






Using Model Cards for ethical reflection on machine learning models: an interview-based study

José Luiz Nunes   [FGV Direito Rio; Department of Informatics, PUC-Rio | jose.luiz@fgv.br]

Gabriel D. J. Barbosa  [Department of Informatics, PUC-Rio | gabrieldjb@gmail.com]

Clarisse Sieckenius de Souza  [Department of Informatics, PUC-Rio | clarisse@inf.puc-rio.br]

Simone D. J. Barbosa  [Department of Informatics, PUC-Rio | simone@inf.puc-rio.br]

 Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Rua Marquês de São Vicente, 225 RDC, Gavéa, RJ, 22451-900, Brazil.

Received: 02 June 2023 • Accepted: 11 December 2023 • Published: 01 January 2024

Abstract

How do tools designed for documenting machine learning models contribute to developers' ethical reflection? We set out to answer this question regarding Model Cards, a tool proposed for such purpose. We conducted a thematic analysis of eight semi-structured interviews based on speculative design sessions. Each participant assumed the role of developer of an artificial intelligence model in one of two scenarios: loan applications or university admissions. We found evidence that designers may have been selective about which ethical issues – from among those they had reflected on – they recorded in the Model Cards. While participants were hesitant to grant full autonomy to the artifact to be developed, we identified they still tended to rely on a third party (outside the design team) to mediate the relationship between the system and other stakeholders. Our findings contribute to our understanding of documentation tools, their epistemic value, and how they can be leveraged to engage in a more ethically informed design process of artificial intelligence systems.

Keywords: Ethical Reasoning, Model Cards, Model Reporting, Reflective Practica

1 Introduction

Machine Learning (ML) models have been increasingly used to mediate various decision-making processes, leading to significant social impacts. Given this role, it has become even more important for developers to reflect on their potential ethical connotations and document their reflections and decisions alongside technical information, so as to be more transparent. This process of recording design decisions might also enhance the reflections themselves. However, there is not enough evidence, currently, on the impacts of documentation tools on ethical reflection.

Various representations for documenting ML systems have been proposed both in academia and in industry. They focus on different parts of the ML development process: on trained models [Mitchell et al., 2019]; on datasets [Bender and Friedman, 2018; Gebru et al., 2018; Holland et al., 2018]; or even on the development process itself [Hutchinson et al., 2021]. By using multiple tools, developers can reach a more holistic understanding of the artifact being built and the process behind its construction. Despite the various tools being proposed, discussions surrounding their use are still ongoing, as is empirical work investigating their usage.

Our work follows a broader view proposed by de Souza et al. [2016] regarding what they refer to as a *semiotic perspective on human-centered software development*. Very briefly, they not only embrace the fact that all computer representations are the result of prior human, value-laden interpretations, but also engage in creating “epistemic tools” to help developers understand the nature, the extent, and the implications of their interpretive activities.

We seek to contribute to the current understanding of the relations between design documentation tools and developers' ethical reflections. Our emphasis is on the sociotechnical relationship between designers and technology. Thus, our approach is directed towards developing a critical view of ethically relevant aspects of machine learning design.

The study reported in this paper is part of a broader comparative investigation covering different types of tools: Model Cards [Mitchell et al., 2019] and the Extended Metacommunication Template [Barbosa et al., 2021]. Our focus is directed to the ethical considerations supported and stimulated by such tools [Barbosa et al., *ming*].

We carried a qualitative study based on speculative design sessions, in which – regarding Model Cards, specifically – eight participants were asked to fill out cards with their design vision, which they constructed based on a design brief they were provided in each session. These covered two different development scenarios for decision-support systems with salient ethical connotations: (i) loan applications at a financial institution and (ii) a university admissions process. We conducted a Thematic Analysis [Braun and Clarke, 2012] of the collected data to identify relevant themes that emerged from the ethical reflection participants engaged in when using the tool. Our main research question in this paper is “*How do software developers reason about ethical issues in machine learning systems development when using Model Cards?*”

This paper is organized as follows.¹ We begin with a brief presentation of tools proposed for the documentation of AI, with a special focus on the Model Card, the conceptual tool in focus – analyzing its potential for ethical reflection. Next, we

¹This paper is a revised and extended version of Nunes et al. [2022].

present our study’s design and methodology, followed by our findings. We then discuss the results in the context of related work, as well as some epistemological aspects of our research orientation and strategy. Finally, we present our concluding remarks and point to interesting future work.

2 Handling transparency in Artificial Intelligence systems

In this section we present the current state of relevant literature, focusing on value-oriented tools, documentation for datasets, and tools and guidelines.

2.1 Value-oriented tools

Here we address proposals which have the common thread of aiming to highlight different values that influence the development of AI systems. This is especially relevant when we consider the different social values at play for the distinct stakeholders of intelligent systems, and decisions will usually imply a trade-off between them. We skip Model Cards [Mitchell et al., 2019], which is discussed in Section 3.

One proposal in this direction is 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 initially seen as technical decisions and metrics (e.g., maximizing accuracy).

They proposed and conducted an initial study in an educational setting, testing three different cards with unique purposes, and highlighted the epistemic value of the cards in their proposal.

- Model cards –focus on different ML models and should capture trade-offs between choices in development of AI applications;
- Persona cards –depict perspectives and values of different stakeholders;
- Checklist cards –enumerate social and technical considerations, which should be used to guide the deliberation and decision process

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. Akin to 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.²

²We include this work here because they frame the entire audit process from values and principles to be defined by the organization, which ought to guide and be verified using the proposed framework.

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, e.g., in the medical and financial industry. They leverage the proposals of Seck et al. [2018] and Mitchell et al. [2019] to establish a framework that includes not only the production of datasheets and Model Cards by the development team, but also a series of other artifacts produced through the auditing process.

The final stage of the process, labeled as Reflection Stage, should culminate in 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.

Barbosa et al. [2021] offer an epistemic tool based on the Semiotic Engineering theory of human-computer interaction [De Souza, 2005; de Souza et al., 2016]. This tool aims to promote designers’ and developers’ reflection throughout the development process, especially related to moral responsibility and ethical reasoning regarding other stakeholders. It is structured around the different persons in a discourse, highlighting the role of designers (1st person) and the effects they may bring on users (2nd person) and other affected parties (3rd person).

We label as “value oriented” the proposals that emphasize values and ethical principles of stakeholders of the development process. However, they may also have wider scope and rely on the interaction among multiple members of design teams, and even other parties, as is the case of auditing teams. We directed the focus of our research to tools that can be directly used by developers and can be integrated into these frameworks, as was done by Hutchinson et al. [2021] with Model Cards [Mitchell et al., 2019].

2.2 Documentation for Datasets

An array of works addresses the issue of documenting datasets, and making this information available to other stakeholders. These efforts are especially relevant to the issue of model and data reuse [Hutchinson et al., 2021; Brandão et al., 2019].

The proposals we describe in this subsection aim to standardize this aspect of the development of AI systems, documenting the datasets’ characteristics, how the data is analyzed, how it is distributed, and diagnosis of possible biases in the data. Among all the works we describe in Section 2, these were the first to appear in the literature. They include standardized and automated framework for the analysis of datasets and documents with characteristics of their data, which would be manually produced by developers.

Holland et al. [2018] proposed *Nutrition Label*, standardized labels created to convey metadata and information about datasets, and to try to reflect a portion of standard exploratory analyses conducted by developers when deciding whether to use a dataset. They include seven different modules to make distinct aspects of information regarding the data available to other parties; each module requires different manual effort and reveals specific elements of the dataset, composing the

‘nutritional’ ingredients of the dataset.

Data Statements [Bender and Friedman, 2018] targeted a specific kind of dataset: natural language datasets. They focus on dealing with issues identified in natural language processing tasks. They argue that new datasets in this area tend to be published with lengthy discussions about how they were annotated, but there is an informational gap regarding the profiles of the people who produced the data (speakers or writers) and those responsible for annotating it.

Data Statements were 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. These aspects highlight the authors’ identification of the relevance of context in language and its construction through interpretation. They argue the proposed tool can mitigate different types of bias, better highlighting what the data represents and what it does not.

Gebru et al. [2018] presented a similar proposal, but for all types of datasets. Their work on *Datasheet for Datasets* aims to address specific goals for different stakeholders. For dataset creators and curators, they wish to promote reflective practices, 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 at hand. 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. Ros-tamzadeh et al. [2022] propose an extension of *Datasheet for Datasets* specifically focused on healthcare.

Miceli et al. [2021] conducted a study in the field of computer vision to compare these general proposals and 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 who have made use and requested this data. Their work identified 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.

Hutchinson et al. [2021] proposed a distinct framework, which identifies different stages in the development process of a dataset. Each step should include specific documentation to include information considered necessary for appropriate accountability of the actors involved. 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, Hutchinson et al. [2021] 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.

Pushkarna et al. [2022] aimed to complement other artifacts by focusing on issues which cannot be directly inferred from the data. This concerns the context regarding the dataset, such as how it was collected/processed, or previous uses and how it performed on them.

2.3 Toolkits and Guidelines

In this section we address different methods and tools that have been proposed to identify issues in algorithms or datasets, coming both from academia and tech companies. Common topics have been the following: 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]. Proposals include toolkits, guidelines, and even service offerings. It is worth mentioning that much of the work cited in this section includes contribution from industry research centers. These works are not necessarily in the same line as the other ones cited, but they share the broad goal of improving fairness and accountability in the use of AI by providing practices and methods to be adopted during the design stage by developers.

Arnold et al. [2019] proposed *FactSheets* to create AI documentation, focusing on transparency issues of technical aspects and evaluations conducted during design and developments. Their proposal differs from others as it focuses on disclosing information and shortcomings not of a specific ML model, such as Model Cards (discussed in Section 3), but of the system as a whole. The questions in *FactSheets* concern 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] described a methodology for creating *FactSheets* that builds upon this artifact.

Also within IBM Research, Hind et al. [2020] presented two interview studies with developers to evaluate challenges in creating documentation for AI systems, and proposed recommendations on how to collect and report relevant information to improve efforts with *FactSheets*.

Another 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

³<https://www.microsoft.com/en-us/haxtoolkit/>

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] defended the use of checklists as a practical and simple way for developers to evaluate their work and whether they have steered away from ethical values. They offered a checklist to be used in data projects and placed 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.

Deng et al. [2022] provided an initial empirical exploration into the use of *Fairness Toolkits*, focused on bias issues, by conducting interviews with practitioners. They highlight the relevance of further study into the possibilities for collaboration and communication.

As will be highlighted in our discussion, we do not ascribe to the understanding that tools such as toolkits or checklists can appropriately address the whole of ethical reasoning that should happen. Regardless, they can be an useful tool to quickstart designers considerations. Thus their mention is important here.

3 Model Cards

Model Cards [Mitchell et al., 2019] are “*short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions [...]*”. They aim to increase transparency of artificial intelligence (AI) models by reporting their use cases, performance metrics, known shortcomings, and other considerations made by developers, as well as standardizing ethical practice and reporting. Such documentation may advise other developers’ on adopting of the model, thus reducing its deployment in unintended scenarios. Furthermore, they could also be used by other stakeholders, each with their own expectations and interpretations of the topics explored in the Model Cards. According to Mitchell et al. [2019], the main use by model developers will be for benchmarking and comparing the performance of their system to other models. In addition, policymakers could understand cases where an ML system may succeed or fail, and how it may impact people.

The Model Cards structure comprises nine sections:

- **Model Details:** general information regarding the software and its developers, including organization, type, and academic reference.
- **Intended Use:** the intended users of the software, the use cases envisioned by the developers, and explicit out-of-scope use cases, where the system could, in theory, be applied but are out of its application range for some reason, practical, functional or ethical.
- **Factors:** details about the model’s performance according to different factors, *e.g.*, the environment the model is deployed and the method for capturing the data. An important discussion in this section is related to the “groups” factor, which ought to include different characteristics that can be used to categorize and analyze 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.

- **Metrics:** information about the performance and tuning of the algorithm, including the reason for choosing those specific metrics of performance, the decision threshold, and how uncertainty and variability are dealt with and estimated by the software.
- **Evaluation Data:** the datasets used for evaluation, processing steps, and motivation for using that dataset.
- **Training Data:** the datasets used for training. However, the authors recognize this might not be feasible in many cases due to other interests, such as the data being proprietary.
- **Quantitative Analyses:** details on how the model performs with respect to each factor and their combined intersections, providing confidence intervals when possible.
- **Ethical Considerations:** the considerations that 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 are to be informed about them. The authors suggest the following questions be explored here:
 - *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.
- **Caveats and Recommendations:** included in order to address any concerns that were not covered in previous sections and are considered relevant to be disclosed.

Even with this predefined structure and the goal of standardization, the authors recognize the usefulness of the proposal and reliability in representing the model depend on the card’s creator. Furthermore, there is no guidance on how to fill Model Cards, including any benefit for reflection in following the determined order.

Despite arguing that Model Cards can aid developers to be forward-looking in analyzing their models, their creators suggest the tool should be used as a means to “[d]isclose information about a trained machine learning model”. Hence, they were thought to be completed once models were already developed, reflecting the options made previously during development. While developers may go back and modify their system after eventual reflections that occurred when creating the card, this tool was not intended for designers to engage with it *during* the development process.

Recent work has built upon Model Cards to generate related tools with distinct goals. Shen et al. [2022] provide a Toolkit for authoring simple model cards which focus at enabling deliberation and collaborative process around the use of AI in communities. However, they do not discuss the

ethical side provided in the original tool. They conducted interviews regarding its initial use in the context of Wikipedia edit reviews.

Crisan et al. [2022] propose a direct expansion to the tool. Building upon the creation of a Natural Language Processing model, the authors explore integration of Interactive aspects to Model Cards. They integrate aspects of Geburu et al. [2018] to expand the datasets information, as well as Goel et al. [2021] to address models' performance variation on different data.

For instance, they provide new sections that would allow other developers, who intend to use the model, to explore its performance on slices of new data. Then, the dynamic nature of the artifact would dynamically generate and compare certain aspects of the model performance, as well as the training set and new data.

4 Study design

We begin this section with some broader considerations about the purpose and object of our inquiry. We intend to contribute to the design of high-quality ML applications. Therefore, all of our beliefs about the meaning of **high-quality** affect and determine the path of our research. It follows that viewing research as a value-free, context-free quest for universal truths and laws is incompatible with our stance. Hence, we choose **qualitative research** as a necessary (albeit not exclusive) part of our overall inquiry design.

In qualitative research, we are not looking for causal explanations in the sense adopted by natural sciences (e.g., "A rise of the temperature of water to 100 degrees Celsius causes it to boil, changing from liquid to gas"⁴). Rather, as clearly stated by Gabriel [Gabriel, 2018, p. 138]: "*Unlike the quantitative researcher who seeks to understand the particular as an instance of the general, the qualitative researcher is seeking to discover the meaning of particular events and experiences, aiming to understand these phenomena as outcomes, intended or unintended, of meaningful human actions, emotions and intentions.*"

Considering the importance of biases in ML models and the applications they are built into, **understanding** subjective factors at play in the development of technology seems to be a crucial component for any useful theory in the field.

The study reported in this paper is part of an overarching investigation on how different representations affect ethical reflections during the conceptual design of AI systems. The selected strategy was to compare how a group of competent ML application developers used different representations in a realistic case, inquiring about their thoughts, choices, and decisions. Participants were therefore required to have some level of previous experience with developing such applications so that their reflections were comparable to those of everyday developers. However, we did not require them to have previous experience with any of the representations used. Moreover, given the exploratory character of the study and the scarcity of professionals from the industry who can dispose of their time and freely talk about their rationale and practices, at this stage of our project our participants were

chosen among student and non-student professionals working in academic R&D laboratories.

4.1 Study Procedure

To analyze their design decisions without being excessively constrained by the practical impediments of real-world development, we opted for a speculative design approach [Auger, 2013]. Figure 1 illustrates our study procedure, in which we used the Model Cards (MC) and the Extended Metacommunication Template (EMT).⁵

We randomly split our participants into four different groups of equal size, varying the order of (i) the scenarios and (ii) the tools used to reduce ordering effects. The systems they were asked to design should: (i) provide a risk assessment of people applying for loans at a financial institution; and (ii) assign a score to a university applicant during an admissions process to help a jury decide whom to admit into the program. We chose scenarios where the automated decisions could have a high impact on the subject's life. They were derived from examples explored in work that discusses existing issues in ML systems, such as Cathy O'Neil's Weapons of Math Destruction [O'Neil, 2016] and Virginia Eubanks's Automating Inequality [Eubanks, 2018]. By having our design scenarios be based on real-world situations where ML is used, participants ought to be able to draw from their own previous experiences in order to improve their final design.

The study comprised two individual sessions, at least one day apart. All sessions were conducted by one of the authors, and in two of them another one was present but only as a spectator in the video call, but did not say anything.

At the start of each session, we asked a few questions about the participant's background and whether they consented to the study's terms. We then presented the design brief of one of the scenarios and a summary of the Bioethical Principles [Beauchamp and Childress, 2019], to provide all participants with a baseline ethical framework that could support their reflections alongside the tools. We noted that participants were not *required* to use those principles, but *might* use them as a source of inspiration for their reflection. Our choice of this specific framework was guided by Floridi and Cowl's work [Floridi and Cowl, 2019], in which they claim that most ethical frameworks for AI could be reduced to the Bioethical principles, with the addition of an "explicability" principle. Thus, we did not opt for specific frameworks as proposed by Vakkuri et al. [2021]; Siqueira De Cerqueira et al. [2022].

After discussing the design brief and the ethical framework, the researcher conducting the session provided a brief explanation of the tool participants were going to use in that session so that they would have sufficient understanding to be able to use it correctly. This constituted allowing them to read the sections present in the Model Card, as well as their descriptions as originally provided in Mitchell et al. [2019]. The researcher would only read the text, if it was preferred by the participant or answer any potential questions they had.

⁴Note the evident universal, context-free character of the explanation embedded in this statement.

⁵This study was submitted and approved by the Pontifical Catholic University of Rio de Janeiro's Institutional Review Board, process #101/2020, protocol #125/2020.

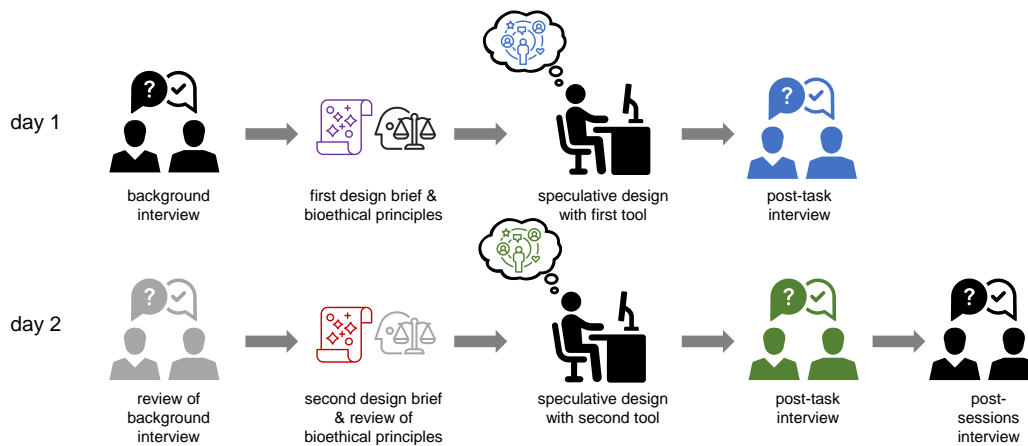


Figure 1. Study procedure

Then, we asked the participant to speculate about their design solution, recording their reflections and decisions in a design document with the relevant tool. We also instructed them to speak their thoughts out loud throughout the process, so that their rationale could also be captured in audio.

After the participant completed the task, we conducted a post-task semi-structured interview about their view of the scenario, of the generated document, and of how they believed it might have influenced their ethical reflections. At the end of the second session, we asked an additional set of questions that were meant to probe specific points of their experience, allowing us to compare their answers then with what they communicated while using each tool.

Each session’s audio and interactions with the document created were captured for later analysis. We then transcribed them *verbatim*, including any pauses and linguistic mistakes participants may have made. We included tags to represent any missing contextual information that would be needed to understand the corresponding statements, such as cues to when participants pointed us to a certain element of the document but did not say it out loud.

All eight interviews were conducted through video conference software, having occurred during social distancing. They lasted on average 01:22 (*max* = 01 : 39 and *min* = 01 : 00). Even though participants were in the same program, they were reached through social media channels (e.g. instant messages groups), and were not part of any specific discipline. Given the circumstances, we have no reason to believe they communicated between or before sessions.⁶

4.2 Coding and analysis

The transcripts and document generated by each participant composed the dataset for our analysis. We followed, as expected, the orientation of qualitative research. In this regard, it is useful to invoke the work of Bryant, one of Grounded Theory’s (GT) most eminent figures [Bryant, 2002, 2017, 2019, 2021], who has repeatedly warned the research community about some severe misunderstandings when using GT. One of these is the idea of *theory emerging from data*, or

using the logic of *quantitative induction* when dealing with meanings, not numbers.

With the transcribed data at hand, we conducted a Thematic Analysis [Braun and Clarke, 2012], looking for relevant themes that emerge from the participants’ experiences with the MC. Coding-wise, we took an inductive approach, allowing our codes to emerge from our initial contacts with the data, while also recognizing that abductive reasoning played a role, as discussed by Tavory and Timmermans [2014]. Since our study’s focus was on the participants’ ethical reflections, most of our final codes had significant ethical connotations, while other potential codes were set aside in our analysis.

Two of the authors coded the data. Each of them coded the data related to two participants (P3 and P4) individually and arrived at an initial set of independent codes, initially without a dedicated data analysis tool. These codebooks were unified in a meeting, and the documents of both participants recoded with the new codebook in a qualitative data analysis assistance software. One of the coders had also analysed the data of 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.

After this step, we maintained a shared codebook in case any changes arose from coding the remaining data. During this stage, the authors involved in the analysis directly contacted each other to define the scope of certain codes. This procedure was used mainly to define boundary cases of some of the codes, including whether some instances should be included or not. This was applied to around 10 excerpts.

After coding all of our corpus with this set, we made a modification in our codebook, splitting some codes to better represent the observed phenomena, e.g., denoting the bioethical principle mentioned in the coded excerpt when appropriate. This final change did not require to reanalyze all of the data, only identifying the relevant bioethical principle in the excerpts already identified. The final coding step was to consolidate the codes and annotations of both coders into a final codebook and re-code the data. After this last round of independent coding, we resolved any divergences via negotiated agreement [Campbell et al., 2013], discussing each case and reaching a consensus between coders.

⁶This is further supported by the fact that when asked, they mostly replied having no knowledge regarding the tools we used.

Table 1. Final codebook

Code name	Description
* considering ethical principles	
DIAGNOSIS OF ELEMENT OF ETHICAL FRAMEWORK	Diagnoses the existence of an ethical issue, and relates it to an element of an ethical framework (e.g., beneficence, non-maleficence, autonomy, or justice) <i>a posteriori</i> .
FRAMING BASED ON ELEMENT OF ETHICAL FRAMEWORK	Implicitly considers an element of an ethical framework (e.g., beneficence, non-maleficence, autonomy, or justice) as a lens to frame their reflection on some subject (context).
SCAFFOLDING AROUND ELEMENT OF ETHICAL FRAMEWORK	Explicitly starting from an element or principle of an ethical framework (e.g., beneficence, non-maleficence, autonomy, or justice), tries to find an issue that is related to it.
* reflecting upon the artifact use	
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.
UNDESIRE CONSEQUENCES	Defines possible consequences of the artifact's use they want to avoid.
SYSTEM'S AUTONOMY	Defines that the artifact's autonomy should be increased (or limited) in certain conditions.
*reflecting upon the artifact construction	
GUIDING VALUES	Identifies or asserts a personal value as a guide for designing the artifact.
ETHICS OF DEVELOPMENT PROCESS	Considers the (un)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.
* owning or delegating moral responsibility	
RESPONSIBILITY FOR ARTIFACT	Remarks about their feeling of responsibility for the artifact being developed.
AGENCY TO ARTIFACT	Attributes agency to the artifact they were developing.

Next, we analyzed the coded excerpts to better understand the meanings we saw in the data. From this understanding we were able to relate the emergent themes into a network of concepts that sheds light into how the participants took into account ethical issues when designing the systems proposed in the scenarios provided to them. In working this way, we sought to engage in the three practices Gabriel [Gabriel, 2018] described as fundamental to qualitative researchers: interpretation, reflexivity, and imagination.

We present our findings and analysis regarding the MC in the next section. We leave the analysis of the use of the EMT, as well as the comparison between the two, for future work.

5 Results

This section reports the results of our study: the profile of the participants who took part in it; the codes and categories that emerged from the analysis, and their relationships; the association between those categories and the Model Card sections and utterance modes; and an interparticipant analysis.

5.1 Participants

Participants' profiles were fairly homogeneous. Although this may be limiting in terms of the diversity of perspectives included in this study, it does reflect the current paradigm in ML development, where most professionals identify as male and hail from the areas of Computer Science, Computer Engineering, and Statistics. Our main constraint is the predominant academic profile, with lower industry experience, although this does not mean they did not have experience in ML development. Given this homogeneity, we were able to freely allocate them to each of our study's scenarios.

Eight participants took part in our study. Seven of them self-identified as men and one as a woman (P3). Four of them have bachelor's degrees in Computer Science (P4, P5, P6, P8), two in Computer Engineering (P1, P3), one in Statistics (P7), and one in Information Systems (P2). All had completed, or were pursuing, graduate degrees in Computer Science. Seven had previous professional experience with developing ML systems, with the other only having academic experience (P3), having been taught about it in an undergraduate course. One of the participants also had experience teaching undergraduate courses in ML (P7).

In terms of their familiarity with MC, only one participant (P1) had previous knowledge of MC but had never used it during development. Concerning their understanding of Semiotic Engineering (which grounded EMT, the other tool used in the broader study), two of them reported a moderate amount of understanding (P1, P4); five at least some knowledge on the topic (P2, P3, P5, P6, P8); and one had never heard of it (P7).

Most participants reported interest in the ethical issues involved in ML development, with only one (P6) stating that they had no interest in the topic. Despite their interest, most had never had any experience discussing these issues in real projects, with only one (P1) having more in-depth discussions in an academic setting.

5.2 Emerging Categories and Relationships

Table 1 presents the high-level categories derived from aggregating the consolidated codes. Ten categories emerged from the data: concerns with the *ethics of the development process*, the role of the developer's own *guiding values* and their *responsibility for the artifact*, the need to adjust the *system's autonomy*, attributing *agency to artifact*, and the *undesired consequences* for the *impacted individuals*.

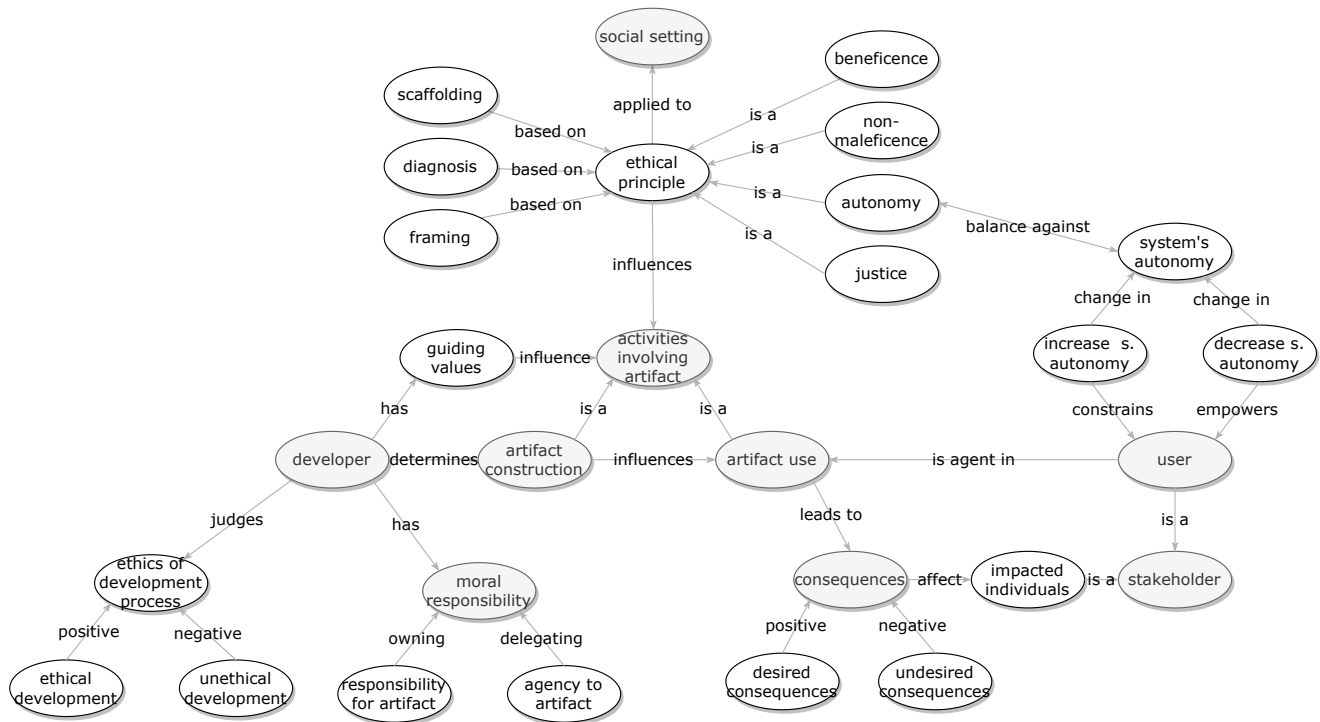


Figure 2. Themes that emerged from our interpretation of participants’ meanings

Moreover, participants made use of the ethical framework not only to explicitly provide a *a priori scaffolding* to their reflection or to enable an *a posteriori diagnosis*, but also implicitly, when *framing* some of their reflection upon an ethical issue.

Figure 2 presents the categories that emerged from the analysis. The white ovals depict the elements present in the codebook, and the shaded ovals depict additional categories that tie them together into a coherent structure to better represent participants’ meanings. These represent relevant aspects of the design and development of ML systems to which our categories were connected, e.g., the focus on stakeholders or the construction of the artifact. We also represent the relations between them as labeled directed edges identifying the connections between related entities.

The concepts and relations illustrated in this diagram can be seen in some of the quotes contained in our data. As we proceed with our discussions on the study’s results, we bring up relevant quotes to provide some grounding for our assessments. For example, the relation between the category “Decrease System’s Autonomy” and the ethical principle of “Autonomy,” which requires we respect our users’ abilities to make decisions for themselves, can be seen in the following statement “*It is not recommended to use the model in a fully automated system of credit analysis, since it has the potential of presenting unexpected and undesired behavior in certain cases. It is recommended to use the score provided by the model only as an additional parameter for the analysis made by a human.*” (P2, written). P2 highlights his reason to limit the artifact’s autonomy: to preserve the user’s decision-making capability.

It is these types of entities and connections that make up our understanding of the meanings within the analyzed data. Of course, we have also expanded beyond the occurrences

in the data via analytical reasoning. Complementarity, for example, is crucial when building a robust theory. Even if there were only occurrences of participants restricting the system’s autonomy, our theory should also be able to account for cases where the developer may want to expand it. So even if there were no occurrences in the data, it would be warranted to include the concept of autonomy expansion within our theory. These are the types of analytical expansions we have made in order to build a more robust and coherent theory.

5.3 Categories Associated with MC Sections

To better understand the role of the tool in evoking the meanings represented in the categories, we investigated the associations between the coded excerpts and the sections of the Model Card (MC).

As our codebook focused on ethical issues and concerns, it is no surprise that the MC section with the most coded excerpts was Ethical Considerations, followed by Caveats and Recommendations. This may have occurred for at least two reasons: (i) most participants filled out the Model Card template sequentially (despite having been told this was not a requirement), so they kept reflecting on ethical issues after having reflected upon Ethical Considerations; and (ii) the Caveats and Recommendations section explored a wider range of ethical issues as it pertained to the model’s appropriation.

It is noteworthy that, although the MC is supposed to document different aspects of the models “in writing”, most codes emerged from spoken remarks, either spontaneous (during the task of filling out the MC) or prompted (by the post-task interview). Although this is not surprising *per se*, it behooves us to examine possible patterns in what is included or omitted in each utterance mode. Some topics they were willing to discuss verbally but not write down for posterity.

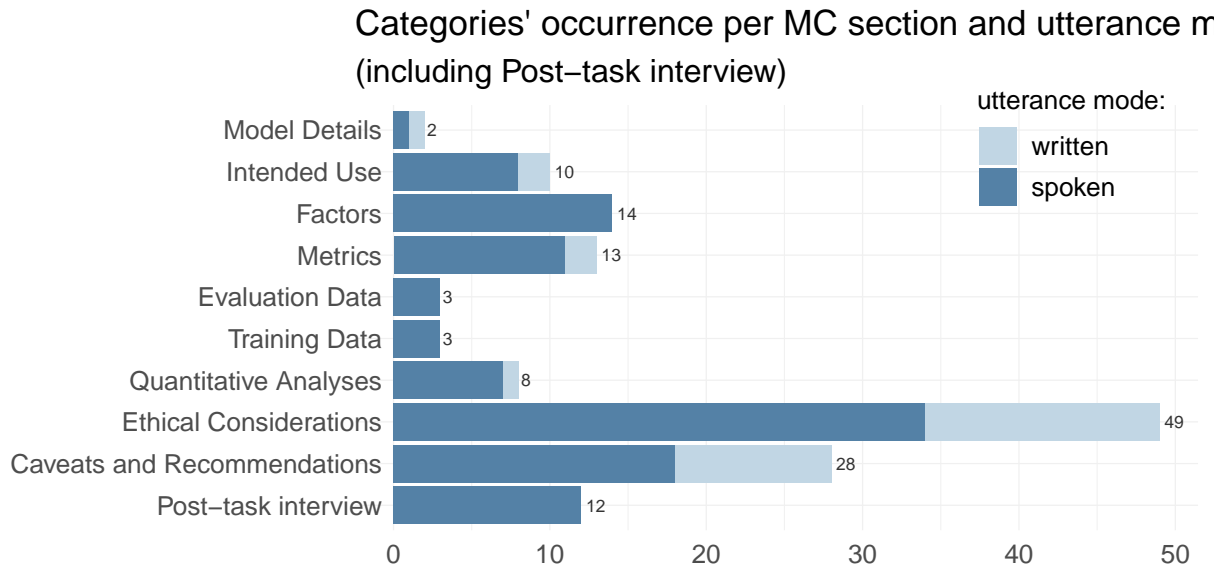


Figure 3. Distribution of coded excerpts per Model Card (MC) section, including the post-task interview, and per utterance mode (spoken vs written). The *written* documentation in some sections did not present any coded excerpts.

Figure 3 shows the distribution of coded excerpts per section of the MC, in the order they are presented in the tool, as well as those referring to the post-task interview. It includes utterances both in spoken and in written form, transparent blue. We see that the distribution is concentrated towards the last sections of the MC, and in the other categories rarely occurred in written form.

5.4 Categories Across Utterance Modes

The frequency of each code in our data is displayed in Figure 4, including its utterance mode. We found participants most frequently expressed their desire to **Decrease the System's Autonomy**, as well as that other ethically relevant categories were predominant in spoken mode. Critically, **Unethical Development** never occurred in written form.

Figure 5 adds to this, distinguishing the category codes in each section, both in spoken and written utterances. In particular, we see a large difference in the category **ethics of development**, which the code unethical development occurred only in spoken form while ethical development occurred in both forms.

The most frequent category was **System's Autonomy - decrease**, which characterized instances where participants stated they believed the autonomy of the artifact under development should be limited in certain decision-making processes. For instance, P1 first stated that the final grade attributed to candidates in the education scenario should be an aggregate of scores for different attributes, but later recorded in their Model Card: “*At a second moment, the grades would be issued and a committee would ~~make a manual attribution~~⁷ calculate the final grade.*” (P1, written). Other participants were even more explicit about this issue, stating that the output of the artifact should not be used without supervision:

⁷In our analysis, we format in strikethrough text fragments that the participant inserted and later erased.

“*The model will aid the decision-making process of the [financial] institution by offering a value related to the risk of each client, but the model should not decide whether the loan will be granted or not*” (P3, written); and “*[t]he result of the selection should not depend solely on the grade given by the model*” (P4, spoken & written).

Another common topic was the **Ethics of Development Process**. These were concerned with the ethics of certain decisions made during the development process. Regarding *ethical* choices, P8 declared “*OK, I believe 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 another*” (P8, spoken). In contrast, they also recognized some *unethical* choices regarding the use of traits that may be used to discriminated against individuals, affirming “*These[social class, race and sex] are information I should be careful because I could be reinforcing existing prejudices that already exist*” (P8, spoken).

We noticed a discrepancy in occurrence between the two utterance modes (spoken vs written). Even for some of the most frequent categories, their occurrences were completely limited to the spoken remarks, such as **Unethical Development**, **Impacted Individuals**, **Agency to artifact**, and **Responsibility for artifact**. Although participants did delve into these issues in their own reflection during the design task, they did not document the results of said reflection in the MC. This is especially interesting considering that all participants, when asked at the end of the interview, answered they did not exclude from the MC document any information they believed was relevant, indicating the exclusion was a conscious choice made by participants based on what they found relevant to record. Naturally, we did expect a larger number of spoken than written utterances, as participants orally explored design alternatives in their reasoning *process* as it unfolded. However, we also expected that at least the final *product* of such reasoning would be recorded in the MC, but this was often not the case.

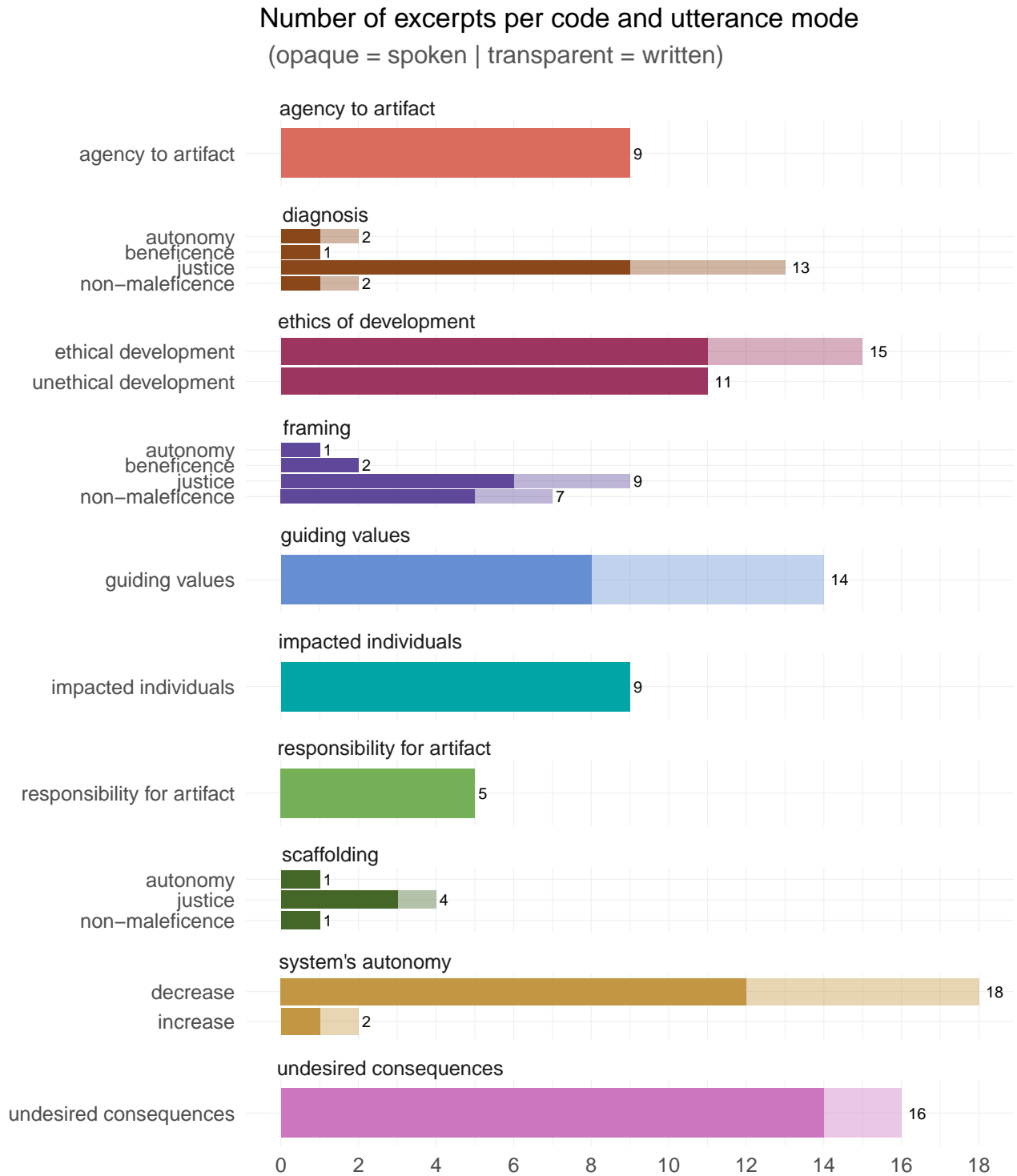


Figure 4. Distribution of coded excerpts per code and utterance mode.

An example of the **Unethical Development** category that occurred during participants' reflection, but was not transferred to the MC, was: *"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"* (P2, spoken).

Later, when filling in the MC template, the participant raised the possibility of the artifact using a spurious relation between the address and income of the loan requester to deny future clients, without making explicit in the document any ethical issue related to the use of this piece of information. Therefore, only the spoken excerpt was coded, as the resulting written documentation did not include an ethical assessment of that piece of information.

The same occurred with P3, who made the following statement during the interview regarding the inclusion of age in the artifact input: *"When I was in this part here [Quantitative Analyses] I removed it, because I saw that it was not something... fair."* (P3, spoken). Despite this realization during the interview and the reflection guided by the Model Card, P3 did not explicitly mention any problem related to the use of this feature. In fact, it was left among the variables of the dataset in the Model Card.

The fact that no instance of **Impacted Individuals** was documented in the MC is also interesting. It may indicate that its sections, including the section on Ethical Considerations, did not draw attention to the individuals that could be impacted by the artifact under development. P6 made two spoken remarks that considered other actors that could influence the use of the artifact under development: *"maybe family members that have a bad financial history and this end up... influencing the answer of the system to that person"* (P6, spoken) and *"[i]magine that a client has a, perhaps their parents have a bad financial history, but as they are just entering the market now, the thing is being used by his parents in order to receive the loan, do you understand?"* (P6, spoken). This code was most frequent during our post-task interviews, where participants were explicitly asked about the potential benefits of the filled-out document to anyone that might be impacted by the artifact described in the MC. This indicates a lack of attention on the part of the developers to the individuals (users, stakeholders, and others) who will be impacted by the product of their work. However, when prompted to think about them, developers were able to identify noteworthy topics.

Another category that presented a similar pattern was **Undesired Consequences**, only occurring twice in writing despite having the most frequent occurrence across all categories. This code was attributed to excerpts where participants identified possible consequences of their system that they deemed undesirable, regardless of the reason why. An example would be P1's utterance during the session, which was also registered in their MC: *"Other information not mapped may be... may be inferred, it is necessary [typing into the MC]. 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 they publish, which are the main vehicles in which they publish, 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" (P1, spoken). The equivalent excerpt in the MC received the codes of **Undesired Consequences** and **Framing based on the non-maleficence principle**, since, in writing, P1 framed that remark around not causing damage by misusing the collected data.

In contrast, P8 verbally expressed a potential consequence of including certain information into their model when thinking aloud, but did not register it in the final document. Reflecting on possible discriminatory results, they 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 their education, because they may need to work for example, and had lower grades. It is not direct, but it is more indirect."* (P8, spoken).

We hypothesize a few different reasons why some instances of participants' reflection were not documented in the MC. For instance, participants may have found the tool unsuitable to include such information, or participants are more reluctant to commit, in writing, to decisions involving sensitive ethical issues. Further investigation on this matter is required for us to better understand the underlying causes of this phenomenon. With this knowledge in hand, we may then devise strategies and tools to better capture developers' ethical reflection and related decisions.

5.5 Interparticipant Analysis

Almost all participants expressed the desire **limit the System's Autonomy**. All of them discussed the **Ethics of their Development** process, while only P2 found no ethical decision, he concentrated most instances of discussing problems of unethicity –characterizing a atypical participant in our analysis. In only one participant we found no instance of . The distribution can be seen in Figure 6 shows each code frequency per participant and mode.

All but one of our participants (P5) expressed an intention to **limit the System's Autonomy**. They generally stated that the artifact should not have final say over the decision-making process, e.g., *"[t]he result of the selection should not depend solely on the grade given by the model"* (P4, spoken & written); *"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."* (P6, written).

On the flip side, only one participant (P5) expressed an interest in **increasing the System's Autonomy**, albeit conditionally: *"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."* (P5, spoken). Despite this, they also expressed their general belief in limiting the autonomy of the artifact: *"I am always a bit uneasy to make something completely autonomous, completely automated"* (P5, spoken). It seems that this participant interpreted the design brief as asking them to create a fully automated system, which they then stated was against their own personal beliefs.

Categories' occurrence per MC section and utterance mode (including Post-task interview)

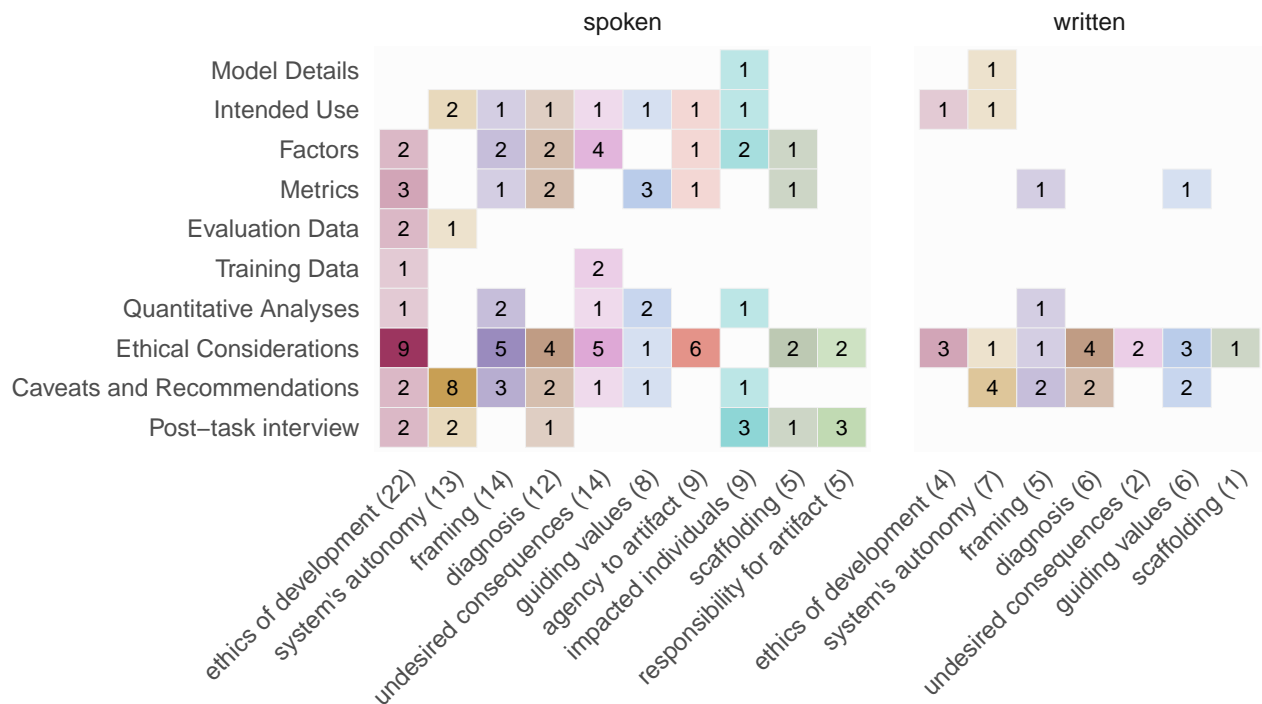


Figure 5. Distribution of categories per Model Card (MC) section, including the Post-task interview, and utterance mode (spoken aloud or actually written in the document). Colors used were the same as in Figure 4, and represent the category.

Another notable aspect is that, excluding P2, all participants analyzed the ethics of certain development choices (**Ethical Development** or **Unethical Development**). Being mindful of possible consequences of their design choices was a recurrent theme across participants, but strongly emphasized by P1 and P8 who, incidentally also emphasized the need to **limit the System’s Autonomy**.

P1 reflected intensely on how to avoid possible harm or damages that arise from the use of the artifact under development (**Framing based on element of ethical framework - non maleficence**). This focus on the consequences of the artifact use stood in contrast to P8, who adopted a more personal stance by explicitly describing their values that guided their design (**Guiding Values**). These also indicate that they were reflecting on some of the ethical connotations of the development process by taking a value-based approach. An exemplary instance of this last category was: “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 to those who, 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.” (P8, spoken)

Regarding 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 P7.⁸ Conversely, codes based on the non-maleficence principle were only present for P1, P3, P6, and P8, and based on the beneficence principle in P3, P4, and P6, while also having lower absolute frequency.

Examining differences across utterance modes, P7 stood out for having no coded excerpts in their writing, only in their speech. This suggests that this session was exceptional in a way. One characteristic of this interview, which corroborates its peculiarity, was the choice made by P7 to define their scenario thinking mainly of corporations, and not individuals, which may have differentiated their reflection from other participants. This was showcased in the following excerpt: “It was... an option, I suppose, I flipped a coin. It could be a model for individuals or corporations. Then I chose in the scenario to be focused on corporations.” (P7, spoken), which was coded as **Impacted Individuals**.

5.6 Co-occurrence Analysis

We also explore the co-occurrence of codes in our data. We found that within the most common pairs, frequently one would be related to the **Ethics of Development Process**, which may have served as a starting point for further reflection.

We consider each interview and the corresponding filled Model Card as separate documents. We only considered codes as co-occurring when excerpts were identified in the

⁸Here we include excerpts identified as related to the system’s autonomy, since the more common option to restrict it directly implies the preservation of the autonomy of humans involved in the process.

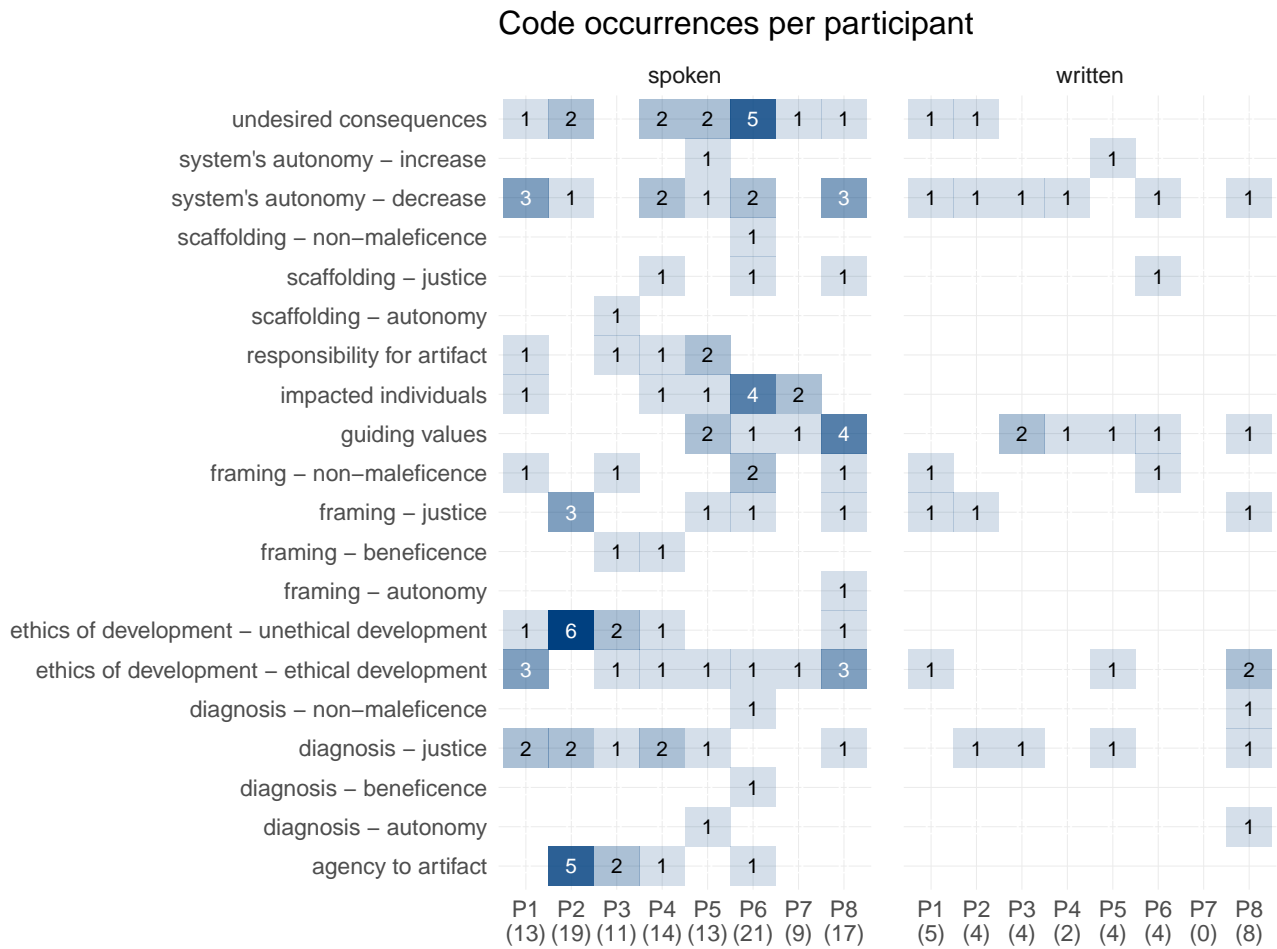


Figure 6. Code frequency per participant and utterance mode

same Model Card section following the structure described in Section 3. Using this information, we were able to explore relations between codes that might have been shared between different participants.

Regarding different participants, the most common pair of codes was **Diagnosis of Element of Ethical Framework - Justice** and **Ethics of Development Process - Unethical Development**, which occurred with five participants. This is a very high rate, considering these codes were present for six and five participants, respectively.

Thus, 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. This pairing suggested that participants were relating identified characteristics of development they considered unethical, and basing this assessment on an issue they identified as a problem of justice.

P2 provides excellent examples of this situation. First, they stated during the interview that “if [the company] is going to maximize its profit, it will exclude people with lower education levels, because they know they would have more losses. And, overall, people with lower levels of education are more humble. Thus, it turns out that this more humble group of people is gonna remain without credit, if compared with people with more money. It is an ethical discussion.” (P2, spoken).

Later, they expanded on the issue. “Well... with basis on the address there is this issue... [writing] ‘there is a chance the model, when using the address as a training data point, will have the tendency [to present discriminatory patterns]’” (P2, spoken).

Another participant for which this occurred was P3. First, they 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.”. Then, they immediately followed by identifying it as potentially causing discrimination, a problem of justice: “Exactly. Thus, I think it could lead to a certain... in a 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 interpret this as participants opting to take actions in the development process they view as ethical, with the goal of avoiding harmful results.

We observed this relation in the Ethical Consideration section of P1’s Model Card document. The participant first wrote, regarding the dataset used, “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." (P1, written). However, despite the effort to avoid identification, they state that some information 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.*"

P8 had a similar sequence of thoughts: "*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 others.*" (P8, spoken). Following: "*I believe I also need to ensure anonymization of the data, for issues like data leaks, or for example to keep information like academic history and family income from individuals.*" 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.

6 Discussions in view of Current Research

We now discuss our findings and contextualize them with relevant related work and our own positionality. We focus on the MC's impacts on ethical reflection, the participants' views on limiting AI autonomy, the perceived need for human supervision of AI performance, and the developers' stance toward their own responsibility.

6.1 Researchers' positionality

When reflecting on our findings, it is essential for us to take our positionality into account. How we perceive the data is influenced by our past experiences and the worldview they create. Two researchers were responsible for the coding and it is their background that is most important for us to accurately frame our analysis.

In terms of their identity, both of them are white men from a high socioeconomic class. This can indicate a possible blind spot towards issues related to the experiences of other genders and socioeconomic classes. They were both educated in private schools and universities, with one of them having a bachelor's in Computer Engineering and the other one in Law. Both of them were graduate students in Computer Science at the time of the analysis. This difference in undergraduate backgrounds increased the diversity of considerations made while looking at the data. In addition, although the computer engineer had greater experience with qualitative methods, the lawyer was also experienced in the qualitative consideration of data, albeit not in a formal research setting. The topic of responsible design of AI systems is at the core of both researchers' work. Besides, both of them took part of several academic discussions on MC in multiple graduate courses.

Another factor that may have had some influence in our analysis was the coders' familiarity with the study's participants. All of them were graduate students in the same Computer Science department. In some cases, they took the same courses or even worked together on research projects. As such, the researchers may have been influenced by their own preconceived judgments about the participants when coding

their transcripts. This effect may have been reduced by the anonymization of the transcripts, but since the coders were also the interviewers, it is entirely possible that they were able to recognize who the transcripts were from during the coding process.

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

6.2 Ethical reasoning

The main goal of this study was to analyze how the MC might impact participants' ethical reasoning. As they proceeded through the speculative design session, they had to make several decisions, many of them with significant ethical connotations. Looking at our resulting theory and examples from the transcripts, we can make a few observations as to how this sort of reasoning took place.

A first finding was the prevalence of reflections on the ethics of the development process, looking at both ethical and unethical actions. This indicates that they were taking a more critical approach as to which steps to take while developing the tool, and not only looking at the impacts of the artifact after it is constructed. However, these were not equally distributed across participants, with their backgrounds potentially playing a role in this trend.

Most of the utterances concerning deliberations over the ethics of development choices occurred in spoken form and did not make it into writing. Furthermore, positive (*i.e.*, ethical) stances were much more likely to make it into the written document than those considered to be unethical. This is relevant because it runs counter to the purpose of the MC, which is to provide reliable documentation of ML models, including any significant ethical issues. Lacking this information, other developers that depend on the MC to understand the models they will be using may end up opting for these decisions that were previously deemed to be unethical, leading to consequences that the model's original creators sought to avoid.

Disclosing and communicating this information is essential for embracing and committing to ethical practices. It increases the development process' transparency, showcasing designers' decisions. Recognizing what development options were actively avoided due to their unethical nature or negative potential results can be as important as noting what these final decisions were. As creators of a model, developers should strive to ensure that they are integrated into different systems appropriately. The MC tries to support this practice by not only having a section for Ethical Considerations but also one for Caveats and Recommendations.

Documenting these aspects of the system's design rationale could also be seen in existing guidelines and toolkits for the development of AI systems. Let us consider the HAX Toolkit, proposed by researchers at Microsoft, as an exam-

ple.⁹ It includes the Guidelines for Human-AI Interaction, which provide recommendations for the development of systems that use AI models and have users interact with them in some manner. Many of these, especially those related to transparency, might be satisfied via the use of a document like the Model Card. For example, users trying to understand why the system is behaving in a certain way may gain greater understanding by reading about the developer's design choices and how they believed the system should and should not behave. Another recommendation in the HAX toolkit is to make it clear to stakeholders why the system may have behaved inappropriately. If they have access to the system's design documentation, including all of the identified risks, they might be able to locate the sources of observed behaviors. For instance, it could be that developers identified the possibility that certain inputs could lead to mistaken predictions by the model. Being aware of this, stakeholders in a problematic situation could then look at the model's inputs and see if they fall within that scope. Communicating risks presented by the design choices can be useful; however, as we have seen in our study's results, developers may not actually include them in their written documentation, even if they did consider them. This is especially notable, as the MC provided them with a field for Caveats and Recommendations, where these sorts of statements should reside.

Another way in which documentation tools, such as the MC, can be used is less prescriptive and more reflective. Toolkits and guidelines are often used to facilitate software development by prescribing fixed solutions. However, they may become unable to deal with the variety of contexts for which these systems are developed. Dealing with the incredible amount of variability that developers face may require deep reflection on multiple options, instead of seeking ready-made answers. Documentation tools, via their structure, can nudge developers to engage in these sorts of reflections, registering their conclusions as they proceed. From what we have seen in our results, the Model Card by itself, as it is presented in its original paper, does not seem to be sufficient to promote this sort of reasoning. However, it is possible that by presenting it in a different way, these incentives for greater reflection can occur.

In the case of our study, results show that even when ethical principles, which are not entirely prescriptive but still propose some guidance, were presented alongside the Model Card it was not enough for participating developers to engage in deeper ethical reflection. This supports a view of ethics not only as a procedure, with a list of principles to be considered at every turn, but also as a continuous deliberative process about actions being taken and decisions being made. Taking a similar approach, Guillemain and Gillam [2004] discuss the relevance of reflection in a research setting, which we believe also highlights its relevance for development situations.

This reflective process may be better served by tools that question and stimulate developers to reflect on the choices they make in building their system. As such, we, as researchers, should investigate how developers make use of tools like the MC and others [Mitchell et al., 2019; Barbosa et al., 2021; Shen et al., 2021; Raji et al., 2020] and how they

impact their own forms of reasoning. Ethical deliberations have become more and more crucial as AI systems take up a greater role in social mediation in today's world. Tools may not be enough for developers to be more concerned with the ethical connotations of their work, with organizational changes also being required, but they can still play a significant role.

6.3 Limiting AI autonomy

A large majority of participants expressed views on limiting the autonomy of the artifact being developed (**system's autonomy - decrease**). Even P5, who was the only one to comment that the system could have greater autonomy under certain conditions, expressed their distrust in giving it complete autonomy. The sentiments expressed by participants in our study are in line with what was seen in Brandão et al. [2019]'s study, when researchers were explicitly asked about the potential impacts of the model and the importance of its social context. Only when prompted did they start to recognize the social dimension of the algorithm they were supposed to develop, including issues such as how the decision process could be interpreted, how it could impact society, and how they ought to communicate with affected individuals. P2 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. [...]*" (P2, spoken). This is an interesting contrast to the opinions of the participants in Brandão et al. [2019], who initially expressed a high degree of trust in the artifact they were developing, the metrics they chose to evaluate its performance, and the data used to train it.

While participants in [Brandão et al., 2019] initially appeared to trust the models they were developing, participants in our study opted to preserve the individuals' autonomy given the possibility of the model's improper behavior. We posit this difference may have been caused by two different factors. First, the use of the tool in conjunction with our initial questions may have shifted their focus to these risks, making ethical considerations more salient to them. Second, it may have been due to differences in both studies' participants' backgrounds. While the interviewees in [Brandão et al., 2019] were part of a private research center, participants in our study were graduate students in an Informatics Department that offers courses that touch on some of these ethical issues present in computing. Moreover, only one of their participants acknowledged algorithmic bias, or other related issues, as a topic of study, whereas at least half of our participants noted their prior interest in this topic.

6.4 Human Supervision of AI Performance

In addition to deciding that the autonomy of the artifact they were building ought to be limited, participants often expressed that this should be achieved by having another person validate the decisions. This person, designated to supervise the AI system's output and the consequences of its use, would therefore be responsible for the final result of the decision-making process. This finding matches those in Brandão et al. [2019]

⁹<https://www.microsoft.com/en-us/haxtoolkit/>

at the final stage of their study. Analyzing their participants' statements, they noticed this sentiment, which they expressed in the following sentence: “[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 deep learning-based technology for applications like BackSys”. Similar concerns can be observed in P2’s comments quoted in the previous subsection, and P8’s comment that the system should include a way to “[...] go through someone’s manual monitoring” (P8, spoken). We can interpret this as participants acknowledging their limited knowledge about the system’s potential behaviors and trying to share responsibility for its results with others, despite being the ones who made crucial decisions about its design. Seven of our participants had at least one comment in this direction. Ultimately, they were still wary of allowing the model to make decisions on its own, relying on the stakeholders involved to validate them.

6.5 Developers’ Responsibility

Participants tried to distance themselves from the consequences of their design choices. They did not seem to acknowledge their level of agency in their statements. An indication of this hesitation was that, even though participants often spoke from a first-person perspective, they seldom wrote in first person, rendering their role as designers in shaping the artifact ambiguous. The only exception occurred in the Intended Uses section: “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)[...]” (P1, written). All other considerations written in the final documents appeared to hide their role in some way.

When filling in the MC, participants mostly used the passive voice or wrote sentences with the model as the agent. For example, P3 wrote in Ethical Considerations: “The model will treat all clients equally, considering only economic factors.” (P3, written) and P8 wrote in Caveat and Recommendations: “It is recommended to have a step for manual revision/monitoring/evaluation” (P8, written). Since using passive voice is a strategy for dodging responsibility using language [Lakoff, 2001], this discrepancy between the spoken (usually active) and written (usually passive) utterances may indicate a limitation of the use of written documents over capturing spoken remarks.

Ascribing responsibility depends on recognizing who has agency and knowledge in a given situation [Talbert, 2019]. By using the passive voice, developers are able to hide their own role in making the decisions and building the system. However, they do have agency and are, therefore, responsible. In terms of their understanding of the situation, we can also see that they mentioned certain possibilities in their speech but did not write them down for others to see. This could also be a sign of an unwillingness to admit that they are aware of some of the risks that their systems present, which would also add to their responsibility if they caused harm. We did not instruct participants to write their considerations in the MC in any specific way, as Mitchell et al. [2019] provided no such guidance. Before the design task, we informed them it was a document for recording relevant decisions about the

AI model’s development. However, the lack of use of first person in the documentation was surprising to us, given how often they adopted this framing when speaking.

Holding AI systems accountable requires that we understand the responsibilities of those who built them [Wieringa, 2020]. If harm is being done because of an individual’s choices, they should be made aware of it and encouraged to change their decision-making. This could not only improve the quality of the system in question but also help them grow as professionals, acknowledging the impacts of their actions. There is also a cultural aspect that we need to consider. If developers expect to be held accountable for the impacts of their choices, they may reflect more before taking action.

The tools offered to AI developers should contribute to their ethical reflections on the potential consequences that may arise from using the AI system under development. One of the virtues of the first-person perspective is that it emphasizes the role played by developers and their responsibility for most outcomes of the system’s use, as discussed in [Barbosa et al., 2021]. By having developers adopt this framing in documentations, such as the MC, it is possible that their increased sense of responsibility will also lead to deeper reflection.

Another issue we found with the language used by participants in this study was that it de-personified users and other individuals affected by the algorithm. It was evidenced by the high frequency of the code of **Impacted Individuals** in our overall data, but with a single occurrence in the written document. Since ethical and moral responsibility rely on the relationship between the agent (developers) and the patient (users and stakeholders), by erasing their own role in the development and hiding those that might be impacted, participants were able to hide both sides of the equation, making it even more difficult to recognize the responsibilities involved.

7 Conclusion

In this paper, we presented the results of a speculative design study with eight participants. By conducting a Thematic Analysis of the data, we were able to identify trends and categories that illustrate how their ethical reflections were impacted by the Model Card (MC) [Mitchell et al., 2019]. Sessions were based on two hypothetical development scenarios with significant ethical implications. We presented our analysis of not only the written documents but also the transcripts of each session’s audio, focusing on signs of ethical reflection.

One of our main findings was a contrast between what our participants chose to include (or not) in the MC. While participants deliberated on the ethics of their development decisions in spoken form, they only recorded those that they considered to be ethical. They would not include options that they deemed unethical and thereby rejected. 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 can be equally valuable for the goal of expanding our knowledge about AI systems and promoting their fairer use, especially considering potential reuse and re-purposing of these algorithms, as identified by Brandão

et al. [2019, p.24].

Furthermore, we found evidence that reinforces the perceived relevance of human mediation of algorithmic performance, in line with Brandão et al. [2019]’s findings. While participants in our study were aware of the potential impact and meaning their artifact could have, they still appealed to a third person to mediate their system. This was the case whenever they expressed that the model they were building should not have full autonomy over the decision process, and that a third person should be responsible for validating or checking the output.

8 Future Work

In terms of future work, our study may serve as a first step that leads to broader empirical studies into how different tools and documents can impact developers’ ethical reflections during the development of AI systems. We sought to understand how MC contributes to this form of reasoning, as well as identify which ethical issues it helped participants focus on, but various other tools remain unexplored.

An important aspect of technology development also remains unexplored: how can tools and documents contribute to collective reflection, a. If our participants may have been unwilling to express unethical points in written form, might that play a role in individual expression in ethical discussions? This is an important questions which we did not address.

We cannot directly compare our results with those reached in Barbosa et al. [ming] regarding the use of the Extended Metacommunication Template (EMT), it seems each tool provided distinct contribution to the ethical reflection in the scenarios. These results may help developers choose and adopt the MC, the EMT or other available tools into their development process while taking greater advantage of its potential to assist in ethical deliberations. Of course, ethical practice goes beyond design tools and documents, requiring an adequate environment and organization for it to take place, but, if used well, tools can assist in those deliberations.

On another side, the research for other tools and with the Model Card itself ought to be expanded in the involved stakeholders. One important extension of our work is verifying how developers with different profile might behave, such as industry experts, as well as other stakeholders. These results might help developers adopt the Model Card into their development process, making the most of its potential to assist in ethical deliberation.

9 Limitations

This directly connects to an important limitation of our work: the profile of our participants. Most of them had a relatively similar background, acting as developers of system that included machine learning models but within an academic environment, and being graduate student themselves.

Furthermore, compared to other published studies, participants were familiar with ethical considerations in machine learning scenarios, even if they had not themselves engaged in a critical exercise of this nature. Even though some of them

expressed having read the original Model Cards paper (or an instance of the tool in a documentation online) they did not consider themselves familiar with it.

Another aspect that might have influenced our results was our choice of ethical framework to be presented to participants, which also represented our own frame of result analysis. To minimize this we provided information and examples in the domain of medical decisions, coupled with the generic framework provided by bioethical principles. In many instances these principles were not explicitly mentioned by participants, although they probably influenced their analysis by preemptively raising their attention to such issues.

Finally, the Model Card itself includes a section on Ethical Considerations, which may have played a role with other factors. We cannot dissociate them, but inasmuch as this prompted the behavior of our participants this might be a desirable epistemic behavior prompted by the tool. We consider especially relevant to include and prompt further study into the divergence we observed between spoken and written remarks: behavior seen as unethical was rarely registered into the tool, not even as an alert to other stakeholders.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Simone Barbosa thanks CNPq for the grant # 313049/2021-1.

References

- 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.
- Auger, J. (2013). Speculative design: Crafting the speculation. *Digital Creativity*, 24(1):11–35.
- Barbosa, G. D. J., Nunes, J. L., de Souza, C. S., and Barbosa, S. D. J. (forthcoming). Investigating the extended metacommunication template: How a semiotic tool may encourage reflective ethical practice in the development of machine learning systems. In *Proceedings of the 22nd Brazilian Symposium on Human Factors in Computing Systems (forthcoming)*, IHC ’23, pages 1–12, New York, NY, USA. Association for Computing Machinery.
- 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.
- 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., Mo-

- jsilovic, 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.
- 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*.
- 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, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, pages 57–71. American Psychological Association, Washington.
- Bryant, A. (2002). Re-grounding grounded theory. *The Journal of Information Technology Theory and Application (JITTA)*, 4(1):25–42. The last issue of JITTA appeared by the end of 2018.
- Bryant, A. (2017). *Grounded Theory and Grounded Theorizing: Pragmatism in Research Practice*. Oxford University Press, New York.
- Bryant, A. (2021). Continual permutations of misunderstanding: The curious incidents of the grounded theory method. *Qualitative Inquiry*, 27(3-4):397–411.
- Bryant, Antony, C. K. (2019). *The SAGE handbook of current developments in grounded theory*. Sage, Thousand Oaks, California.
- 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.
- Crisan, A., Drouhard, M., Vig, J., and Rajani, N. (2022). Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, pages 427–439, New York, NY, USA. Association for Computing Machinery.
- De Souza, C. S. (2005). *The semiotic engineering of human-computer interaction*. MIT press.
- de Souza, C. S., de Gusmão Cerqueira, R. F., Afonso, L. M., and Ferreira, J. S. J. (2016). *Software Developers as Users. Semiotic Investigations on Human-Centered Software Development*. Springer International, Cham, Switzerland.
- Deng, W. H., Nagireddy, M., Lee, M. S. A., Singh, J., Wu, Z. S., Holstein, K., and Zhu, H. (2022). Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, pages 473–484, New York, NY, USA. Association for Computing Machinery.
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, Inc., USA.
- Floridi, L. and Cowl, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 1(1):1–15.
- Gabriel, Y. (2018). Interpretation, reflexivity and imagination in qualitative research. In Ciesielska, M. and Jemielniak, D., editors, *Qualitative Methodologies in Organization Studies: Volume I: Theories and New Approaches*, pages 137–157. Springer International Publishing, Cham.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2018). Datasheets for Datasets.
- Goel, K., Rajani, N. F., Vig, J., Taschdjian, Z., Bansal, M., and Ré, C. (2021). Robustness Gym: Unifying the NLP Evaluation Landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.
- Guillemin, M. and Gillam, L. (2004). Ethics, reflexivity, and “ethically important moments” in research. *Qualitative Inquiry*, 10(2):261–280.
- 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*.
- 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.
- Lakoff, P. o. L. R. T. (2001). *The Language War*. University of California Press, Berkeley, first edição edition.
- 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 *Proceed-*

- ings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pages 220–229, New York, NY, USA. ACM. event-place: Atlanta, GA, USA.
- Nunes, J. L., Barbosa, G. D. J., de Souza, C. S., Lopes, H., and Barbosa, S. D. J. (2022). Using model cards for ethical reflection: A qualitative exploration. In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems, IHC '22*, pages 1–11, New York, NY, USA. Association for Computing Machinery.
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group, USA.
- Pushkarna, M., Zaldivar, A., and Kjartansson, O. (2022). Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, pages 1776–1826, New York, NY, USA. Association for Computing Machinery.
- 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.
- Richards, J., Piorkowski, D., Hind, M., Houde, S., and Mojsilović, A. (2020). A Methodology for Creating AI Fact-Sheets. *arXiv:2006.13796 [cs]*.
- Rostamzadeh, N., Mincu, D., Roy, S., Smart, A., Wilcox, L., Pushkarna, M., Schrouff, J., Amironesei, R., Moorosi, N., and Heller, K. (2022). Healthsheet: Development of a Transparency Artifact for Health Datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, pages 1943–1961, New York, NY, USA. Association for Computing Machinery.
- Seck, I., Dahmane, K., Duthon, P., and Loosli, G. (2018). Baselines and a datasheet for the Cerema AWP dataset. *arXiv: 1806.04016*.
- 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.
- Shen, H., Wang, L., Deng, W. H., Brusse, C., Velgersdijk, R., and Zhu, H. (2022). The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, pages 440–451, New York, NY, USA. Association for Computing Machinery.
- Siqueira De Cerqueira, J. A., Pinheiro De Azevedo, A., Acco Tives, H., and Dias Canedo, E. (2022). Guide for Artificial Intelligence Ethical Requirements Elicitation - RE4AI Ethical Guide. In *Hawaii International Conference on System Sciences*.
- Talbert, M. (2019). Moral Responsibility. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Online, winter 2019 edition.
- Tavory, I. and Timmermans, S. (2014). *Abductive Analysis: Theorizing Qualitative Research*. University of Chicago Press, Chicago, illustrated edition edition.
- Vakkuri, V., Kemell, K.-K., Jantunen, M., Halme, E., and Abrahamsson, P. (2021). ECCOLA — A method for implementing ethically aligned AI systems. *Journal of Systems and Software*, 182:111067.
- 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.
- Wieringa, M. (2020). What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*, pages 1–18, Barcelona, Spain. Association for Computing Machinery.