Toward a Goal-Oriented Argumentation Approach for Fair ML Measures Using i*

Rohith Sothilingam[1], Eric Yu[1,2]

[1]Faculty of Information, University of Toronto, 140 St George St, Toronto, ON M5S 3G6
[2]Department of Computer Science, University of Toronto, 40 St George St, Toronto, ON M5S 2E4

Abstract

In the Fair Machine Learning (ML) literature, there are two well-known distinct classes of measures: Group and Individual Fairness. It is argued in a recent study by Binns [1] that, while past works in Fair ML assert that these concepts are in conflict, from a political and legal philosophical perspective, these concepts are not fundamentally in conflict. Using goal-oriented reasoning, we complement the argument made by [1], whereby this apparent conflict occurs due to early assumptions made rather than as a result of choosing between these two categories of Fair ML measures. We demonstrate the ability of i* modeling to support the design of Fair ML solutions which accommodate both Individual and Group Fairness measures. We explore the i* concept of *Belief* [2], which is based on that of Claim in the NFR framework [3], which in turn draws on argumentation frameworks. The i* concept of *Belief* is used to represent assumptions made. We use i* to link the Fair ML measures of Individual and Group fairness to philosophical paradigms using Softgoal refinement and justification of *Beliefs*.

Keywords

Fair ML, Requirements Engineering, Responsible AI

1. Introduction

Recognition of opportunities in Data Science is quickly reshaping many discipline areas, and is driven not just by data itself but all other aspects that could be created and transformed by understanding, exploring, and using data. While rapid advancements have been made in the area, there has been a growing focus on the social impact of Data Science, from a social responsibility perspective. This focus has led to the growing advocacy toward the notion of Responsible Data Science.

As one area of Responsible Data Science, Fair Machine Learning (ML), of which recent studies have emerged (e.g. [4]). Measuring and assessing the impact and "fairness" of ML-based systems is central to responsible recommendation efforts. However, the complexity of fairness definitions and the proliferation of fairness metrics in research literature have led to a complex decision-making space [5]. It is challenging for practitioners to operationalize Fairness objectives and pick metrics that work within their unique context. This challenge suggests that practitioners require more decision-making support, but it is not clear what type of support would be beneficial. As one approach to support argumentation, Requirements Engineering

iStar'23: 16th International Workshop at the 31st IEEE International Requirements Engineering Conference, September 04–08, 2023, Hannover, Germany
rohith.sothilingam@mail.utoronto.ca (R. Sothilingam); eric.yu@utoronto.ca (E. Yu)

(RE) and Goal-Oriented RE (GORE) enable systematic reasoning for decision-making processes. There has been recent work on argumentation in RE and GORE, which has been applied to other areas of Software engineering (e.g. [6] [7]).

RE, and more specifically, GORE offer promising approaches to support decision-making and reasoning for designing Responsible Data Science solutions. In this paper, we have adopted a greatly simplified approach to argumentation, following the NFR Framework [3]. We use i* modeling to support the argumentation from a recent case study [1]. This case study draws upon Fair ML literature to argue that Individual Fairness and Group Fairness are two distinct measures which are not fundamentally in conflict. This paper contributes to i* with respect to demonstrating its ability as a qualitative reasoning approach as a step towards an argumentation framework for Responsible Data Science. The main contribution of this paper is its demonstration of the ability of i* to support such argumentation in Fair ML and utilize the i* concept of *Belief*, to convey early assumptions in Fair ML decision processes and the conflicts which can occur due to such assumptions and philosophical ideals.

The concept of *Belief* was introduced in the original i* 1.0 [8]. The existence of assumptions is typical during the design of ML models. An example of an assumption commonly held in ML is that disentanglement is useful for downstream tasks, for example through a decreased sample complexity of learning [9]. Similarly, the concept of *Beliefs* is part of both GRL [10] and the NFR framework [3]. In GRL, *Beliefs* aim at capturing reasons behind selecting certain goals or tasks. For example, Murukannaiah et al. [11] proposed a GRL-based approach for capturing inconsistencies between stakeholders' goals and *Beliefs*, and resolving goal conflicts.

2. Apparent Conflict between Individual and Group Fairness

There are two approaches discussed as alternatives of criteria to achieve the goal of "Fair ML": Individual Fairness and Group Fairness [1] (Fig. 1). On the one hand, group fairness ensures some form of statistical parity (e.g. between positive outcomes, or errors) for members of different protected groups (e.g. gender or race) [12]. On the other hand, individual fairness ensures that people who are 'similar' with respect to the classification task receive similar outcomes [12] [13]. It is argued by [1] that this apparent tradeoff exists due to assumptions made early in the decision-process, rather than directly between Individual and Group Fairness.

In Fig. 1 we convey this apparent tradeoff between Individual and Group Fairness using an i* Goal Model, to illustrate and analyze the constituent factors behind this tradeoff. A tradeoff between Individual and Group Fairness occurs at the following level of goal refinement: Primary differences in goals between Individual Fairness vs. Group Fairness is the Goal of the ML Model to obtain *similar error rates* vs. *lowest error rates* across *groups* vs *individuals*.

In the case of Group Fairness, if chosen, a conflict occurs with the underlying Goal of ***Similar error rates across groups***, which *helps* the Softgoal of ***Overall Accuracy be maximized*** while it simultaneously *helps* the Softgoal of ***Statistical Parity be achieved between protected groups in each outcome class***.

In the case of Individual Fairness, if chosen, a conflict occurs with the underlying Goal of ***Lowest error rates across individuals***, which *helps* the Softgoal of ***Differences between group accuracies be minimized*** while it simultaneously *hurts* the Softgoal of ***Statistical Parity be***

Figure 1: i* Goal Model of Apparent Conflict between Individual and Group Fairness as two Opposing Options for Fair-ML Criteria

achieved between protected groups in each outcome class.

Borrowing from BIM [14], Indicators are used to measure the *performance* of each of these Goals: ***Task-relevance Similarity (distance)*** and ***Accuracy of judgements in each case***.

Further tradeoffs exist with regard to the extent to which Individual Fairness can be achieved. In order to achieve the goal of ***lowest error rates of individuals***, this can be done by two alternative operationalizations: (i) ***Use individual metric*** or (ii) ***Use task-relevant distance metric***.

In the case of Individual Metric, true known labels must already be known and mapped in the dataset (i* task: ***Map to existing true known labels***) (which occurs infrequently) and future true labels must be assumed (i* task: ***Assume future true labels***).

3. Balancing Individual and Group Fairness

Binns [1] maps each of Individual and Group Fairness to the following philosophical concepts for fairness: Egalitarianism and Consistency. Based on this mapping exercise, it is argued that there are many different ways to put together a set of measures to achieve each strand of Egalitarianism and Consistency. However, a systematic, consistent approach is not provided for

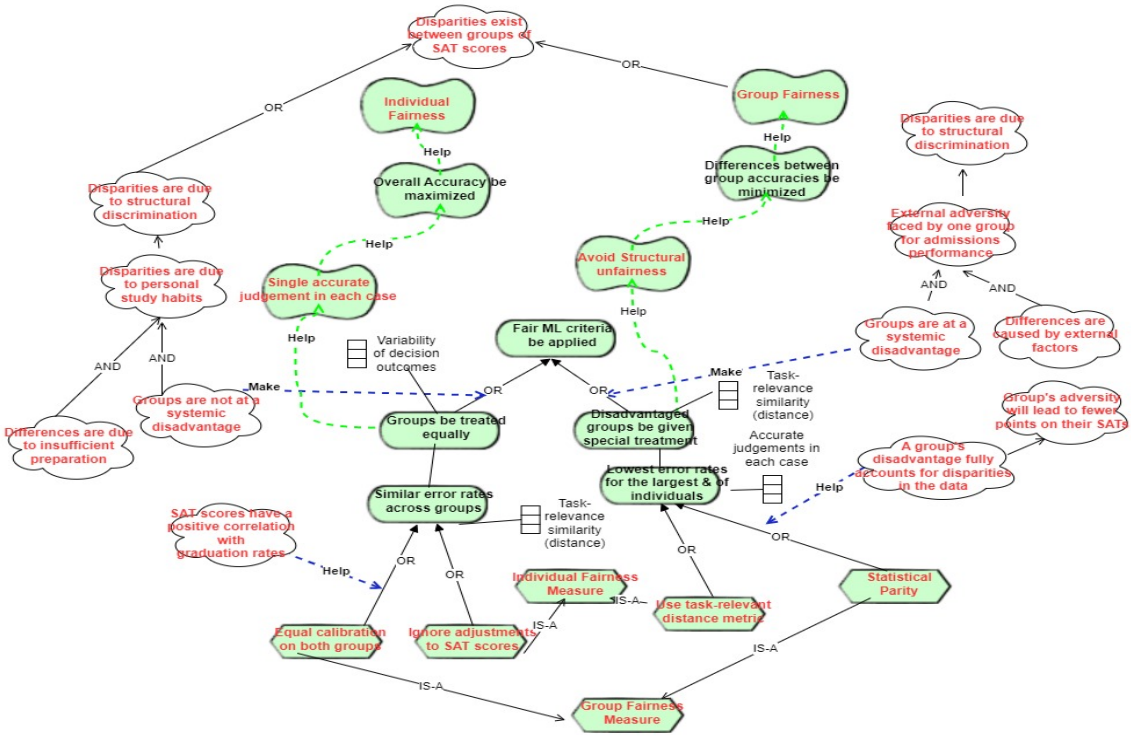

Figure 2: i* Goal Model of College Admissions Case Study [1]

choosing among that set of measures.

The goal model in Fig.2 is a goal model which supports the argumentation presented in a case study [1], of which each of Individual and Group Fairness measures are considered for the fair consideration of College Admissions. The goal model is specific to this case study and aims to highlight how alternatives for Individual or Group Fairness measure can both be used, dependent on higher-level assumptions. New modeling elements (Goals, Softgoals, and Beliefs) which did not appear in Fig. 1 are highlighted in orange text.

Each of the goals of ***Groups be treated equally*** and ***Disadvantaged groups be given special treatment*** contribute to Softgoals which eventually, through Softgoal refinement, contribute to each of ***Individual Fairness*** and ***Group Fairness*** respectively. Measures (i.e. i* tasks) which can be grouped as either Individual or Group Fairness are available as options for both sets of goals. This grouping is done by mapping these measures as i* tasks to the task of either ***Group Fairness Measure*** or ***Individual Fairness Measure*** using an **IS-A** link. For example, the task "Equal calibration on both groups" is linked to the task ***Group Fairness Measure*** by using an **IS-A** link. Ultimately, this relationship complements the argument from [1]. We can see that the apparent tradeoff between Individual Fairness and Group Fairness are in fact caused by a refinement of several assumptions which are made in the decision process.

Assumptions can be seen to cause the apparent tradeoff between using measures which

would typically be categorized as either Individual Fairness vs. Group Fairness measures. Once an assumption has been agreed upon by the decision-maker(s), appropriate measures can then be chosen. For example, the goal of "Fair ML criteria" can be achieved depending on which assumption is agreed upon (i.e. conveyed by i* *Beliefs*): ***Groups are not at a systemic disadvantage*** OR ***Differences be caused by external factors outside of their control***, of which the relationship is conveyed using the **Make** link.

In i*, *Beliefs* can be further refined to further substantiate. For example, the Belief ***Groups are not at a systemic disadvantage*** is supported by the *Belief* ***SAT*** [1] ***and graduation rate disparities are due to personal choice***, which is the ultimate justification given by the decision-makers to substantiate their position, as described by [1]. This refinement graph represents how the decision-making process led to this final judgement determination, allowing the reader to understand the constituent factors which could lead to particular assumptions, which then ultimately affects the type of measure to be used (i.e. which type of either Individual Fairness or Group Fairness measure and the specific reason as to why). *Beliefs* ultimately affect goal model evaluation. If a *Belief* is overturned, then the alternative it supports would also be invalidated. For example, the *Belief* ***Groups are not at a systemic advantage*** is supported by three *Beliefs* using **AND** links: ***Disparities are not due to structural discrimination***, ***Disparities are due to personal study habits***, and ***Differences are due to insufficient preparation***. The satisfaction of these *Beliefs* together will lead to the full satisfaction of the *Belief* ***Groups are not at a systemic disadvantage***.

The qualitative reasoning used in this model supports structured argumentation, whereby we can determine whether elements of the i* model in Fig. 2 determine whether the elements of the goal model presented are acceptable given stakeholders' *Beliefs* as well as the potential contradictory evidence (i.e. tradeoffs from the *Beliefs*).

4. Conclusions and Ongoing Work

In this work, we explored and demonstrated how i* goal modeling can be used to support argumentation of and identify where the apparent tradeoff about conceptions of fairness (individual vs. group fairness) occurs and "diagnose" the validity of this tradeoff (Fig. 2). Future work will investigate potential incorporation of advances from more recent work in argumentation [6] [7]. This paper is a part of ongoing work towards larger PhD thesis research [15] [16] [17] with objectives, which include the following: (1) a requirements-driven framework which deals with conflicting goals at design decision points throughout Responsible AI; (2) compilation and codification of design knowledge from pertinent literature on Responsible AI to be available during design decision in the form of knowledge catalogs; (3) tool support for the proposed framework.

References

[1] R. Binns, On the Apparent Conflict Between Individual and Group Fairness, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020) 514–524.

[1]The SAT ("Scholastic Aptitude Test") is a standardized college admissions test in the United States.

[2] E. Yu, P. Giorgini, N. Maiden, J. Mylopoulos, Social Modeling for Requirements Engineering, MIT press, 2011.
[3] L. Chung, B. A. Nixon, E. Yu, J. Mylopoulos, Non-functional requirements in software engineering, volume 5, Springer Science & Business Media, 2012.
[4] M. Veale, R. Binns, Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data, Big Data & Society 4 (2017) 2053951717743530.
[5] J. J. Smith, L. Beattie, H. Cramer, Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective, Proceedings of the ACM Web Conference 2023 (2023) 3648–3659.
[6] Y. Elrakaiby, A. Borgida, A. Ferrari, J. Mylopoulos, CaRE: a Refinement Calculus for Requirements Engineering based on Argumentation Theory, Software and Systems Modeling 21 (2022) 2113–2132.
[7] M. van Zee, F. Bex, S. Ghanavati, Rational GRL: A Framework for Argumentation and Goal Modeling, Argument & Computation 12 (2021) 191–245.
[8] E. Yu, Towards Modelling and Reasoning Support for Early-Phase Requirements Engineering, Proceedings of ISRE'97: 3rd IEEE International Symposium on Requirements Engineering (1997) 226–235.
[9] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem, Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, International Conference on Machine Learning (2019) 4114–4124.
[10] J. Castro, M. Kolp, J. Mylopoulos, A Requirements-Driven Development Methodology, Advanced Information Systems Engineering: 13th International Conference, CAiSE 2001 Interlaken, Switzerland, June 4–8, 2001 Proceedings 13 (2001) 108–123.
[11] P. K. Murukannaiah, A. K. Kalia, P. R. Telangy, M. P. Singh, Resolving Goal Conflicts via Argumentation-based Analysis of Competing Hypotheses, 2015 IEEE 23rd International Requirements Engineering Conference (RE) (2015) 156–165.
[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness Through Awareness, Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (2012).
[13] P. Lahoti, K. P. Gummadi, G. Weikum, ifair: Learning Individually Fair Data Representations for Algorithmic Decision Making, 2019 IEEE 35th International Conference on Data Engineering (ICDE) (2019) 1334–1345.
[14] J. Horkoff, D. Barone, L. Jiang, E. Yu, D. Amyot, A. Borgida, J. Mylopoulos, Strategic Business Modeling: Representation and Reasoning, Software & Systems Modeling 13 (2014) 1015–1041.
[15] R. Sothilingam, Analyzing Organizational Processes in Machine Learning Projects: Exploring Modeling Approaches, Master's thesis, Faculty of Information, University of Toronto, 2020.
[16] R. Sothilingam, V. Pant, E. Yu, Using i* to Analyze Collaboration Challenges in MLOps Project Teams, Proceedings of the 15th International i* Workshop 2022: Hyderabad, India, October 2022. (2022).
[17] R. Sothilingam, E. Yu, Modeling Agents, Roles, and Positions in Machine Learning Project Organizations., Proceedings of the 13th International i* Workshop 2020: Zurich, Switzerland, September, 2020, Vol. 2641. (2020) 61–66.