




MoReXAI - A Model to Reason About the eXplanation Design in AI Systems

Niltemberg de Oliveira Carvalho^(✉), Andréia Libório Sampaio,
and Davi Romero de Vasconcelos

Universidade Federal do Ceará, Campus de Quixadá, Ceará, Brazil
niltemberg@gmail.com
<https://www.quixada.ufc.br/>

Abstract. The interest in systems that use machine learning has been growing in recent years. Some algorithms implemented in these intelligent systems hide their fundamental assumptions, input information and parameters in black box models that are not directly observable. The adoption of these systems in sensitive and large-scale application domains involves several ethical issues. One way to promote these ethics requirements is to improve the explainability of these models. However, explainability may have different goals and content according to the intended audience (developers, domain experts, and end-users. Some explanations does not always represent the requirements of the end-users, because developers and users do not share the same social meaning system, making it difficult to build more effective explanations. This paper proposes a conceptual model, based on Semiotic Engineering, which explores the problem of explanation as a communicative process, in which designers and users work together on requirements on explanations. A Model to Reason about the eXplanation design in Artificial Intelligence Systems (MoReXAI) is based on a structured conversation, with promotes reflection on subjects such as Privacy, Fairness, Accountability, Equity and Explainability, aiming to help end-users understand how the systems work and supporting the explanation design system. The model can work as an epistemic tool, given the reflections raised in the conversations related to the topics of ethical principles, which helped in the process of raising important requirements for the design of the explanation.

Keywords: Semiotic Engineering · Ethics · Explanations · Artificial Intelligence

1 Introduction

In recent years, interest in Machine Learning (ML) systems has grown. One of the major challenges for the adoption of these systems in some contexts is that

This work is partially supported by the FUNCAP projects 04772314/2020.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
H. Degen and S. Ntoa (Eds.): HCII 2022, LNAI 13336, pp. 130–148, 2022.
https://doi.org/10.1007/978-3-031-05643-7_9

many of the mathematical models implemented in these intelligent systems hide their fundamental assumptions, input information and parameters in black box models that are not directly observable, requiring specific techniques to improve the interpretation of the generated outputs [31].

EXplainable Artificial Intelligence (XAI) refers to a field of study of Artificial Intelligence (AI) that investigates techniques to improve the interpretability or explainability of machine learning models [28]. An interpretable system is one whose operations are understandable to humans, either through inspection of the system or some explanation produced during its operation. [6].

Current XAI tools do not fully capture the types of explanations many people want due to the complexity of the algorithms behind AI systems. Furthermore, there is a variety of audiences for which this explanation is intended. AI experts, domain experts and end-users need different levels of explanation and have different goals regarding the explainability of AI models [8]. For example, an AI expert seeks to improve explainability to obtain better performance from the algorithms, while domain experts seek explanation to improve confidence and gain greater knowledge about how the input data relates to the generated outputs [30]. In the context of end-users, explanations have several objectives: to improve transparency, reliability, trust, identify discriminatory bias and improve privacy awareness, given that the explanation helps users to assess the privacy of their data, revealing which of these data are being used in algorithmic decision making [27].

In addition to these objectives, the right to an explanation has already been regulated in several countries. These laws provide for the right to explanation, in the case of fully automated decisions that may have an impact on the life of the data subject.

Although the field of research on explainability of AI algorithms is not new, the opacity of machine learning algorithms brings new challenges in the quest to unravel the interior of these black boxes, or even to look for relationships between the input data and the outputs generated by these models. A lot of research has focused on technologies to visualize or expose the structures, features or decisions of these algorithms, or even the large data sets on which they are trained. These “explainable” systems usually present a simplified version of complex computational architecture, without providing evidence that justifies the use of the system or that it is effective [29]. According to [10], there is a gap between theories of construction of computing and theories of use of computing, and they propose the use of Semiotic Engineering Theory to explore the problem of explanations as a communicative process, in which designers and users are integrated into this process.

Semiotic Engineering (EngSem) is a theory based on communication. The Designer communicates to the user, through the system interface, who the user is, what problems he can solve, and how he should interact with the system [12]. For this communication to be effective, the designer needs to include in this communication the social meaning of the systems being designed, taking into account objectives, contexts of use, cultural and ethical aspects of the target

audience. Often this social meaning is not thought of by AI developers, requiring a mediated process by another field of study, such as HCI. Reflect on the social meaning of AI systems, and include pragmatic thinking in the process of developing these systems and thinking about how they can affect end-users directly or indirectly [7].

The goal of this paper is to propose a model to support the design of explanations in AI systems. To this end, we promote reflection among stakeholders about the social meaning of AI systems, based on conversation topics on ethical principles and explanations, during the process of developing these AI models. The idea is that this reflection works as an epistemic tool for the design of explanations.

In Sect. 2 we present the theoretical foundations, the bibliography that support the construction of the model and related works. In Sect. 3 we present how the model was defined, specifying each of its elements. In Sect. 4 we present a case study, in which we apply the model in the context of developing a recommender system, and finally, in Sect. 5, some important points of discussion regarding the model and the conclusion in Sect. 6.

2 Foundations and Related Works

In this section we briefly introduce the key points of the foundation theory and background work that, has provided us with insights on either the problem we want to address, or the solution we propose.

2.1 Ethical Principles and Explainable Artificial Intelligence (XAI)

The growing advances in the area of AI and its increasing use in people's daily lives, has brought a lot of ethical discussions, in view of its increasing influence, in the most varied contexts. Thus, governments, intergovernmental organizations, private companies, professional associations, advocacy groups and multistakeholder initiatives that are related to these technologies have created, updated, or adopted a set of unethical principles for AI [17].

The goals for each of these sectors of society are varied. Civil society and multistakeholder documents [2, 23, 32, 38] can serve to set a supporting agenda or set a direction for ongoing discussions, as well as establish a code of ethics and conduct for computing professionals. In government [13], these principles are often presented as part of an overall national AI strategy. In the private sector [20, 21, 25], the intention is to apply better governance for the organization's internal development and use of AI technology, as well as communicating its objectives to other relevant stakeholders, including customers and regulators.

Some authors carried out bibliographic reviews on ethical principles for AI, and defined sets of key themes, or main dimensions for these principles [9, 17, 18]. Explainability appears as one of the very important dimensions, from this it becomes easier to verify the adequacy of other ethical principles such as: privacy, accountability, fairness, reliability and safety.

Reflecting about these principles can happen within certain stages of the ML model development process. The development life cycle of ML models is divided into two major stages: one data-centric (data collection, data preparation and feature engineering) and another model-centric (training, testing and inferences) [37].

An example of data-centric development can be seen in [19]. They proposed data sheets for data sets, in which each data set is accompanied by a fact sheet that documents its motivation, composition, collection process, recommended uses, etc. Data sheets facilitate communication between data set creators and consumers and encourage the machine learning community to prioritize transparency and accountability, mitigate unwanted biases, facilitate greater reproducibility of results and help researchers and professionals to select the most appropriate data sets for the chosen tasks, considering that the characteristics of these data sets will influence the behavior of the model [19]. In this sense, data sheets help in the process of reflection, assessment of risk or potential harm and usage implications, as well as they can be valuable for policymakers, consumer advocates, individuals within the data set and those who may be affected by trained machine learning models [19].

In the model-centric stage, the model cards proposed by [26], are used to record reports and evaluate ML models, in addition to traditional evaluation metrics. In the structure of the model cards, it is possible, from a human-centered perspective [34], to record evaluations carried out in the construction of the model taking into account population, cultural, demographic and phenotypic groups. Model cards allow stakeholders to assess ethical, bias and fair issues, bringing varied perspectives to serve everyone involved in the project. Including group analysis as part of the reporting procedure prepares stakeholders to begin assessing fairness and including future machine learning system outcomes. Thus, in addition to supporting decision-making processes to determine the suitability of a given machine learning model in a given context, model reporting is an approach to transparent and trusted responsible practices in machine learning.

In the context of end-users, explanations may seek to achieve certain goals such as: (i) improving transparency to help users understand how the model works; (ii) improve reliability from assurance that the model will act as intended, generating reliable outputs in real-world scenarios, thus improving users' confidence in the system and its predictions; (iii) help end-users inspect whether systems are biased or have any discriminatory biases; and (iv) can also improve privacy awareness, given that explanation helps users evaluate the privacy of their data by revealing which of that data is being used in [27]. In addition to these goals, improving explainability can assist with auditability and accountability for damages caused by predictions generated in the context of users, and verification of compliance with regulatory standards. This list of goals is not exhaustive, nor are they exclusive to these types of users, and intersections can be found between the objectives of explainable AI and other types of users.

The above set of end-user explainability goals served as a basis for building the proposed model, presented in Sect. 3.

2.2 Semiotic Engineering

Semiotic Engineering (EngSem) is a theory that allows us to understand the phenomena involved in the design, use and evaluation of an interactive system [12]. The user's interaction with the system is seen as a conversation between the designer and the user, through the interface, at the moment the user makes use of it. The interface communicates to the user the designer's vision regarding who it is intended for, which problems it can solve and how to interact with it.

EngSem is grounded in Semiotics, and its ontology comprises the processes of signification and communication, the interlocutors involved in this process, and the HCI design space [12]. Signification is the process through which the expression and content of signs are established based on social and cultural conventions known to the people who will use them. Communication is the process through which people, using these signs, produce messages in order to express certain contents [12]. In this sense, culture influences human communication, considering the common sharing of signs and meanings that converge in the form of patterns of representation, used in the production and exchange of messages.

The HCI design space is structured in: context, sender, receiver, message, code and channel [22]. To design the meta-message, the designer must make decisions about each element of this model in order to identify: who are the interlocutors (receiver and sender) and what aspects of limitations, motivations, beliefs, and preferences should be taken into account for the benefit of meta-communication; what is the context of communication and what elements of interaction (psychological, sociocultural, technological, etc.) must be processed by the system; what is the communication code and how can or should be used to support efficient metacommunication; what is the available channel for designer-user metacommunication and what is the message that the designer must tell users, that is, what is the designer's communicative intention.

The designer has an active role in the interaction, considering that he/she is the interlocutor and must help users understand the meta-message contained in the interface. For this, he must reflect about the types of strategies he should use, the signs he can project on the interface, and the consequences that the limitations of the computational meanings bring to the interaction [3, 11]. For this, he/she uses epistemic tools [1, 5, 33, 35] that allow him/her to reflect on issues related to metacommunication artifacts and compare different proposed solutions.

Therefore, the process of interface design is a communicative act, in which the designer must make decisions about the solution that will be used to compose the interface (definition of signs and signification systems), it is necessary to have a better understanding of who the users are, their activities, experiences, values, and expectations, to allow a better transmission of the meta-message, from the interaction with the system. The designer has an active role in the interaction, considering that he/she is the interlocutor and must help users understand the meta-message contained in the interface. For this, it should reflect on the types of strategies it should use, the signs it can project onto the interface and

the consequences that the limitations of computational meanings bring to the interaction [3, 11].

2.3 Related Works

Design Based on EngSem. The work [4], is an example of the use of Semiotic Engineering to address ethical and social responsibility issues in the production of digital artifacts. They proposed an extension of the metacommunication template, an artifact used in Semiotic Engineering, to support human-centered design, in order that directly address moral responsibility and ethical issues. This extension works as an epistemic tool and can be used to create and elaborate knowledge related to these issues. The idea is to bring into the construction of the metacommunication message the vision not only of the designer, but of all those involved in the process of developing digital technology. The main actor is no longer “I” and becomes “We”. The questions used to build the meta-message are now answered by all stakeholders, adding questions at each design step that bring an ethical reflection about the product they develop and how this technology can affect users.

Explanation Design with an End-Users Focus. There is recent work on explanation design with an end-users focus, for example in [15] they propose guidelines to improve the transparency of AI algorithms. The content of an explanation (what to explain) is elicited from the following steps: capturing the mental model of AI experts and what they consider ideal for users; capture of users’ mental model; and target mental model synthesis, which the main components of the expert’s mental model that are most relevant to users, and the level of detail preferred by them, are selected. However, there is no promotion of a reflection on the social aspects involved in AI systems. Furthermore, MoReXAI promotes a conversation so that decisions about what and how to explain are made together with end-users, AI experts and HCI experts.

In [29] the authors point out that there is a need for use-inspired human-focused guidelines for XAI. They propose a “Self-Explanation Scorecard”, which can help developers understand how they can empower users by enabling self-explanation. In addition, they present a set of empirically-grounded, user-centered design principles that may guide developers to create successful explainable systems. In this work, we also use an approach involving end-users, however, we use communication focused design, where users and developers discuss explanations in a previously structured discussion model.

In [24] the authors present the development process of a multiperspective, user-centric tool for machine learning interpretability called Explain-ML. The tool was designed to implement a workflow in which the user can interactively perform the lifecycle steps of an ML model. For each project, it can create multiple runs, changing the model’s definition and optimization settings of hyperparameters, and generating a set of views that convey aspects of the model, the model’s training data set and also instance-specific information. These views act

as explanations for the model as it is designed to provide different perspectives that complement each other (global, database and local), which help the user to interpret the results of the model. A qualitative study was carried out to analyze in depth the users' perspective and perceptions of the tool. Based on an analysis of the results obtained in the evaluation of users' experience with Explain-ML, they observed potential relevance to meet the principles for designing Interactive Machine Learning interfaces [14], as well as consolidating them.

Users are the primary audiences for explanations in recommender systems. Explanations in this context usually reflect the goals that the designer wants to achieve with that explanation, such as: transparency, trust, scrutiny, persuasion, efficiency, effectiveness, and user satisfaction [36]. In [30], the authors propose to evaluate explanations of Facebook advertising recommendations using the semiotic inspection method, grounded in EngSem. They observe that although the explanations use good user-centered design practices, there are disruptions in interface communication due to the lack of meaning sharing of signs used in the explanation. Furthermore, users and designers have different goals with the explanations presented. Users are often concerned with ethical issues.

Explanation Design with an End-Users and EngSem Focus. Like our research, [16] is about XAI for end-users of AI systems. They argue that is need to discuss XAI early in the AI-system design process and with all stakeholders. They aimed at investigating how to operationalize the discussion about XAI scenarios and opportunities among designers and developers of AI and its end-users. They took the Semiotic Engineering as the theoretical background and the Signifying Message as our conceptual tool to structure the different dimensions that should be considered for XAI scenarios discussion.

3 The Model for Reasoning About AI Explanation Design

3.1 Model Questions

From the literature review related to the sets of unethical principles, those that are related to explainability were selected, based on the users' objectives, they are: **Privacy and Human Control (T1):, Responsibility and Accountability (T2):, Reliability and Security (T3):, Transparency and Explainability (T4): e Fairness, equity and non-discrimination (T5):.**

After this definition, we mapped them with the question sets addressed in the data sheets proposed in [19], and with the structure of the template cards proposed in [26], and the questions suggested by [7]. Then, we propose the following set of questions that act as a guide for MoReXAI. They are:

- **P1:** For what purpose was the data set created? Was there a specific task in mind? Was there a specific gap that needed to be filled? (T1)

- **P2:** What do the instances that make up the data set represent? Are there different types of instances? Are relationships between individual instances made explicit (e.g., user’s movie ratings, social media links)? (T1, T3, T5)
- **P3:** Does the data set represent all instances, or is it a sample of a larger set? If it is a sample, what was the sampling strategy? (T3, T5)
- **P4:** Is there any information missing from individual instances? What strategy was used to balance the data set? How was this strategy validated? Are there errors, noise sources, or redundancies in the data set? (T3, T2, T5)
- **P5:** Do data sets remain constant or can they be modified or deleted over time? (T3)
- **P6:** Is the data set related to people? Does it contain data that could be considered confidential or sensitive? Does it contain data that, if viewed directly, could be offensive, insulting, threatening, or can cause anxiety? (T1, T2, T5)
- **P7:** Does the data set identify any sub populations (by age, gender)? If yes, how are these sub populations identified and how are they distributed in the data set? (T1, T3, T5)
- **P8:** Is it possible to identify individuals, directly or indirectly, from the data set? (T1, T2)
- **P9:** How was the data collection done? What mechanisms or procedures were used? How were these mechanisms or procedures validated? Who was involved in the data collection process? (T1, T2, T3, T4, T5)
- **P10:** Have the individuals in question been notified of the data collection? Did they consent to the collection and use of their data? How was consent sought? If consent has been obtained, have mechanisms been provided to revoke your consent in the future or for certain uses? (T1, T2)
- **P11:** Over what period of time was the data collected? Does this period of time correspond to the period of creation of the data associated with the instances? (T1, T3, T5)
- **P12:** Have ethical review processes been carried out (e.g. by an institutional review board)? (T1, T2, T3)
- **P13:** Has an analysis of the potential impact of the dataset and its use on data subjects been carried out (e.g. a data protection impact analysis)? (T1, T2, T5)
- **P14:** Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labeled that could affect future uses? Is there anything a prospective user should know to avoid uses that could result in unfair treatment of individuals or groups or other undesirable harm? Is there anything a future user could do to mitigate this unwanted damage? (T2, T5)
- **P15:** Are there tasks for which the dataset should not be used? (T2)
- **P16:** What kind of ML model is being developed? Could you explain how it works? (T4)
- **P17:** Which algorithm is used to train the model? What is the degree of interpretability of the algorithm? Could you explain how it works? (T4)
- **P18:** What is the main internal use of the model? Who are the intended users that the model will serve? Any other usage scenarios outside this scope? (T1, T3)

- **P19:** Can the model affect demographic or phenotype groups in any way? What influence can these factors have on model performance? How is the influence of these factors on model performance evaluated? (T2, T3, T4, T5)
- **P20:** Can the instruments or the data set collection environment somehow influence the model’s result? What influence can these factors have on model performance? How is the influence of these factors on model performance evaluated? (T1, T2, T3, T5)
- **P21:** What model performance metrics are being reported and why were they selected over other performance metrics? How are they calculated? (T3)
- **P22:** What data sets were used to evaluate the model? Why were these data sets used? How was the data pre-processed for evaluation? (T1, T2, T3, T5)
- **P23:** Are quantitative analyzes performed against disaggregated population subgroups of the data set? How did the model perform for each factor? How did the model perform in relation to the intersection of the factors evaluated? (T2, T3, T5)
- **P24:** What risks may be present in using the model? What risk mitigation strategies were used during model development? Are there any intended uses for the model that are ethically worrisome? (T2)
- **P25:** Is the model scalable? How to ensure that the initial model trained and evaluated, when applied in the real context of users, maintains the same results obtained previously? Can the model be transferred from the intended context to another? (T2, T3)
- **P26:** Is there documentation related to the data sets? And what about the model? (T2)
- **P27:** What is explainability? What is the purpose of an explanation? Who is this explanation for? How is it presented? When should it be presented? Where should it be presented? (T4)

3.2 MoReXAI Structure

The conceptual model proposed in this research aims to extract requirements for the design of explanations, based on structured conversations between stakeholders. It was based on Jakobson’s communication model [22], thus having all the elements previously proposed by him:

- **Context:** involves the application domain, the topics of conversation (ethical principles and explanation) and how they can be impacted during the machine learning model development process. Thus, the conversation takes into account the collection and pre-processing of data, training, testing and evaluation of the ML model.
- **Interlocutors:** are all the stakeholders of the ML system under development. They can play the role of programmers, data scientists, academics, end-users, etc. It is important to capture the roles of the interlocutors in the conversation, in order that, in the analysis, it is possible to identify the perspective of each stakeholder on the topic discussed. Another important interlocutor in this process is the mediator, as it is, he/she who will initiate and lead the

conversation between the stakeholders, based on a set of pre-defined questions in the planning of application of the model. As it is a multidisciplinary conversation, it is up to the mediator to make interventions, in order to verify that the understanding of what is being talked about is understandable for everyone and to ask the sender to seek new ways to explain the subject being addressed.

- **Channel:** corresponds to the hardware and software where the conversation will run. As this is a conceptual model, it has not yet been defined which tools will best support the conversation within the model, taking into account that it can happen synchronously or asynchronously.
- **Message:** The message and model component that contains the communicated content. It is related to the typical conversations (ethical principles and explanation) that are proposed in the model. The set of messages exchanged from the answers to the questions raised by the mediator, bring a multidisciplinary perspective of the stakeholders, regarding how they think about ethical principles within the application domain, bringing to this conversation ways to explain the approaches used and that are understandable to those involved in the project. The message has the following structure:
 - **Sender:** identifier of the person sending the message. This could be, for example, the name of the person speaking and their role within the model development process.
 - **Receiver:** identifier of the person receiving the message. This message can be directed to everyone in the discussion, or to specific people.
 - **Date/Time:** date and time when the message was sent.
 - **Question:** This element is related to the development of the ML system, which is related to the application context and the typical conversation (ethical principles and explanation) as defined in Sect. 3.
 - **Answer:** text written in natural language that corresponds to the answer to a question in the model.
- **Code:** natural language is used.

3.3 MoReXai Use

Using the MoReXAI model involves 3 main steps:

1. **Planning:** This step is carried out by the person who will mediate the discussion and provides for the following actions:
 - (a) meetings with the development team to obtain a vision of the project being developed, to know the domain, the context of use, which step of development it is in, which algorithms will be used, which database will be used to train the algorithm and who will be the users of the system.
 - (b) selection of questions that will be used in the discussion.
 - (c) Selection and recruitment of discussion participants.

- (d) Interview with the participants: interviews are carried out with the recruited participants to learn about, in the case of the technical team, the time of development of ML applications, experience on the ML explanation project, what their role within the project. About users, what they use in this app and what they know about AI.
2. **Conversation - Exchange of Messages:** The mediator starts the conversation by introducing everyone in the group, presents the context of the application to be discussed and the dynamics of the conversation. Then the mediator leads the conversation, following the script of questions until all are exhausted. It is suggested that this step occurs in at least two moments in order that the conversation does not become tiring for the participants. This step can happen in synchronous meetings (online or face-to-face) or in an asynchronous conversation through online discussion tools (e.g. WhatsApp, Telegram or email). At the end, we seek to hold a focus group to capture general perceptions about what was discussed. The mediator should summarize what was discussed and close. These activities must be recorded (with everyone's permission) for further data analysis.
3. **Analysis of the results:** in this step, the collected data are organized, analyzed and a report is written with the directions obtained from the conversations for the design of explanations.

4 Case Study

We carried out an exploratory case study to observe the use of the model in a real context of developing an AI system.

The study was carried out during the development of a service recommendation system of the Government of the State of Ceará that will work within the Ceará App application. Ceará App is a mobile application with the objective of offering the main government services quickly and remotely in a single location. Through it, users can access services such as: 24-h online service for healthcare professionals, scheduling Covid-19 tests and vaccinations, issuing negative and regularity certificates, applying for a driver's license (CNH) and requesting 2nd via or renewal of CNH, advertisement, search and purchase of family farming products, among others.

4.1 Case Study Planning

The case study aimed to explore the use of the model in the real context of developing an AI system. In addition, to observe how the model can contribute to the design of explanations in the context of the application from the reflections on ethical principles within the development process. Evaluate the set of questions used in the model and their relationship with ethical principles. Also evaluate the process of applying the model, taking into account the roles of the interlocutors and the synchronous way in which the model was applied. The

service recommendation system is in the initial step of development. Developers have a database of user accesses to government services, collected over a period of 3 months, with more than 130,000 data records. The database contains service access date, service name, user identifier, device brand-model, device type (smartphone, tablet), device platform (Android, iOS), operating system version, Cear App on the day of access and SDK version. Developers are using Apple’s Turicreate framework and mentioned that they are testing some algorithms like DBScan, K-nodes, K-prototype, K-means. The model questions were selected based on the application context and the development step it is in (Table 1). The conversation within the model was divided into two steps, one using questions focused on the datasets, and the other more focused on the training, testing and evaluation steps of the ML model. The two researchers involved in the development of the recommender system and three users of the App were recruited.

Table 1. Set of questions addressed in each meeting according to the step of the application development cycle

Development stage	Questions
Data-centric	(P1), (P2), (P3), (P4), (P5), (P6), (P7), (P8), (P9), (P10), (P11), (P12), (P26)
Centered on the ML model	(P16), (P17), (P18), (P19), (P20), (P21), (P24), (P26), (P27)

4.2 Model Application

From the interviews, it was verified that the researchers have experience in the area of development of ML systems, being a Ph.D. in computing, and an undergraduate student of the software engineering course, which we will refer to as (D1) and (D2) respectively. Developers are working together at all stages of the proposed system’s development cycle and reported that they had not yet thought about explanations to users, or even about ethical principles. The three users (U1, U2 and U3) had experience in using the app, using services such as: scheduling exams, registering for Covid-19 vaccination, in addition to scheduling a second license. As for knowledge about AI, U1 and U2 said they had little knowledge, U3 knows a little more about AI concepts.

The model application conversations took place in two meetings through Google Meet. In addition to the mediator, the conversation had the participation of an HCI professional, who assisted in the information collection process, writing down data that he thought was relevant during the conversation. A focus group was held in order to evaluate the model. On that occasion, questions were asked related to the importance of reflecting on ethical principles in the context of machine learning models, and whether this reflection can influence the process of development and use of these systems and assist in the design of explanations.

4.3 Results: Talking About Explanations

Analyzing the case study with MoReXAI, we performed a discourse analysis on the conversations. We extract requirements on the 6 points that should compose a good explanation project (why? what? when? how? where? for whom?) The collected data and the full report are available at the following link: <https://bit.ly.com/jRXhT>.

Por Que Explicar? Questions from the MoReXAI model, in various topics, led stakeholders to realize that explanations are important because: (i) they guarantee rights (U1: *“You use an application and you no longer want your data to be used. If the law guarantees that, you will want it, then the system will have to find a way to solve it. I know there are technical limitations but I was thinking about the whole thing as a user”*); (ii) improve satisfaction in using the recommender system (D1: *“...if the user is satisfied with that (the explanation), then a duty for the system, providing this joy to him/her, would be an additional source of information...”*); (iii) they help to increase trust (U1: *“...I think it is the issue of credibility, trust, as I just said so you don’t think it’s just marketing...”*) and (iv) allow greater control over the data that is used (U1: *“...depending on the type of data they request, I don’t even continue and end up changing platforms, I go to another environment that does not have so much of my data”*).

Os usuários estavam interessados em saber o que foi levado em consideração para gerar a recomendação, quais dados foram usados e se esses dados estão protegidos, por exemplo: U3 disse: *“como foi chegado a esse resultado, com base em que, quais dados estão sendo usados pra lidar com isso?”*; Apesar disso um dos usuários não sentiu necessidade em conhecer como é o funcionamento interno do modelo de ML, U1 disse: *“Se tiver uma orientação geral talvez eu leia, mas se tiver algo muito técnico, mais aprofundado, eu como usuária que não sou da TI, eu acredito que não iria aprofundar a leitura da explicação não.”*

What to Explain? As for the content that explanations should contain, there were several suggestions. One suggestion was to explain how the recommendation was made and based on what data, or even why the user is viewing a particular recommendation. In the case of informing the data used to generate the recommendation, emphasize those that had the most relevance in this prediction. In this sense, U1 said: *“I thought of indicators (...) The basic elements that make up that relevance and recommendation calculation. (...) if we take Spotify, the most listened, the most downloaded, the most played. I think it is possible to have simple indicators that are easy to understand, in order that people who do not even know that there is an indicator, but understand that it was from what generated a recommendation”*

An important point raised is in the case of applications that are not public, they must inform users if the recommendation is something related to marketing, if it is a sponsored recommendation, and say why, and based on what, they are recommending that product/service.

It was also suggested to insert in the explanation which date the recommended one was generated within the model. Because there may be recommendations based on old data. Furthermore, in the case of recommendations that consider these profiles similar, the importance of explaining the profile of the user group that is being used to associate with the profile of the person receiving the recommendation was discussed.

Still on what to explain, the model provided a conversation about inserting examples known to users in the explanations. In this sense U3 said: *“Yes, and similar to that recommendation system that we already have in some applications, based on what I always do, more videos will appear, for example on YouTube, similar to what I always see, why will they interest me, in this case will Cear  app be based on my history of use of the Cear  app, will there always be some function related to this use, similar, which is the next step, right?”*

How to Explain? There was talk about the explanation project: participants recognize that explanations should not be long and should be presented gradually, whenever possible using examples, and could also be presented in conversational forms, for example U3 said: *“I think, me as a user, having a little conversation, like a bot, a conversation with a little robot, some animation, as if it were a really informal conversation, could be more playful, even for lay users... and not that it was a boring reading that they most of the time they will not read”.*

Where Should These Explanations Be? Some places were suggested where the explanations should appear, being able to come along with the recommendations, staying in a specific place of the application where users can make this query whenever they feel the need. The place where the explanations should be presented came from questions related to the terms of authorization for the use of the data. For example, D1 said: *“I think it should be shown along with the recommendation, have a link on the side, understand more how this recommendation was generated, for example.”* and D2 said: *“(...) to be available in some session of the application in case he/she curiously wants to go read it again”.*

When to Explain? It was also discussed when to explain, the participants had a consensus that the explanations should be presented close to the recommendations (in the case of a recommendation system). In this sense, D2 said: *“it has to appear right away to the user, at the moment he/she starts the application and also be available in some session of the application in case he/she curiously wants to read it again.”* and U3 said: *“it could be a step by step when initializing the system, like a tutorial on how to move and within that tutorial, it is saying that this data will be used, what it will be used for and how it will be used, and it will be stored there for when he/she wants to look, or on some screen, right”.*

The idea of configurable controls for viewing explanations also came up. In this sense, U3 said: *“(...) I think it should be a configuration, like, when we are*

going to make a configuration, we allow it or not, for example on the cell phone, we allow or not the use of mobile data, in a certain situation, I think it should be in the profile, when the user goes to see the configuration. Because by law it has to be available to the user at any time, all information (...)”

To Whom to Explain? During the conversation, it became very clear among all the participants that the explanations in the context of the recommender system should focus on the users who use this system.

5 Discussion

5.1 About the Epistemic Character of the MoReXAI

The case study carried out allowed us to reflect on elements of the model as well as its use. Leading developers to reflect on the artifact they are developing together with users is quite rich to build more effective explanations. The developers brought testimonials about the model’s questions and how they can influence their development process, as observed in D2’s statements: “...*I confess that most of the questions expanded my vision a little. I think a mistake I make a lot as a developer is to think only as a developer, and I forget that what I am doing is for a user, it is for a person, then these questions you asked helped me a lot to reflect on what I am working on and other concerns that you need to have*”, and in the speech of D1: “*I think these meetings were very relevant, it changes people’s view of some things that were usually already in a cast, so I am going to change the way I did things based on these meetings, because I am going to try to facilitate, or at least try to leave a framework of how to do these things that we were talking about here, since I agree that they are good things*”.

In addition, the model brought a new look to users about the systems they use. For example, U1 said: “*I found it super interesting, I had not stopped to think about how much recommender systems are present in the applications I use. I found this bias of the ethical issue super interesting, generally speaking only about data confidentiality, but the ethical framework, until reaching this data, before having this data, should have an ethical concern for its use. Often we only worry about the end when the data is already there and we do not have this worrying about ethics before*”.

Still on the epistemic character of the model, we noticed that the explanations given by developers to users, in a technical format, brought terms that are not suitable to be used in explanation, such as: “...*predictive model, database, relationship of a matrix...*”. At this time, the mediator had the role of helping to translate the developers’ explanation to the users and also to check with the users if it was understood. We understand that it was a rich moment to know which terms should or should not be used in the explanations of the system under development.

5.2 Improvements to the MoReXAI

New step identification for the model: In the case study we inserted the Focus Group to get feedback on the model. However, we realized that it was a space for gathering important information that helped to elaborate requirements for the explanation project. Therefore, we decided to insert the Focus Group at the end of the conversation in order that there is this moment to summarize what was discussed and to make a closing.

We noticed that some questions of the model need to be better explained to users. One way to do this is to use general examples, preferably from another system known to the group. For example, in the case study, some conversations took place using the Netflix movie recommendation system as an example. Therefore, we realize that it is important to guide the mediator to add examples related to the questions and the unethical principles they support.

We had anticipated that the model would be mediated by someone familiar with the elements of the model and who would organize the conversation. Among the roles we envisage are: defining the scope of the application to be discussed, defining and inviting discussion participants, scheduling and conducting the discussion, analyzing and compiling the data collected. However, in the case study we realized that an important role of the mediator is to assist in the communication between developers and users. During the experiment, we noticed that the developer used technical terms a few times. In these cases, the mediator must carry out a “translation” of what was said, or even intervene in order that the developers seek other ways to explain it to the users, and the mediator must follow what is being said by the technicians and check if the users are understanding. This process is interesting to capture the meaning system shared by the group. At first, we thought the mediator was an expert in Human-Computer Interaction, but we realized that he/she also needs to have basic knowledge of AI systems.

6 Conclusão

We propose a conceptual model to support the elicitation of explanations in ML projects. We use an approach that involves user participation and is based on communication-centered design.

We conclude that users’ statements related to ethical principles topics (privacy, security, responsibility, reliability, transparency, explainability, justice, equity and non-discrimination) generated important requirements for the explanation design. The model promoted the conversation about these principles and then suggested ideas for explanations for the interface, given by the user himself/herself. It was possible to talk about What, Why, How, When and Whom to explain.

In addition, we noticed the epistemic character of the model, as all the participants in the conversation said they had changed their view on the points that were addressed. The case study brought us the opportunity to reflect on the synchronous or asynchronous use of the proposed model. The fact that the

two meetings were synchronous was quite rich, as the contact between the stakeholders allowed for greater involvement and engagement in the conversation. In addition, the mediator had the opportunity to provoke the participants in order that everyone participated by giving their opinion. On the other hand, the main advantage of asynchronous conversation is giving people time to reflect on the questions.

Although the case study did not allow more time for participants to reflect on the questions of the conversation, as the conversations were synchronous, we realized that the fact that we had an interview days before starting the conversations already led the participants to think about what we were talking about. In addition, as the conversations took place in two sections, there was time between one section and the other, in this case it was two days, for those involved to reflect on the issues.

In further studies, we intend to explore the use of an asynchronous tool, or even a mixed methodology with synchronous and asynchronous moments, which allows those involved to have time to reflect on the model's questions. Regardless of whether the conversations are synchronous or asynchronous, the mediator will have the role of maintaining the group's engagement in the conversation, through targeted messages, ensuring that everyone participates.

We imagine the use of the proposed model in an AI system construction scenario where there is an interest in designing explanations. In this context, we envision an HCI expert interacting with the AI team to work together on this challenge of designing explanations. This team will invite users to join the conversations. These conversations can happen multiple times, with different users. In this context, the model works as an epistemic tool that generates knowledge, considering that with each application by this team, even in different and varied contexts, stakeholders will be adding knowledge about the design of explanations.

References

1. de A. Barbosa, C.M., Prates, R.O., de Souza, C.S.: Identifying potential social impact of collaborative systems at design time. In: Baranauskas, C., Palanque, P., Abascal, J., Barbosa, S.D.J. (eds.) INTERACT 2007. LNCS, vol. 4662, pp. 31–44. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74796-3_6
2. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, 1st edn. IEEE (2019). <https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html>
3. Barbosa, S., Silva, B.: *Interação humano-computador*. Elsevier, Brasil (2010)
4. Barbosa, S.D.J., Barbosa, G.D.J., de Souza, C.S., Leitão, C.F.: 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 2021, pp. 363–374. Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3442188.3445900>

5. Barbosa, S.D.J., de Paula, M.G.: Designing and evaluating interaction as conversation: a modeling language based on semiotic engineering. In: Jorge, J.A., Jardim Nunes, N., Falcão e Cunha, J. (eds.) DSV-IS 2003. LNCS, vol. 2844, pp. 16–33. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39929-2_2
6. Biran, O., Cotton, C.: Explanation and justification in machine learning: a survey. In: IJCAI-17 Workshop on Explainable AI (XAI), vol. 8 (2017)
7. Brandão, R., Carbonera, J., de Souza, C., Ferreira, J., Gonçalves, B., Leitão, C.: Mediation challenges and socio-technical gaps for explainable deep learning applications (2019)
8. Brennen, A.: What do people really want when they say they want “explainable AI?” we asked 60 stakeholders. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, CHI EA 2020, pp. 1–7. Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3334480.3383047>
9. Burle, C., Cortiz, D.: Mapping principles of artificial intelligence (November 2019)
10. Carbonera, J., Gonçalves, B., de Souza, C.: O problema da explicação em inteligência artificial: considerações a partir da semiótica. TECCOGS: Revista Digital de Tecnologias Cognitivas (17) (2018)
11. De Souza, C.S., Leitão, C.F.: Semiotic engineering methods for scientific research in HCI. Synth. Lect. Hum. Centered Inf. **2**(1), 1–122 (2009)
12. De Souza, C.S., Nardi, B.A., Kaptelinin, V., Foot, K.A.: The Semiotic Engineering of Human-Computer Interaction. MIT Press (2005)
13. (DIB), D.I.B.: AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. Department of Defense (DoD) (2019). https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI-PRINCIPLES_PRIMARY_DOCUMENT.PDF
14. Dudley, J.J., Kristensson, P.O.: A review of user interface design for interactive machine learning. ACM Trans. Interact. Intell. Syst. (TiiS) **8**(2), 1–37 (2018)
15. Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., Hussmann, H.: Bringing transparency design into practice. In: 23rd International Conference on Intelligent User Interfaces, IUI 2018, pp. 211–223. Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3172944.3172961>
16. Ferreira, J.J., Monteiro, M.: Designer-user communication for XAI: an epistemological approach to discuss XAI design. arXiv preprint [arXiv:2105.07804](https://arxiv.org/abs/2105.07804) (2021)
17. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M.: Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication (2020-1) (2020)
18. Floridi, L.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**(4), 689–707 (2018)
19. Gebru, T., et al.: Datasheets for datasets. arXiv preprint [arXiv:1803.09010](https://arxiv.org/abs/1803.09010) (2018)
20. Google: AI at Google: our principles (2018). <https://www.blog.google/technology/ai/ai-principles/>
21. IBM: Everyday ethics for artificial intelligence (2019). <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>
22. Jakobson, R.: Linguistics and poetics. In: Style in Language, pp. 350–377. MIT Press, MA (1960)
23. Future of Life Institute, F.: Asilomar AI principles (2017). <https://futureoflife.org/ai-principles/>

24. Lopes, B.G., Soares, L.S., Prates, R.O., Gonçalves, M.A.: Analysis of the user experience with a multiperspective tool for explainable machine learning in light of interactive principles. In: Proceedings of the XX Brazilian Symposium on Human Factors in Computing Systems, pp. 1–11 (2021)
25. Microsoft: Microsoft AI principles (2019). <https://www.microsoft.com/en-us/ai/our-approach-to-ai>
26. Mitchell, M., et al.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, pp. 220–229. Association for Computing Machinery, New York, NY, USA (2019). <https://doi.org/10.1145/3287560.3287596>
27. Mohseni, S.: Toward design and evaluation framework for interpretable machine learning systems. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, pp. 553–554. Association for Computing Machinery, New York, NY, USA (2019). <https://doi.org/10.1145/3306618.3314322>
28. Molnar, C.: Interpretable Machine Learning. Lulu.com (2020)
29. Mueller, S.T., et al.: Principles of explanation in human-AI systems. arXiv preprint [arXiv:2102.04972](https://arxiv.org/abs/2102.04972) (2021)
30. de O. Carvalho, N., Sampaio, A.L., Monteiro, I.T.: Evaluation of Facebook advertising recommendations explanations with the perspective of semiotic engineering. In: Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems, IHC 2020. Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3424953.3426632>
31. O’Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, 1st edn. Crown, New York (2016)
32. ACM Code of Ethics and Professional Conduct: ACM Code of Ethics and Professional Conduct. Association for Computing Machinery (ACM) (2018). <https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf>
33. Sampaio, A.L.: Um Modelo para Descrever e Negociar Modificações em Sistemas Web. Ph.D. thesis, PUC-Rio (2010)
34. Shneiderman, B.: Human-centered artificial intelligence: reliable, safe & trustworthy. *Int. J. Hum. Comput. Interact.* **36**(6), 495–504 (2020)
35. Silveira, M.S., Barbosa, S.D.J., de Souza, C.S.: Model-based design of online help systems. In: Jacob, R.J.K., Limbourg, Q., Vanderdonckt, J. (eds.) *Computer-Aided Design of User Interfaces IV*, pp. 29–42. Springer, Dordrecht (2005). <https://doi.org/10.1007/1-4020-3304-4.3>
36. Tintarev, N., Masthoff, J.: Explaining recommendations: design and evaluation. In: Ricci, F., Rokach, L., Shapira, B. (eds.) *Recommender Systems Handbook*, pp. 353–382. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_10
37. Toreini, E., et al.: Technologies for trustworthy machine learning: a survey in a socio-technical context. arXiv preprint [arXiv:2007.08911](https://arxiv.org/abs/2007.08911) (2020)
38. UNI Global Union: Top 10 principles for ethical artificial intelligence. Nyon, Switzerland (2017)