Investigation of Artificial Mental Models for Healthcare AI Systems

Sabine Janzen[1], Wolfgang Maass[1,2], and Prajvi Saxena[1]

[1] German Research Center for Artificial Intelligence, Saarbrücken, Germany
{sabine.janzen,wolfgang.maass,prajvi.saxena}@dfki.de
[2] Saarland University, Saarbrücken, Germany

Abstract. In the evolving landscape of healthcare, personalized Artificial Intelligence (AI) systems are vital for patient-centered care. However, patients facing health challenges often struggle with cognitive limitations, leading to incomplete or biased data that hinders their decision-making abilities. To address this issue, this research in progress explores the concept of Artificial Mental Models (AMM) within healthcare AI systems. AMMs are meta-representations of patient mental models, capturing their understanding and assumptions about therapy and rehabilitation processes. We present a research design for investigating AMMs in healthcare AI systems that adopts a Design Science Research (DSR) approach consisting of four iterative phases: elicitation, individualization, action, and transfer. In the elicitation phase, discrimination-free basis models are generated through web scraping and synthetic patient data. The individualization phase fine-tunes AMMs for individual patients by incorporating diverse data sources. The action phase integrates AMMs into AI systems and evaluates their real-world impact. The transfer phase applies the resulting framework to support therapy decisions for patients with compromised decision-making abilities. This research aims to enhance therapy outcomes and patient care while advancing the understanding of mental models in healthcare.

Keywords: Artificial mental models · Healthcare AI systems · Design Science research · Patient-centered care.

1 Introduction

In the rapidly evolving domain of healthcare and well-being, adaptive and personalized Artificial Intelligence (AI) systems have emerged as pivotal enablers of patient-centered care. These systems promise to deliver support tailored to individual patient needs by harnessing detailed insights into user behaviors and situational contexts [21, 34, 4, 25]. But, patients grappling with illness, pain, or therapeutic risk decisions, exhibit cognitive limitations. Such limitations often lead to the generation of incomplete, inaccurate, or biased data, severely undermining patients' ability to engage in informed decision-making, comprehend complex medical narratives, or articulate their symptoms and concerns effectively [33, 26, 9, 8]. Acknowledging the critical impact of these cognitive barriers,

it becomes imperative to leverage AI healthcare systems designed to mitigate these gaps, thereby ensuring the provision of care that is optimally aligned with each patient's unique needs and circumstances [21, 34]. Understanding the dynamics of mental models [29] in therapeutic contexts and their application in related AI systems presents a novel avenue for enhancing patient care and therapy outcomes. Mental models serve as cognitive frameworks, enabling individuals to interpret their surroundings and anticipate system behaviors [18]. Unlike static knowledge repositories, these models are fluid, continually shaped by experiences and interactions. Related work emphasizes the importance of accurately capturing and understanding these models, particularly within therapeutic settings [15, 11, 28, 2]. The challenge lies in eliciting and conceptualizing these mental models to create a meta-representation that encompasses the patient's understanding and assumptions about their therapy and rehabilitation process [19, 23, 5, 17].
Our research in progress aims to explore the utility of those meta-representations of mental models called artificial mental models (AMM) within healthcare AI systems to bolster patient support in making informed decisions under conditions characterized by uncertainty and risk. To address the aforementioned challenges, this paper presents a comprehensive research design for eliciting, individualizing, and integrating AMM into healthcare AI systems, with a focus on improving therapy outcomes and patient care. The research design adopts a Design science research (DSR) methodology [31], outlining a multi-phase research design with iterative build and evaluate cycles, each aimed at refining the conceptualization and application of AMM within healthcare AI systems. By systematically reviewing mental models in therapeutic contexts and employing advanced AI and machine learning techniques, this research intends to contribute to the evolving field of AI in healthcare [21, 34, 4, 25]. The iterative design process, informed by both technical and subject-based experiments, aims to create bias-free, patient-centric models that reflect the complex realities of rehabilitation and therapy.
We will focus on two use cases: (1) enhancing post-knee/hip surgery rehabilitation outcomes and (2) providing decision support for therapy options to patients with compromised decision-making capabilities, such as those suffering from dementia. As a result, we intend to enhance the efficacy of AI-supported therapies but also to advance the understanding of mental models in healthcare, paving the way for more personalized and effective treatment strategies.

2 Mental Models

Mental models are dynamic 'working models' [6, 18] that serve as cognitive frameworks that individuals construct to understand and interact with complex, dynamic systems. They are not static repositories of knowledge but are continually evolving as they are shaped by experiences and interactions [19]. In therapeutic contexts, recognizing the plurality of stakeholder's perceptions, values, and goals is a key aspect of effective therapy. Each stakeholder brings a unique mental model to the table, influenced by their background, values, and personal objectives. Understanding these various mental models is crucial as it can inform the

development of AI systems that enhance effective therapies, support improved decision-making processes, and help to identify and rectify patients' knowledge limitations and misconceptions. According to Norman (1983) the complex, dy-

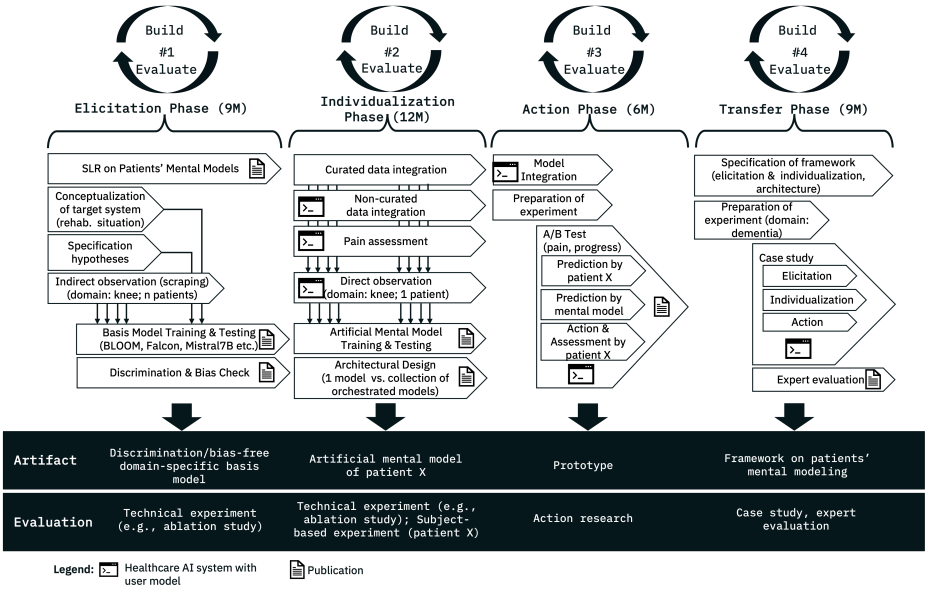

Fig. 1. Research design for investigating Artificial Mental Models (AMM) in healthcare AI systems

namic systems patients are interacting with can be defined as the target system (t), i.e., the surrounding physical system a patient is facing in a rehabilitation situation. This system does not exist in isolation; rather, it is ensconced within the broader context of the patient's environment [29]. The conceptual model of target system ($C(t)$) is an appropriate representation of the physical system (accurate, consistent, complete). The unknown patient's mental model of the target system ($M(t)$) reflects her beliefs about the physical system and allow the user to understand and to anticipate the behavior of t. The mental model naturally evolves through interaction with the target system. As $M(t)$ is unknown, we need to anticipate the patient's mental model in form of a conceptualization of that mental model $C(M(t))$; means a model of the $M(t)$. This meta-representation strives to encapsulate the relevant components of a patient's belief system about the target system. To elicit the constructs of these models, it is imperative to engage directly with patients. This necessitates conducting psychological experiments and meticulous observations to capture the nuances of the patient-system interaction. But, the process of eliciting this internal representation $C(M(t)$ is a critical initial challenge. Most current procedures operate under the premise that a meta-representation of an individual's mental model

can be visualized as a network of concepts and relationships. For instance, the Conceptual Content Cognitive Map (3CM) method [20] is designed to directly elicit a network representation of a mental model from an interviewee through a diagrammatic interview. Indirect elicitation techniques, conversely, involve the researcher reconstructing the network from oral or written responses; e.g., by applying fuzzy cognitive maps [12, 13, 3]. This endeavor raises several methodical questions, particularly concerning the relative strengths and weaknesses of direct versus indirect $C(M(t))$ elicitation and the specific techniques used, e.g., missing handling of the dynamic and evolving character of mental models [6, 18]. Furthermore, the distinction between 'espoused theories' — what patients claim they believe—and 'theories in use'—how they actually behave—is crucial. Understanding whether the elicited meta-representation $C(M(t))$ reflects the patient's stated beliefs or their practical application is vital [1]. Discrepancies between these can explain the often-observed conflicts between mental and behavioral models [19]. Since mental models are assumed to underpin reasoning, decision-making, and behavior, it is often the 'theory-in-use' that is of most interest.

3 Research Design for Investigating Artificial Mental Models

For handling the aforementioned challenges in elicitation of artificial mental models $C(M(t))$, we specified a research design according to Design Science Research (DSR) [16, 32] covering four iterative build & evaluate cycles framing four main phases - **elicitation, individualization, action and transfer** – in form of a multi-phase process model indicating input and outpoint points for the healthcare AI system as well as planned publications (cf. Fig. 1). DSR is an approach commonly used in various fields such as information systems, engineering, and applied sciences to create and evaluate artifacts intended to solve identified problems. The design process is iterative, involving cycles of development, testing, and refinement. Within each cycle, an artifact is created and evaluated (cf. Fig. 1) [30]. This artifact can be a model, a method, a framework, or a technology (prototype). For evaluation of artifacts, we will employ various methods, such as technical experiments, subject-based experiments, prototyping, action research or case studies. Emphasizing the generation of new knowledge, the research design includes not only insights about the specific artifact created but also contributions to theories and practices in the field, providing valuable insights for future research and practice (publications) (cf. Fig. 1). The outcome of this work is thus twofold: practical, problem-solving artifacts and advancements in scientific knowledge through results gained in the build and evaluation cycles.

3.1 Elicitation Phase

Objective of the elicitation phase is the generation of a discrimination- and bias-free domain-specific basis model $C(M(t))$ in the domain of knee rehabilitation

(use case 1) (cf. Fig. 1). This phase involves a systematic literature review (SLR) on patients' mental models. For eliciting such a meta-representation of a patient's mental model, indirect observation of patients will be applied by using a twofold scraping approach for building a large dataset. Once open user forums, social media and channels with related topics like rehabilitation, knee surgery, physical therapy etc. will be scraped (focus: Germany according to special features of healthcare system). Here, we will use libraries such as tweepy[3], beautifulSoup[4], scrapy[5] for web scraping. Additionally, synthetic patient communication data will generated by usage of Large Language Models (LLM) (ChatGPT); *"Take the role of a patient after knee surgery: How do you feel today?"* Driven by specified hypotheses and a conceptualization of the target system ($C(t)$), means a conceptual model of rehabilitation situations, the resulting dataset is used for training of a basis model. For building this model, an open foundation model like BLOOM[6], Mistral7B[7] or Falcon[8] will be used. For ensuring fairness and mitigating bias in the basis model, discrimination and bias checks are planned. Here a combination of methods will be investigated, for instance, counterfactual fairness testing [22], adversarial debiasing [24], fairness metrics such as equality of opportunity [14] or demographic parity [36], and interpretability [10]. The resulting artifact – a discrimination- and bias-free domain-specific basis model - will be evaluated in a technical experiment, e.g., an ablation study.

3.2 Individualization Phase

The individualization phase intends to use the basis model for building and fine-tuning an artificial mental model (AMM) for an individual patient X (cf. Fig. 1). For training the model, curated data like medication, rehabilitation therapy plans, data on injury, surgical procedure and duration, complications, as well as non-curated data like movement data, sleep, fitness status etc. partially provided by the AI system have to be integrated. A further data component will be pain assessment by patient X. Here a multimodal approach will be employed to measure pain, combining both subjective and objective measures to gain a comprehensive understanding of the patient's experience, e.g., self-reported pain scales (e.g., Visual Analog Scale (VAS) [7], Wong-Baker FACES Pain Rating Scale [35]); behavioral and physiological indicators like body movements, vocalizations, heart rate, sweating; and pain diaries. When training the AMM diverse hypotheses on optimal architecture design will be investigated in this phase, namely the decision for training one model versus a collection of orchestrated models [27]. The resulting artifact – an AMM for patient X - will be evaluated in a technical experiment (e.g., ablation study) as well as a subject-based exper-

[3] https://docs.tweepy.org/en/stable/api.html
[4] https://pypi.org/project/beautifulsoup4/
[5] https://github.com/scrapy/scrapy
[6] https://bigscience.huggingface.co/blog/bloom
[7] https://huggingface.co/mistralai
[8] https://huggingface.co/tiiuae/falcon-40b

iment in form of a pre-study involving patient X and the AMM in preparation for the upcoming action phase.

3.3 Action Phase

Objective of the action phase is the integration of the model into a prototype of an AMM-powered AI system in healthcare and its evaluation by means of an action research approach (cf. Fig. 1). Means, the use of the prototype in the real-world rehabilitation situation of patient X as part of a research intervention, evaluating its effect on the real-world situation. For the experimental setting a kind of A/B test is planned with focus on the anticipation respectively prediction of (1) pain in a therapy unit, and (2) the therapy progress. The predictions of (1) and (2) are generated in two variants; (A) by patient X, and (B) by the AMM. For ground truthing, predictions are mirrored with a subsequent actual pain assessment by patient X in a therapy unit as well as actual assessment of therapy progress by patient X and the therapist. Objective of the study is to evaluate the overlapping of predictions of patient X and the AMM in concrete cases as well as their accuracy with respect to ground truth.

3.4 Transfer Phase

Within the transfer phase results of previous phases are transferred to use case 2 describing therapy decision support for patients with compromised decision-making abilities (e.g., dementia) (cf. Fig. 1). Therefore, a framework on patients' mental modeling is specified including elicitation and individualization processes and the architectural design of AMM determined from the build & evaluate cycles 1-3. The framework is evaluated by means of a case study supported by an expert evaluation. In a planned cooperation with the German Center for Neurodegenerative Diseases (DZNE), the experiment is prepared with respect to constraints of the new domain with respect to elicitation and individualization of AMMs. Within the case study, the framework is applied in the domain of therapy decision support for dementia patients (use case 2). Results and implications are cross-checked in an expert evaluation for assessing the framework by one or more experts.

4 Contribution and Limitations

The research in progress presented in this paper contributes to the evolving field of healthcare AI systems by addressing critical challenges in personalized patient care. In contexts characterized by uncertainty and risk, e.g., therapy decision-making for patients with compromised cognitive abilities, improved decision support can be provided by using AMMs. The adoption of a DSR methodology facilitates the stepwise refinement of conceptualization and application of AMMs in healthcare AI systems through multiple build and evaluate cycles and multimodal data integration. Thus, the creation of bias-free, patient-centric

models is enabled that reflect the complexities of rehabilitation and therapy. Nonetheless, there are limitations. The process of eliciting and individualizing AMMs from patients presents significant challenges as the accuracy and completeness of elicited models may be influenced by factors such as patient variability and data quality. While the action phase includes real-world evaluation of AMM-powered AI systems, the scope of evaluation may be limited to specific use cases such as post-surgery rehabilitation and therapy decision support for dementia patients. Generalizability to broader healthcare contexts or diverse patient populations may require additional validation and testing. Here, implementing AMM-powered AI systems in broader healthcare settings necessitates interdisciplinary expertise and resource-intensive efforts from data collection and model training to system deployment and evaluation. Last, challenges related to algorithmic fairness, interpretability, and privacy in AMM development remain. Besides acknowledging the importance of fairness metrics and bias checks, ongoing monitoring and adaptation in AMM development, training and inferencing is required to address emerging ethical concerns.

5 Conclusion and Future Work

This research in progress paper explored the concept of Artificial Mental Models (AMM) within the realm of healthcare AI systems, emphasizing their potential to enhance patient support and improve therapy outcomes. We presented a research design based on a systematic Design Science Research (DSR) approach covering phases of elicitation, individualization, action, and transfer for developing and evaluating AMMs tailored for specific healthcare use cases. The approach starts with the elicitation of domain-specific basis models, devoid of discrimination and bias, and advanced towards the individualization of these models towards concrete patient needs. The action phase intends to demonstrate the practical application and efficacy of AMMs in real-world rehabilitation scenarios. The transfer phase is designed to further validate the adaptability and scalability of the resulting framework in further therapy contexts.Looking ahead, the proposed research design will be applied in a research project (2024 - 2026) covering a multitude of directions for future exploration and refinement. A primary focus will be on expanding the scope of AMMs to encompass a broader spectrum of healthcare domains, thereby amplifying their impact across diverse patient demographics and conditions. Additionally, we aim to explore the enhancement of AMM interpretability and transparency. As these models become more intricate, ensuring that they remain understandable and accountable to healthcare professionals and patients alike is imperative. This includes developing methods for explaining model decisions and predictions in a manner that is accessible and meaningful to non-technical stakeholders. Finally, collaborative efforts with healthcare practitioners, patients, and regulatory bodies will be essential in refining and validating AMM frameworks.

References

1. Argyris, C., Schon, D.A.: Theory in practice: Increasing professional effectiveness. John Wiley & Sons (1992)
2. Barber, T., Crick, K., Toon, L., Tate, J., Kelm, K., Novak, K., Yeung, R.O., Tandon, P., Sadowski, D.C., Veldhuyzen van Zanten, S., et al.: Gastroscopy for dyspepsia: Understanding primary care and gastroenterologist mental models of practice: A cognitive task analysis approach. Journal of the Canadian Association of Gastroenterology **6**(6), 234–243 (2023)
3. Bernard, D., Cussat-Blanc, S., Giabbanelli, P.J.: Fast generation of heterogeneous mental models from longitudinal data by combining genetic algorithms and fuzzy cognitive maps. In: Proceedings of the Hawaii International Conference on System Sciences. pp. 1570–1579 (2023), https://api.semanticscholar.org/CorpusID:256902783
4. Bollos, L.A.C.L., Zhao, Y., Soriano, G.P., Tanioka, T., Otsuka, H., Locsin, R.: Technologies, physician's caring competency, and patient centered care: A systematic review. The Journal of Medical Investigation **70**(3.4), 307–316 (2023)
5. Borders, J., Klein, G., Besuijen, R.: Mental model matrix: Implications for system design and training. Journal of Cognitive Engineering and Decision Making p. 15553434231226317 (2024)
6. Craik, K.J.W.: The nature of explanation, vol. 445. CUP Archive (1967)
7. Crichton, N.: Visual analogue scale (vas). J Clin Nurs **10**(5), 706–6 (2001)
8. Dildine, T.C., Amir, C.M., Parsons, J., Atlas, L.Y.: How pain-related facial expressions are evaluated in relation to gender, race, and emotion. Affective Science pp. 1–20 (2023)
9. Dildine, T.C., Necka, E.A., Atlas, L.Y.: Confidence in subjective pain is predicted by reaction time during decision making. Scientific reports **10**(1), 21373 (2020)
10. Doshi-Velez, F., et al.: Towards a rigorous science of interpretable machine learning (2017), https://ar5iv.org/abs/1702.08608
11. Gabbas, M., Kim, K.: Gamified user interface design for dysphagia rehabilitation based on common mental models. In: DRS2022, Bilbao, Spain (2022)
12. Gray, S.A., Gray, S., Cox, L.J., Henly-Shepard, S.: Mental modeler: A fuzzy-logic cognitive mapping modeling tool for adaptive environmental management. In: Proceedings of the 46th Hawaii International Conference on System Sciences. pp. 965–973. IEEE (2013). https://doi.org/10.1109/HICSS.2013.399
13. Gray, S.A., Zanre, E., Gray, S.: Fuzzy cognitive maps as representations of mental models and group beliefs. In: Papageorgiou, E.I. (ed.) Fuzzy Cognitive Maps for Applied Sciences and Engineering: From Fundamentals to Extensions and Learning Algorithms, pp. 29–48. Springer, Berlin, Heidelberg (2014)
14. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems. vol. 29 (2016), https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html
15. Hermans, A., Muhammad, S., Treur, J.: You feel so familiar, you feel so different: A controlled adaptive network model for attachment patterns as adaptive mental models. In: Mental Models and Their Dynamics, Adaptation, and Control: A Self-Modeling Network Modeling Approach, pp. 321–346. Springer (2022)
16. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Quaterly **28**, 75–105 (2004), https://api.semanticscholar.org/CorpusID:13553735

17. Im, J., Evans, J.M., Grudniewicz, A., Boeckxstaens, P., Upshur, R., Steele Gray, C.: On the same page? a qualitative study of shared mental models in an interprofessional, inter-organizational team implementing goal-oriented care. Journal of Interprofessional Care **37**(4), 549–557 (2023)
18. Johnson-Laird, P.N.: Mental models. The MIT Press (1989)
19. Jones, N.A., Ross, H., Lynam, T., Perez, P., Leitch, A.M.: Mental models: an interdisciplinary synthesis of theory and methods. Ecology and Society **16**(1), 46 (2011), https://api.semanticscholar.org/CorpusID:38976887
20. Kearney, A.R., Kaplan, S.: Toward a methodology for the measurement of knowledge structures of ordinary people. Environment and Behavior **29**, 579–617 (1997), https://api.semanticscholar.org/CorpusID:143280871
21. Khera, R., Butte, A.J., Berkwits, M., Hswen, Y., Flanagin, A., Park, H., Curfman, G., Bibbins-Domingo, K.: Ai in medicine—jama's focus on clinical outcomes, patient-centered care, quality, and equity. Jama (2023)
22. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in Neural Information Processing Systems. vol. 30 (2017), https://papers.nips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
23. LaMere, K., Mäntyniemi, S., Vanhatalo, J., Haapasaari, P.: Making the most of mental models: Advancing the methodology for mental model elicitation and documentation with expert stakeholders. Environmental Modelling & Software **124**, 104589 (2020)
24. Lemoine, B., Zhang, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the ACM Conference (2018), https://research.google/pubs/mitigating-unwanted-biases-with-adversarial-learning/
25. Li, W., Ge, X., Liu, S., Xu, L., Zhai, X., Yu, L.: Opportunities and challenges of traditional chinese medicine doctors in the era of artificial intelligence. Frontiers in Medicine **10** (2023)
26. Meyer, A.N., Giardina, T.D., Khawaja, L., Singh, H.: Patient and clinician experiences of uncertainty in the diagnostic process: current understanding and future directions. Patient Education and Counseling **104**(11), 2606–2615 (2021)
27. Mohammed, A., Kora, R.: A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University-Computer and Information Sciences (2023)
28. Naik, A.D.: Collaborative decision-making: Identifying and aligning care with the health priorities of older adults. In: Geriatric Medicine: A Person Centered Evidence Based Approach, pp. 1–21. Springer (2023)
29. Norman, D.A.: Some observations on mental models, pp. 241–244. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1987)
30. Peffers, K., Rothenberger, M., Tuunanen, T., Vaezi, R.: Design science research evaluation. In: Design Science Research in Information Systems. Advances in Theory and Practice: 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012. Proceedings 7. pp. 398–410. Springer (2012)
31. Peffers, K., Tuunanen, T., Niehaves, B.: Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research. European Journal of Information Systems **27**(2), 129–139 (2018)
32. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A design science research methodology for information systems research. Journal of Management Information Systems **24**, 45 – 77 (2007), https://api.semanticscholar.org/CorpusID:17511997

33. Timm, A., Schmidt-Wilcke, T., Blenk, S., Studer, B.: Altered social decision making in patients with chronic pain. Psychological Medicine **53**(6), 2466–2475 (2023)
34. Wan, T.T., Wan, H.S.: Predictive analytics with a transdisciplinary framework in promoting patient-centric care of polychronic conditions: Trends, challenges, and solutions. AI **4**(3), 482–490 (2023)
35. Wong, D.L., Baker, C.M.: Wong-baker faces pain rating scale. Pain Management Nursing (2012)
36. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 962–970. PMLR (2017), https://proceedings.mlr.press/v54/zafar17a.html