Towards Responsible AI Code Development: A Six-Value Process Model for Junior and Novice AI Programmers

Author(s)
van Kersbergen, Rick; Robben, Saskia

Publication date
2023

Towards Responsible AI Code Development: A Six-Value Process Model for Junior and Novice AI Programmers

Kersbergen van, Rick
Dept. Digital Media and Creative Industry
University of Applied Sciences,
Amsterdam, The Netherlands
Rickvkersbergen@gmail.com

***Abstract*—This paper presents a comprehensive study on assisting new AI programmers in making responsible choices while programming. The research focused on developing a process model, incorporating design patterns, and utilizing an IDE-based extension to promote responsible Artificial Intelligence (AI) practices. The experiment evaluated the effectiveness of the process model and extension, specifically examining their impact on the ability to make responsible choices in AI programming. The results revealed that the use of the process model and extension significantly enhanced the programmers' understanding of Responsible AI principles and their ability to apply them in code development. These findings support existing literature highlighting the positive influence of process models and patterns on code development capabilities. The research further confirmed the importance of incorporating Responsible AI values, as asking relevant questions related to these values resulted in responsible AI practices. Furthermore, the study contributes to bridging the gap between theoretical knowledge and practical application by incorporating Responsible AI values into the centre stage of the process model. By doing so, the research not only addresses the existing literature gap, but also ensures the practical implementation of Responsible AI principles.**

***Index Terms*—Responsible AI, Process Models, Fairness, Extension**

I. Introduction

The last few years have seen a huge growth in the capabilities and applications of AI [4], as evident from the recent surge in the development of systems using advanced large language models like ChatGPT [30]. As AI becomes increasingly integrated into various domains, a larger segment of the population becomes impacted by its influence, such as the implementation of AI in smartphone functionalities [31]. However, the consequential nature of these systems cannot be overlooked, as their flawed implementation can lead to severe adverse outcomes for individuals and society. Instances of systemic racism resulting from AI in governmental systems [17, 28] and instances of fatalities caused by algorithmic errors in self-driving cars [29] highlight the grave consequences of misapplied AI technologies.

The growing responsibility bestowed upon AI systems, coupled with their expanding presence in societal frameworks, accentuates the potential for harm they possess. Importantly, harms from such systems arise mostly unintentionally, as researchers and developers who are mostly responsible for how AI systems behave [4], typically strive to avoid such negative outcomes [2].

The objective of this project is to support AI developers in cultivating Responsible AI (RAI). The primary focus of this research will centre on junior or novice AI programmers, as they represent the forthcoming generation of AI practitioners. Hence, the central research question arising from this investigation is as follows:

How can one assist new AI programmers make responsible choices while programming?

This paper provides a comprehensive description of research conducted to answer the aforementioned main question. The subsequent section delves into an exploration of related work. Next, the methodology used is described, drawing upon the existing literature for guidance. The effectiveness of the methodological approach is then tested through an evaluation, followed by a detailed discussion of the evaluation results. Last, the paper concludes with a summary of the findings, subsequent areas for future research, and expressions of gratitude through acknowledgements.

II. Related work

A. Responsible Artificial Intelligence in practice

Numerous works have been dedicated to the subject of Responsible Artificial Intelligence, and an analysis of the literature reveals a consistent theme emphasizing the significance of upholding a set of values to ensure the responsible development and deployment of artificial intelligence. Coeckelberg delves into the ethical implications of artificial intelligence, examining its impact on society

and providing valuable insights for navigating the ethical challenges presented by AI, highlighting the importance of transparency, explainability and privacy besides warning for bias [1]. These are topics that also appear in Piersma's book [2]. Furthermore, she provides tips and tasks to consider in order to prevent harm done by intelligent systems and make them sustainable [2].

Dignum provides a comprehensive examination of the ethical challenges and societal implications of AI, offering practical frameworks and guidelines for the responsible development and deployment of AI systems [4]. An example of this is her ART model [4]. Besides this, existing initiatives [6-14] on Responsible AI were analysed and upon conducting an analysis, there could be concluded that all the initiatives prioritize the centrality of human well-being in the development of artificial intelligence (AI), as well as the basic principles from Dignum's ART model [4].

Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges toward Responsible AI" provides a comprehensive exploration of explainability in AI, including taxonomies, opportunities, and challenges, with the aim of promoting responsible AI development, touching on other topics as well, namely fairness, privacy, transparency and privacy [49]. Benjamins, Barbado and Sierra emphasize the significance of these values and introduce a novel methodology that incorporates these values as essential components [35]. Their approach entails posing relevant questions and conducting specific tasks to address these values, as presented in their work. This perspective aligns with the discussions presented by Piersma [2] and Dignum [4].

Furthermore, Steen highlights the existence of distinct categories of values [3]. Specifically, he draws a distinction between instrumental values and intrinsic values [3]. Instrumental values serve as means to achieve or realize intrinsic values, which hold value in and of themselves [3].

The examination of the Responsible AI field revealed a pronounced overlap of values, albeit with instances where the same value was named differently or provided with varying definitions. For instance, Dignum incorporates the value of Accountability in her ART model [4], which Coeckelbergh and Benjamins et al. refer to as explainability and transparency [1] [35]. Despite these naming differences, the authors were essentially referring to the same fundamental concept or idea.

Dignum, for example, defines her Accountability value as "the requirement for the system to be able to explain and justify its decisions to users and other relevant actors." and "decisions should be derivable from, and explained by, the decision-making mechanisms used" [4], while Coeckelbergh refers to transparency as well as explainability as "[...] not necessarily and certainly not only about disclosing the software code. The issue is mainly about explaining decisions to people" [1].

This observation recurs multiple times throughout the reviewed literature, indicating a pressing need for consistency in defining and labelling Responsible AI values.

B. Process models and design patterns

Research was conducted on whether process models and design patterns are effective in teaching programming best practices. In this research, the definition for design patterns from the Gang of Four and Alexander has been used [33][34].

A mapping study [36] found little evidence to support claims made about software design patterns in general, but some qualitative indication that they can provide a framework for maintenance. However, this research may not be conclusive due to limited studies on the topic and lower validity. Other studies [37, 39, 40] suggest that design patterns can have positive effects on software quality attributes, but their effectiveness is dependent on the specific application and the problem being addressed. Mangalaraj et al. found that design patterns can directly impact the quality of code that novice programmers write positively [38]. Overall, the effectiveness of design patterns in improving code quality and teaching coding is still a grey area, and more research is needed to provide a definitive answer. As no study implied that design patterns or process models had a negative effect, the choice was made to continue with the idea of using design patterns to teach the values to developers.

An analysis has been conducted on several sources encompassing the TensorFlow Responsible AI Guidebook [32], Google's Responsible AI Design Patterns [21], Google's Responsible AI Process Model [41], and CRISP-ML(Q) as proposed by Studer et al. [23]. These sources were examined to identify relevant components and insights that can be incorporated into the development of the Responsible AI process model.

Both CRISP-ML(Q) [23] and Tensorflow's Guidebook [32] primarily describe steps that need to be completed while developing AI. The steps from both sources come down to defining the problem, preparing the data, building and training the model, evaluating the model and deployment of the system. CRISP-ML(Q) adds monitoring as a last step.

Another observation is that, except Google's Responsible AI Process Model [41], the prominence of Responsible AI values within process models and patterns appears to be relatively limited. While these values are mentioned, they do not seem to hold a pivotal role within the framework or process model. Overall, the sources tend to prioritize techniques over values. This observation appears to contradict the research conducted

on Responsible AI, where values are considered crucial in the development of responsible AI practices. The scarcity of process models and methodologies that place values at the core signifies a gap in the existing literature, which the novel process model proposed in this paper aims to address.

C. Plug-ins and extensions

Research was conducted to find out whether process models and design patterns within extensions are a feasible solution to teach the principles from a design pattern of process model. Visual Studio Code had been chosen as IDE, as this code environment and its possibilities for extension development was already well understood by the researcher.

The papers [43] and [44] suggest that using design pattern documentation in the form of pattern comment lines (PCLs) can be beneficial for programmers. PCL's are references to a design pattern or process model in the form of text within the assignment and/or environment. Although the results from these papers may not be directly applicable to responsible AI programming, using design pattern documentation may be beneficial in teaching programmers responsible AI programming.

Plug-ins and tool found in the research that helps the programmer actively while programming were Snippet [45], JavaScript Code Snippets [46], Pro [47] and the Responsible AI Toolbox by Microsoft [48]. The conclusion of the research was that available plug-ins primarily focus on general coding, offering auto-completion features that help programmers write code faster but do not necessarily teach responsible coding. AI-based auto-completion plug-ins like [47] or [45] could be helpful in teaching programming techniques. In addition, plug-ins like the Responsible AI Toolbox [48] can aid programmers in retrospectively assessing the explainability of their code after loading data and training a model.

Currently, there appears to be an absence of an extension that offers support to programmers during the coding process in the form of an assistant or aid, providing explanations and assistance rather than solely delivering code. This gap is precisely what the plug-in developed in this project endeavours to address. By leveraging Pattern Comment Lines within an interactive environment, the proposed plug-in aims to assist programmers in the development of responsible AI code, effectively filling this gap in the existing landscape.

III. Methodology

This chapter delineates the systematic approach undertaken to define RAI values, the development of the process model, its accompanying plug-in and the evaluation of both. The goal is to provide a comprehensive understanding of the various components involved in the development and their individual procedures. The methodology is structured in multiple phases: identifying responsible AI values, the process model, the plug-in and the evaluation.

A. Identifying Responsible AI Values

To gain an in-depth understanding of the practical applications of Responsible AI (RAI) beyond existing literature, we conducted qualitative interviews with five experts in the field of RAI. These experts hold relevant doctoral and master's degrees and are associated with the University of Applied Sciences of Amsterdam. Each expert was individually interviewed in a dedicated meeting room on the university premises.

The interview employed open-ended questions to extract their perspectives on the practical manifestations of responsible artificial intelligence. The objective was to identify the key concepts and considerations they deemed critical within the context of RAI. Following the interviews, the qualitative data gathered was thoroughly analysed. The insights were then compared with the current literature on RAI. This comparison was a vital step in refining and interpreting the values essential to developing a process model.

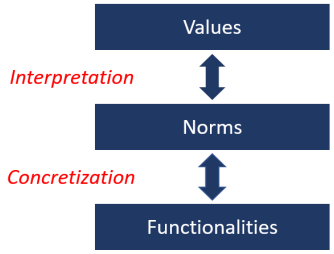

Fig. 1. Design for Values process as described by Dignum [4]

To effectively translate abstract, high-level values into specific norms, a process known as Design for Values, proposed by Dignum [4], which builds upon the foundation of Value Sensitive Design (VSD) proposed by Friedman et al. [15], was utilized. This process facilitates the interpretation and concretization of values, thereby enabling their incorporation into a process model (Figure 1).

As described in the related work, there seems to be a need for consistency in defining and labelling Responsible AI values. First, to start the Design for Values process, the values were selected based on description of said values which were named the most within the literature and the interviews with the experts. Names for these values were selected based on what seemed logical and representative of the meaning. Descriptions and interpretations of these values from the related work and the experts have been combined

and used as norms for these values to complete the next step of the Design for Values process.

Based on Steen's theory on the differences between intrinsic and instrumental values [3], the aforementioned values could be assigned varying degrees of importance. First, Steens theory was applied to identify instrumental and intrinsic values. Second, instrumental values that seemed to intersect with only one intrinsic value were subsumed within the intrinsic value, keeping the name of the intrinsic value.

B. The Process Model

To shape the process model, the final step of the Design for Values approach, namely concretization into practical functionalities, was utilized [15]. This approach aligns with the works of Dignum [4], Piersma [2], and Benjamins [35], as well as TensorFlow's model [32], all of which employ the technique of posing value-related questions to uphold ethical principles.

These questions could be linked to specific instrumental values [2, 4, 35], and subsequently translated into tasks that must be executed to address them, as demonstrated by Benjamins [35]. It was deemed necessary to only link questions to the instrumental values, as through proving one upholds the instrumental values, one could also prove to uphold the intrinsic values by extension. Suggestions for code implementation could be provided to programmers to fulfil these tasks, enabling the development of Responsible AI code while adhering to the values of the process model, completing the final step of Design for Values process [4].

Since programming follows a defined process with steps outlined in the related work chapter on existing process models and patterns, tasks and associated questions related to instrumental values could be assigned to specific steps within the programming process.

C. The Plug-In

The plug-in development phase initially intended to incorporate auto-completion or code suggestions, similar to existing Visual Studio Code plugins discussed in the related work. However, several constraints regarding the use of a Large Language Model (LLM) necessitated an alternative approach, as such an approach was deemed impractical due to lack of data, (hardware) resources and IDE optimisation issues.

Instead, development began on an alternate approach. This version of the plug-in would enable the user to initiate a new Responsible AI notebook project when creating a new VSC project. This process generated markdown cells within a Jupyter notebook, each related to a specific programming step question defined in the process model. Alongside this, a new window would open on the right side of the screen, displaying, explanations, and code suggestions relevant to the active programming step.

The initial concept aimed to incorporate a button with the code suggestions that, upon activation, would automatically insert the code into the designated Jupyter notebook cell. However, this concept was abandoned due to potential ethical concerns that could have arisen. The underlying hypothesis was that if the extension excessively assisted the programmer, their learning process would be hindered, as automation would eliminate the need for active engagement.

By maintaining certain inconveniences within the plug-in, the hypothesis was that programmers are compelled to continually interact with the system, thereby introducing a slight challenge and fostering engagement. This aspect further substantiates why auto-completion may not have been the optimal choice for addressing this specific problem, despite its prevalent usage in existing literature. Consequently, the button's functionality was modified to solely copy the code to the clipboard, necessitating the programmer to manually locate the appropriate notebook cell and insert the code themselves. There had been consultation with the stakeholders and this approach was approved.

D. The Evaluation

In this chapter, the evaluation methodology for assessing the functionality of the plug-in and the effectiveness of the Responsible AI Process Model will be outlined. This approach is based on similar experiments done using references to patterns with the use of Pattern Comment Lines [43, 44], with both experiments showing the success of this approach.

1) ***Selection of the participants****:* A group of students from the target audience was selected to participate in the evaluation process. Specifically, students from the Bachelor of the Artificial Intelligence program at the University of Amsterdam and Master students of the University of Applied Sciences were invited to assist with the evaluation. The students were divided into two groups: a control group and an experiment group. The control group would complete the assignment without using the plug-in, while the experiment group would utilize the plug-in.

2) ***The evaluation assignment****:* Participants were given the task to make an assignment in Visual Studio Code. In this assignment, the participant was tasked with preparing a dataset, training a model and evaluating the model created, which are the steps of the process model. As it's unpreferable for the participants to let the assignment take excessively long, most of the code in the assignment was already written. This way, the assignment could be done within an acceptable amount of time.

The dataset chosen for this assignment was the Adult Loan dataset from OpenML [51]. This dataset was chosen specifically, as there is clear bias in the dataset, and it contains features where their meaning is not clear. Furthermore, this dataset is publicly available and is somewhat older, as this dataset was created in 1996. The hypothesis was that the older age of the dataset makes it unlikely that any of the participants, who are of younger age, already have knowledge of the existing biases and explainability issues within the dataset.

The question were formed specifically to relate to the programming steps. These specific questions were asked, as participants from both the control and experiment group would be able to answer them. The hypothesis is that these specific questions could still give insight into whether the extension is used to answer it, by making the questions more theoretical than practical.

Within each process step, one or more questions were asked referring to that step, expecting different answers from the experiment group than the control group, especially in terms of references to RAI values and the thought process behind upholding them. In Table 1, an overview is given of the question asked and their expected answer per group.

Following the completion of the assignment, each participant would undergo an interview. The purpose of the interview was to ascertain whether the responsible AI plug-in successfully achieves its intended purpose, which is to teach thinking about programming AI responsibly. Also, it was expected that the elaboration of the answers within the assignment could be lacking and thus not fit for interpretation. This interview provides an additional, qualitative way to assess the functionality of the plug-in and process model. The interview consisted of one question:

- Do you believe you succeeded in completing the assignment and creating a good model? Please explain your reasoning.

Further questions were asked based on the answer given to this main question. By combining the findings from the code review and the interview, an overall assessment of the effectiveness of the Responsible AI Process Model and plug-in could be made.

IV. Results

The first section of this chapter describes the results of the interview conducted with the responsible AI experts and the final result of the responsible AI values identification. The second section describes the final process model, incorporating the defined values. The following section shortly describes the final plug-in iteration and the last section described the results of the evaluation process of both the process model and the plugin.

A. Final definition of Responsible AI values

After the completion of the interviews, the gathered data was analysed. The findings revealed a significant degree of convergence with the existing literature on RAI examined in the related work section of this paper. Moreover, it was observed that several researchers placed notable emphasis on 'resource allocation' or efficiency as an additional aspect of interest within the field [50]. This would be interpreted as the allocation of appropriate resources for specific tasks.

Furthermore, during the interviews, certain experts expressed considerable focus on aspects preceding the actual coding phase of RAI development. Specifically, they underscored the significance of a concept referred to by one expert as 'moral imagination'. This concept revolves around the preliminary stages of problem definition and project feasibility assessment, thus encompassing the ethical considerations associated with project initiation. These experts recognized the crucial role of clarifying the problem statement and evaluating whether the proposed project is viable and ethically sound. [50].

Following the methods described in section A of the methodology chapter first resulted in the following values with their corresponding interpretations:

1) **Fairness:** No groups involved should be disadvantaged by design using equal opportunity for all. (As Fairness can have widely different normative interpretations [4] [34], the interviewees that named fairness were asked how they would define it. They concluded equal opportunity would suffice [50].)
2) **Sustainability:** Being careful with resources and designing systems for the longest term possible. [2] [3] [50]
3) **Human Centrism:** Humans should stay centric in the design. Humans are not a resource or tool to be used or to be a means to an end. [2] [3] [4] [35]
4) **Efficiency:** Use the right resources for the right task. [2] [50]
5) **Transparency:** Be clear, open about and know what you're using. [1] [2] [3] [4] [35] [49] [50]
6) **Explainability:** Be able to explain your choices and of your system. [1] [2] [4] [35] [49] [50]
7) **Privacy:** Making sure people can not be identified within the data. [1] [2] [3] [4] [35] [50]

The concept of moral imagination, as articulated by certain experts interviewed, diverges from being categorized as a distinct value since it represents more of a task or obligation

TABLE I
QUESTIONS FOR THE EVALUATION, HOW THEY'RE EVALUATED AND EXPECTED OUTCOMES

Question/Task	Variables used for measurement	Expected answer control group	Expected answer experiment group
Construct and Prepare Data			
Look at the feature-names within the dataset. Given these features, can you already say anything about which features you'd rather not use for training?	The amount of sensitive variables and non-explainable in the answer with argumentation of them being sensitive or non-explainable	No or few sensitive and non-explainable features	Contains the sensitive features named in the plug-in that apply and non-explainable features
Plot the correlation matrix of the dataset. Which correlations do you see, and do you notice (other) features that you might want to use for training the model or not?	The amount of features named and dropped with the reason being the feature is sensitive or a proxy	Fewer right features named and dropped with the right reason.	More or all right features named and dropped with the right reason
See the uneven distribution of the classes within the target feature. Why would we want to change this?	Inclusion of reference to bias and/or fairness problems	Naming of performance. Lack of references to bias and/or fairness	Besides performance, contains references to bias and/or fairness
Build and Train Model			
You can choose between using the Logistic Regression or the Dense Neural Network below for training. Choose one and explain your choice.	Choice made for Logistic Regression. Number of supported by references to Explainability, Efficiency, Problem Complexity and Sustainability.	Either choice made for a DNN with only reference to performance or choice for LR with lacking reference to Explainability, Efficiency, Problem Complexity or Sustainability	Choice made for Logistic Regression, supported by references to Explainability, Efficiency, Problem Complexity and Sustainability.
Evaluate Model			
What do the False Negative Rates for gender and race tell you?	Knowledge of FNR's and argumentation on the fairness of the model.	Lacking knowledge of FNR, wrong argumentation on fairness	Comprehension of FNR's and their meaning. Logical argumentation on the outcome.
In the previous question, we looked at Fairness and the Performance of the model. Which other indicators could be used to evaluate the model and why?	Amount of references to Resource Demand, Model Complexity, Robustness and Explainability	No or little references to the other indicators.	Most or all references to indicators present.

that one should undertake, rather than a specific value to be upheld. Nevertheless, the insights and perspectives shared by these experts regarding moral imagination will be integrated into the process model. By doing so, their ideas and recommendations will contribute to shaping the stages and activities of the process model, ensuring that the critical task of moral imagination is appropriately addressed and incorporated into the overall framework.

After applying Steen's theory on intrinsic and instrumental values [3], as described in section A of the methodology, the aforementioned values could be assigned varying degrees of importance. Taking into account the broad descriptions provided in the literature, privacy had been considered an instrumental value that supports the intrinsic value of Human Centrism and had therefore been subsumed by that intrinsic value.

An additional set of instrumental values can be identified, namely Transparency, Explainability, and Efficiency. These values play a crucial role in ensuring that one can effectively assess and uphold the intrinsic values of Fairness, Sustainability, and Human Centrism. Transparency enables an understanding of the purpose and actions involved, while Explainability provides the necessary insights into the reasons

behind these actions. Efficiency ensures the optimal allocation of resources to fulfil tasks accurately. It is worth noting that these instrumental values are not merged with the intrinsic values, as each instrumental value contributes to multiple intrinsic values, as noted by Arietta in the case of explainability [49]. Therefore, the final collection of values encompassed in the framework comprises Fairness, Sustainability, Human Centrism, Transparency, Efficiency, and Explainability (Figure 2). A visualisation of all values, their relationships and the inclusion of moral imagination can be seen in figure 3.

Main Values	Definition
Fairness	Equal opportunity for all / no groups involved should be disadvantaged by design
Sustainability	Being careful with resources and designing systems for the longest term possible.
Human Centrism	Humans should stay centric in the design. Humans are not a resource or tool to be used or to be a means to an end.
Supporting Values	**Definition**
Efficiency	Use the right resources for the right task
Transparency	Be clear, open about and know what you're using.
Explainability	Be able to explain your choices and of your system.

Fig. 2. The final collection of values for the process model

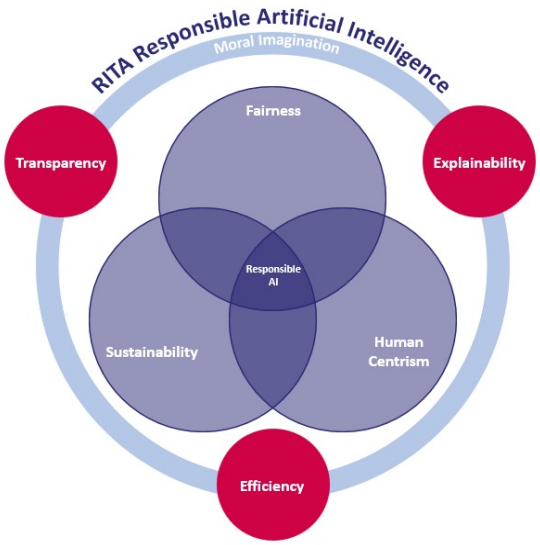

Fig. 3. The core of the process model, concerning Responsible AI

B. The Final Process Model

Conducting the methods described in section B of the methodology chapter resulted in a process model encompassing four programming steps: Define Problem, Construct and Prepare Data, Build and Train Model, and Evaluate Model. The monitoring step, which appeared in some existing models described in the related work, had been excluded from this process model, as its focus is solely on the programming of AI, not the overall system. Each step identifies the values that should be upheld, and corresponding questions are stated, which simultaneously correlate with the tasks to be performed within each programming step (Figure 4). The final process model. including explanations and questions, can be viewed in the appendix (Appendix A).

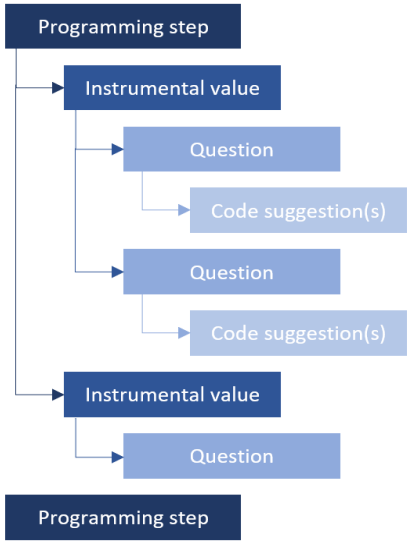

Fig. 4. Architecture of the process model

In Figure 4, it is evident that not every question is accompanied by a code suggestion. This is because some questions are theoretical in nature, lacking a definitive best practice solution. The questions associated with the instrumental values included in the final process model are directly referenced or inspired by the relevant literature on responsible AI, existing process models and design patterns, and insights gained from interviews conducted with RAI experts.

The code examples used as suggestion for the programmer to complete a task and answer a question are based on the best practices from *Deep Learning* by Goodfellow et al. [42] and Responsible AI: Implementing ethical and unbiased algorithms by Agarwal et al. [34].

C. The Final Plug-In

The development process commenced with the creation of a paper prototype intended for students enrolled in the Minor Applied Artificial Intelligence program at the University of Applied Sciences of Amsterdam. The prototype aimed to simulate a Visual Studio Code (VSC) extension tailored specifically to this target audience, and their feedback was encouraging.They found the AI programming extension to be user-friendly and helpful in learning about responsible AI. The participants appreciated the guidance provided by the extension and felt that the interaction between the extension and the user was effective. They could envision themselves using the extension to learn and experiment with AI models,

acknowledging its potential to help them program more responsibly.

While the initial feedback was positive, some inconveniences in the user interface became apparent during the development process. One limitation was the presence of only one code block for each question, despite some questions having multiple code suggestions. This restriction seemed counterintuitive, since Jupyter Notebooks typically allow code to be split into multiple blocks for separate execution. Additionally, the compact layout of the extension's screens on smaller laptop screens proved to be inconvenient. Another observation was that the code suggestions lacked explanations, unlike the process model. Furthermore, there was a lack of connection between the extension and the important values of the process model, which were crucial for teaching junior programmers about responsible AI programming.

Taking these inconveniences into account, a second version of the extension was developed, addressing the feedback and introducing improvements. This version included a walkthrough tutorial, clearer explanations for questions and code suggestions, and the incorporation of the values of the process model. However, during the testing of this second prototype iteration, it was noted that one of the initial testers had limited familiarity with Visual Studio Code, which posed challenges in understanding the tutorial. The testers did not thoroughly read the text within the walkthrough tasks, leading to confusion later on. Once the importance of clicking on a programming step was clarified, the interaction felt relatively intuitive.

Additionally, a stakeholder's feedback raised concerns about the clickability of markdown cells, as they did not appear visually clickable. To address this issue, it was proposed to decouple all plugin functionality from the Jupyter Notebook and allow scrolling through the questions within the side-panel. Recognizing the recurring nature of this feedback and the need for further improvements, testing on the current prototype iteration was halted.

Based on the received feedback, the decision was made to improve the product by modifying the side-panel to encompass all functionalities and completely decouple it from the Jupyter Notebook. This iterative process led to the final prototype iteration, representing the culmination of the valuable feedback and insights gathered throughout the testing and development phases (Figure 5).

Fig. 5. The final version of the extension panel

D. The Evaluation Results

In this chapter, the results from the evaluation are described. First, the results from the questions within the assignment are described with visual assistance. Second, the results from the interviews taken with the participants of both groups are described.

After completing the assignment, the assignment of each participant was reviewed and analyzed whether the answers to the questions align with the expected result for each group using the intended variable as measurement for each question. (Figure 6). From the results, these general observations were made for each question:

1) **Question 1:** Both experiment group participants seem to have good knowledge of sensitive features on first sight, one control group participant as well.
2) **Question 2:** Both experiment group participants seem to see more proxies in the data.
3) **Question 3:** Both experiment group participants name bias and/or fairness, in control group only one.
4) **Question 4:** All participants seem to have chosen LR. All name at least one good reason.
5) **Question 5:** Both experiment group participants understand the meaning of FNR's, reference equal opportunity. Both control group members seem to lack this.
6) **Question 6:** One participant in the control group names explainability. the others didn't understand the question

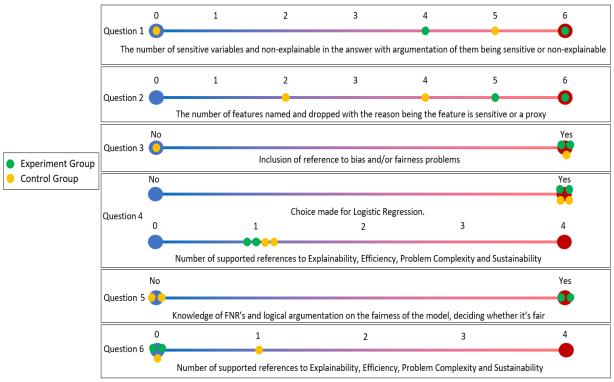

Fig. 6. Results from the questions of the assignment

After the completion of the assignment, each participant was interviewed. Both participants from the experiment group seem to support their definition of a good model with references to Responsible AI values, primarily explainability and fairness. Both were especially able to explain well why they chose the features to train on with references to the terms proxy's and sensitive variables. With argumentation on the model itself, clear reference to the terms black box and explainability.

There are hints to Responsible AI values in the argumentation of the control group participants, but described vaguely. One participant outright stated to have no idea on how to decide whether this is a good model. He stated it's bad, but with reference to performance, but no reference to Responsible AI values. The other participant primarily focussed on accuracy and named the term bias, but wasn't able to properly explain the role of bias in the quality of the model.

V. Discussion

In this chapter, first the results will be analyzed further interpreted. Second, the broader implications of these results will be noted, making a connection to existing literature. Last, the limitations of the research that might have affected the results will be discussed. Here, the reliability and validity of the research will be described.

A. Results interpretation

The research question that guided this study was how one could assist new AI programmers make responsible choices while programming. To address this question, an experiment was conducted using a novel RAI process model and a VSC extension incorporating this model to discover whether this approach succeeds in teaching new AI programmers about the concept of Responsible AI and how they could incorporate this in their code.

The hypothesis was that participants of the experiment that used the plug-in would have a better understanding of Responsible AI after using the plug-in and process model to solve an AI problem within an assignment. The analysis of the data provided insights into the effect of using the extension and process model on the ability of the programmer to make responsible choices while programming.

In line with the hypothesis, the results show that using the extension referencing the RAI process model has a positive effect on the ability of the programmer to make responsible choices while programming and afterwards being able to argue why their choices were responsible. This indicates that the extension and process model succeed in their purpose of equipping new AI programmers with knowledge of the principles of responsible AI programming, which is based on the existing research on responsible Artificial Intelligence.

B. The Implications

These results further add to the support that using process models and patterns may have a positive effect on the code development capabilities of its users, as described in [37, 39, 40] and contradicting [36], assisting with clarification of the grey area surrounding the effectiveness of patterns. These results also support existing evidence on the effectiveness of using references to patterns and process models within an extension, as described by Prechelt et al [43] and Vokáč et al. [44].

The results further confirm the theory that by asking questions related to Responsible AI values, one can uphold them as outlined by Dignum [4], Piersma [2], and Benjamins [35], which had barely been incorporated in existing models used in practice. Therefore the results show that the process model succeeded with making a connection between the literature and practice, filling the gap by incorporating RAI values in the centre stage of a process model.

C. The Limitations

Despite the valuable insights gained from this study, it is important to acknowledge the methodological limitations that may have influenced the results. Regrettably, on the day scheduled for the evaluation, the intended target audience, as planned, failed to attend due to unforeseen circumstances, such as a sudden change in their availability. This unexpected turn of events necessitated an adjustment in the participant selection process.

In order to proceed with the evaluation and gather valuable feedback, the decision was made (after consultation and approval of the coach of the researcher) to include alternative participants who already possessed some knowledge of the project and its goals, but were still part of the target audience, namely Master Applied AI students. This resulted in a lower

amount of participants, challenging the validity of the results, besides introducing certain biases.

While it was a concern that the alternative participants, who possessed prior knowledge of the project, might introduce bias or skew their responses, several precautions were taken to ensure the integrity of the evaluation process. It was made clear to the participants that honest and unbiased feedback was crucial for the success of the evaluation. We assured them of the confidentiality of their responses and stressed the importance of providing their genuine opinions.

Furthermore, as the sole researcher conducting this study on Responsible AI (RAI), it is important to acknowledge the potential bias that may have influenced the selection and interpretation of the collected opinions. Recognizing the significance of diverse perspectives in understanding the concept of Responsible AI, deliberate efforts were made to incorporate a range of viewpoints and minimize the potential impact of personal biases on the research process.

Multiple data collection methods were employed, which included conducting interviews and reviewing existing literature from various reputable sources. While these methods allowed for the collection of diverse opinions, it is important to note that the scope and extent of the research were inevitably influenced by the limitations of being a single researcher. The sample size, although diverse, might not have fully represented all perspectives within the field of Responsible AI.

VI. Conclusion

This research aimed to develop a way to assist new AI programmers make responsible choices while programming, within the context and scope of the project as described in chapter 1.3. Research was conducted on the possible benefit of using process models, design patterns and using an IDE-based extension to incorporate these, resulting in a concept of responsible Artificial Intelligence, a process model showing how to achieve it and a Visual Studio Code extension incorporating this, which had been through multiple prototype tests with the target audience.

An experiment was conducted with the target audience, evaluating the effectiveness of the process model and extension's intended purpose. These results indicate that using the process model and extension developed, one is more capable of making responsible choices while programming artificial intelligence.

The findings of this study provide additional support to the existing literature, suggesting that the use of process models and patterns can positively impact the code development capabilities of users. This aligns with previous research [37, 39, 40], contradicting earlier claims [36], and contributes to clarifying the effectiveness of patterns in addressing the grey areas in software development. Furthermore, the results reinforce the existing evidence on the effectiveness of incorporating references to patterns and process models within an extension, as demonstrated by Prechelt et al. [43] and Vokáč et al. [44].

Moreover, the results of this study confirm the theory proposed by Dignum [4], Piersma [2], and Benjamins [35], which emphasizes the significance of asking questions related to Responsible AI values. These findings highlight the successful integration of these values into the center stage of a process model, bridging the gap between existing literature and practical application. By incorporating Responsible AI values, the process model not only fills a crucial gap but also ensures the upholding of these values in practice.

In summary, the outcomes of this research contribute to the body of knowledge by providing further support for the positive impact of process models and patterns on code development capabilities. Additionally, they validate the importance of incorporating Responsible AI values into process models, aligning with existing theories and enhancing the practical application of these values, which resulted in a solution that assists new AI programmers make responsible choices.

VII. Future Work

The present study has shed light on the integration of Responsible AI (RAI) values into process models and extensions, providing valuable insights into its benefits and implications. However, there are still exciting avenues for future research that can build upon these findings and advance the understanding of RAI in practice.

The limited sample size of participants in this study poses a potential threat to the validity of the findings. Future research should aim to conduct evaluations on a larger scale, involving a more substantial number of participants. By increasing the sample size, we can enhance the statistical power of the study and ensure more robust and reliable results. Additionally, expanding the participant pool to include diverse backgrounds and industries would provide a more comprehensive understanding of the effectiveness and applicability of the proposed process model.

Furthermore, while this study provided valuable insights into the integration of RAI values into process models, it is important to note that it was not a comprehensive mapping study. As the sole researcher, time constraints limited the extent of the research. Future research could dedicate more time and resources to conduct an extensive mapping study on what Responsible AI truly means.

Together, these future research directions will pave the way for a more comprehensive and effective integration of Responsible AI into process models, ultimately leading to the development and deployment of AI systems that align with ethical principles and societal values.

VIII. Acknowledgements

Throughout the writing of this paper, I have received a great deal of support and assistance.First, I would like to thank my coaches and supervisors Dr. Saskia Robben and Dr Michelangelo Vargas Rivera for assisting me throughout the whole research process, whose expertise was invaluable and feedback enormously helpful.Second, I would like to thank the participants from my target audience, who assisted my through the prototype iterations and evaluation. Hereby special thanks to Taner Özgüner and Dennis van Schie, who assisted me with the first iterations and whose feedback was crucial for the development of the prototype. Last, I would like to thank my fellow Responsible AI lab mates for their collaboration and amazing time at the lab this semester.

References

[1] M. Coeckelbergh, Ai ethics. ©2002: The MIT Press, 2020.
[2] N. Piersma, System error, please restart: Hoe we verantwoorde it-systemen Kunnen Bouwen. Amsterdam: Hogeschool van Amsterdam, 2022.
[3] M. Steen, Ethics for people who work in Tech. Boca Raton ; London ; New York: CRC Press, Taylor et Francis Group, 2023.
[4] V. Dignum, Responsible artificial intelligence. SPRINGER, 2020.
[5] "Universal declaration of human rights," United Nations. [Online]. Available: https://www.un.org/en/about-us/universal-declaration-of-human-rights. [Accessed: 10-Mar-2023].
[6] "Autonomous and intelligent systems (AIS)," IEEE Standards Association, 09-Mar-2023. [Online]. Available: https://ethicsinaction.ieee.org/. [Accessed: 10-Mar-2023].
[7] "High-level expert group on artificial intelligence," Shaping Europe's digital future. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai. [Accessed: 10-Mar-2023].
[8] "Home," Partnership on AI, 03-Mar-2023. [Online]. Available: https://partnershiponai.org/. [Accessed: 10-Mar-2023].
[9] AI for humanity. [Online]. Available: https://www.aiforhumanity.fr/en/. [Accessed: 10-Mar-2023].
[10] "Governance of Artificial Intelligence (AI) - committees - UK parliament," UK Parliament. [Online]. Available: https://committees.parliament.uk/work/6986. [Accessed: 10-Mar-2023].
[11] "Ai Principles," Future of Life Institute, 10-Mar-2023. [Online]. Available: https://futureoflife.org/open-letter/ai-principles/. [Accessed: 10-Mar-2023].
[12] "Barcelona Declaration for the proper development and usage of artificial intelligence in Europe," AIdeclaration. [Online]. Available: http://www.iiia.csic.es/barcelonadeclaration. [Accessed: 10-Mar-2023].
[13] "Montreal declaration for a responsible development of artificial ..." [Online]. Available: https://nouvelles.umontreal.ca/en/article/2017/11/03/montreal-declaration-for-a-responsible-development-of-artificial-intelligence/ [Accessed: 10-Mar-2023].
[14] "JSAI ethical guidelines" [Online]. Available: http://www.ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf. [Accessed: 10-Mar-2023].
[15] B. Friedman and D. G. Hendry, Value sensitive design: Shaping technology with moral imagination. Cambridge, MA: The MIT Press, 2019.
[16] C. Bowles, Future ethics. East Sussex, United Kingdom: NowNext Press, 2018.
[17] J. Buolamwini and T. Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', in Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Jan. 2018, pp. 77–91. Accessed: Mar. 10, 2023. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html
[18] How to Conduct Interviews in Qualitative Research — Rev. (n.d.). Retrieved February 13, 2023, from https://www.rev.com/blog/education/how-to-conduct-interviews-in-qualitative-research
[19] H. Washizaki et al., "Software-Engineering Design Patterns for Machine Learning Applications," Computer, vol. 55, no. 3, pp. 30–39, Mar. 2022, doi: 10.1109/MC.2021.3137227.
[20] Kolamanvitha, "Design Patterns for Machine Learning," Medium, Jul. 19, 2021. https://towardsdatascience.com/design-patterns-for-machine-learning-410be845c0db (accessed Mar. 12, 2023).
[21] "People + AI Guidebook." https://pair.withgoogle.com/guidebook (accessed Mar. 12, 2023).
[22] "Design Patterns and Refactoring." https://sourcemaking.com (accessed Mar. 12, 2023).
[23] S. Studer et al., "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology." arXiv, Feb. 24, 2021. Accessed: Mar. 12, 2023. [Online]. Available: http://arxiv.org/abs/2003.05155
[24] "Your First Extension." https://code.visualstudio.com/api/get-started/your-first-extension (accessed Mar. 12, 2023).
[25] "UX Guidelines." https://code.visualstudio.com/api/ux-guidelines/overview (accessed Mar. 12, 2023).
[26] S. Agrawal, "Design the Nudges," Medium, Feb. 24, 2020. https://uxplanet.org/digital-nudge-design-process-48086f16595c (accessed Mar. 12, 2023).
[27] "What's a 'Nudge' in Product Design?," Delve. https://www.delve.com/insights/whats-a-nudge-in-product-design (accessed Mar. 12, 2023).

[28] 'Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms', Amnesty International, Oct. 25, 2021. https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/ (accessed Mar. 15, 2023).

[29] A. J. Hawkins, 'Tesla didn't fix an Autopilot problem for three years, and now another person is dead', The Verge, May 17, 2019. https://www.theverge.com/2019/5/17/18629214/tesla-autopilot-crash-death-josh-brown-jeremy-banner (accessed Mar. 15, 2023).

[30] A. Agrawal, J. Gans, and A. Goldfarb, 'ChatGPT and How AI Disrupts Industries', Harvard Business Review, Dec. 12, 2022. Accessed: Mar. 15, 2023. [Online]. Available: https://hbr.org/2022/12/chatgpt-and-how-ai-disrupts-industries

[31] 'On-device AI: Mobile AI', Samsung Semiconductor Global. https://semiconductor.samsung.com/insights/topic/ai/on-device-ai (accessed Mar. 15, 2023).

[32] "Responsible AI Toolkit," TensorFlow. https://www.tensorflow.org/responsible_ai (accessed Mar. 15, 2023).

[33] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design patterns: Elements of reusable object-oriented software. Beijing Shi: Ji xie gong ye chu ban she, 2021.

[34] C. Alexander, A pattern language. Munchen: Fachhochsch., Fachbereich Architektur, 1990.

[35] R. Benjamins, A. Barbado, and D. Sierra, 'Responsible AI by Design in Practice'

[36] C. Zhang and D. Budgen, 'What Do We Know about the Effectiveness of Software Design Patterns?', *IIEEE Trans. Software Eng.*, vol. 38, no. 5, pp. 1213–1231, Sep. 2012, doi: [10.1109/TSE.2011.79]https://doi.org/10.1109/TSE.2011.79.

[37] P. Sfetsos, A. Ampatzoglou, A. Chatzigeorgiou, I. Deligiannis, and I. Stamelos, 'A Comparative Study on the Effectiveness of Patterns in Software Libraries and Standalone Applications', in *2014 9th International Conference on the Quality of Information and Communications Technology*, Guimaraes, Portugal, Sep. 2014, pp. 145–150. doi: [10.1109/QUATIC.2014.26]https://doi.org/10.1109/QUATIC.2014.26.

[38] G. Mangalaraj, S. Nerur, R. Mahapatra, and K. H. Price, 'Distributed Cognition in Software Design: An Experimental Investigation of the Role of Design Patterns and Collaboration', *MIS Quarterly*, vol. 38, no. 1, pp. 249–274, 2014, Accessed: Mar. 24, 2023. [Online]. Available: https://www.jstor.org/stable/26554877

[39] 'Research state of the art on GoF design patterns: A mapping study — Elsevier Enhanced Reader'. https://reader.elsevier.com/reader/sd/pii/S0164121213000757?token=CA6B8111F3416323A27223EDAC53DE20F10977FB15CBA4CAD3A7859B2CA788FA223428C438D2C09A5E672CE4E0F83FC3&originRegion=eu-west-1&originCreation=20230323143520 (accessed Mar. 23, 2023).

[40] M. Vokáč, W. Tichy, D. I. K. SjØberg, E. Arisholm, and M. Aldrin, 'A Controlled Experiment Comparing the Maintainability of Programs Designed with and without Design Patterns—A Replication in a Real Programming Environment', *Empirical Software Engineering*, vol. 9, no. 3, pp. 149–195, Sep. 2004, doi: [10.1023/B:EMSE.0000027778.69251.1f]https://doi.org/10.1023/B:EMSE.0000027778.69251.1f.

[41] 'Google Responsible AI Practices', *Google AI*. https://ai.google/responsibility/responsible-ai-practices/ (accessed May 11, 2023).

[42] I. Goodfellow, Y. Bengio and A. Courville, 'Deep Learning'. https://www.deeplearningbook.org/](https://www.deeplearningbook.org/ (accessed Mar. 31, 2023).

[43] L. Prechelt, B. Unger, M. Philippsen, W. Tichy, 'Two controlled Experiments Assessing the Usefulness of Design Pattern Documentation in Program Maintenance', Accessed: Mar. 24, 2023. [Online]. Available: http://www.ipd.uni-karlsruhe.de/tichy/uploads/publikationen/147/patdoc_tse2002.pdf](http://www.ipd.uni-karlsruhe.de/tichy/uploads/publikationen/147/patdoc_tse2002.pdf

[44] M. Vokáč, W. Tichy, D. I. K. SjØberg, E. Arisholm, and M. Aldrin, 'A Controlled Experiment Comparing the Maintainability of Programs Designed with and without Design Patterns—A Replication in a Real Programming Environment', *Empirical Software Engineering*, vol. 9, no. 3, pp. 149–195, Sep. 2004, doi: [10.1023/B:EMSE.0000027778.69251.1f](https://doi.org/10.1023/B:EMSE.0000027778.69251.1f).

[45] 'Snippet - Visual Studio Marketplace'. https://marketplace.visualstudio.com/items?itemName=vscode-snippet.Snippethttps://marketplace.visualstudio.com/items?itemName=vscode-snippet.Snippet (accessed May 14, 2023).

[46] 'JavaScript (ES6) code snippets - Visual Studio Marketplace'. https://marketplace.visualstudio.com/items?itemName=xabikos.JavaScriptSnippets (accessed May 14, 2023).

[47] 'Pro', *tabnine*. https://tabninesite.wpengine.com/pro (accessed May 14, 2023).

[48] 'Responsible AI Toolbox'. Microsoft, May 13, 2023. Accessed: May 14, 2023. [Online]. Available: https://github.com/microsoft/responsible-ai-toolbox

[49] A. B. Arrieta et al, 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI — Elsevier Enhanced Reader'. https://reader.elsevier.com/reader/sd/pii/S1566253519308103?token=04EF558C38965E8CE5328583FAE6272BBA254F6131A9916B0EE09188E28395A7E5F0012309124073C465902BC2ACCBC2&originRegion=eu-west-1&originCreation=20230516105057 (accessed May 16, 2023).

[50] R. van Kersbergen. 'Personal Interview with Responsible Artificial Intelligence experts', University of Applied Sciences Amsterdam, (2023)

[51] "OpenML." https://www.openml.org/search?type=data&sort=runs&id=179&status=active (accessed May 30, 2023)

[52] Verordening (EU) 2016/679 van het Europees Parlement en de Raad van 27 april 2016 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens en tot intrekking van Richtlijn 95/46/EG (algemene verordening gegevensbescherming) (Voor de EER relevante tekst), vol. 119. 2016. Accessed: Mar. 31, 2023. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj/nld

Appendix A: The Final Process Model

Programming Step	Instrumental Value	Question	Explanation
Define Problem	None, this is the moral imagination	Which problem am I trying to solve? [50]	By identifying the specific problem, you can ensure that the project addresses a real and relevant issue, avoiding the development of AI systems without a clear objective.
		What are the potential costs of the project? [50]	Understanding the potential costs associated with the AI project is crucial for responsible AI design. It helps to assess the resources required, such as financial investments, infrastructure, data collection, and maintenance. Considering the costs upfront enables informed decision-making and ensures that the project is financially viable and sustainable.
		What are the potential benefits of the project? [50]	Evaluating the potential benefits of the AI project is essential to determine its value and impact. Identifying the positive outcomes helps stakeholders assess whether the project aligns with their goals and whether it will bring meaningful improvements to society or individuals.
		Are the costs and benefits of the project in balance? [50]	Assessing whether the potential benefits justify the costs involved helps ensure that resources are allocated wisely. It also helps avoid situations where the costs outweigh the benefits or where unintended negative consequences may arise.
		For who am I creating this application? [50]	By identifying the target audience, you can tailor the system to their needs, ensuring that it provides value and does not inadvertently harm or exclude certain groups. Considering diverse perspectives helps mitigate biases and promotes inclusivity.
		Which problems does my project solve? [50]	This question helps clarify the specific pain points or challenges the project intends to overcome, allowing for a more focused and effective solution. It ensures that the AI system provides tangible benefits and contributes to meaningful problem-solving.
		Which problems does my project not solve? [50]	Identifying what the project cannot achieve helps prevent false assumptions and ensures that stakeholders are aware of its boundaries. Clear communication about the project's scope and limitations fosters transparency and trust.
		Which new problems arise when doing this project? [50]	By considering the potential new problems that may arise, you can proactively identify and address risks or ethical concerns. This question encourages a holistic perspective on the project's impact and promotes responsible AI development by minimizing harm and maximizing positive outcomes.
Construct and Prepare Data	Efficiency	Am I using the right amount of data? [2,50]	Don’t load in your whole dataset immediately. Try your idea first on less data if possible. Loading more data which might not be needed costs more resources. This isn’t efficient and will not contribute to a sustainable way of programming. [2]
		What type of data am I using? [3]	Data can be numerical, categorical, ordinal, or temporal. It is important to know the type of data you are using, because different types of data require different preprocessing steps. For example, categorical data needs to be encoded before it can be used in a model. Also, different types of data require different types of models. For example, you cannot use a linear regression model on categorical data. It comes down to an exploratory data analysis to understand the data you are working with. [23]
		Am I missing data? [2, 4, 50]	Missing data refers not only to missing values in the data like NaN values, but also to features you might miss that should be included when building a model for certain predictions. Missing data can be handled in different ways. For example, you can drop the rows with missing values, or you can impute the missing values with the mean or median of the feature. It is important to know how to handle missing data, because it can have a big impact on the performance of your model. If you miss important features, your model will not be able to make accurate predictions. [23]
		Have I transformed the data correctly? [23]	Data transformation is the process of converting data from one format or structure into another format or structure. This is done to prepare the data for the model. For example, you can transform the data by scaling it, or by encoding categorical data. It is important to know how to transform the data correctly, because it can have a big impact on the performance of your model. If you transform the data incorrectly, your model will not be able to make accurate predictions. Categorical data needs to be encoded before it can be used in a model. Numerical data needs to be scaled before it can be used in a model, because the model will otherwise give more weight to the features with the highest values. [23]
		Am I using the right data? [4]	Possibly the data could contain excessive disturbances or may lack the necessary inputs required to anticipate the intended results. When choosing features, it is advisable to exclude unproductive ones because they contribute little to no value to the model, while also posing the risk of introducing errors such as instability during the operation of the machine learning system. It is a prudent practice to choose only the essential features. It is also advised to avoid using features that are highly correlated with each other, as this may lead to overfitting. Also, try not

			to use sensitive features, as this may lead to bias and thus fairness issues. [23, 42]
		Does the data contain sensitive variables? [35]	Know whether the data you use contains sensitive variables: variables referring to certain characteristics of people, like gender, age, ethnicity etc. This is important, because using these variables without further knowledge of them may result in bias. The literature on this topic defines a list of widely recognized sensitive variables/protected features: -Race -Colour -Sex including gender, pregnancy, sexual orientation, and gender identity -Religion or creed -National origin or ancestry -Citizenship -Age -Pregnancy -Familial status -Physical or mental disability status -Veteran status -Genetic information [34]
		Do any of the variables strongly correlate with the sensitive variables? [2, 35]	If the data contains sensitive variables, it is important to know whether any of the other variables strongly correlate with the sensitive variables. These variables are called proxy variables. If there are proxy variables, it is important to know whether they are necessary for the model. Proxy variable scan be used to unearth sensitive information about the data subjects, even if the sensitive variables are not used in the model. This can result in bias and thus fairness issues. If the proxy variables are not necessary for the model, it is best to remove them from the data. [2]
		Are groups within features over-/underrepresented? [2]	Over- or underrepresentation of groups or classes within features of the data may result in bias and thus fairness issues. If groups or classes are over- or underrepresented, the model will not be able to make accurate predictions for these groups or classes. To fix this, you can try to balance the data by oversampling or undersampling. Oversampling means that you add more data to the underrepresented groups or classes. Undersampling means that you remove data from the overrepresented groups or classes. [23]
	Transparency	Are people in the dataset aware they're in it? [1, 2, 4, 35, 50]	Using information about people without their consent is not responsible and can even be illegal according to - for example - the general data protection regulation [52]. Not abiding by this rule ignores the Human Centrism value
		Does the training data resemble the context of use? [4]	When you're working with data, it's important to define the requirements carefully to make sure that you're only including data that makes sense in the real world. Data that doesn't fit the expected conditions could be considered "anomalies," and you may need to evaluate them manually or exclude them automatically. To help you spot anomalies, you can use visualization tools like bar charts, scatterplots, or line charts, depending on the kind of data you're working with. Ideally, you should try to identify any potential anomalies before you start programming, when you're collecting the data. [23]
		Where and how was it collected, by whom, how is it updated? [4]	You should be aware about whether the data was collected legally with the consent of the people involved and by whom. If data needs to be updated, the same rules apply. [4]
		Is the data available for replication studies?	To be fully transparent, you should be able to publish your data and be open about what you have used to train your model [4]
	Explainability	Can I explain every choice I have made in this step?	-
Build and Train Model	Efficiency	Am I using the right model for the type of task? [50]	It is important to choose the right model for the type of task. At a high level, there are three types of tasks: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is the most common type of task. In supervised learning, the model learns to predict labels from labeled data. In unsupervised learning, the model learns to find patterns in unlabeled data. In reinforcement learning, the model learns to make decisions from trial and error experience. The type of task determines the kind of model you should use. First, determine the type of task you are dealing with and for supervised learning, determine whether it is a regression or classification task. Then, choose the model that is suitable for the type of task. Be sure you understand the assumptions of the model and whether they are valid for your task. Assuming that your task belongs to an "AI-complete" domain such as object recognition, speech recognition, machine translation, and similar ones, it would be advisable to start with a suitable deep learning model, as it is likely to yield good results. Otherwise, you can start with a simple model such as logistic regression or a decision tree. [42]
		Am I using the right optimization algorithm? [50]	One good way to optimize your machine learning model is to use an algorithm called SGD with momentum, which should include a learning rate that decreases

			over time. You can choose different ways to decrease the learning rate, like decreasing it linearly, exponentially, or dividing it by a certain factor each time the validation error plateaus. Another good option is to use a different algorithm called Adam.[42]
		Are you using the right hyperparameters? [50]	When setting up your machine learning model, you can choose the hyperparameters either manually or automatically. If you choose them manually, it's important to understand how they affect things like training error, generalization error, and computational resources (memory and runtime). This means you need to have a good grasp of the key concepts related to how the learning algorithm works. In short, be explainable and transparent. When choosing them automatically, you need to do hyperparameter tuning. When there are three or fewer hyperparameters, choose Grid Search. The obvious problem with grid search is that its computational cost grows exponentially with the number of hyperparameters. With more hyperparameters, use Random Search. [42]
		Am I using the right loss function? [50]	When you're working on a machine learning problem, the kind of loss function you use will depend on whether it's a classification or regression problem. There are many different loss functions to choose from and there's no one "right" function. Usually, people use Mean Squared Error for regression problems and Cross Entropy for classification problems. [42]
		Am I using the right model for the complexity of the task? [50]	When you're working on a machine learning problem, it's usually a good idea to start with simpler models and then gradually make them more complex. Depending on how difficult your problem is, you might even start with a basic statistical model like logistic regression instead of jumping straight to deep learning. If you do use deep learning, make sure you check to see whether your model is either overfitting (memorizing the training data but not generalizing well) or underfitting (not learning enough from the training data), and adjust the complexity of the model accordingly. Be aware that a higher complexity model will need more resources and will become less explainable. Try a simpler model first. [23, 42]
	Transparency	Do I know what I have done and what I am using?	-
	Explainability	Can I explain every choice I have made in this step?	-
Evaluate Model	Efficiency	Have I found balance between the model complexity, explainability, performance and resource demand? [23]	Models should be evaluated by metrics besides performance, namely complexity, explainability and resource demand. These four metrics are all intertwined. A higher model explainability and lower resource demand might cost a few percentage of accuracy or performance on a test set. The model should perform well on all groups included in the data and provide equalized odds while trying to maintain the lowest resource demand and be able to explain it's decisions. Try to find a balance between these metrics. What 'performing' well means for you depends on the project and you should discuss with your stakeholders [23]
		Is the model robust? [23]	Robustness is the model's resiliency to inconsistent inputs and to failures in the execution environment. [23].
		Is the model fair? [2, 4, 23, 35, 50]	No people included in the dataset should be disadvantaged by the decisions made by the model. Everyone should have equal opportunity. [4, 35, 50]
	Transparency	Do I know what I have done and what I am using?	-
	Explainability	Why did the model come to a certain decision?[1, 2, 4, 35, 49]	When you're working with a machine learning model, it's important to be able to understand how it's making its predictions. To do this, it's a good idea to look at the features that have the biggest impact on the predictions and make sure they make sense in the context of the problem you're working on. [23]