Managing Trade-offs in the Nested Iterative Cycles of Responsible AI*

Rohith Sothilingam[1], Vik Pant[1], and Eric Yu[1]

[1] *Faculty of Information, University of Toronto, 140 St George St, Toronto, ON M5S 3G6*

Abstract

This paper addresses the challenge of managing decisions in machine learning (ML) development, where choices in one iterative cycle affect subsequent cycles, each with varying evaluation results. The research objective is to evaluate how well our proposed modeling constructs—Sensors, Actuators, and Iterative Loops—enhance existing goal-oriented conceptual modeling to better analyze decisions in Responsible AI, particularly within nested iterative cycles. We evaluate the efficacy of our proposed goal modeling constructs in analyzing trade-offs among business, technical, and Responsible AI goals using these constructs. Our findings suggest that these constructs improve upon current goal modeling methods, offering more effective decision-making support for Responsible AI outcomes.

Keywords

Goal-Oriented Modeling, Machine Learning, Responsible AI

1. Introduction

Bias and other social responsibility challenges in AI arise from both the underlying Machine Learning (ML) model and the context in which it is used. AI systems, due to inherent model biases, can propagate these issues at scale, affecting numerous user applications [6] [15] [23] [24]. As AI systems are increasingly deployed for critical tasks, concerns about safety and security also escalate.

ML system design involves multiple stages, each with multiple decision points, and iterative cycles, including data gathering, feature engineering, ML model training, deployment, and user output. Responsible AI is an approach within ML-based that integrates fairness, transparency, and ethical considerations at each stage, ensuring that decisions are evaluated not only for technical effectiveness but also for their societal and ethical impact.

Supporting decision-making in iterative ML and Responsible AI cycles is crucial for refining models and improving accuracy by quickly identifying and addressing issues like overfitting or data drift. It ensures that each iteration adds value, ultimately leading to more reliable and robust outcomes. Goal-oriented conceptual modeling is a well-established technique to support systematic decision-making processes [2] [7] [8]. This approach argues that the rationale for system development lies outside the system itself, in the enterprise context. It enables modelers to evaluate goal satisfaction, compare design alternatives, inform requirements, validate design reasoning, and facilitate communication. Through goal refinement, business and Responsible AI goals are broken down into sub-goals and alternative tasks that can achieve goals. Quality objectives are treated as softgoals.

As goals are operationalized in terms of tasks, current goal modeling approaches do not consider how each task alternative may contribute differently to various softgoals, across iterative cycles. To deal with this problem, this paper proposes 3 new goal modeling constructs and examines how they

ER2024: Companion Proceedings of the 43rd International Conference on Conceptual Modeling: ER Forum, Special Topics, Posters and Demos, October 28-31, 2024, Pittsburgh, Pennsylvania, USA

* Corresponding author.

† These authors contributed equally.

rsothilingam@mail.utoronto.ca (R. Sothilingam); vik.pant@mail.utoronto.ca (V.Pant); eric.yu@utoronto.ca (E. Yu)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3849/forum3.pdf

aid in analyzing tradeoffs and conflicts in Responsible AI that are distributed throughout the ML lifecycle. Design decisions at each stage interact and contribute to goals at different stages, with issues like concept drift causing ML processes to evolve over time. Recognizing where tradeoffs occur is crucial, as focusing solely on technical decision points can lead to oversimplified solutions that do not meet other objectives. We explore how tradeoffs at the operationalization level can eventually impact those at the business level, emphasizing the importance of modeling processes and decisions. To illustrate relevant aspects of Responsible AI and their tradeoffs, we draw upon the literature [4] [22]. We use these sources to demonstrate the challenges of dealing with iterative cycles in ML model development and how our proposed Goal Modeling approach can help. We focus on Explainability, Fairness, Privacy, and Accuracy to analyze and demonstrate conflicts at different ML process stages. Our main contribution is enhancing current goal modeling approaches to improve decision-making in Responsible AI design by addressing conflicts between goals in iterative ML cycles.

In Section 2, we first consider tradeoffs between business goals and technical ML goals. Then in Section 3, we introduce the proposed goal modeling notation. In Section 4, we apply the modeling notation to deal with tradeoffs between responsible AI and technical ML goals in various stages of ML development. We discuss related work in Section 5 and conclude with future work in Section 6.

2. Analyzing Tradeoffs in ML Development

In the design of ML-based systems, technical ML objectives can often conflict with business objectives due to competing priorities, leading to tradeoffs [27]. For instance, in the development of a customer feedback system, a technical ML objective might be to prioritize the accuracy of sentiment analysis, which could be achieved by using a Support Vector Machine (SVM) algorithm that maximizes the margin between positive and negative feedback [28]. However, this could conflict with the business objective of improving the ease for customers to provide feedback, as the SVM algorithm might require a large amount of labeled training data to achieve high accuracy, which could be time-consuming and costly to obtain. Additionally, the SVM algorithm might be sensitive to noise and outliers in the data, which could lead to a poorer user experience if customers are required to provide precise and detailed feedback to be understood. In this case, the team may need to make tradeoffs, such as using a simpler ML algorithm that balances accuracy with ease of use, or implementing additional features that help customers provide more effective feedback, such as natural language processing or sentiment analysis tools. This tradeoff would allow the team to meet the business objective of improving the customer experience while still achieving a reasonable level of performance in the ML model.

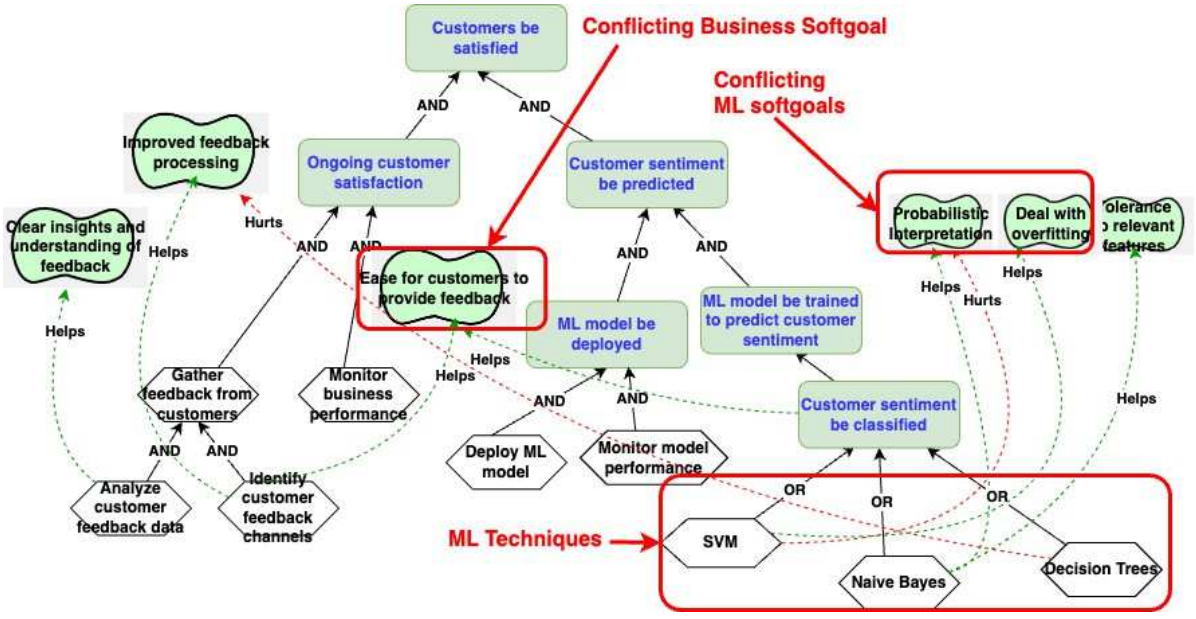

Figure 1: Goal Model conveying an example of tradeoffs that can occur between Business and Machine Learning Goals, to achieve customer satisfaction.

While one can describe tensions and conflicts between various aspects of ML using narrative text, goal modeling supports decision-making, to help solve problems systematically, through incremental steps. For instance, there are known conflicts between explainability and accuracy [22]. Specific techniques, such as ad-hoc methods for explainability, can impede accuracy. But why? By utilizing goal modeling, we can see that specific techniques contribute to one or more softgoals, elucidating why the conflict occurs and at which point in the ML process.

The Goal model in Figure 1 demonstrates how we analyze the above example of conflicting priorities between technical ML and business objectives. The Goal “Customers are satisfied” is refined into two goals: “Ongoing customer satisfaction” (a business goal) and “Customer sentiment be predicted” (an ML goal).

Thus far, these goals have been business goals. To achieve the latter goal, ML goals are now required: "ML model be deployed" and "ML model be trained to predict customer sentiment". We can see how these goals are refined in Figure 1. As we conduct goal and task refinement on the ML goal, we analyze options for ML model techniques (SVM, Naive Bayes, and Decision Trees).

As we attach tasks to achieve the refined goals in Figure 1, the modeling techniques of SVM, Naive Bayes, and Decision Trees are represented as alternative Tasks. The tradeoff explained above is illustrated in this goal model using the positive and negative softgoal contributions among these Tasks. By conducting goal refinement, then identifying and analyzing conflicting contributions of task alternatives to softgoals, we can identify tradeoffs which occur during design decisions when choosing techniques, by deciding to prioritize between specific softgoals (e.g. interpretability of the model and improving the ease for customers to provide feedback).

Though the examples of goal modeling presented in this section is useful and allow us to identify simple tradeoffs, it does not allow us to identify tradeoffs that occur at different stages of the ML lifecycle. Specifically, the tradeoffs and decisions in these examples occur at the same stage of model training. In Figure 1, this can be seen as the tradeoff occurs at the same task refinement level, as we refine the goal of "ML model be trained to predict customer sentiment". This can lead to wrong or poor decisions because, through the goal modeling, we cannot trace the positive and negative effects that the softgoal contributions have to other stages in the ML lifecycle, such as data preparation, feature engineering, or business decisions such as cost effectiveness.

Decisions get made at different iterative cycles in the ML lifecycle. For example, in Figure 1, the primary goal of “ML model be trained to predict customer sentiment” would involve iterations where model training is continuously done until the stopping criteria is met. To achieve the success of “ongoing customer satisfaction”, this goal will involve iterations of continuous monitoring. When we drill down and expand into these goals, further tradeoffs appear with respect to how computational, business, and Responsible AI goals must be simultaneously achieved. Traditional goal modeling does not allow us to analyze the distribution of goals and tradeoffs at different iterative cycles, as well as how they interact with each other. In the remainder of this paper, we take the goal modeling a step further by analyzing conflicts between technical ML and Responsible AI goals at decision points distributed through various stages of the ML lifecycle.

3. Introducing the Proposed Goal Modeling Approach for Responsible AI

We propose three modeling constructs to help us improve our design of ML processes concerning appropriate decisions at each iterative cycle: Sensors, Actuators, and Iterative Loops (Figure 2). A metamodel is shown in Figure 3. In each iteration, based on the most recent information, decisions

are made and actions taken, to incrementally get closer to meeting the objectives. Since there are multiple decisions aiming to meet multiple interacting objectives, it is

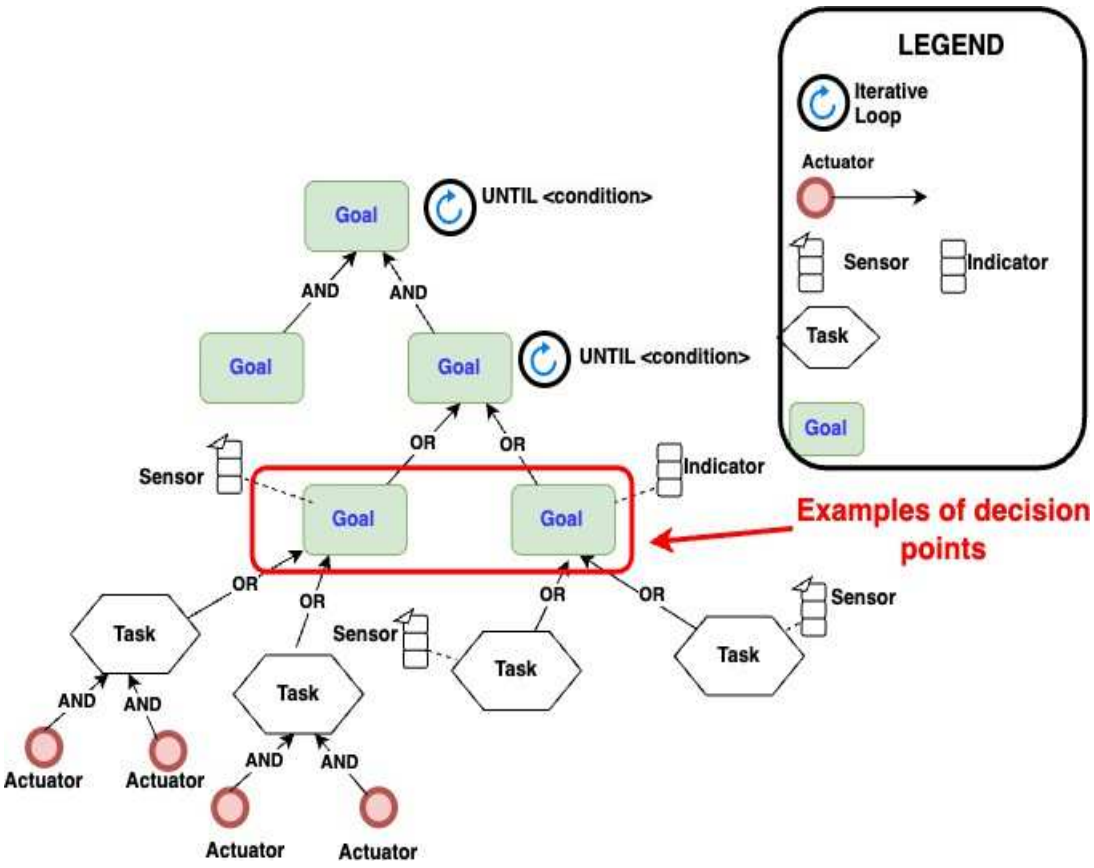

Figure 2: Proposed Modeling Notation conveying Sensors, Actuators, and Iterative Loops.

important to position the decision points and their associated information collection and actions appropriately in the nested iterative structures. Together, the three modeling constructs aim to facilitate systematic reasoning that ensures that decisions at each iteration are purposeful and aligned with the overarching goals of ML model development.

The concepts of goals and tasks are drawn from i* [8] [33]. Sensors and Actuators are used in an abstract conceptual sense and do not refer to physical devices. Sensors collect information from the environment. Information can be collected through tasks, or when pursuing goals. Sensor variables are used as input for decisions. An Indicator [12] is one type of Sensor. They are associated with goals so as to indicate how well the goals are achieved.

Actuators are used to manipulate the environment through Tasks. Actuator variables are settings for parameters in tasks. They are outputs of decisions, and can be thought of as levers or knobs for adjusting values. A Task may manipulate multiple Actuators.

Iterative Loops are associated with goals. They repeat until a condition, the stopping criteria, is reached. When a goal that has an Iterative Loop is refined into a subgoal that also has an iterative loop, the latter loop is said to be nested inside the former loop. The latter loop is the inner loop and the former the outer loop. The inner loop is iterated multiple times for each iteration of the outer loop.

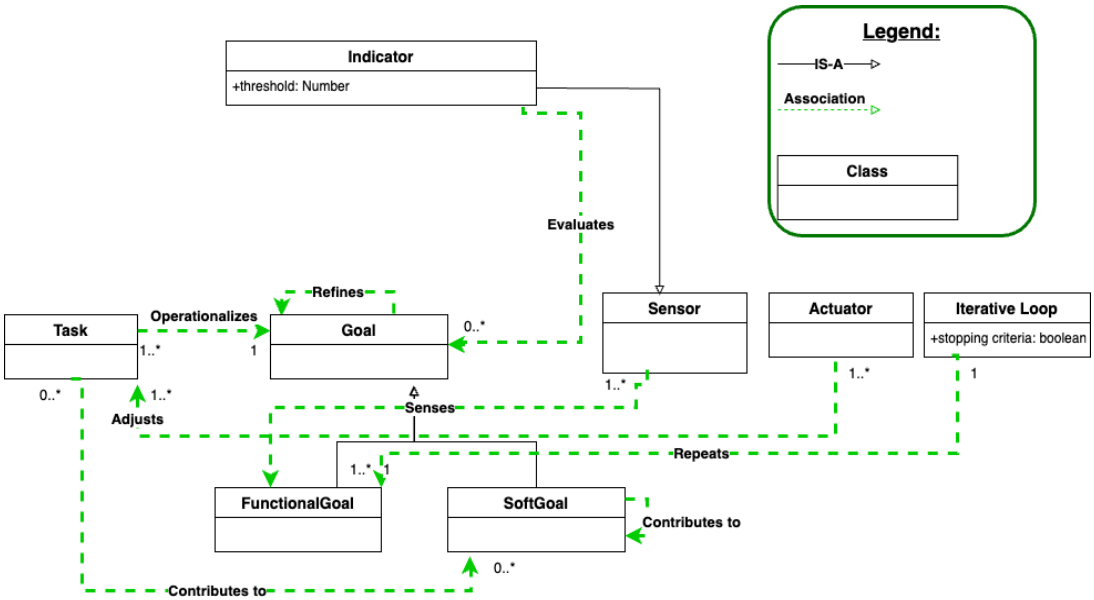

Figure 3: Metamodel for proposed Goal Modeling Framework, conveying Sensors, Actuators, and Iterative Loops.

Consider the following example. In the initial goal model below (Figure 4), we provide a detailed illustration of the various conflicts that can arise between different aspects of Responsible AI, specifically focusing on privacy and fairness at a high level. Below, we break down the initial set of useful features in this goal model.

In Figure 4, Fairness and Privacy are conveyed as separate goals with their own set of alternative tasks which can achieve the respective goal. For each of the goals of Fairness and Privacy, there is an Indicator which is used as a gauge to determine the success of the goal. To achieve the "***Privacy Methods be Set***" goal, the Indicator "***Composite Privacy Score***" is calculated to meet its desired threshold. To gauge whether this Indicator can be met, each of these alternative techniques, have a Sensor which *senses* a specific value to determine if the Indicator threshold has been met. For example, if the task T-closeness is chosen, the Sensor of "***T-closeness***" collects the ***T-closeness*** data value, which is used to gauge the success of the Indicator "***Composite Privacy Score***". Task refinement allows for representing conditional softgoal contributions to Responsible AI goals based on the choice of alternative techniques. For instance, K-value (a technique to achieve privacy) can negatively impact the softgoal "Balanced Accuracy."

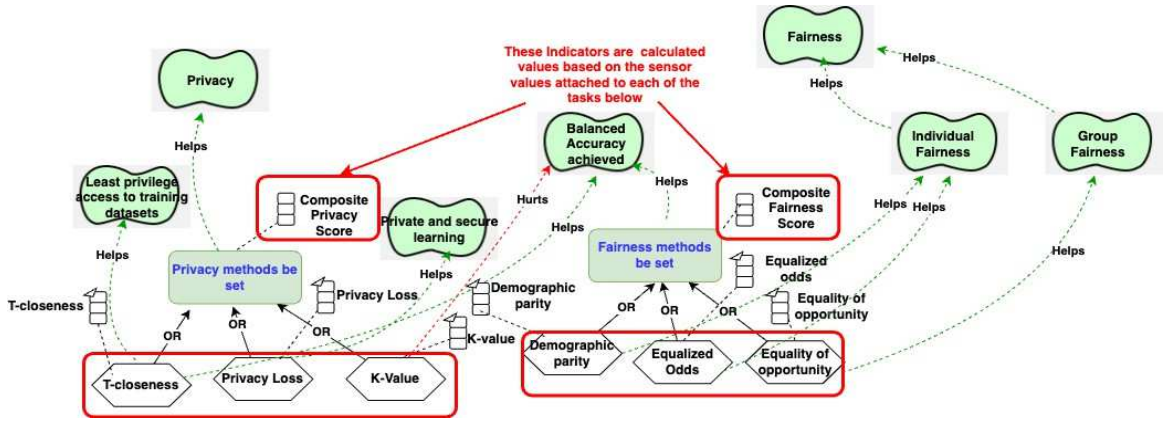

Figure 4: Goal Model illustrating tradeoffs between aspects of Responsible AI: Fairness, Privacy, Explainability, and Accuracy.

4. Responsible AI Decisions and Tradeoffs along different Iterative Cycles in the ML Process

4.1. General Goal Model of the ML Lifecycle

Let us consider the following challenge of Principal Component Analysis (PCA), which is a technique option that can help with feature dimensionality reduction but affects fairness. Traditional Principal Component Analysis (PCA) is not designed with fairness in mind and may perpetuate biases, leading to unequal reconstruction errors across different demographic groups, resulting in potentially harmful and unfair outcomes [20].

Let us consider the goals involved. In Figure 5, we present the following parent goals: Model algorithms be set and features be transformed. Together, these goals would eventually support the producing the prediction. Upon identifying the parent goals, refine the parent goals into further sub-goals until we can identify alternative techniques (tasks) for accomplishing those goals. When refining the goals, we ensure that the topic is consistent. We refine the goal "Features be transformed" into the following sub-goals: "Features be normalized" "Features be encoded", and "Feature Dimensionality Reduction". We do not yet refine the goal Model algorithm(s) be set because we can identify alternative techniques for this goal.

Next, we identify the alternative tasks that can accomplish each of the sub-goals. Upon identifying these tasks, we attach softgoal contributions (help and hurt) to each softgoal. By identifying the softgoal contributions, we can visualize tradeoffs that arise as a result of choosing one alternative technique over another.

Finally, we attach the Actuators and Sensors to each Goal where they apply in Figure 5 to identify specifically where the tradeoff occurs and why. Toward the right of the model, we can see that we can see that the tradeoff between PCA and LDA can negatively affect the success of Balanced Accuracy, which in turn eventually affects Fairness, through softgoal contributions. This is helpful because it gives us a visual breakdown of why choosing PCA can eventually hurt fairness while providing technical benefits in feature generalization and noise reduction. However, how does this conflict then affect the larger ML development process? At what point in the iterative loops involved does this conflict occur and how does it affect other stages?

The conflict occurs at the "feature transformation" stage, where the PCA technique can be chosen as a dimensionality reduction technique for achieving feature transformation. This technique then negatively affects the group fairness softgoal "mortality prediction output be fair across groups". The Sensors "Explained Variance Ratio" and "Discriminant Power" are used as inputs to determine when

Dimensionality Reduction has successfully converged (the Iterative loop stopping criteria for this goal). These Sensors serve as inputs for consideration to adjust the "number of components" Actuator for each of the PCA and LDA techniques.

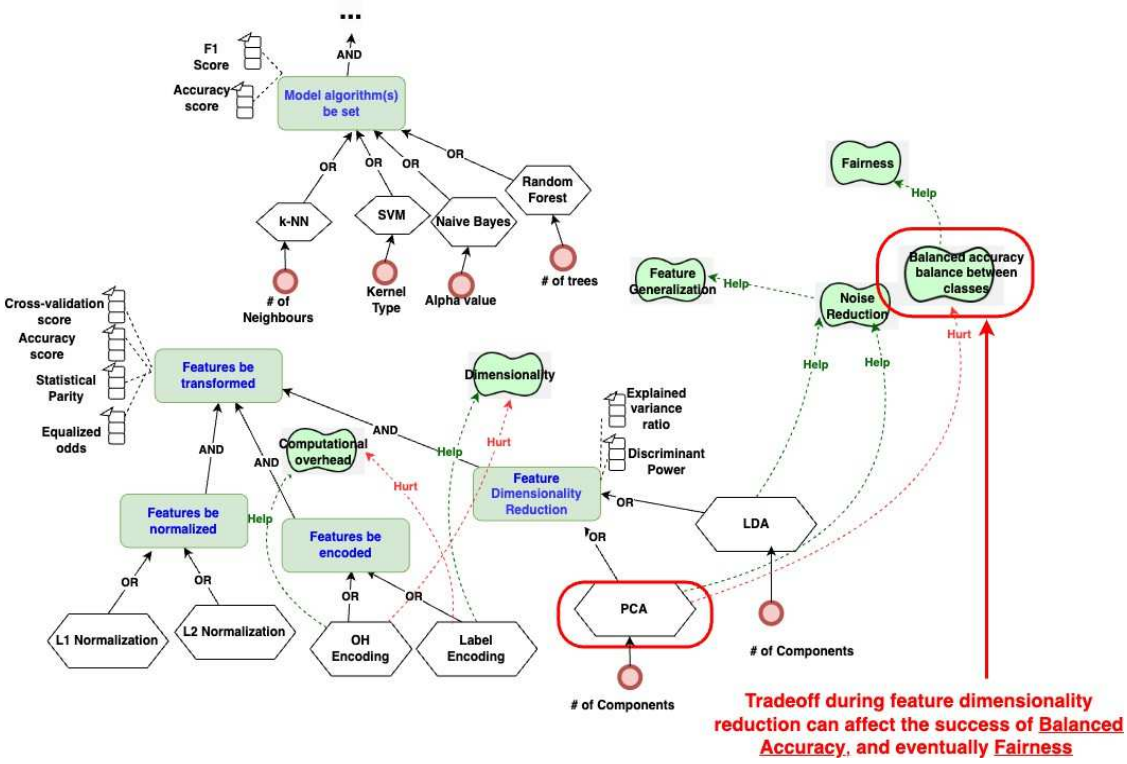

Figure 5: Goal Model illustrating a tradeoff between Fairness and Feature Generalization

4.2. Conflicting Responsible AI Goals in a Case Study

In this section, we introduce the context of a recent case study [4] that we will draw upon to build on the previous model, toward a comprehensive goal model that captures conflicts between interpretability, explainability, accuracy, and fairness, using a case study example. In this case study, the authors conducted empirical research on conflicts arising between healthcare stakeholders due to ethical concerns with ML applications in healthcare. The authors map the relationships between stakeholders and potential "values-collisions," identifying several themes of conflict. For our purposes, we focus on the following themes:

- Bias and perpetuation of bias (Bias)
- Conflicting values and perspectives on death and end-of-life care (Fairness)
- Transparency and evaluation of efficacy (Transparency)
- Determining the recipients of ML output (Explainability)
- Patient consent and involvement (Privacy)

The ML model evaluated in the case study aims to identify what individuals might benefit from advance care planning by addressing a proxy problem: predicting the probability of a given patient passing away within the next 12 months, to aid in palliative care consults. Based on the outcome of the mortality prediction, patients will have the option of being notified if they ought to be considered for advance care planning based on the mortality prediction.

In the following subsections, for the purpose of our goal modeling in this paper, we will focus on the prediction of mortality rates, and build upon the initial goal model (Figure 4). This model will use the case study [4] to illustrate specific aspects of fairness, accuracy, privacy, and explainability appearing in various goals. Each of these Responsible AI goals has been further refined, and conflicts are observed at different goal refinement points, representing various stages of the broader ML process. This approach enables us to illustrate how different Responsible AI challenges are distributed and interact throughout the ML lifecycle, using an empirical case study.

4.3. Conflicts within Responsible AI: Interpretability, Explainability, and Fairness

Conflicts often arise not only between technical ML priorities and Responsible AI objectives but also among different aspects of Responsible AI itself [22]. For instance, a model prioritizing fairness may compromise transparency, as fairness metrics might involve complex calculations that are challenging to interpret. Regarding explainability and interpretability, a model emphasizing explainability might sacrifice interpretability, as explanations could necessitate simplifications or approximations that obscure the original model's nuances.

In the preceding sections, we explored (1) conflicts between Responsible AI and technical ML objectives (e.g., feature generalization vs. fairness) and (2) conflicts between distinct Responsible AI goals (e.g., explainability vs. fairness). In Figure 6, we illustrate the broader ML lifecycle as it relates to the case study. In this goal model, the nested iterative stages are represented in Loop 1.1 and Loop 1.2 nested within Outer Loop 1. This initial goal model provides us with a breakdown of where each functional goal and set of tasks exist with respect to the nested iterative loops they are a part of.

Building on this model, Figure 6 below introduces softgoal contributions and illustrates a goal model encompassing the broader context of technical ML, business, and Responsible AI goals throughout the ML lifecycle. This figure provides a comprehensive view, analyzing the interactions among three critical aspects of Responsible AI: interpretability, explainability, and fairness.

In this example, we identify two primary conflicts: (1) between fairness and interpretability and (2) between fairness and explainability. Regarding the first conflict, depicted on the left side of the goal model, fairness and interpretability are at odds because interpretability can enhance "Tolerance to outliers," which adversely affects the goal of "mortality prediction output being fair across groups," thereby undermining group fairness. Through softgoal refinement, it becomes evident that interpretability indirectly impacts fairness by enhancing tolerance to outliers. This figure allows us to visualize conflicts that can occur across different stages in the ML development process.

The second conflict, between fairness and explainability, arises because the Fairness softgoal of "mortality prediction output being fair across groups" may compromise decision trustworthiness and, consequently, the softgoal of "explainable prediction result." When considering these tradeoffs among different Responsible AI aspects represented as softgoals, the model designer must evaluate and prioritize these objectives accordingly.

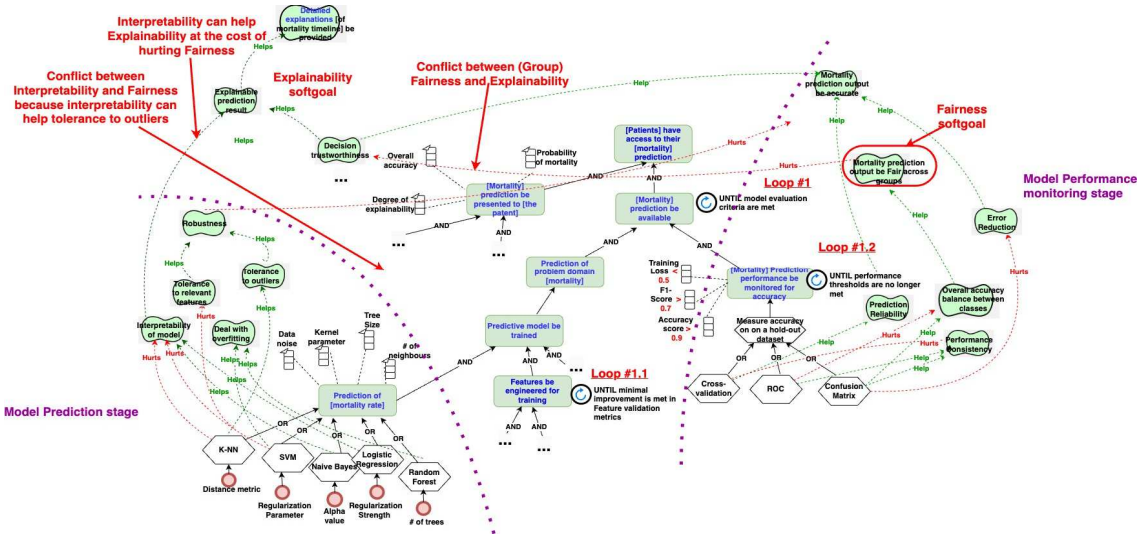

Figure 6: Goal Model conveying tradeoffs in case study between various aspsects of Responsible AI: Interpretability, Explainability, and Fairness.

5. Related Work

5.1. Checklists, Guidelines, and Principles

Principle-based approaches are utilized within specific guidelines and ethical frameworks for Responsible AI. These approaches are often prescriptive to specific contexts and issues, rather than being universally applicable to the broader spectrum of Responsible AI. Current methodologies are constrained to addressing a finite set of ethical concerns, such as explainability, fairness, privacy, and accountability. They lack inclusivity regarding the various sub-concepts, perspectives, and interpretations of Responsible AI. Translating a list of ethical objectives into actionable steps poses significant challenges, including determining the most appropriate metric or technique for each use case.

Principle-based approaches, standards, and guidelines (e.g., [13]) are designed to be universal, aiming to apply to all projects. However, requirements are inherently project-specific. Often, principles may conflict with one another, and some may not be relevant or meaningful within the project's specific context. Through Goal Modeling, principles (represented as softgoals) can be refined according to specific Responsible AI contexts, rather than adhering to a finite set of principles applied uniformly across all contexts.

Goal modeling facilitates the clarification and operationalization of vague or ambiguous requirements through goal refinement. Our approach extends the advantages of goal modeling by offering a reasoned and systematic methodology for making design decisions at various stages of the ML lifecycle.

5.2. Computational Techniques for Responsible AI

Initial research on fairness predominantly concentrated on formulating quantitative definitions of fairness (see, e.g., [9], [11], [31]) and developing technical methods for 'debiasing' AI models in accordance with these mathematical formalizations (see, e.g., [1], [3], [34]).

As the application of computational techniques proves valuable in addressing challenges within this domain, the notion of Responsible AI is increasingly recognized as contextual. This necessitates greater attention to the varying definitions and needs of Responsible AI, alongside the specific practices and requirements of practitioners. The inherent complexities and contextual nuances of fairness make it impractical to fully de-bias an AI system or guarantee its fairness [14], [21]. The primary objective, therefore, is to mitigate fairness-related harms and other adverse outcomes to the greatest extent possible ([16], [19]).

It is crucial to approach ML as a holistic process, actively considering the diverse social perspectives, stakeholders, and interactions involved. For example, Srivastava et al. [30] discovered that competing definitions of fairness often do not align with established mathematical definitions. Current computational techniques (e.g. [10]) and tools (e.g. [9]) provide conceptual frameworks that facilitate decision-support for data-driven applications. However, these tools lack critical reasoning capabilities, such as tradeoff mechanisms, goal refinement processes, and the operationalization of those goals.

5.3. Inadequacies of Current GORE Approaches

Kuwajima and Ishikawa [13] proposed a goal-oriented conceptual modeling approach that adheres to the Ethics guidelines for trustworthy AI set forth by the European Commission. While this approach is methodical, it is constrained by its narrow focus on a singular dimension of Responsible AI. It does not encompass the diverse interpretations of Responsible AI, such as fairness, explainability, security, and privacy. Consequently, it is ill-equipped to address the conflicting goals and priorities that arise from these varied interpretations. In contrast, our proposed approach is designed to be versatile and adaptable, accommodating multiple lenses and perspectives to suit any specific context within Responsible AI cases.

GR4ML is another related framework that employs a goal-oriented approach to link analytics and business goals [18]. However, GR4ML falls short in addressing the interrelationships and trade-offs between these goals, particularly within the scope of Responsible AI. To our knowledge, our approach represents the first goal-oriented conceptual modeling framework specifically tailored for Responsible AI.

Existing goal-oriented modeling languages exhibit limited capabilities in integrating Sensors, Actuators, and nested Iterative Loops. Although awareness requirements and adaptive systems in goal modeling address some aspects of sensing, they remain inadequate. For instance, Morandini et al. (2008) present a goal-oriented approach for designing self-adaptive systems, emphasizing the engineering of self-adaptive software.

Awareness Requirements [25] are defined as requirements that reference other requirements or domain assumptions, monitoring their success or failure at run-time. This type of reasoning facilitates adaptability by supporting the monitoring, diagnosis, planning, and execution of requirements. Our proposed Sensor modeling construct extends this concept by enabling inputs from the causal world to inform decisions based on sensed variable values, thereby facilitating interaction with the non-intentional world.

6. Conclusions and Future Research

The design of Responsible AI solutions necessitates a systematic approach to accommodate the dynamically evolving decision points inherent in ML processes. This involves the alignment of both business and Responsible AI objectives, alongside the meticulous analysis of conflicts and trade-offs that emerge throughout the nested stages of the ML lifecycle. While contemporary goal modeling approaches offer potential value for designing Responsible AI solutions, they fall short in effectively supporting the analysis of nested iterative cycles in ML development specific to Responsible AI. This paper introduces three novel modeling constructs as part of an innovative goal modeling methodology aimed at systematically designing Responsible AI solutions. Given the benefits of the approach presented, we acknowledge the added complexity. Modelers would have to weigh whether the problem context warrants the added expressiveness and analytical capabilities from using the proposed approach.

In future work, we will augment our proposed goal modeling framework by integrating Agent-Oriented (AO) modeling. Specifically, we will explore how conflicting stakeholder goals might impact the modeling process or the resulting AI solutions. The various stages of the ML life- cycle often involve distinct individuals, where conflicts arising at each stage can be more localized than the current goal models suggest, requiring acknowledgment of the interests and cultural contexts of

these individuals. Understanding how humans and AI co-evolve as a hybrid learning system within organizations is a critical area of exploration. Academic discourse has advocated for viewing human-AI systems as collaborative and co-creating rather than merely co-existing systems [32]. In this context, we propose the application of Agent- Oriented conceptual modeling to dissect and analyze conflicts and trade-offs among stakeholders during Responsible AI projects, thereby guiding the design of Responsible AI solutions in a manner that systematically balances diverse values, goals, and interests.

To demonstrate the utility of our modeling approach, we will focus on enhancing the initial analysis and results of the study by identifying the following:

- Strategic interests (i.e., values) of actors involved and the conflicts arising (1) between the interests of each actor and (2) among the subsequent goals in which they are involved.
- Specific points in the ML process where these actors are engaged.
- Extension of the goal modeling to examine how conflicts within nested iterative cycles in the ML lifecycle interact with the interests and priorities of actors.

Subsequent development of our conceptual modeling framework will involve its application and validation through an empirical case study to assess its practical relevance in real-world settings. The framework will incorporate knowledge catalogs to aid in the design of Responsible AI solutions, and this research will identify the necessary catalogs. A comprehensive methodology and detailed guidelines will be formulated for the use of the new framework, encompassing phases such as Modeling, Evaluation, Exploration, and Implementation. We will also explore options for tool development to support our proposed approach.

References

[1] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H. (2018). A reductions approach to fair classification. In International conference on machine learning (pp. 60-69). PMLR.

[2] Amyot, D. (2003). Introduction to the user requirements notation: learning by example. Computer Networks, 42(3), 285-301.

[3] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., Kalai, A. T. (2016). Man is computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

[4] Cagliero, D., Deuitch, N., Shah, N., Feudtner, C., Char, D. (2023). A framework to identify ethical concerns with ML-guided care workflows: a case study of mortality prediction to guide advance care planning. Journal of the American Medical Informatics Association, 30(5), 819-827.

[5] Castro, J., Kolp, M., Mylopoulos, J. (2001). A requirements-driven development methodology. In Advanced Information Systems Engineering: 13th International Conference, CAiSE 2001 Interlaken, Switzerland, June 4–8, 2001 Proceedings 13 (pp. 108-123). Springer Berlin Heidelberg.

[6] Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180080.

[7] Chung, L., Nixon, B. A., Yu, E., Mylopoulos, J. (2012). Non-functional requirements in software engineering (Vol. 5). Springer Science Business Media.

[8] Dalpiaz, F., Franch, X., Horkoff, J. (2016). istar 2.0 language guide. arXiv preprint arXiv:1605.07767.

[9] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214- 226).

[10] Hajian, S., Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. IEEE transactions on knowledge and data engineering, 25(7), 1445-1459.

[11] Hardt, M., Price, E., Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.
[12] Horkoff, J., Barone, D., Jiang, L., Yu, E., Amyot, D., Borgida, A., Mylopoulos, J. (2014). business modeling: representation and reasoning. Software Systems Modeling, 13, 1015-1041.
[13] Kuwajima, H., Ishikawa, F. (2019). Adapting square for quality assessment of artificial intelligence systems. In 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW) (pp. 13-18). IEEE.
[14] Kleinberg, J., Mullainathan, S., Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.
[15] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689.
[16] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.
[17] Morandini, M., Penserini, L., Perini, A. (2008). Towards goal-oriented development of self-adaptive systems. In Proceedings of the 2008 international workshop on Software engineering for adaptive and self-managing systems (pp. 9-16).
[18] Nalchigar, S., Yu, E. (2020). Designing business analytics solutions: a model-driven approach. Business Information Systems Engineering, 62(1), 61-75.
[19] Norval, C., Cornelius, K., Cobbe, J., Singh, J. (2022). Disclosure by Design: Designing information disclosures to support meaningful transparency and accountability. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 679-690).
[20] Pelegrina, G. D., Duarte, L. T. (2023). A novel approach for Fair Principal Component Analysis based on eigendecomposition. IEEE Transactions on Artificial Intelligence.
[21] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K. Q. (2017). On fairness and calibration. Advances in neural information processing systems, 30.
[22] Sanderson, C., Douglas, D., Lu, Q. (2023). Implementing responsible AI: Tensions and trade- offs between ethics aspects. In 2023 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.
[23] Scheuerman, M. K., Paul, J. M., Brubaker, J. R. (2019). How computers see gender: An evaluation of gender classification in commercial facial analysis services. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-33.
[24] Singh, J., Cobbe, J., Norval, C. (2018). Decision provenance: Harnessing data flow for accountable systems. IEEE Access, 7, 6562-6574.
[25] Silva Souza, V. E., Lapouchnian, A., Robinson, W. N., Mylopoulos, J. (2011). Awareness requirements for adaptive systems. In Proceedings of the 6th international symposium on Software engineering for adaptive and self-managing systems (pp. 60-69).
[26] Sothilingam & Yu, "A A Goal-Oriented Approach for Modeling Decisions in ML Processes", *2024 IEEE 32nd International Requirements Engineering Conference Workshops (REW),* Reykjavik, Iceland, 2024, pp.321-325, doi: 10.1109/REW61692.2024.00048.
[27] Sothilingam, R.,, "A Requirements-Driven Conceptual Modeling Framework for Responsible AI," *2023 IEEE 31st International Requirements Engineering Conference (RE),* Hannover, Germany, 2023, pp. 391-395, doi: 10.1109/RE57278.2023.00061.
[28] Sothilingam, R., Pant, V., and Yu, E., "Using i* to Analyze Collaboration Challenges in MLOps Project Teams", *Proceedings of the 15th International i* Workshop 2022*, 2022.
[29] Sothilingam, R., and Yu, E., "Modeling Agents Roles and Positions in Machine Learning Project Organizations", *Proceedings of the 13th International i* Workshop 2020*, vol. 2641, pp. 61-66, 2020.
[30] Srivastava, M., Heidari, H., Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery data mining (pp. 2459- 2468).
[31] Verma, S., Rubin, J. (2018). Fairness definitions explained. In Proceedings of the international workshop on software fairness (pp. 1-7).

[32] Waardenburg, L., Huysman, M. (2022). From coexistence to co-creation: Blurring boundaries in the age of AI. Information and Organisation, 32(4).
[33] Yu, E., Giorgini, P., Maiden, N., Mylopoulos, J. (2011). Social Modeling for Requirements Engineering (Cooperative Information Systems Series). MIT Press.
[34] Zafar, M. B., Valera, I., Gomez Rodriguez, M., Gummadi, K. P. (2017). Fairness beyond disparate treatment disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th international conference on world wide web (pp. 1171-1180).