results, or the mistake would be an expected behavior. This general trend may represent a culture of the area, where we give a lot of attention to positive actions, and not enough to negative or what should we should refrain from doing.

This view is also supported by the fact that, when asked, none of the interviewees expressed they wanted to include any issues or topic they had reflected but not already included in the Model Card, nor that they included something that they deemed as not being appropriate to the tool. In fact, Participants 4 and 6 expressed they believed there should be more space for technical aspects of the artifact they were developing, while Participant 8 stated he thought there could be more space for aspects of the design, such as how the model will be used. On the other hand, Participant 7 expressed he felt compelled to include more topics than he would initially think about.

What we observed was that, even if the Model Card is fit for these negative considerations, our participants decided to not include them. The disclosure of these scenarios is especially relevant when we consider the challenge of model reuse, raised by Brandão et al. (2019, p. 24) and other related work (Hutchinson et al., 2021): such reuse may ignore the social meaning and consequences of the model in use if we do not carry these findings from the model that is being transferred to another scenario.

This contrast is not exclusive of development discussions, and is a common issue in academic publications across a diversity of disciplines, for example, where there is a strong publication bias towards studies that find positive results, thus also biasing what becomes public information. However, with the goal of increasing our knowledge of any area, it is also important to understand and acknowledge what does not work, so we can build upon this and also invest in studies according to previous data points from both sides.

This result is also an indication of the insufficiency of only adhering to guidelines or checklists for considering ethical issues. This is supported by the

view of ethics not only as a procedure, with a list of formalities or checks that ought to be made, but as a continuous reflective process about actions and choices being taken. In other words, ethical concerns should go beyond a merely prescriptive list of actions and results to be achieved or avoided. Guillemin and Gillam (2004) discusses the relevance of these process in the research setting, which we believe highlights the relevance of reflection also in the development setting.

This reflective process is better served through the use of tools that question and stimulate developers to reflect on the system they are developing, and on their choices. This view accentuates the importance of investigating how developers make use of tools like the Model Card, as well as other tools cited in Chapter 3.

7.2 Limitation of AI Autonomy

Another finding of our study was that the absolute majority of participants expressed views on limiting the autonomy of the artifact under development. Even Participant 5, the only one to comment on conditions to increase the autonomy of the system, expressed his distrust over allowing an AI system to act with complete autonomy.

The following quote by Participant 2 expresses well what we found to be the general sentiment of our participants regarding the autonomy of the artifact they were developing: “Caveats and Recommendations regarding possible uses of the model... Have a person to evaluate, do not trust a computer. [...]” (47). This is an interesting contrast to the reaction of participants in Brandão et al. (2019), who initially expressed a high degree of trust in the artifact they were developing in the scenario presented, in the MNIST dataset they were supposed to use, and the evaluation metrics they had to evaluate the performance of their model.

The sentiment expressed by participants in our study is more aligned to what was seen in the study after the researchers explicitly prompted the possible impact, and social context for the system, in a second interview. At this phase, participants started recognizing the social aspect of the algorithm they were supposed to develop: how the decision process could be interpreted, how it could impact society, and how they ought to communicate with affected individuals.

This is a significant difference between the two findings. While participants in Brandão et al. (2019) initially expressed their trust in the artifact they were developing, and its role in the decision-making process, participants in

our study opted to preserve the autonomy of the individuals, aligned with the bioethical principle of autonomy, usually also expressing some concern about unfair results. We believe this may be caused by two different factors. The first one is the possibility that the use of the tool, and our initial questions, already shifted their focus to these issues, making ethical considerations more salient to them, even if not explicitly indicated by us during the scenario.

The second one is the difference in participants' background between both studies. While the interviewees in Brandão et al. (2019) were part of a research center in a company, participants in our study were graduate students in a department that has been offering courses that cover these issues. Moreover, only one of their participants acknowledged algorithmic bias, or other related issues, as a topic of study, whereas half of our eight participants had engaged with research on these topics previously, as mentioned in Section 7.1.

7.3 Third Person and Artifact Mediation

In addition to expressing their view that the autonomy of the artifact they were developing should be limited, in many cases participants expressed this should be done by having another person validate decisions. This person, encumbered with supervising the AI system's output, would therefore be responsible for the final result of the decision process.

We have this finding in common to what was observed by Brandão et al. (2019) at the second stage of their interview. After analyzing what was said in all steps of their interview, they stated "[w]e see that participants said they would rely on team members, project managers, or someone to help them deal with social meaning considerations that necessarily arise when developing DL-based technology for applications like BackSys". This was the second mediation challenge identified in their research.

This can also be observed, for instance, in Participant 2's quoted comment above, or in Participant 8's excerpt 160, when he said the system should include a way to "[...] go through someone's manual monitoring.". We can interpret this as participants transferring part of the responsibility for the result of the system created. Despite being the ones responsible for its development, seven of our participants had at least a comment in this regard.

Another aspect of this relation was the fact that we found a lot of sentences structured around the first person in the interviews, which highlights the role of the designer in shaping the artifact in the scenario. However, there was only one instance participants expressed themselves this way in the Model Card itself: in the "Intended Uses" section Participant 1 wrote: "I receive the

previous achievements from a candidate, split it into three dimensions, for each achievement I attribute a score (based on the knowledge extracted from the model)[...]".

In other instances, the text was written either using the passive voice, or with the model as the subject. For instance, Participant 3 wrote into his "Ethical Considerations", in excerpt 50: "The model will treat all clients equally, considering only economic factors.". Participant 8 wrote in "Caveat and Recomendations", excerpt 143: "It is recommended to have a step for manual revision/monitoring/evaluation".

We did not instruct participants to write their considerations in the Model Card in any specific way, since none was specified by Mitchell et al. (2019). We informed them it consisted of a document with the goal of documenting the AI model developed. However, the fact that there was no use of first person –even for participants who made the first interview with the Extended Metacommunication Template (EMT), which explicitly requires the use of the first person –is a bit surprising.[2]

Moreover, it is desirable that the tools offered contribute to the goal of highlighting ethical relations and consequences that arise from the AI system under development. One of the virtues of the first person structure is exactly that it emphasizes both the role played by developers through the system (first person), and also the user of the system (second person).

Another downside of the results we identified in the model card is that this language de-personifies users and other individuals affected by the algorithm. An evidence of this is the fact that, despite the high frequency of the code of "Impacted Individuals" in our overall data, it only occurred once in the Model Cards. Thus, despite being a common thread in the interviews, it was rarely transposed to the documents with the same focus.

[2]Due to how we structured interviews, participants 3, 4, 7 and 8 had their first interview using the EMT and the second one using the Model Card. This is discussed in Chapter 4.

8
Conclusion and Future Work

In this dissertation, we conducted a qualitative analysis of the data collected through an interview-based study with eight participants. We included in our study two tools proposed by the literature to aid on the development of Machine Learning models: the Model Card and Extended Metacommunication Template. We have presented our results related to participants use of the Model Card, and focus on their ethical reflection and how that was documented.

Our research contributes to increasing the transparency of ML models through the improvement of documentation about the development process and the models themselves. This can be done through internal changes to the development process, by aiding and informing the ethical reasoning the responsible team, and/or externally, by making relevant information public to other stakeholders.

Our work is a first step that can be used to lead to a series of broader studies into the ethical side of the development of Artificial Intelligence systems. We intended to understand how Model Cards contribute to developers' ethical reasoning, and what ethical issues it helps identify.

One of our main findings was a contrast between what our participants opted to include and not include in the Model Card. We found that, while participants evaluated whether certain decisions they took during the development was ethical or not, they would only reflect that decision in the Model Card when they considered it ethical.

We believe this contrast is related to a general culture where we focus on what our systems should do, or how it should be used, but usually not on what should not be done or scenarios where we believe our system should not be used. However, both types of information are equally desirable to be recorded, with the goal of expanding our knowledge about AI systems and promoting a fairer use of such systems, including the reuse and repurposing of these algorithms, as identified by Brandão et al. (2019, p.24).

Furthermore, we found evidence that reinforces the relevance of algorithm mediation, as pointed by Brandão et al. (2019). While participants in our study were aware of the potential impact and meaning their artifact could

have, they also appealed to a third person to mediate their system. This was especially the case when they expressed the model they were building should not have autonomy over the decision process, and that a third person should be responsible for validating or checking the output.

While our methodology does not allow us to determine what causes this difference, we identified three factors that may have influenced this finding. The first is the background of some participants, and familiarity with some of the research cited in this work. The second is that our initial questions and summary of the bioethical principles (even in their original context) may have directed participants questions to these issues. Finally, the Model Card itself, asking for potential ethical issues related to the system, may play a part in this contrast.

Future work should focus on analyzing the remaining data we collected, and identifying the similarities and differences in ethical reflection motivated by the Model Cards and the Extended Metacommunication Template, *i.e.*, whether and to what extent our findings are reproduced in an analysis of the use of the Extended Metacommunication Template. Another relevant inquiry is how different stakeholders may interact with each of these tools, and whether they can attribute relevant meaning to the information included in each of them, especially those without a technical background.

Bibliography

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., and Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6:1–6:13.

Barbosa, S. D. J., Barbosa, G. D. J., de Souza, C. S., 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.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *SSRN Electronic Journal*.

Bathaee, Y. (2017). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology (Harvard JOLT)*, 31(2):889–938.

Beauchamp, T. L. and Childress, J. F. (2019). *Principles of Biomedical Ethics*. Oxford University Press, New York, 8th edition edition.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv:1810.01943 [cs]*.

Bender, E. M. and Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of Machine Learning Research 81*, pages 149–159.

Brandenburg, G. (2011). Dcps administrators won't or can't give a dcps teacher the impact value-added.

Brandão, R., Carbonera, J., de Souza, C., Ferreira, J., Gonçalves, B., and Leitão, C. (2019). Mediation Challenges and Socio-Technical Gaps for Explainable Deep Learning Applications. *arXiv:1907.07178 [cs]*. arXiv: 1907.07178.

Braun, V. and Clarke, V. (2012). Thematic analysis. In Cooper, H., Camic, P. M., Long, D. L., Panter, A. T., Rindskopf, D., and Sher, K. J., editors, *APA handbook of research methods in psychology*, volume 2, chapter 4. American Psychological Association, Washington, DC, US.

Campbell, J. L., Quincy, C., Osserman, J., and Pedersen, O. K. (2013). Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement. *Sociological Methods & Research*, 42(3):294–320.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Coeckelbergh, M. (2020). *AI Ethics*. The MIT Press, Cambridge, MA.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023.

De Souza, C. S. (2005). *The semiotic engineering of human-computer interaction*. MIT press.

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.

Desai, D. R. and Kroll, J. A. (2017). Trust but verify: A guide to algorithms and the law. *Harvard Journal of Law & Technology (Harvard JOLT)*, 31(1):1–64.

Floridi, L. and Cowls, J. (2019). A unified framework of five principles for ai in society. *Harvard Data Science Review*, 1(1).

Galhotra, S., Brun, Y., and Meliou, A. (2017). Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering - ESEC/FSE 2017*, page 498–510. ACM Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2020). Datasheets for datasets. *arXiv:1803.09010 [cs]*. arXiv: 1803.09010.

Guillemin, M. and Gillam, L. (2004). Ethics, reflexivity, and "ethically important moments" in research. *Qualitative Inquiry*, 10(2):261–280.

Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. (2019). A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 181–190. Association for Computing Machinery.

Hind, M., Houde, S., Martino, J., Mojsilovic, A., Piorkowski, D., Richards, J., and Varshney, K. R. (2020). Experiences with Improving the Transparency of AI Models and Services. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–8, New York, NY, USA. Association for Computing Machinery.

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.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. (2021). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 560–575, New York, NY, USA. Association for Computing Machinery.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174.

Kroll, J., Huey, J., Barocas, S., Felten, E., Reidenberg, J., Robinson, D., and Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3):633.

Loukides, H., Mason, M., and Patil, D. (2018). Of oaths and checklists. https://www.oreilly.com/radar/of-oaths-and-checklists/.

Miceli, M., Yang, T., Naudts, L., Schuessler, M., Serbanescu, D., and Hanna, A. (2021). Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 161–172, New York, NY, USA. Association for Computing Machinery.

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. event-place: Atlanta, GA, USA.

Mittelstadt, B., Russell, C., and Wachter, S. (2019). Explaining explanations in ai. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 279–288. Association for Computing Machinery.

O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 33–44, New York, NY, USA. Association for Computing Machinery.

Rawls, J. (1971). *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, Massachussets, 1 edition.

Richards, J., Piorkowski, D., Hind, M., Houde, S., and Mojsilović, A. (2020). A Methodology for Creating AI FactSheets. *arXiv:2006.13796 [cs]*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(55):206–215.

Saldaña, J. (2009). *The Coding Manual for Qualitative Researchers*. The Coding Manual for Qualitative Researchers. Sage Publications Ltd, Thousand Oaks, CA.

Seck, I., Dahmane, K., Duthon, P., and Loosli, G. (2018). Baselines and a datasheet for the Cerema AWP dataset. *arXiv:1806.04016 [cs, stat]*.

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.

Sunstein, C. R. (2019). Algorithms, correcting biases. *Social Research: An International Quarterly*, 86(2):499–511.

Swartout, W. R. and Moore, J. D. (1993). Explanation in second generation expert systems. In David, J.-M., Krivine, J.-P., and Simmons, R., editors, *Second Generation Expert Systems*, page 543–585. Springer Berlin Heidelberg.

Wachter, S., Mittelstadt, B., and Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2):76–99.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., and Wilson, J. (2019). The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1.

A
Coded Excerpts

Table A.1: Final set of coded excerpts information.

	source	text	section	code
1	P1 - MC Doc	O candidato teria três notas, uma para cada dimensão (publicações acadêmicas, histórico profissional e formação acadêmica) e a nota final seria uma média ponderada das três, em que o peso seria um valor indicado pelo comitê de seleção — imagino que o fator acadêmico seja mais relevante que o profissional.	detalhes do modelo	other
2	P1 - MC Doc	em que o peso seria um valor indicado pelo comitê de seleção	detalhes do modelo	system's autonomy - decrease
3	P1 - MC Doc	O objetivo dessas métricas seria identificar possíveis indicativos de bias no comportamento do sistema.	metricas	framing based on element of ethical framework - justice
4	P1 - MC Doc	Não há uso de dados que permitam a identificação direta de características como, gênero, cor da pele, religião, renda, endereço ou qualquer outra informação que não esteja relacionada às realizações do candidato.	consideracoes eticas	ethics of development process - ethical development
5	P1 - MC Doc	O fator origem do candidato (de qual estado ou cidade) poderá ser inferida com base nas informações disponibilizadas. Outras não informações não mapeadas poderão ser inferidas, é necessário estar atento ao uso indevido deste dataset.	consideracoes eticas	framing based on element of ethical framework - non-maleficence
6	P1 - MC Doc	O fator origem do candidato (de qual estado ou cidade) poderá ser inferida com base nas informações disponibilizadas. Outras não informações não mapeadas poderão ser inferidas, é necessário estar atento ao uso indevido deste dataset.	consideracoes eticas	undesired consequences

7	P1 - Model Card	Sendo que aqui eu tentei evitar aquelas informações que possam ser... possam ajudar na identificação, que possam ser consideradas sensíveis, do tipo idade, raça, cor da pele, essas coisas que possam depois gerar alguma polêmica. Principalmente religião e inclusive gênero.	dados de treino	ethics of development process - ethical development
8	P1 - Model Card	Entendi. Então a distribuição de notas por... É, uma das principais questões quando você fala da contratação de pessoas é que existe uma... um problema histórico que é sempre o homem é sempre mais preterido (parece que deveria ser preferido) que a mulher né	metricas	diagnosis of element of ethical framework - justice
9	P1 - Model Card	Então, o problema do gênero quando se fala do mercado de trabalho, ele é extremamente forte	metricas	diagnosis of element of ethical framework - justice
10	P1 - Model Card	Ah, mas aí poderia extrapolar, por grupos de candidatos	metricas	other
11	P1 - Model Card	Notas de grupos do candidato por gênero. Como é que você consegue ser, garantir que não tá tendo segregação, sem segregar. Porque toda vez que eu divido um grupo de candidatos por gênero, de alguma forma eu to segregando. Porque um outro fator aqui seria por gênero, cor	metricas	ethics of development process - ethical development
12	P1 - Model Card	Cor da pele é uma coisa que varia muito, só que eu acho... ah... não sei se é correto pedir essa informação sobre a cor.	metricas	ethics of development process - unethical development
13	P1 - Model Card	Mas é... (escrevendo) verificaria. É mas cor não é uma informação que eu iria colocar, acho que isso aqui já é um primeiro passo. Ah, cor, idade talvez... Não, mas a academia mais experiente, mais velho a pessoa é, então não faz muito sentido.	metricas	ethics of development process - ethical development
14	P1 - Model Card	(Lendo um ponto em voz alta) Em um segundo momento, as notas seriam emitidas e um comitê (apagado) faria a atribuição manual... o cálculo da nota final.	dados de avaliacao	system's autonomy - decrease

15	P1 - Model Card	Tá considerações éticas é que não há dado que permita uma identificação direta de características não relacionadas ao candidato. Até destacar aqui que é identificação direta, porque a origem do candidato, de onde ele vem, poderá ser identificada com base nas informações disponibilizadas. Se ele fez graduação, mestrado, doutorado provavelmente ele é daquele lugar e mora lá a vida toda. O que pode permitir cenários de xenofobia, coisas do tipo	consideracoes eticas	framing based on element of ethical framework - non-maleficence
16	P1 - Model Card	Outras informações não mapeadas poderão ser... poderão ser inferidas, é necessário (digitando apenas). Não to conseguindo explicar isso aqui muito bem, mas é o seguinte que esse dataset pensando la q ele tenha as realizações do candidato, as publicações então outras informações elas podem ser inferidas. Por exemplo, onde é que ele publica, quais são os principais canais que ele publica, dai se eu tirar algumas outras informações que podem ser utilizadas pra perfilar esse usuário, identificar interesses e gostos deles. Isso não seria necessário pra essa seleção, mas poderia ser uma coisa.... é um produto possível olhando pra esses dados, então (inaudível) certo cuidado.	consideracoes eticas	undesired consequences
17	P1 - Model Card	Teriam três notas uma para cada dimensão, a nota final seria uma média ponderada das 3 notas em que o peso seria um valor indicado pelo comitê de seleção.	revisao e perguntas	other
18	P1 - Model Card	o peso seria um valor indicado pelo comitê de seleção	revisao e perguntas	system's autonomy - decrease
19	P1 - Model Card	Nessa parte das considerações éticas acho que vários fatores. É... porque assim, primeira coisa que eu pensei foi no dataset, como é que eu poderia construir ele. Tá a segunda é minhas preocupações é evitar cenários em que tenha presença de bias e e que haja discriminação ou alguma forma de segregação de candidatos.	revisao e perguntas	responsibility for artifact

20	P1 - Model Card	Tá, um órgão regulador... talvez não um órgão regulador, talvez um... um sistema de, alguém que fosse auditar o sistema, ela seria impactada porque uma das coisas que eu coloquei aqui é que o sistema não só daria nota, mas também o resultado da avaliação	revisao e perguntas	impacted individuals
21	P1 - Model Card	Ah pensei que ele iria... voltaria para o comitê para uma avaliação, com um resultado de avaliação e essa avaliação, e provavelmente iria, ele poderia receber um... as avaliações ranqueadas, do melhor avaliado até o menos avaliado, até a menor nota. Mas esse rank, essa avaliação, o ideal seria uma tabela, uma tabela com todas as notas.	revisao e perguntas	system's autonomy - decrease
22	P2 - MC Doc	o que faz com que o modelo tenha uma performance melhor no cálculo de score para pessoas idosas.	metodos analise quantitativa	framing based on element of ethical framework - justice
23	P2 - MC Doc	Assumindo que a grande maioria das pessoas que não honraram seus empréstimos no dataset de treinamento foram pessoas com baixa escolaridade, o modelo tenderá a dar baixo escores para pessoas mais pobres, visto que esse grupo em geral possui baixa escolaridade.	consideracoes eticas	undesired consequences
24	P2 - MC Doc	Existe a chance do modelo ao utilizar o endereço como atributo de treinamento, tenha a tendência de apresentar padrões preconceituosos a respeito de determinadas localidades.	consideracoes eticas	diagnosis of element of ethical framework - justice
25	P2 - MC Doc	Não é recomendado o uso do modelo em um processo totalmente automatizado de análise de crédito, visto que este tem o potencial de apresentar comportamentos inesperados e inadequados em alguns casos. Recomenda-se o uso do score dado pelo modelo apenas como um parâmetro adicional para análise de crédito feita por um ser humano.	cuidados e recomenda-coes	system's autonomy - decrease
26	P2 - Model Card	O desbalanceamento do dataset pode afetar, né	fatores	other
27	P2 - Model Card	Que informações eu ia pedir? "dois pontos... idade..." o sexo? será que sexo influencia na análise de crédito? o endereço... (voltando) acho que sexo não faz sentido.	dados de avaliacao	ethics of development process - unethical development

28	P2 - Model Card	Calma aí, profissão é importante. Profissão, escolaridade. Aí eu posso ser um pouto antiético e pedir o endereço. Já vou levantar uma questão ética aqui. Vou botar lá, endereço.	dados de avaliacao	ethics of development process - unethical development
29	P2 - Model Card	Sim, é porque eu falei aqui que tem modelos que podem prejudicar o modelo[destacando fatores].	metodos analise quantitativa	other
30	P2 - Model Card	Aham... Calma aí, deixa eu pensar um pouquinho. Tipo, eu vou pegar... vê se eu entendi direito. Por exemplo, eu falei que o desbalanceamento dos dados históricos dos clientes pode afetar a performance do modelo. Por exemplo, tenho muito mais idosos na minha base. Então meu modelo tem um viés para, vai acertar muito mais pra idosos do que pra não idosos.	metodos analise quantitativa	framing based on element of ethical framework - justice
31	P2 - Model Card	o que faz com que o modelo tenha uma performance melhor no cálculo para pessoas idosas do que para jovens	metodos analise quantitativa	framing based on element of ethical framework - justice
32	P2 - Model Card	É tem uma questão, eu sempre tenho uma questão ética sobre esse lance dos dados na seguinte hipótese, por exemplo digamos que meu modelo, os dados né, o modelo só vai refletir o padrão que ele encontrou nos dados.	consideracoes eticas	other
33	P2 - Model Card	Suponhamos que os dados, a maioria das pessoas com baixa escolaridade dão calote, sabe. O grupo lá no cluster de caloteiras, a galera com baixa escolaridade domina né, então acaba que o meu modelo vai ter um viés de oferecer melhores créditos, ou oferecer o crédito mais pra pessoas com alta escolaridade. Então ele meio que vai limar essa galera aí dos empréstimos	consideracoes eticas	framing based on element of ethical framework - justice
34	P2 - Model Card	O grupo lá no cluster de caloteiras, a galera com baixa escolaridade domina né, então acaba que o meu modelo vai ter um viés de oferecer melhores créditos, ou oferecer o crédito mais pra pessoas com alta escolaridade. Então ele meio que vai limar essa galera aí dos empréstimos.	consideracoes eticas	undesired consequences
35	P2 - Model Card	Mas foi só uma constatação que o modelo enxergou nos dados, não é culpa do modelo em teoria, é culpa do dado que tá ali.	consideracoes eticas	agency to artifact

36	P2 - Model Card	Mas até que ponto, sei lá, é antiético da empresa fazer isso, sabe?	consideracoes eticas	ethics of development process - unethical development
37	P2 - Model Card	se ela for maximizar o lucro ela vai deixar a galera de baixa escolaridade de fora, porque sabe que vai ter mais prejuízo ali. E a galera de baixa escolaridade, no geral, são pessoas mais humildes. Então acaba que a galera mais humilde vai ficar sem crédito, do que a galera mais abastada. É uma discussão ética.	consideracoes eticas	diagnosis of element of ethical framework - justice
38	P2 - Model Card	se ela for maximizar o lucro ela vai deixar a galera de baixa escolaridade de fora, porque sabe que vai ter mais prejuízo ali. E a galera de baixa escolaridade, no geral, são pessoas mais humildes. Então acaba que a galera mais humilde vai ficar sem crédito, do que a galera mais abastada. É uma discussão ética.	consideracoes eticas	diagnosis of element of ethical framework - justice
39	P2 - Model Card	É não deixar de ser, mas aquela situação, não é culpa do modelo, é culpa do raio do dado que enfiaram dentro do modelo tadinho.	consideracoes eticas	agency to artifact
40	P2 - Model Card	Mas vamos lá, vou botar... "assumindo que a grande maioria das pessoas que não pagaram seus (reescrevendo) honraram, pra ficar bonito, honraram seus empréstimos foram (escrevendo anteriormente: no dataset de treinamento) pessoas com baixa escolaridade, o modelo tenderá a dar baixo escores para pessoas com... mais pobres, visto que esse grupo em geral possui baixa escolaridade"	consideracoes eticas	undesired consequences
41	P2 - Model Card	É... Com base no endereço tem a questão... "Existe a chance de o modelo ao utilizar o endereço como atributo de treinamento, tenha a tendência de...	consideracoes eticas	ethics of development process - unethical development
42	P2 - Model Card	É... Com base no endereço tem a questão... "Existe a chance de o modelo ao utilizar o endereço como atributo de treinamento, tenha a tendência de...	consideracoes eticas	diagnosis of element of ethical framework - justice
43	P2 - Model Card	O modelo tem como ser preconceituoso?	consideracoes eticas	agency to artifact

44	P2 - Model Card	Eu ia botar preconceituoso, mas eu achei muito forte. O modelo não é preconceituoso. "a tendência do modelo apresentar padrões preconceituosos a respeito de determinadas localidades"	consideracoes eticas	agency to artifact
45	P2 - Model Card	O modelo não é preconceituoso.	consideracoes eticas	agency to artifact
46	P2 - Model Card	Cuidados e recomendações sobre os possíveis usos do modelo... Bota uma pessoa pra avaliar, não acredita no computador.	cuidados e recomenda-coes	system's autonomy - decrease
47	P2 - Model Card	Mas eu incluindo ele aqui eu já posso meio que estigmatizar uma certa área da cidade que, pode ser sei lá, uma área comunidade, é uma favela, que por ter maioria negra eu vou acabar excluindo os negros da análise de crédito, só por causa disso.	revisao e perguntas	ethics of development process - unethical development
48	P3 - MC Doc	O modelo utilizado será do tipo supervisionado para prever o risco de cada cliente relacionado ao empréstimo.	detalhes do modelo	other
49	P3 - MC Doc	O modelo ajudará o processo de decisão da instituição apresentando um valor referente ao risco de cada cliente, não devendo este decidir se acontecerá o empréstimo ou não.	consideracoes eticas	system's autonomy - decrease
50	P3 - MC Doc	O modelo tratará todos os clientes de maneira igual, levando em consideração apenas fatores econômicos	consideracoes eticas	guiding values
51	P3 - MC Doc	O modelo não deve ser usado de maneira discriminatória	cuidados e recomenda-coes	guiding values
52	P3 - MC Doc	O modelo não deve ser usado de maneira discriminatória (inserindo novas características para treino que não tratariam os clientes de maneira justa).	cuidados e recomenda-coes	diagnosis of element of ethical framework - justice
53	P3 - Model Card	Aí ele iria colocar esses dados no modelo pra assim obter uma resposta que o ajudasse na decisão	usos pretendidos	framing based on element of ethical framework - beneficence
54	P3 - Model Card	O modelo poderia julgar que todas as pessoas, tipo assim. É bem um pouco fora da realidade, mas é só pra dizer, ele poderia julgar que todas as pessoas com, naquela faixa etária, não teriam, tipo assim.... Como é que eu explico isso? Tipo assim, é porque existe uma feature que não é tão relevante, mas que ela pode tipo, acabar fazendo uma... É, eu estou esquecendo as palavras.	fatores	agency to artifact

55	P3 - Model Card	Eu não sei se tá claro, é só pra falar que não vai levar em consideração o nome da pessoa, sexo, essas coisas assim.	consideracoes eticas	ethics of development process - ethical development
56	P3 - Model Card	Porque se esses dados não estão corretos, eu não sei se entra na minha parte, do princípio da não-maleficência, porque imagina que ela colocou dados incorretos, e acaba que um empréstimo seja aprovado pra alguém que não seria, então....	consideracoes eticas	framing based on element of ethical framework - non-maleficence
57	P3 - Model Card	Ela poderia responsabilizar o modelo por ter, mas os dados estavam incorretos	consideracoes eticas	agency to artifact
58	P3 - Model Card	(Sobre primeiro tópico em Cuidados e Recomendações, sobre não usar o modelo de forma discriminatória, especialmente incluindo novas variáveis) O que eu estou querendo dizer aqui, porque é claro que ele não deve ser usado.	cuidados e recomenda-coes	ethics of development process - unethical development
59	P3 - Model Card	Então é, me influenciou mais nessa parte dos fatores porque eu sei que, tipo assim, eu fui me lembrar de algumas coisas e eu sei que algumas empresas levam em consideração a idade, por exemplo, só que tipo, quando eu tava fazendo eu não fiquei confortável em colocar a idade. Quando eu tava nessa parte aqui[Métodos de Análise Quantitativa] eu até tirei, porque eu vi que não era uma coisa... justa, assim. Tipo, não é uma coisa...	revisao e perguntas	responsibility for artifact
60	P3 - Model Card	Quando eu tava nessa parte aqui[Métodos de Análise Quantitativa] eu até tirei, porque eu vi que não era uma coisa... justa, assim. Tipo, não é uma coisa...	revisao e perguntas	ethics of development process - unethical development
61	P3 - Model Card	Exatamente. Então tipo, eu acho que poderia gerar uma certa, assim, que querendo ou não, talvez pras empresas significa alguma coisa, mas eu não sei se acabaria sendo algo que não discriminaria, sabe. Então...	revisao e perguntas	diagnosis of element of ethical framework - justice
62	P3 - Model Card	Essa parte dos métodos de análise quantitativa, e... Hmm.. aqui em usos pretendidos, porque eu me lembrei, tipo assim, dessa parte de autonomia né, onde que eles próprios né (inaudível). Ai eu coloquei... eu me lembrei, aí eu botei nessa parte	revisao e perguntas	scaffolding around element of ethical framework - autonomy

63	P3 - Model Card	Muito importantes, realmente eu fiquei pensando se eu poderia inserir eles em cada coisa.	revisao e perguntas	scaffolding around element of ethical framework
64	P3 - Model Card	Eu diria que ambos são cenários delicados sabe, por ter que avaliar questão de pessoas, sabe. Então... eles são parecidos nesse aspecto.	revisao e perguntas	impacted individuals
65	P4 - MC Doc	O resultado da seleção não deve depender somente da nota do modelo	usos pretendidos	system's autonomy - decrease
66	P4 - MC Doc	Todos os alunos em circunstâncias semelhantes devem receber de forma igual as notas atribuídas pelo modelo.	consideracoes eticas	guiding values
67	P4 - MC Doc	O modelo pode não refletir o momento presente do aluno	consideracoes eticas	other
68	P4 - Model Card	Na verdade, seriam os usuários do sistema.	detalhes do modelo	impacted individuals
69	P4 - Model Card	Outro cenário que eu quero prever, é que tá. Aqui, na verdade o cenário ta me dizendo que... será utilizado pelo comitê para decidir quais estudantes são admitidos ou não. Eu quero enfatizar nesse meu próximo ponto que a nota do modelo não deveria ser o único ponto de decisão. Deixa eu ver como eu posso escrever "o resultado do modelo...." Na verdade, o resultado da seleção aqui eu to falando do modelo, né. "O resultado da seleção não deve depender somente da nota do modelo". Tá.	usos pretendidos	system's autonomy - decrease
70	P4 - Model Card	Acredito eu que, com alguma justiça isso vai ser alcançado. mas eu não consigo prever, por exemplo, como é que tá o presente desse aluno	usos pretendidos	diagnosis of element of ethical framework - justice
71	P4 - Model Card	Então, eu não consigo embutir essa informação que eu acho justo no modelo, por favor não uso o modelo como o único ponto de decisão.	usos pretendidos	system's autonomy - decrease
72	P4 - Model Card	Não assim, mesmo que ele usasse todos as notas, históricos e depoimentos de professores. Usasse todo tipo de dado né. Eu imagino esse modelo que eu to montando basicamente como um cara que ia chacoalhar esse histórico e ia me dar uma nota né. Ai eu to pensando: eu preciso mesmo do modelo?	metricas	agency to artifact

73	P4 - Model Card	Cara, eu penso muito nesse aqui ó "justiça". Já que eu to tomando uma decisão sobre vida de pessoas. Ó, deixa eu ver se eu consigo trabalhar com isso aqui. Lendo descrição: "Todos os pacientes em circunstâncias semelhantes devem receber de forma igual o melhor tratamento possível." Agora eu tenho medo cara, desse meu sistema não ser capaz de cumprir isso por um motivo. Possa ser que o histórico escolar das pessoas, sejam iguais, notas, revisão de professores, e cartas de recomendação. Mas vai ter uma coisa que eu não vou estar levando em consideração, ou pelo menos de início eu não estava pensando nisso, em dados socioeconômicos das pessoas. A gente sabe que existe N formas de fazer o ensino médio e muitas que são oferecidas no ensino público nem se compara com o que é oferecido no ensino particular, né. A gente sabe que tem grupos sociais que são atingidos mais com esse... com essa disparidade, indígena, quilombola, e N outros que até o Ministério da Educação cita. Aí eu não sei, cara, se só histórico de notas incluiria esse background.	consideracoes eticas	scaffolding around element of ethical framework - justice
74	P4 - Model Card	Agora eu tenho medo cara, desse meu sistema não ser capaz de cumprir isso por um motivo	consideracoes eticas	responsibility for artifact
75	P4 - Model Card	Se for usar só notas, pode ser que pessoas que tem desvantagem continuam em desvantagem né, e eu quero que de alguma forma esse meu modelo, então....	consideracoes eticas	diagnosis of element of ethical framework - justice
76	P4 - Model Card	mesmo que um aluno de um grupo que é historicamente desvantagem, tem desvantagem histórica, ele pode ter uma nota menor, mas o meu modelo teria que levar em consideração isso né	consideracoes eticas	undesired consequences

77	P4 - Model Card	Então, quando eu li aqui o cenário, né, que nem eu te falei. Eu considerei esse realizações anteriores, que o modelo não leva em consideração, pelo menos o pessoal que me pediu né, da universidade, não levou em consideração esse histórico social. Ai na parte de considerações éticas eu to dizendo, que o modelo que me pediram não leva essas considerações, mas que eu como desenvolvedor queria levar essas considerações, ai expliquei o motivo.	consideracoes eticas	ethics of development process - ethical development
78	P4 - Model Card	Aqui eu botei uma consideração meio óbvia, que o modelo pode não refletir o estado presente do aluno. Eu posso ser um aluno horrível durante os 3 anos do ensino médio, e fazer um vestibular fantástico e entrar para a melhor universidade que tem. Então, meu estado atual era ser um aluno bom, mas no passado não foi o caso. O modelo sempre vai considerar casos do passado né, e... é ruim extrapolar isso, dizer que o aluno se o aluno tem uma nota ruim, chamar ele de um aluno ruim no momento. O máximo que eu posso fazer é dizer que ele foi um aluno ruim no passado.	consideracoes eticas	ethics of development process - unethical development
79	P4 - Model Card	Mas pra um aluno, pra uma seleção ai que envolvem gente, eu tenho que dizer você não passou, e você não passou por isso, ai eu mostro o resultado do meu modelo e explico. Se você tivesse uma nota melhor aqui, uma atividade melhor aqui, um review melhor... ai talvez você passaria.	cuidados e recomenda-coes	framing based on element of ethical framework - beneficence
80	P4 - Model Card	isso aqui vai reduzir o meu leque de possibilidades de algoritmos muito	fatores	undesired consequences
81	P4 - Model Card	Como eu te disse, eu acho que leva a uma reflexão mais pessoa. Você usar eu, responsabilidade fica mais ali na sua mão, você, você vê mais empatia por esse você. Então... pode botar leva a reflexão pessoal e leva a empatia.	revisao e perguntas	responsibility for artifact
82	P5 - MC Doc	Remover as features que não deveriam influenciar na nota de um aluno. Por exemplo, nacionalidade, sexo, cor, etc.	usos pretendidos	ethics of development process - ethical development

83	P5 - MC Doc	Deve buscar-se minimizar todo e qualquer viés racial, ideológico que o modelo pode capturar a partir dos dados, de forma que todo candidato(a) seja avaliado de forma justa.	consideracoes eticas	guiding values
84	P5 - MC Doc	Após a implantação do modelo, eu recomendaria que houvesse uma avaliação comparativa entre as notas inferidas pelo modelo para o processo de admissão corrente x notas dadas por avaliadores antes de uma implementação totalmente automática (que exclui a necessidade de um avaliador). Se existe consistência entre as respostas do modelo x respostas dos avaliadores após algumas rodadas de avaliação, acredito que aí sim, o modelo poderia atuar sem muitas preocupações que dizem respeito se aquele modelo está sendo justo para a admissão de novos candidatos(as).	cuidados e recomenda-coes	system's autonomy - increase
85	P5 - MC Doc	Se existe consistência entre as respostas do modelo x respostas dos avaliadores após algumas rodadas de avaliação, acredito que aí sim, o modelo poderia atuar sem muitas preocupações que dizem respeito se aquele modelo está sendo justo para a admissão de novos candidatos(as).	cuidados e recomenda-coes	diagnosis of element of ethical framework - justice
86	P5 - Model Card	Por exemplo, eu acredito que features como nacionalidade, sexo, cor, qualquer coisa que esteja alheia a capacidade de um aluno, elas não devem entrar em um modelo desse	usos pretendidos	guiding values
87	P5 - Model Card	Então, eu acho que... essa coisa de sempre buscar essas features que estão fora, não deveriam ser levadas em consideração, podem colocar outliers, por exemplo, você tem um cara lá que não apresenta notas boas, não tem muitos cursos, mas ele foi aceito ali no processo com uma nota que estava boa comparada com outros alunos que tinham curso, que tinham notas boas e acabaram não entrando. Então tem um motivo ali que não está explícito nos dados, mas eles estão acabando entrando	fatores	framing based on element of ethical framework - justice
88	P5 - Model Card	Então é importante identificar esses casos, saber quais tipos de dados pra tirar esse cara de lá e dizer: modelo isso aqui não é o certo.	fatores	ethics of development process - ethical development

89	P5 - Model Card	porque de repente você tem uma instância ali, mas aquela nota está muito discrepante do que deveria ser	metodos analise quantitativa	undesired consequences
90	P5 - Model Card	Eu acho que o fator aqui é mais buscar diminuir qualquer viés, seja racial, ideológico que o modelo, que é o que acontece na maioria das vezes, o modelo acaba capturando a partir desses dados	consideracoes eticas	diagnosis of element of ethical framework - justice
91	P5 - Model Card	não é a gente que, asism claro que a gente está responsável por deixar esse dado passar, esse negoço ir a frente, o ideal é que a gente busque minimizar durante a concepção e a implementação desse modelo qualquer viés ai que possa ser evidenciado.	consideracoes eticas	responsibility for artifact
92	P5 - Model Card	Pra que o candidato seja avaliado de forma justa, e que não acabe tendo consequências com coisas que não deveriam entrar no processo de avaliação.	consideracoes eticas	guiding values
93	P5 - Model Card	Aqui (seção de Cuidados e Recomendações) eu botei, aqui foi um pouco que numa etapa pós implementação, eu sempre sou meio assim de colocar uma coisa totalmente autônoma, totalmente automática, eu sempre acho que tem que existir ali... uma vez que a gente desenvolveu o modelo tem que ter uma parceria entre especialistas e o modelo pra verificar se aquilo que o modelo ta inferindo ali faz sentido no mundo real, pra gente não deixar que pessoas sejam prejudicadas.	cuidados e recomenda-coes	diagnosis of element of ethical framework - autonomy
94	P5 - Model Card	eu sempre sou meio assim de colocar uma coisa totalmente autônoma, totalmente automática	cuidados e recomenda-coes	system's autonomy - decrease
95	P5 - Model Card	uma vez que a gente desenvolveu o modelo tem que ter uma parceria entre especialistas e o modelo pra verificar se aquilo que o modelo ta inferindo ali faz sentido no mundo real, pra gente não deixar que pessoas sejam prejudicadas	cuidados e recomenda-coes	undesired consequences
96	P5 - Model Card	Vamos dizer que durante os dois próximos anos a gente vai fazer uma admissão mista. A gente vai pegar as inferências do modelo e inferências também dos avaliadores, ver se tá OK. Se tiver OK quem sabe no próximo ano a gente já não usa só o modelo. Então é nesse sentido.	cuidados e recomenda-coes	system's autonomy - increase

97	P5 - Model Card	Eu acredito que quando você direciona um tópico pra isso, você deixa aberto pra isso, você até começa a pensar no que você tá efetuando ali, o que você tá transcrevendo da sua ideia. Então eu acho que é importante ter esse tópico de consideração ética e das recomendações é super importante, porque faz a gente se questionar sobre o que a gente está transcrevendo se ela reflete ali essa preocupação.	revisao e perguntas	responsibility for artifact
98	P5 - Model Card	É como se você estivesse dando voz pra pessoas, você ta dando voz e sumarizando esse pensamento de pessoas que podem estar sendo afetadas por esse modelo e de coisas que... porque as pessoas que estão desenvolvendo isso nem sempre entende o problema, as vezes você vê só a ponta do iceberg, você nem entende o quão profundo é que ele pode ir. Então eu acho que quanto mais isso tá próximo da pessoa que vai ter um impacto ali na mudança, num aluno que tá passando por um processo ali na universidade, ou numa entrevista de emprego. Eu acredito que ele aponta todos esses fatores que deveriam... dá voz mesmo.	revisao e perguntas	impacted individuals
99	P6 - MC Doc	O modelo para determinado grupo de clientes sempre aprova ou nega o empréstimo.	consideracoes eticas	scaffolding around element of ethical framework - justice
100	P6 - MC Doc	Analisar cuidadosamente as respostas do modelo para um grupo de clientes, buscando sempre minimizar o viés de suas respostas.	cuidados e recomenda-coes	guiding values
101	P6 - MC Doc	Analisar cuidadosamente as respostas do modelo para um grupo de clientes, buscando sempre minimizar o viés de suas respostas. Buscar o perfil e histórico de um cliente em diferentes fontes para tentar minimizar os clientes que agem com má fé.	cuidados e recomenda-coes	framing based on element of ethical framework - non-maleficence
102	P6 - MC Doc	Não tomar o modelo como único recurso para aprovar ou negar um empréstimo. O modelo poderia ser usado para auxiliar na decisão de um funcionário da companhia.	cuidados e recomenda-coes	system's autonomy - decrease
103	P6 - Model Card	o modelo poderia meio que sugerir, não o modelo, mas poderia usar essa informação do modelo	usos pretendidos	agency to artifact

104	P6 - Model Card	mas eles meio que teriam que tá correndo um risco, né, aquele cliente tinha um risco um pouco alto.	usos pretendidos	undesired consequences
105	P6 - Model Card	Isso poderia impactar a performance do modelo	fatores	undesired consequences
106	P6 - Model Card	Aí esse cara, ele pode, meio que imaginando num caso extremo, ele poderia meio que tá com uma nova identidade e não ter esse histórico, e ele ser uma pessoa de alto risco, né. Entendi. 37:42 Aí ele meio que poderia se passar por outra pessoa pra poder aceitar o empréstimo, mas o histórico dele na verdade não era aquilo que tava descrito no dado né. Aham. 37:53 (digitando) talvez não pequeno, mas um.. como é que eu posso dizer que ele foi alterado? Um histórico financeiro pequeno ou... um histórico financeiro maquiado, entendeu? adulterado talvez.	fatores	undesired consequences
107	P6 - Model Card	Aí eu to imaginando um cenário que uma pessoa tem, é..., talvez familiares que tenham um histórico financeiro ruim e isso acabe é... influenciando na resposta do modelo praquela pessoa só por ter algum grau de parentesco, entendeu?	fatores	framing based on element of ethical framework - non-maleficence
108	P6 - Model Card	pessoa tem, é..., talvez familiares que tenham um histórico financeiro ruim e isso acabe é... influenciando na resposta do modelo praquela pessoa	fatores	impacted individuals
109	P6 - Model Card	Imagina que um cliente tem um, talvez, os pais dele possuem um histórico financeiro ruim, mas como ele tá meio que entrando agora, o troço está sendo usado pelos pais pra poder aceitar o empréstimo, entendeu	fatores	impacted individuals
110	P6 - Model Card	Aí imagino que esse parentesco aqui seria algo que poderia ser levado em consideração pelo modelo, que aí é... é um pouco ficar nesse histórico financeiro pequeno, o cliente ta meio que se passando por outra pessoa, pedindo em nome de uma pessoa, né, pra ele poder usar o dinheiro, esse dinheiro do empréstimo.	fatores	diagnosis of element of ethical framework - non-maleficence
111	P6 - Model Card	Aí eu acho que o modelo poderia levar isso em consideração, entendeu?	fatores	other

112	P6 - Model Card	Deixa eu ver se eu consigo imaginar uma forma de ser aprovado no modelo e não ter um histórico bom.	fatores	scaffolding around element of ethical framework - non-maleficence
113	P6 - Model Card	Mas não seria a resposta do modelo em si mesmo. Porque pelo que eu entendi o modelo só vai responder ou aprovou ou nego o empréstimo.	metricas	other
114	P6 - Model Card	Aqui, é baseado naquele mesmo sentido, pro meu modelo não aprender a escolher so no caso ideal né	dados de treino	undesired consequences
115	P6 - Model Card	No caso, aqui seria (digitando item) definir "score" de risco para cada cliente analisando seu histórico (reescrevendo), seu perfil e histórico financeiro bem como seus respectivos parentescos ou proximidades. Ai aqui no caso, eu taria meio que definindo, por exemplo, o score de uma pessoa com base no perfil e histórico dela e também das pessoas que estão em sua volta. Nos colegas de trabalho, na relação direta de parentesco como marido, esposa, nesse sentido né. Amigos... amigos não sei.	metodos analise quantitativa	impacted individuals
116	P6 - Model Card	Deixa eu voltar aqui pros princípios e ver o que...	consideracoes eticas	scaffolding around element of ethical framework
117	P6 - Model Card	No caso aqui, por exemplo, imaginando que o modelo, talvez pra determinado tipo de, não tipo de pessoa, mas pra determinado perfil de usuário, ele sempre aceita o empréstimo. Só que por conta talvez de uma variável. Isso seria... eu não sei seu to conseguindo me expressar, mas por exemplo.	consideracoes eticas	scaffolding around element of ethical framework - justice
118	P6 - Model Card	Não se isso tá refletindo o que eu falei, mas alguma coisa nesse sentido, o modelo tá meio que enviesado praquele grupo, ou praquela bolha de pessoas. Seria algo que deveria ter que tomar cuidado.	consideracoes eticas	undesired consequences
119	P6 - Model Card	Então meio que tentar evitar o bias do modelo.	cuidados e recomenda-coes	guiding values
120	P6 - Model Card	Então teria que tomar esse cuidado talvez, como eu posso escrever isso. (escrevendo) Analisar cuidadosamente as respostas do modelo para um grupo de clientes, buscando sempre minimizar o bias, o viés, né, de suas respostas.	cuidados e recomenda-coes	framing based on element of ethical framework - justice

121	P6 - Model Card	Não sei se (inaudível) então (escrevendo segundo ponto) buscar o perfil e histórico de um cliente em diferentes plataformas, não sei se plataformas, em diferentes fontes para tentar minimizar os clientes que agem com má fé, vou colocar nesse sentido que é... que agem, ah falta o com. Aqui no caso eu taria meio que tentando ter uma base maior, um conjunto de dados maior pra cada cliente, pra tentar minimizar aqueles clientes que tentam mascarar de alguma forma o seu histórico né, seja com, pedindo algum familiar pra fazer o empréstimo, ou ele mesmo maquiando o seu histórico pra parecer bem pra empresa.	cuidados e recomenda-coes	framing based on element of ethical framework - non-maleficence
122	P6 - Model Card	Aí eu buscar essas informações sobre aquele cliente em diferentes fontes poderia ajudar nesse sentido.	cuidados e recomenda-coes	ethics of development process - ethical development
123	P6 - Model Card	Aí seria, um cuidado tipo, em não tomar o modelo como único recurso para aprovar ou não empréstimo né. Aí no caso, calma aí (alterando texto) tipo, o modelo não seria utilizado como um único recurso pra aprovar ou negar o empréstimo, o modelo seria mais usado pra auxiliar a tomada de decisão de um funcionário, entendeu.	cuidados e recomenda-coes	system's autonomy - decrease
124	P6 - Model Card	Exatamente, talvez que o modelo poderia ser usado para auxiliar na decisão de um funcionário no caso	cuidados e recomenda-coes	system's autonomy - decrease
125	P6 - Model Card	Aí poderia ter uma satisfação melhor pro cliente né. O cliente poderia sair um pouco mais satisfeito, tendo a intermediação de um funcionário,	cuidados e recomenda-coes	diagnosis of element of ethical framework - beneficence
126	P6 - Model Card	Ah, sim. No caso dos fatores que poderiam impactar a performance do meu modelo né, que eu pude imaginar diferentes clientes, né, que poderiam tentar parecer bem para o modelo. Poderiam meio que tentar maquear seu histórico né. Então acho que essa seção ajudou bastante. Aí imaginar diferentes perfis, não perfis de usuário, mas diferentes perfis de clietnes que poderiam meio que estar tentando burlar o sistema, né.	revisao e perguntas	impacted individuals
127	P7 - Model Card	Ai foi um... uma opção de sei lá, joguei a moeda aqui. Pode ser um modelo pessoa física ou pessoa jurídica né. Ai deixei no caso voltado pra pessoa jurídica.	usos pretendidos	impacted individuals

128	P7 - Model Card	E se o cara já deu default pra uma determinada instituição ela não vai... é difícil emprestar de novo. Quer dizer, precisa primeiro quitar, depois resolver né. Mas enfim.... pode conseguir de forma indireta, quer dizer.	dados de treino	undesired consequences
129	P7 - Model Card	Então, muitas vezes pro consumidor final você não pode usar um modelo caixa preta total, porque é importante você dar o feedback	metodos analise quantitativa	guiding values
130	P7 - Model Card	A aferição e autenticação dos dados, quer dizer não só os históricos que você obteve são os históricos recentes de dívidas do cliente, as informações que o cliente te passa são informações fidedignas, pra você poder tomar a melhor decisão. Essa questão muitas vezes dos dados restritos, que eu coloquei ali também. Que as vezes tem dados que não são declarados, ou não está oinformados e isso é porque não tem o dado ou porque esse dado é ruim e nã oquero informar. É isso.	consideracoes eticas	ethics of development process - ethical development
131	P7 - Model Card	Só uma dúvida, quando você diz operador aí você ta pensando na pessoa que seria só responsável de colocar as informações, ou ela teria algum outro papel? Não, ela seria a pessoa responsável por colocar as informações né. Tranquilo. Quer dizer, como ela pode atuar né. Pega os dados, ela checa os dados com o cliente, ela checa corretamente ou não.	cuidados e recomenda-coes	impacted individuals
132	P7 - Model Card	Bom, eu acho que aqui é mais variado no sentido de vários, várias dimensões que foram selecionadas aqui, isso chama atençao, embora que eu acho assim que, o... o modelo de ontem, o problema de ontem, eu me senti assim, mais é... mais preocupado com a questão de ser justo com todos os elementos, com todas as pessoas, etc, do que... mas é por causa do tema né, desse tema de hoje, sabe.	revisao e perguntas	responsibility for artifact
133	P7 - Model Card	modelo de ontem, o problema de ontem, eu me senti assim, mais é... mais preocupado com a questão de ser justo com todos os elementos, com todas as pessoas, etc, do que... mas é por causa do tema né, desse tema de hoje, sabe.	revisao e perguntas	responsibility for artifact

134	P7 - Model Card	De um modo geral assim, você vai pensando em ser o melhor possível no sentido de ser justo, de não cometer injustiças de qualquer parte ou de qualquer lado. Você	revisao e perguntas	scaffolding around element of ethical framework - justice
135	P7 - Model Card	Você não pode... Sobre o que você está fazendo, sobre a sua responsabilidade sobre o que está fazendo. Você não escreve na verdade que o modelo fará, mas eu farei o modelo que vai fazer isso, né. Entendi. E que é o certo eu acho no final, porque quem tá fazendo o modelo sou eu. Quem tá colocando os dados ali condicionados sou eu. Então eu tenho que ter alguma responsabilidade sobre o que eu to pensando sobre isso e as consequências disso.	revisao e perguntas	responsibility for artifact
136	P8 - MC Doc	Proporção/quantidade de alunos aprovados que não tem condições de ingressar em uma universidade particular	metricas	guiding values
137	P8 - MC Doc	Garantir que os dados sejam bem distribuídos entre as características como classe social/renda familiar, raça, sexo, etc., para, por exemplo, não favorecer mais um grupo do que outro	consideracoes eticas	diagnosis of element of ethical framework - justice
138	P8 - MC Doc	Garantir que os dados sejam bem distribuídos entre as características como classe social/renda familiar, raça, sexo, etc., para, por exemplo, não favorecer mais um grupo do que outro	consideracoes eticas	ethics of development process - ethical development
139	P8 - MC Doc	Anonimização dos dados	consideracoes eticas	diagnosis of element of ethical framework - non-maleficence
140	P8 - MC Doc	Anonimização dos dados	consideracoes eticas	ethics of development process - ethical development
141	P8 - MC Doc	Garantir que um candidato possa apagar os seus dados históricos do modelo/armazenados	consideracoes eticas	diagnosis of element of ethical framework - autonomy
142	P8 - MC Doc	Cuidado para garantir que não haja um favorecimento de um determinado grupo por parte do modelo, seja por uma má distribuição dos dados ou por fatores externos	cuidados e recomenda-coes	framing based on element of ethical framework - justice

143	P8 - MC Doc	Recomendado ter um passo para revisão/monitoramento/avaliação manual Recomendado ter um passo para calibração/ajuste de parâmetros/modelo, ou peso dado para as características utilizadas	cuidados e recomenda-coes	system's autonomy - decrease
144	P8 - Model Card	O meu ponto aqui não é nem pelo dado, é de forma geral se eu tenho pouco dado disponível... então independente do dataset certo, se eu tenho pouco dado talvez possa atrapalhar, se ele tem... por exemplo mal distribuído para uma tal característica também pode me atrapalhar depois, se você quiser eu posso descrever um pouco mais aqui.	fatores	other
145	P8 - Model Card	Acho que de maneira indireta talvez... essas questões como classe social e raça pode impactar de forma indireta, não necessariamente pela raça, mas posso, dependendo do dado, se eu não me ater a olhar, não é só pela raça porque eu vou usar isso como uma característica pra negar a entrada dele, mas outras informações. Então talvez... classe social, raça, botar sexo também. Que é informação que eu tenho que ter cuidado que eu possa só ta reforçando certos preconceitos que já existem.	fatores	ethics of development process - unethical development
146	P8 - Model Card	Eu posso não usar raça, por exemplo, como dado, mas... pode impactar de uma forma mais indireta que é porque que alguém que tem uma classe social menor teve muito mais problema de estudo, porque precisa trabalhar, por exemplo, teve notas mais baixas. Não é direto, mas é de uma forma mais indireta.	fatores	undesired consequences
147	P8 - Model Card	Talvez... quero falar que se a pessoa trabalha pode impactar na hora que ela entra na universidade, porque ela tem uma carga horária muito dividida, ou alguma coisa assim.	fatores	diagnosis of element of ethical framework - justice
148	P8 - Model Card	Talvez aqui uma quantidade de alunos aprovados/reprovados, isso se possível eu quero dividir por certas questões como classe social, raça sexo. Que é pra gente dar uma olhada se eu não to favorecendo muito um lado e ignorando o outro talvez, acho que essa possa ser uma possível métrica.	metricas	framing based on element of ethical framework - justice

149	P8 - Model Card	Eu acho que aqui como eu to considerando que é uma universidade pública, então a meu ver eu deva dar mais oportunidade para quem, por exemplo, não tem a condição financeira de fazer uma universidade, poderia pagar uma universidade particular. Então acho que essa seria uma segunda métrica. Vou pensar como eu escrevo isso.	metricas	guiding values
150	P8 - Model Card	Então eu penso que essa universidade pública deveria ter mais pessoas que não tem condição que quem tem. Entendi. Então pra mim quanto maior é esse número, talvez melhor. O que mais?	metricas	guiding values
151	P8 - Model Card	Talvez alguma métrica, acho que é parecido com esse primeiro ponto, alguma métrica de justiça, equidade. Não sei como eu posso chamar, também não sei exatamente como, mas... Acho que entra no primeiro ponto	metricas	scaffolding around element of ethical framework - justice
152	P8 - Model Card	Eu quero, por exemplo, garantir que não haja, muito... Eu tenho 90% de aprovação pra gente branco e 10% pra gente negra, por exemplo. Mas acho que entra nesse primeiro ponto que eu escrevi, acho que eu queria algo mais meio a meio. Então eu vou ignorar isso	metricas	guiding values
153	P8 - Model Card	Acho que aqui também poderia ter uma análise a respeito de, se eu pegar todos esses dados que eu tenho, então por exemplo, qual a média do ENEM dividido por raça, por classe e por sexo. Então fazer alguma coisa assim, fazer alguma análise em cima pra tentar analisar se... se o dado de fato já tem esse viés talvez, esse viés de problemas sociais.	metodos analise quantitativa	ethics of development process - ethical development
154	P8 - Model Card	Como eu falei anteriormente, acho que o ideal é eu garantir que todo mundo que deveria ser aprovado foi aprovado	metodos analise quantitativa	guiding values
155	P8 - Model Card	Tá, então acho que eu tenho que garantir que os dados sejam bem distribuídos para, por exemplo, características como classe social. Para por exemplo, não favorecer mais um grupo do que outro.	consideracoes eticas	ethics of development process - ethical development
156	P8 - Model Card	Acho que eu também preciso garantir uma anonimização dos dados, para questões desde vazamento, ou por exemplo ficar mantendo informações talvez como histórico escolar e renda familiar das pessoas.	consideracoes eticas	framing based on element of ethical framework - non-maleficence

157	P8 - Model Card	Acho que eu também tenho que deixar possível uma forma de por exemplo, eu deixei meus dados, então eu quero apagar meus dados sempre que eu quiser. Talvez ande junto com esse segundo, mas eu garantir que haja uma forma de apagar os dados... que o um candidato possa apagar seus dados históricos do modelo.	consideracoes eticas	framing based on element of ethical framework - autonomy
158	P8 - Model Card	Então não posso favorecer nenhum grupo, garantir que ta anonimizado, apagar detalhes do modelo armazenado	consideracoes eticas	ethics of development process - ethical development
159	P8 - Model Card	Acho que uma possível recomendação é... Eu acho que tenho que ter algum passo, alguma forma de eu monitorar esses casos, ou até por exemplo, talvez até passar por uma monitoração manual por alguém.	cuidados e recomenda-coes	system's autonomy - decrease
160	P8 - Model Card	Recomendo uma revisão, monitoramento, avaliação manual... ter um passo para calibração e ajusto de parâmetros/modelo. Então acho que eu não posso ficar...	cuidados e recomenda-coes	system's autonomy - decrease
161	P8 - Model Card	Aqui a ideia do modelo é ele fazer de forma automática se o candidato deve ingressar na universidade ou não. Mas talvez deva ter um passo que por exemplo eu posso pedir para fazer uma revisão do meu caso, alguma coisa assim, ou passar por uma avaliação manual, alguém olhar os dados, ver se não houve algum equívoco, alguma coisa assim.	cuidados e recomenda-coes	system's autonomy - decrease

B
Additional Plots

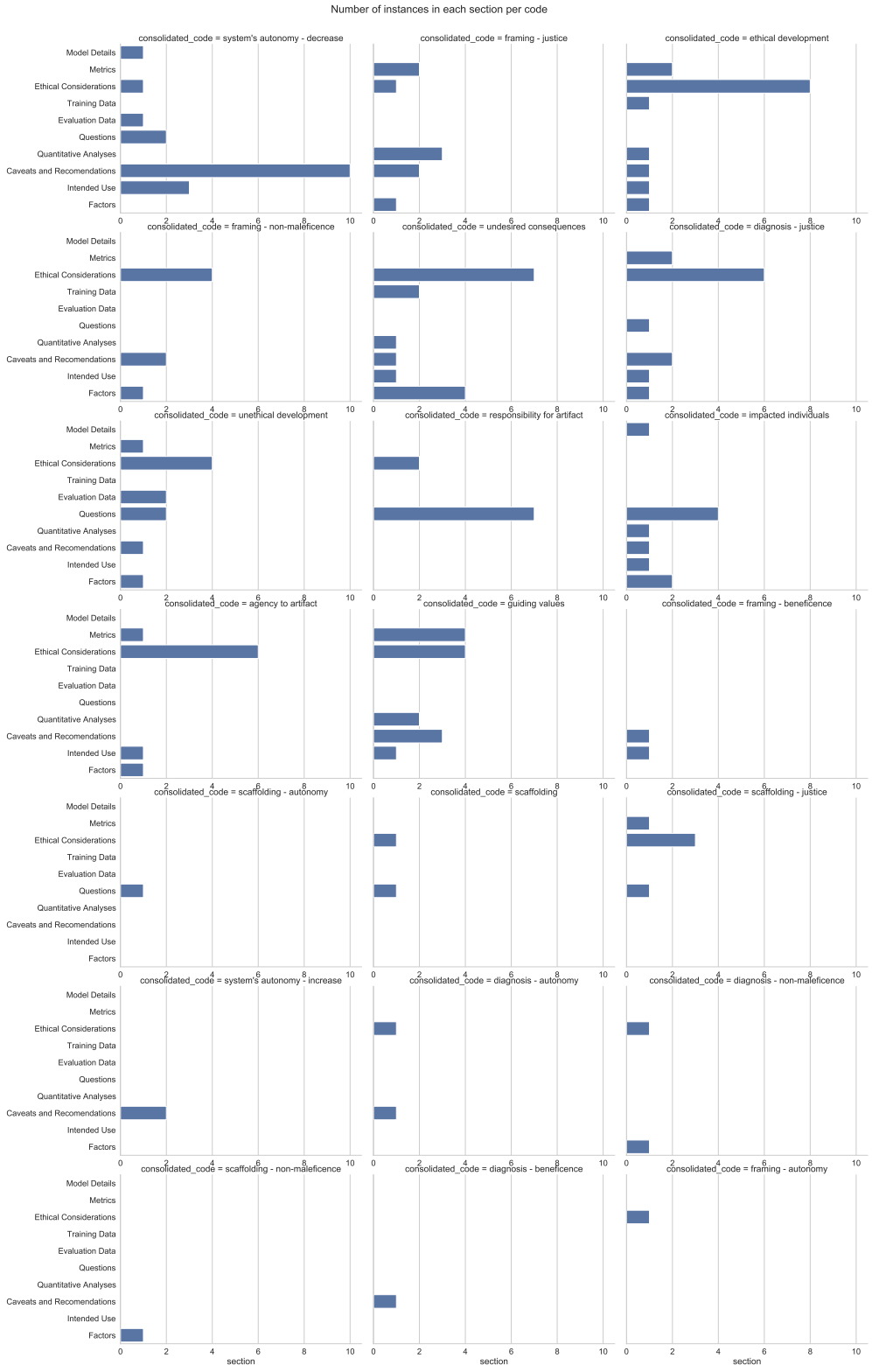


Figure B.1: Frequency of passages in each section per code.

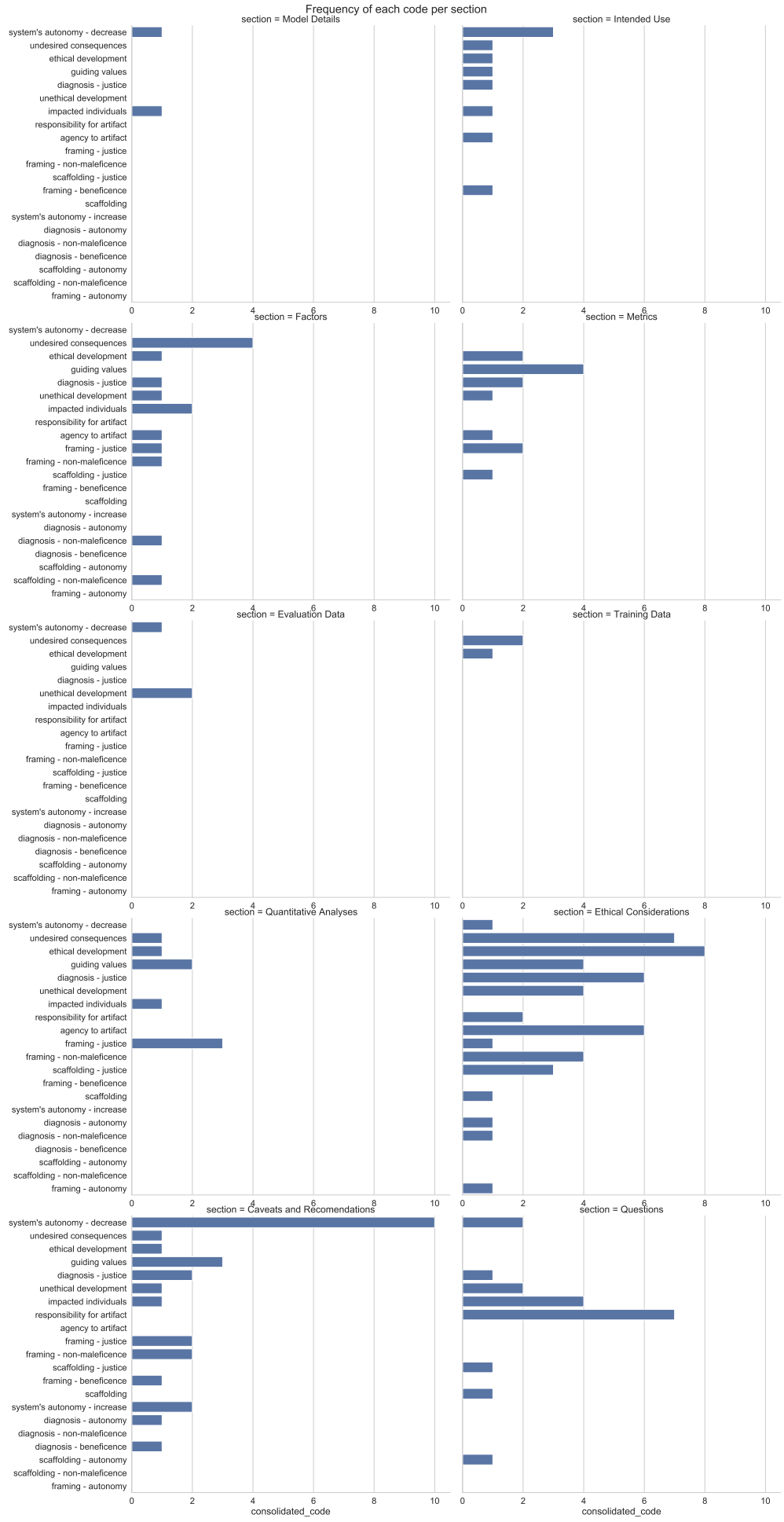


Figure B.2: Frequency of each code in each section.

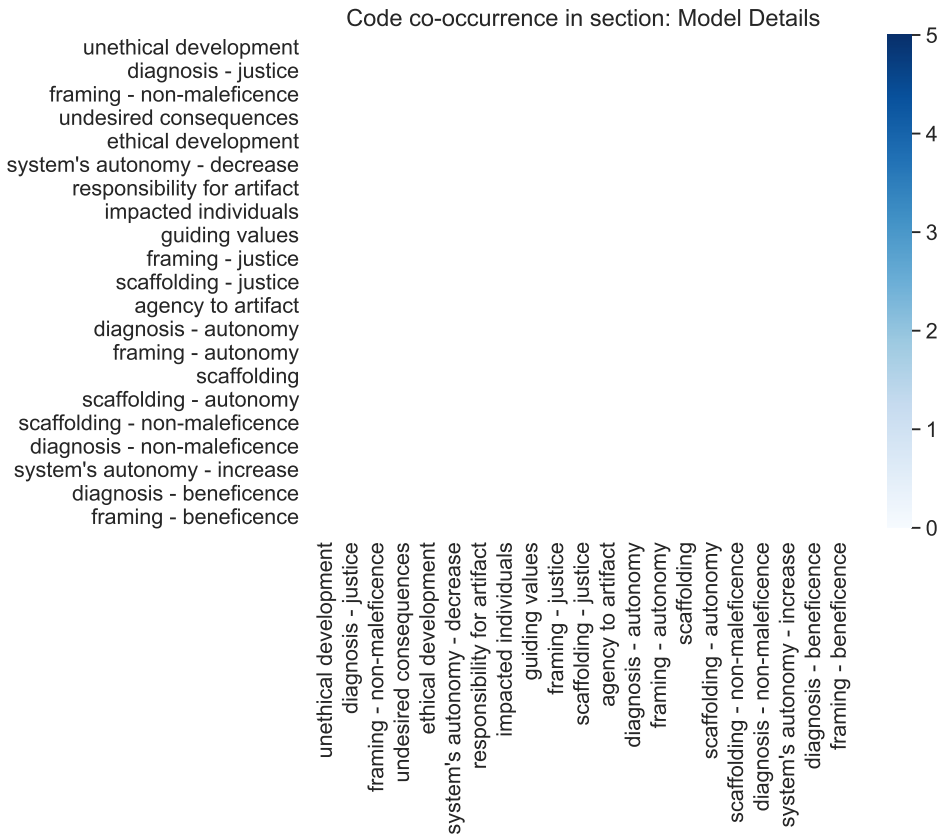


Figure B.3: Code co-occurrence for section Model Details.

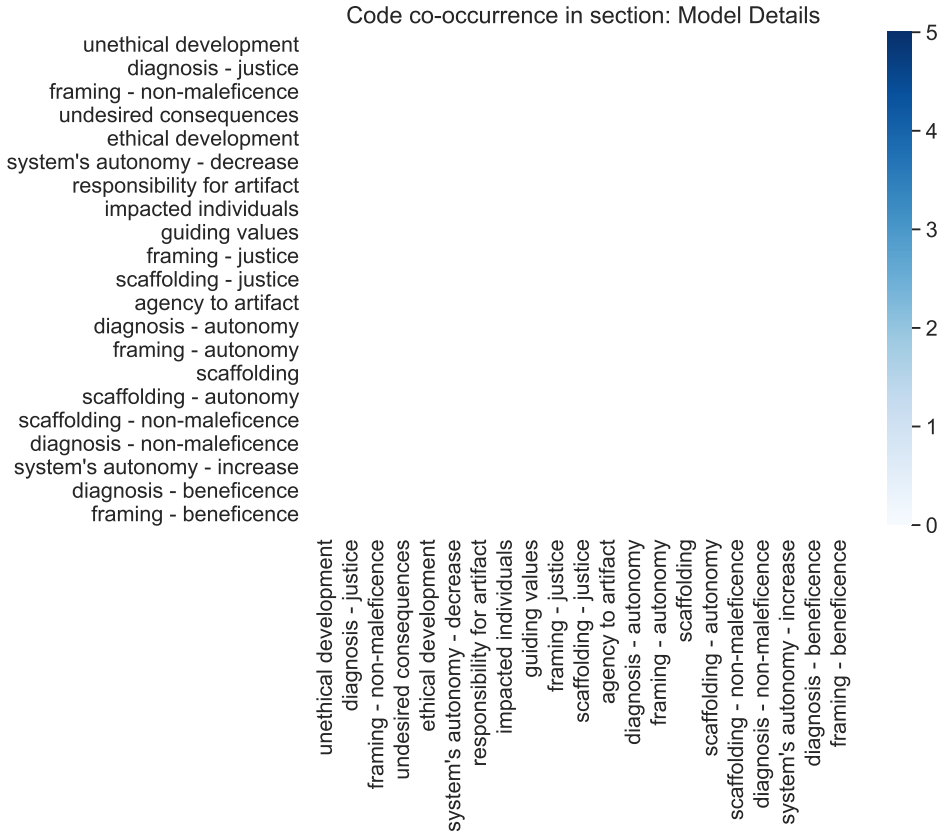


Figure B.4: Code co-occurrence for section Intended Use.

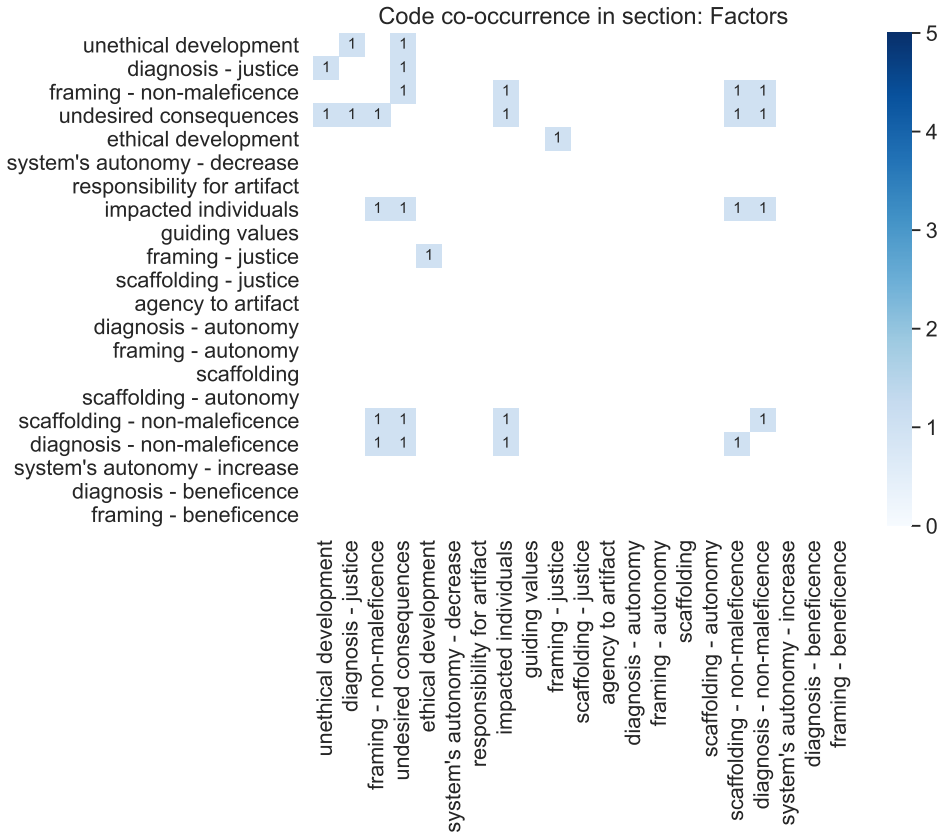


Figure B.5: Code co-occurrence for section Factors.

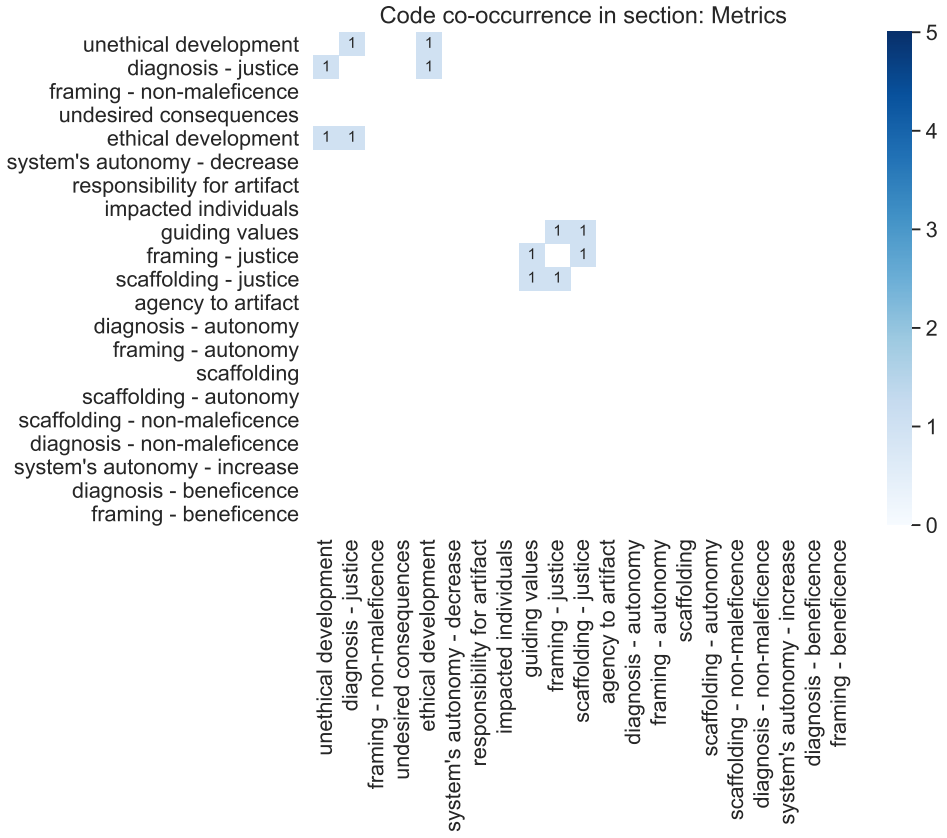


Figure B.6: Code co-occurrence for section Metrics.

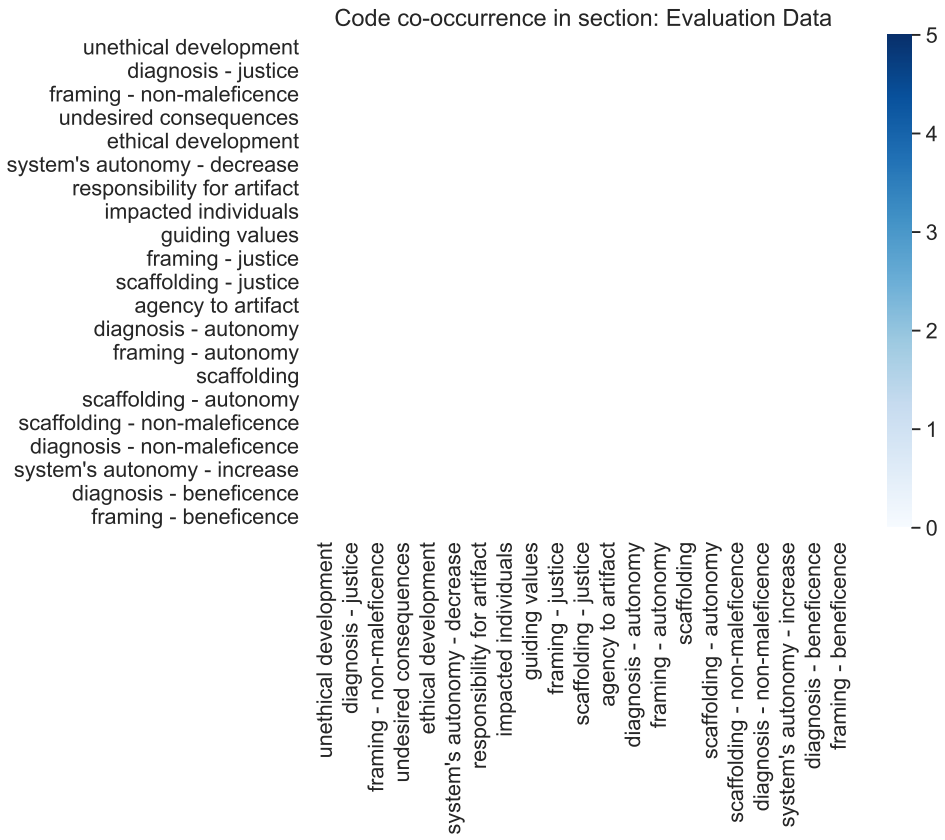


Figure B.7: Code co-occurrence for section Evaluation Data.

Figure B.8: Code co-occurrence for section Training Data.

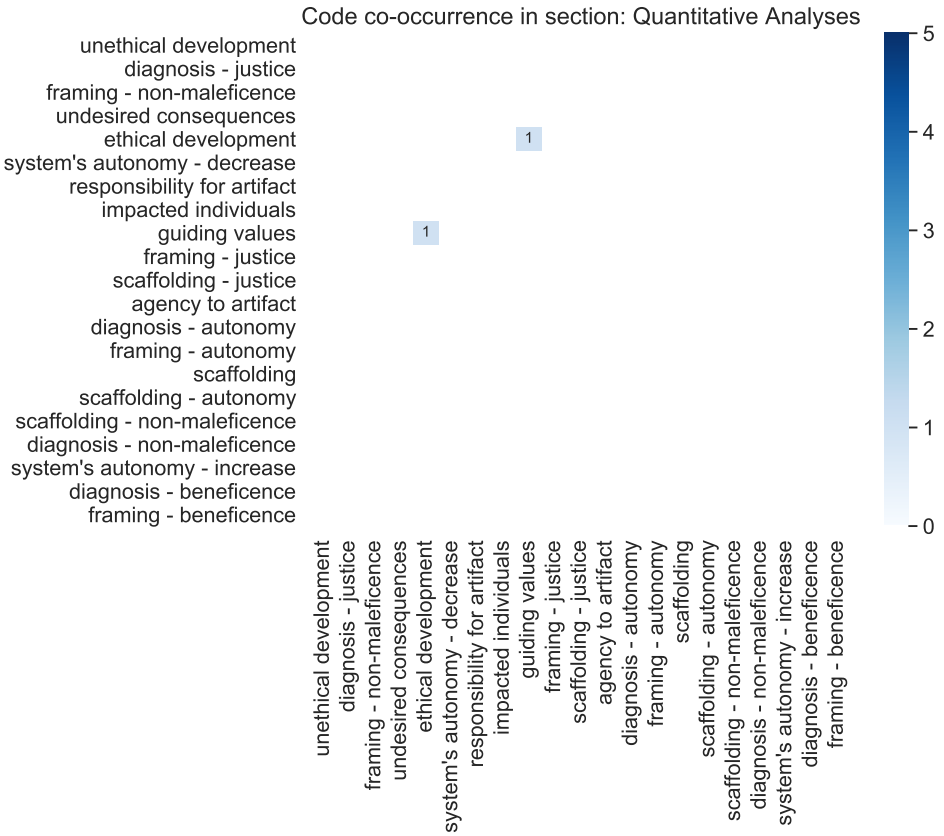


Figure B.9: Code co-occurrence for section Quantitative Analyses.

Figure B.10: Code co-occurrence for section Ethical Considerations.

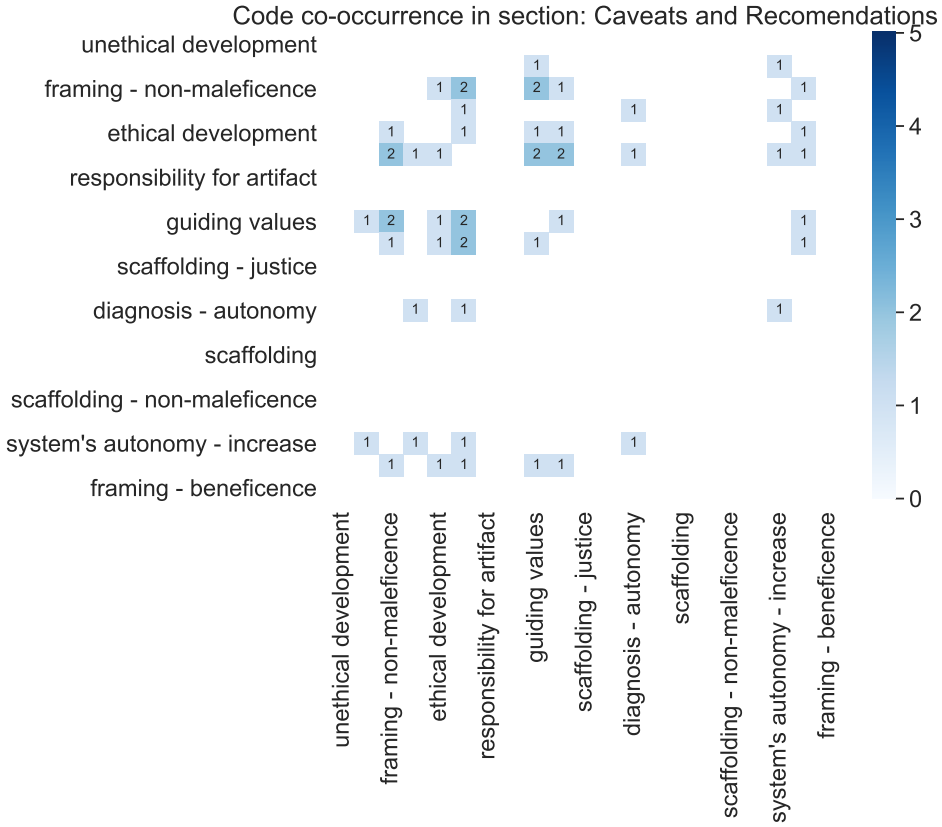


Figure B.11: Code co-occurrence for section Caveats and Recomendations.

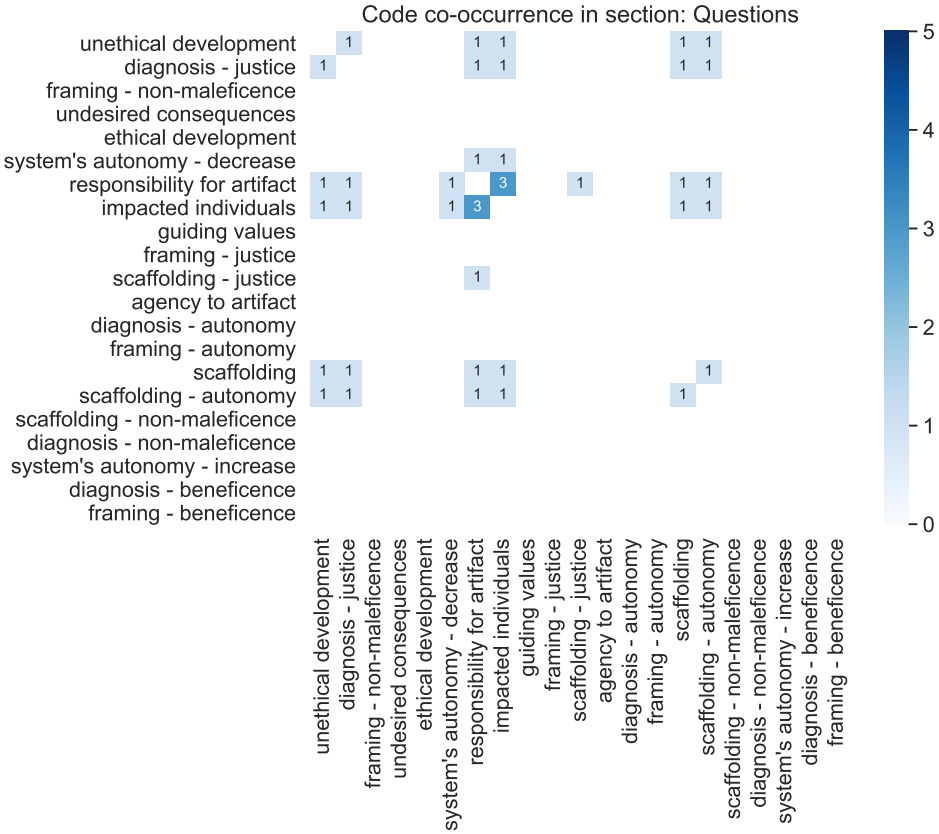


Figure B.12: Code co-occurrence for section containing our questions at the end of the interview.

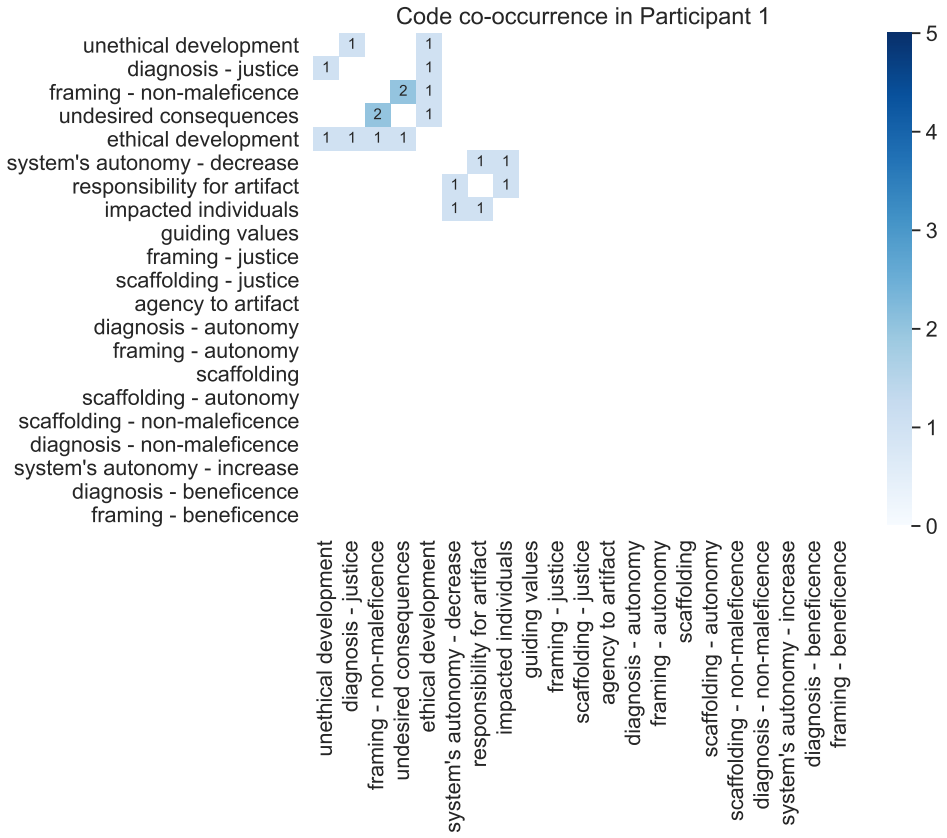


Figure B.13: Code co-occurrence for Participant 1.

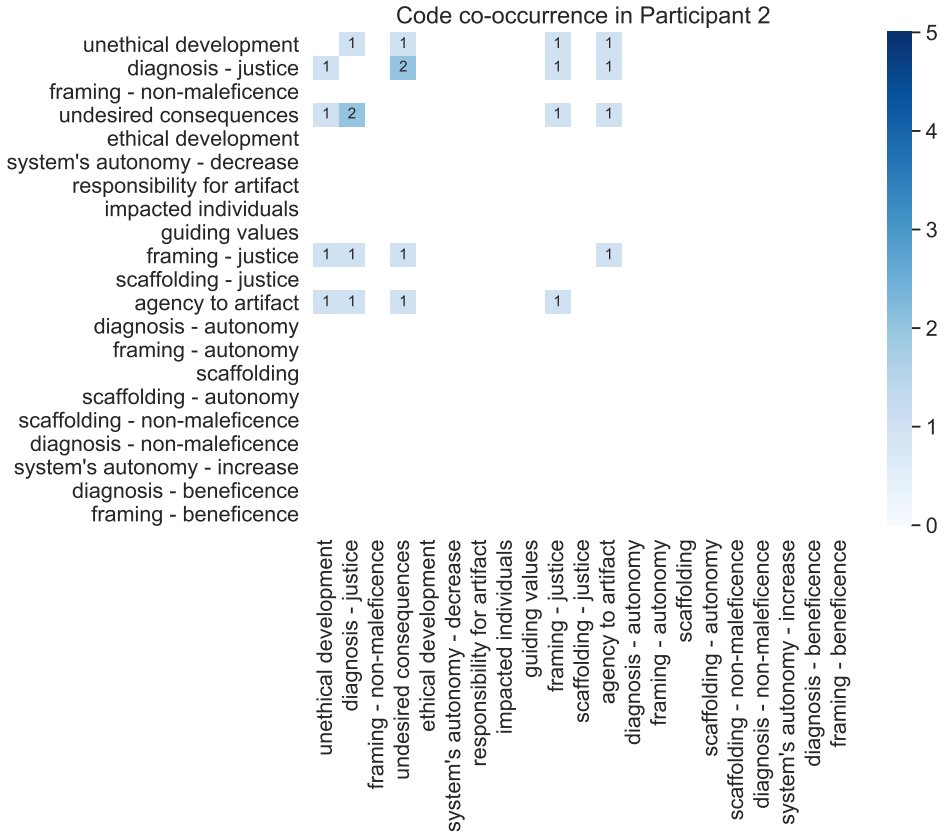


Figure B.14: Code co-occurrence for Participant 2.

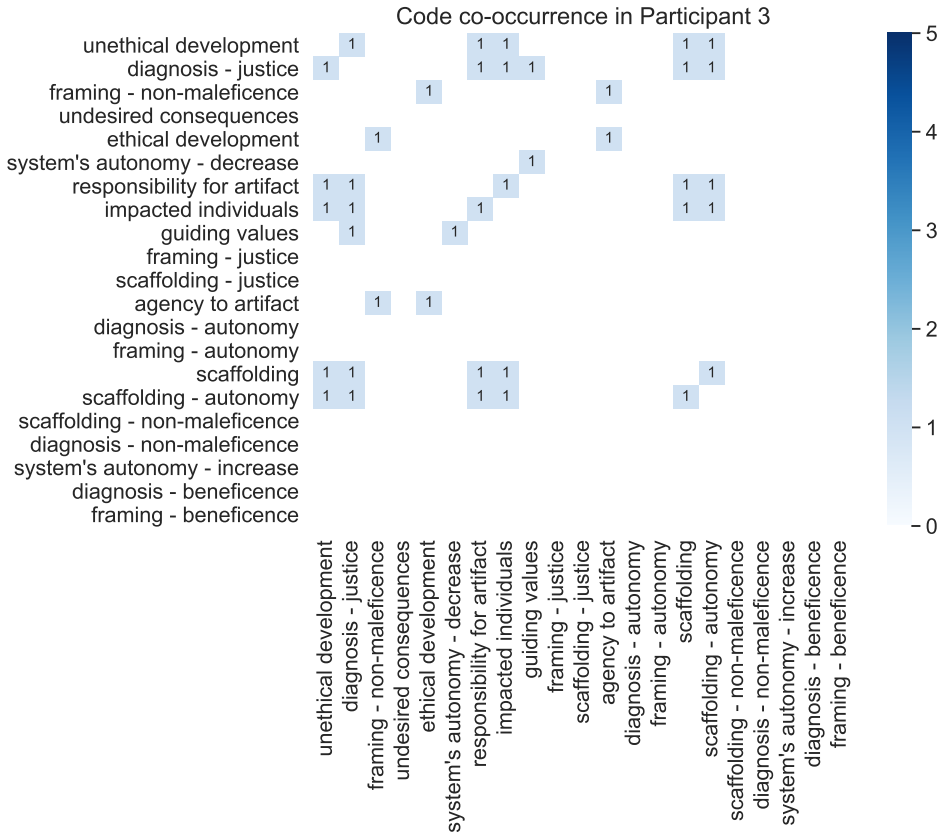


Figure B.15: Code co-occurrence for Participant 3.

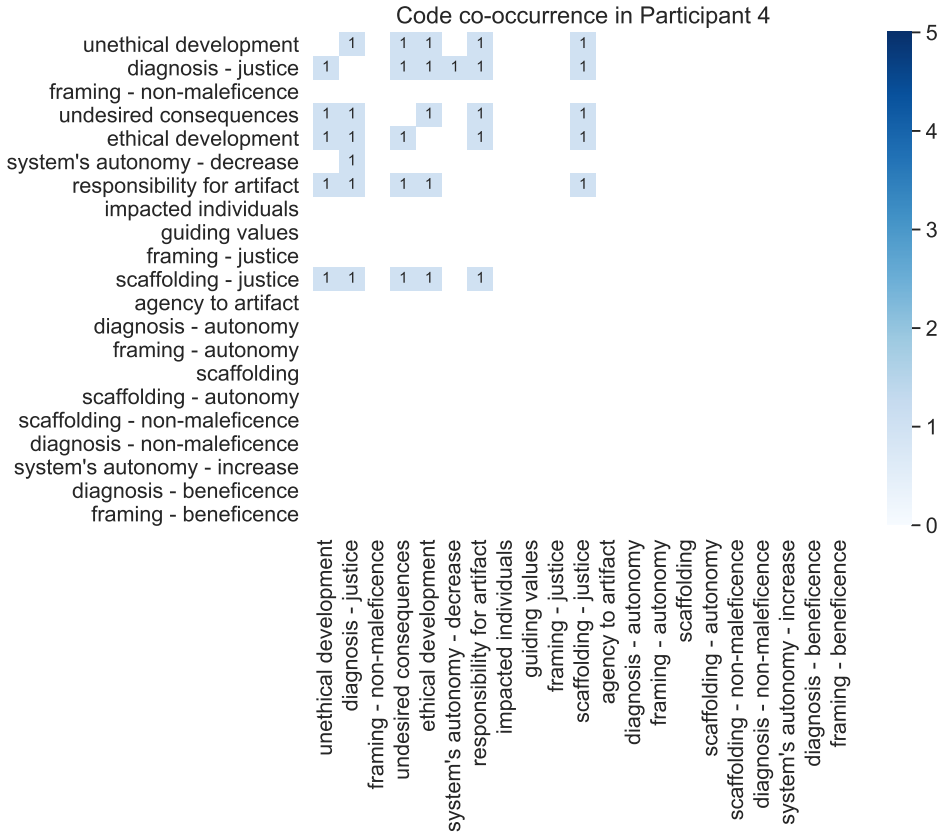


Figure B.16: Code co-occurrence for Participant 4.

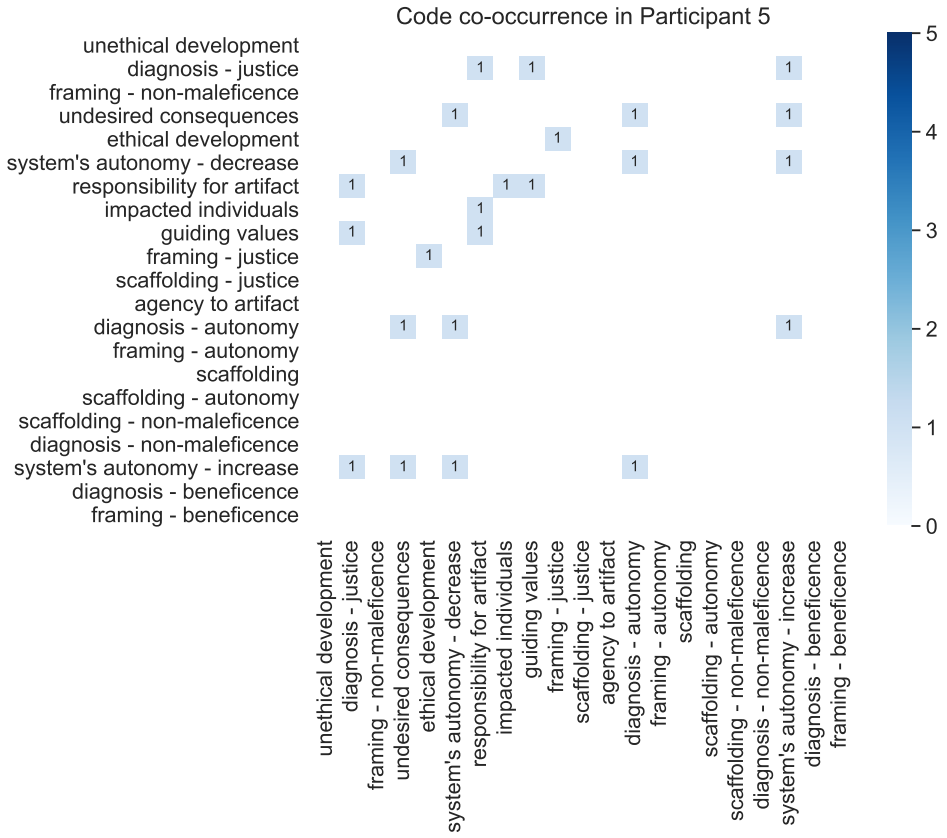


Figure B.17: Code co-occurrence for Participant 5.

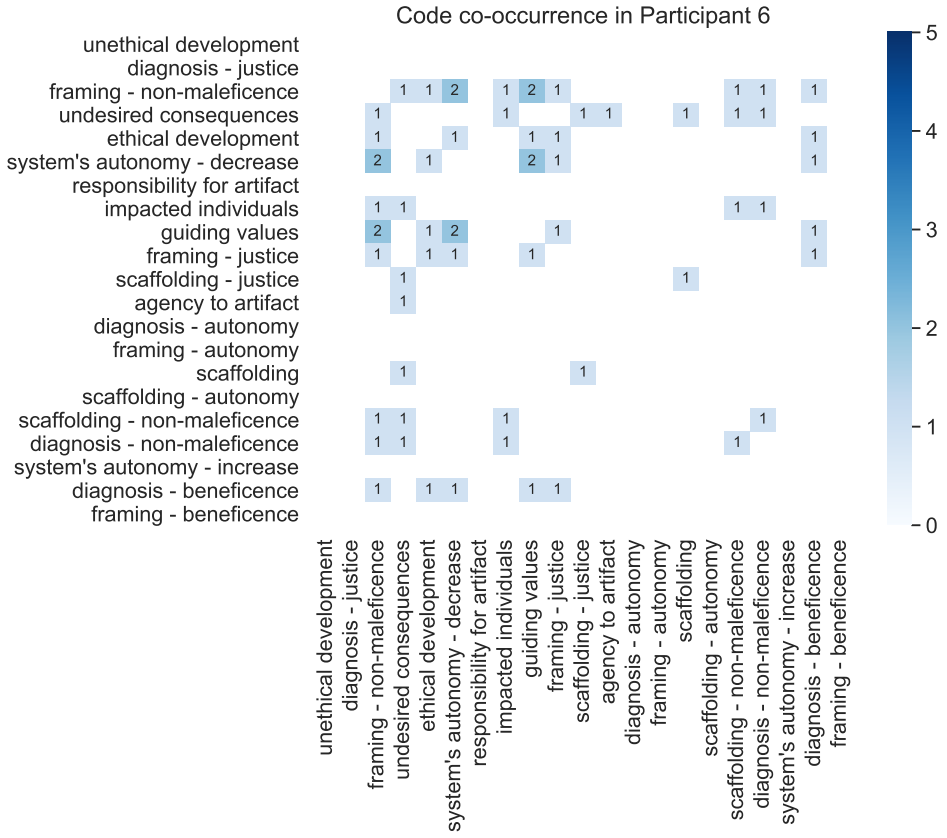


Figure B.18: Code co-occurrence for Participant 6.

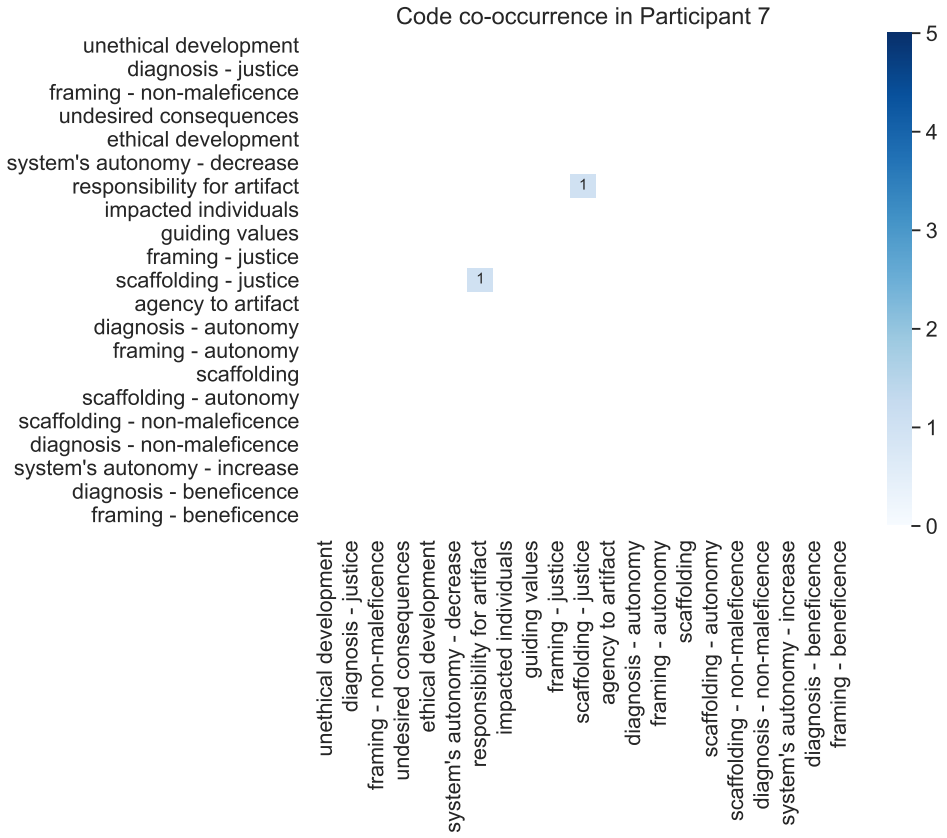


Figure B.19: Code co-occurrence for Participant 7.

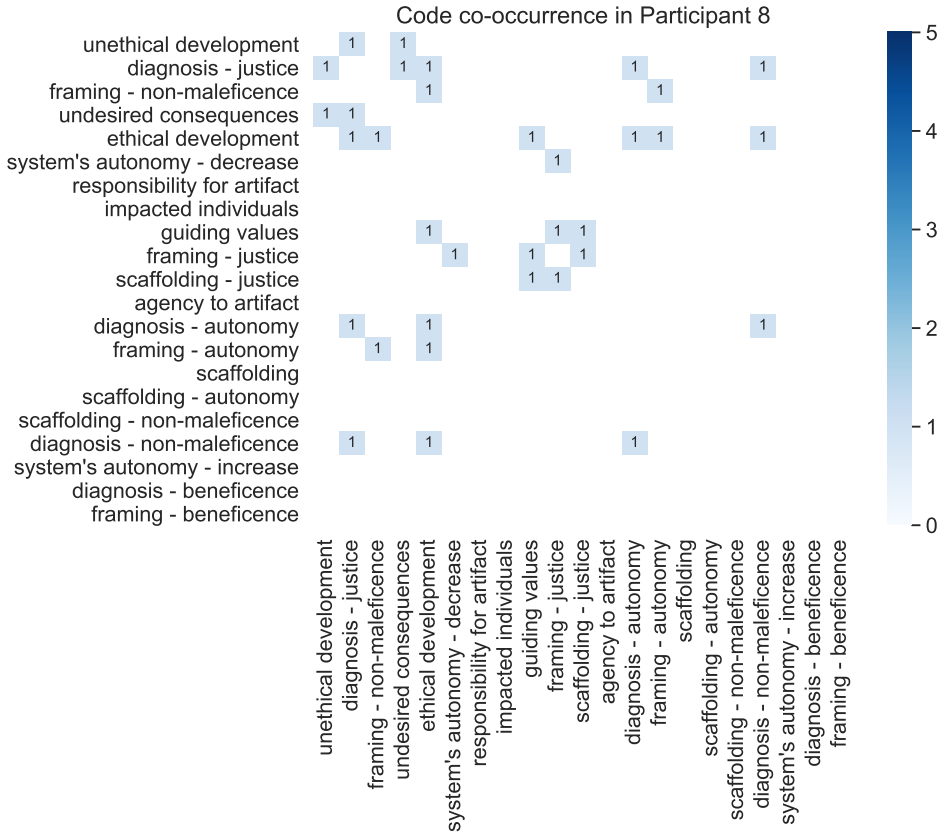


Figure B.20: Code co-occurrence for Participant 8.

C
Code

C.1
Functions used to load data

```
import docx
import numpy as np
import pandas as pd
from pathlib import Path


def get_text_start_index(row, docs):
    '''
    Get index of first instance of text.
    '''
    i = docs[row['Case']].find(row['Text'])

    if i == -1:
        # em alguns casos ele nao acha pq QDA muda quebra de linha p whitespace
        # ao criar o CSV com codificacao
        i_sem_quebralinha = docs[row['Case']].replace('\n', ' ').find(row['Text'])
        if  i_sem_quebralinha != -1:
            i = i_sem_quebralinha
        else:
            pass
    return i


def load_codificacao(fname, path_to_docs, coder=None):
    '''
    Loads csv generated by QDA with codes.
    Includes cleaning of dataframe and adding some information
    fname = name of csv file to be loaded
    path_to_docs = path to directory cointaining all anotated files
```

```
    coder = id of the coderresponsible. If not given column is not created
    '''
    cods = pd.read_csv(fname, encoding='cp1252')
    cods.dropna(axis=1, how='all', inplace=True)
    # removes white spaces at tail of string
    cods['Case'] = cods['Case'].str.strip()
    cods['Code'] = cods['Code'].str.lower()
    # removes categories used during codign from code name
    cods['Code'] = cods['Code'].replace(r'(metareflection - )|(ethical reasoning -
    # cleans one column name
    cods.rename(columns={'%\xa0Words':'% Words'}, inplace=True)

    # loading dict with docs to use on apply
    docs = get_docs(path_to_docs)
    cods['start_code'] = cods.apply(get_text_start_index, axis=1, args=(docs,))
    cods['start_code'] = cods['start_code'].where(cods['start_code'] > -1)
    cods['end_code'] = cods['start_code'] + cods['Text'].str.len()
    cods['len_code'] = cods['end_code'] - cods['start_code']

    # Adds info about section code appeared
    prompts = pd.read_csv('data/prompts.csv', sep=';')
    # set coder
    cods['section'] = [get_prompt(row, prompts) for i, row in cods.iterrows()]

    if coder:
        cods['Coder'] = coder
    return cods


def get_prompt(row, prompt_df):
    '''
    retrieves prompt of coded section
    '''
    cond = (row.Case == prompt_df.Case) & \
           (row.start_code >= prompt_df.ini) & \
           (row.end_code <= prompt_df.end)

    p = prompt_df[cond]['prompt'].values
```

```
    if len(p) == 0:
        return None
    elif len(p) == 1:
        return p[0]
    else:
        raise Exception(
            'There was an error found more than 1 prompt for the code'
        )


def get_text(filename):
    '''
    Loads text for a docx file
    '''
    doc = docx.Document(filename)
    fullText = []
    for para in doc.paragraphs:
        fullText.append(para.text)
    return '\n'.join(fullText)


def get_docs(folder_path):
    '''
    Loads texts for all docx files in a folder
    returns dict with doc id as key and text as value
    '''

    docs = dict()

    for d in Path('projeto_coding').glob('*.docx'):
        text = get_text(d)

        f_name = d.as_posix()

        part = f_name.split('/')[-1].split(' - ')[0]

        case = part.upper() + ' - '

        if 'Doc' in f_name:
```

```
            tipo = 'doc'
        else:
            tipo = 'entrevista'

        if 'EMT' in f_name:
            tool = 'emt'
            case += tool.upper()
            if tipo == 'doc':
                case += ' '
                case += tipo.title()
        else:
            tool = 'mc'
            if tipo == 'doc':
                case += tool.upper()
                case += ' ' + tipo.title()
            else:
                case += 'Model Card'


        docs[case] = text

    return docs

# Before final code names
selected_codes = [
    'agency to artifact',
    'element of ethical framework in ethical issue',
    'element of ethical framework in ethical issue - autonomy',
    'element of ethical framework in ethical issue - beneficence',
    'element of ethical framework in ethical issue - justice',
    'element of ethical framework in ethical issue - non-maleficence',
    'ethics of development process - ethical development',
    'ethics of development process - unethical development',
    'framing based on element of ethical framework',
    'framing based on element of ethical framework - autonomy',
    'framing based on element of ethical framework - beneficence',
    'framing based on element of ethical framework - justice',
    'framing based on element of ethical framework - non-maleficence',
    'guiding values',
```

```
    'scaffolding around element of ethical framework',
    'scaffolding around element of ethical framework - autonomy',
    'scaffolding around element of ethical framework - beneficence',
    'scaffolding around element of ethical framework - justice',
    'scaffolding around element of ethical framework - non-maleficence',
    'scaffolding around element of ethical framework - justice',
    'impacted individuals',
    'responsibility for artifact',
    "system's autonomy",
    "system's autonomy - increase",
    "system's autonomy - limit",
    'undesired consequences'
]
```

C.2
Name replacement for plots and final version

```
#translation of Model Cards sections
sec = {
        'detalhes do modelo': 'Model Details',
        'usos pretendidos': 'Intended Use',
        'metricas': 'Metrics',
        'consideracoes eticas': 'Ethical Considerations',
        'metodos analise quantitativa': 'Quantitative Analyses',
        'fatores': 'Factors',
        'dados de treino': 'Training Data',
        'dados de avaliacao': 'Evaluation Data',
        'revisao e perguntas': 'Questions',
        'cuidados e recomendacoes': 'Caveats and Recommendations',
}

# Order of sections
sec_order = [
     'Model Details',
     'Intended Use',
     'Factors',
     'Metrics',
     'Evaluation Data',
     'Training Data',
     'Quantitative Analyses',
```

```
        'Ethical Considerations',
        'Caveats and Recommendations',
        'Questions'
]

# Dict to replace code names for plotting
# represent final code names
cods_replace = {
    'framing based on element of ethical framework - justice': 'framing - justice',
    'ethics of development process - ethical development': 'ethical development',
    'framing based on element of ethical framework - non-maleficence': 'framing - n
    'element of ethical framework in ethical issue - justice': 'diagnosis - justice
    'ethics of development process - unethical development': 'unethical development
    'framing based on element of ethical framework - beneficence': 'framing - benef
    'scaffolding around element of ethical framework - autonomy': 'scaffolding - au
    'scaffolding around element of ethical framework': 'scaffolding',
    'scaffolding around element of ethical framework - justice': 'scaffolding - jus
    'element of ethical framework in ethical issue - autonomy': 'diagnosis - autono
    'element of ethical framework in ethical issue - non-maleficence': 'diagnosis -
    'scaffolding around element of ethical framework - non-maleficence': 'scaffoldi
    'element of ethical framework in ethical issue - beneficence': 'diagnosis - ben
    'framing based on element of ethical framework - autonomy': 'framing - autonomy
}
```

D
Bioethical Principles Summary

This appendix displays the Summary of the bioethical principles as presented to participants in our study.

D.1
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

D.1.1
Os 4 Princípios da Bioética

D.1.1.1
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.

D.1.1.2
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.

D.1.1.3
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.

D.1.1.4
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.

D.1.2
Resumo

A prática diária requer ponderação desses princípios.

1. Autonomia - direito do paciente de decidir sobre seu tratamento
2. Beneficência - dever de fazer o que for melhor para o paciente
3. Não-maleficência - não causar danos ao paciente
4. Justiça - acesso igual a tratamento