ABSTRACT

ALVAREZ, LAUREN STEPHANIE. Safety, AI Ethics, and Education: A Multidisciplinary Investigation of Responsible AI and Trustworthy ML Development. (Under the direction of Chair Veronica Cateté).

This dissertation is an interdisciplinary investigation and analysis of trustworthy machine learning (TML) and responsible artificial intelligence (RAI) research and development. Each chapter presents a different research approach for examining safety, AI ethics, and TML education. TML and RAI are emerging fields exploring the challenges and effectiveness of traditional machine learning methods applied to socio-technical environments. Socio-technical implementations are technical solutions to real-world problems requiring varying degrees of sociological and ethical knowledge. To appropriately address socio-technical problems and assess the societal risk and harms, practitioners must be trained on AI ethics, safety, responsibility, and trustworthiness early on.

Computer Science education, particularly AI education, is a growing field that has historically focused on algorithms, optimization, and mathematical foundations of AI. RAI and advanced AI ethics courses are not widely available to undergraduates. Currently, RAI is considered an advanced topic offered at limited universities for graduate students and advanced undergraduates. This work connects safety, AI ethics, and education theory to investigate useful methods for teaching RAI to undergraduates early in their academic careers. The contributions of this dissertation are (1) an abolitionist framework for critical engagement in AI ethics, (2) a framework for trustworthy implementation development, and (3) a curriculum to teach RAI for undergraduates. The main conclusions of this work are (1) RAI and TML implementation is inherently challenging and requires varying levels of AI ethics and sociological considerations, difficult to teach active CS practitioners, and (2) a thematic analysis of undergraduate feedback on an introductory course designed to teach RAI and TML concepts without extensive technical experience. As the field grows, more research is needed to determine the most effective pedagogical methods to educate future AI practitioners on the complexities and ethical challenges of socio-technical AI applications.

Safety, AI Ethics, and Education: A Multidisciplinary Investigation of Responsible AI and Trustworthy ML Development

by
Lauren Stephanie Alvarez

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina
2025

APPROVED BY:

Tiffany Barnes	Edgar Lobaton
Thomason Price	Xiaorui Liu

Chair Veronica Cateté
Chair of Advisory Committee

BIOGRAPHY

Lauren Alvarez received her Ph.D. from the Department of Computer Science at North Carolina State University, where she has distinguished herself as a scholar, educator, and leader. Originally from Los Angeles, California, Lauren earned her Bachelor of Science in Applied Mathematics from Loyola Marymount University. As a Ronald E. McNair Scholar and participant in the ACCESS program, she was encouraged early in her academic journey to pursue graduate education and to use her research to address systemic inequities in STEM.

Lauren's doctoral research is an interdisciplinary investigation into the responsible AI development, integrating perspectives from computer security, education, and ethics. Her work applies sociological theories to the evaluation and design of trustworthy AI/ML, with an emphasis on interpretability, safety, and social impact. She has authored eight peer-reviewed publications, including in JAIR, IEEE Software, AAAI-EAAI, SIGCSE, ASEE, Koli Calling, and VL/HCC, and was awarded three prestigious graduate fellowships from IBM, SREB, and GEM. Her 2024 JAIR publication exemplifies her commitment to progressive AI ethics, presenting a primer for embedding foundational sociology theory and abolitionist theory into AI research.

Lauren's graduate career at NC State is equally defined by transformative leadership. She re-established the Underrepresented STEM Graduate Student Association (USGSA), and served as the graduate advisor for the Latine Computer Science Association (LA/CSC), the department's first Latine student group. She has also co-authored the CSC Department 2020 Anti-Racism Initiative Statement, collaborated with the Juntos Program to support Latine and first-generation high school students, and received multiple recognitions for her leadership, including the 2024 CSC Graduate Student Leadership Award. Her commitment to diversity, equity, and inclusion (DEI) is evidenced by her graduate minor in DEI in Education, and longstanding mentorship of over 25 students and teachers across K–12, undergraduate, and graduate levels, many from underrepresented backgrounds.

Inspired by the James Baldwin quote, "The paradox of education is precisely this — that as one begins to become conscious one begins to examine the society in which [they are] being educated... This is the only way societies change," Lauren is dedicated to leadership, mentorship, and civic responsibility. After graduating, Lauren plans to expand her research experience outside of academia and continue to advocate for the value of education.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support, guidance, and encouragement of many individuals and communities who have shaped both my academic and personal journey.

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Veronica Cateté, for her unwavering support and mentorship. Dr. Cateté inspires me personally and professionally by existing in spaces that were designed to systematically exclude people with her identities. Her encouragement to be ambitious and let others "tell you no" has profoundly shaped my thinking and the trajectory of my research. She was a guiding light in the confusing labyrinth of higher academia. She changed my life, and I will forever be grateful to her.

I am thankful to my committee members Dr. Tiffany Barnes, Dr. Thomason Price, Dr. Edgar Lobaton, and Dr. Xiaorui Liu for their feedback and time engaging with my work. Your expertise and feedback made me a stronger researcher.

I am grateful to the SREB Doctoral Scholars Program, Ronald E. McNair Scholars Program, and LMU's ACCESS program for opening the door to higher academia and research. Programs dedicated to advocating for underrepresented, under-resourced, marginalized, low-income, and first-generation students have supported me throughout my entire academic journey. These programs encouraged me to be a leader who catalyzes adversity and my experiences to create change in my life and those around me. They exposed me to knowledge I would have never known, and provided supplemental funding I did not have.

To department friends, Dr. Derek Martin, Dr. Zari McFadden, Dr. Alexis Goslen, and Ally Limke, thank you for the countless conversations, late nights, commiserating, and shared victories. You made graduate school easier to endure. To my fellow Ph.D. friends, Dr. Delecia Utley, Jessica Estrada, and Vivian Mota, it's been a privilege to see other successful women progress through their own graduate experiences. To my best friends, Simone Scott, Isabella Powers, and Eduardo Encina, thank you all for being who you are. The friendships I made during my time as a student made this journey worthwhile.

To Adam and Jessica, it's been a long 6 years, with many challenges I never saw coming. Your consistent support and presence in my life mean more to me than words can describe. To Uncle David and Joe, you have stepped in and shown up without being asked. Your support has been instrumental to my achievements. To my Dad, thank you for showing up in the ways you could. I never once doubted your love or how proud you were of me. To my Mom, you were my first teacher, and the foundation of who I am today. I carry your lessons with me always. A mis abuelos, gracias por su amor, apoyo, y por ser una parte muy important de mi vida. Me alegra mucho saber que ambos estén presente para verme recibir el título de Doctora Alvarez.

Everything I am and everything I've accomplished is because of my family. Thank you all for your love, patience, and belief in me. My family support and deep friendships have sustained me in ways I can never fully express.

As a first-generation student, I could not have anticipated the mental and emotional fortitude academia would demand. More importantly, the qualities that made me successful were not taught to me in a classroom. My resourcefulness, resilience, and authenticity are characteristics that were forged through my unique lived experience. I'm deeply grateful to the younger version of myself who chose to walk down an unknown path. She did not know where she would go or what it would take to get there, but she knew she would figure it out.

My academic experience shaped me into a confident, critically aware researcher. I'm thankful to everyone who played a role in my growth. A Ph.D. may be a terminal degree, but it's only the beginning of my career. As I proudly close this chapter, I look forward with excitement to the next chapter of my life. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

1.1 Motivation

This dissertation is anchored in two real-world events. The first is the 2015 Dieselgate scandal (Ewing 2020; Fracarolli Nunes and Lee Park 2016), where the U.S. Environmental Protection Agency (EPA) sued Volkswagen. Volkswagen was found guilty of using illegal software to lie to EPA evaluators regarding their vehicle emission rates, raising a larger concern regarding preventing lying and hiding malicious behavior. More recently, in 2024, Ford was found guilty of similar emission deception and will pay $2 billion, the largest settlement under the Clean Air Act Electrek (2024). It is crucial to understand that large corporations have used advanced technology to mislead evaluators for credibility and profit. As a well-informed research community, we must consider: **As AI continues to grow, how can practitioners design and implement AI systems that are technically robust, safe, and responsible across diverse sociological contexts?**

Several landmark case studies Buolamwini and Gebru (2018); Angwin et al. (2016); Obermeyer et al. (2019); Keyes (2018) and books Benjamin (2020); Noble (2018); Eubanks (2018); O'Neil (2017) have demonstrated how historical, systemic discrimination has been perpetuated in machine learning (ML) models. The automation of systemic discrimination raises many questions about the source of bias and how fairness interventions can remedy the harmful

impacts. As AI and ML become more pervasive in high-risk domains, these models have the power to detrimentally affect society. Unfortunately, as Ehsan et al. (2022a) discusses, the impacts of algorithmic discrimination can not be undone, and there is an urgent need to stop current harms and prevent future ones. Automating high-risk decision processes indicates that ML applications are socio-technical problems. To address socio-technical problems, ML researchers and practitioners are required to think critically about the technical and sociological aspects of their systems from the early stages of development to the necessary evaluation after deployment. People socialized under Westernized systems of oppression such as patriarchy, white supremacy, colonization, and capitalism are direct products of these systems, and without critical engagement with ethics and ML, these systems will manifest in the automated systems designed by engineers McFadden and Alvarez (2024). With the lack of interdisciplinary connections in computer science education, there is a need for more AI ethics and sociology curricula at all levels of education. There is also a need for stricter ML regulations in the public and private sectors requiring practitioners to analyze their developments with a stricter socio-technical lens.

This dissertation investigates different avenues to improve ML practitioners' lack of knowledge regarding harmful ML outcomes, such as (1) **early education on artificial intelligence (AI) and ethics**, (2) **critical engagement with ML and ethics in project development**, and (3) **trustworthy ML content knowledge and pipeline implementation**. This dissertation deliberately employs methodological diversity to match the inherently interdisciplinary nature of responsible AI. Just as RAI requires both technical and social considerations, understanding its implementation demands multiple research lenses. The strength of this work lies precisely in its methodological triangulation, which reveals insights that would remain hidden through any single disciplinary approach.

1.2 Research Questions & Dissertation Approach

The specific research questions investigated in this dissertation and their corresponding chapters are listed in Table 5.1 below.

This dissertation addresses critical gaps in Trustworthy Machine Learning (TML) and Responsible AI (RAI) research and development, offering both short-term and long-term solutions to ongoing challenges. Following the comprehensive literature review in Chapter 2, the research unfolds across five substantive chapters: Chapters 3 and 4 present complementary studies examining how to create trustworthy AI systems for socio-technical environments, particularly for facing potential malicious exploitation (RQ1). Chapter 5 investigates how practitioners can meaningfully engage with abolitionist framing in AI development (RQ2).

Chapter 6 presents the results of a ethics-integrated AI/ML module camp implementation and the impact on girls (RQ3). Chapter 7 focuses on educational aspects, exploring optimal pedagogical approaches for developing technical competence alongside ethical awareness (RQ4). Finally, Chapter 8 identifies areas where undergraduate students need additional support in bridging foundational computer science topics with advanced concepts like RAI and TML (RQ5 & 6).

Table 1.1: Structure of Dissertation Studies

Chapters	**Overarching Research Question Addressed**
Chapter 3 & 4	**RQ1:** How to create trustworthy AI systems in socio-technical environments, particularly when facing potential malicious deception?
Chapter 5	**RQ2:** How practitioners can meaningfully engage with abolitionist framing in AI development?
Chapter 6	**RQ3:** How does an AI/ML curriculum with integrated social relevance and ethics considerations impact girls' confidence and self-efficacy in AI/ML?
Chapter 7	**RQ4:** How do we explicitly apply Design Based Research (DBR) to design a course that integrates technical AI and AI ethics skills development?
Chapter 8	**RQ5:** How did the course impact student confidence in each of its learning objectives? **RQ6:** After the course, how do students perceive the usefulness of the course and what suggestions do they have for course improvement?

1.3 RAI & TML Terminology Disclaimer

The research domains that work to mitigate harms and create trust in AI/ML are: fair, trustworthy, transparent, interpretable, explainable, accountable, responsible, honest, and safe AI/ML. These words are used interchangeably, but experts in each field detail that there are differences in the motivations behind each domain. For example, Rudin describes interpretable and explainable as different because interpreting a model's behavior is not the same as explaining the model's behavior to a human Rudin (2019a). Molnar describes interpretable ML as interpreting the model's behavior for humans to understand Molnar et al. (2021). For consistency, I will use the terms **Trustworthy ML and Responsible AI**, but I claim that all of these fields are encapsulated under the trustworthy ML research umbrella. To be able to ensure safe and responsible ML and improve reliability, researchers must be able to accurately interpret the behavior of the "black-box model" and estimate any socio-techincal risks and potential harms.

CHAPTER

2

LITERATURE REVIEW

2.1 Literature Review

Specific research is needed in trustworthy ML (TML) and responsible AI (RAI), but requires interdisciplinary scholars to have knowledge from sociology and machine learning. An important trend to note is in the last 3 years, previously popular fields such as bias mitigation and fairness have redirected or rebranded to safety and alignment with particular attention to large language models and generative AI. When looking at this shift in vocabulary it is relevant to remember that although there are new trending research topics, the core of new fields have built off of previous research areas.

Each of the following chapters has its own dedicated literature review orienting the research in the related work. This introductory literature review serves as a summary of foundational knowledge, highlights the gap where these different domains connect, and identifies how this work addresses the knowledge gap. I named each subsection with the representing overarching research domain RAI or TML with the specific sub-domain discussed in the section. I also included chronological history of foundational literature, tracking the development of traditional machine learning literature and its progression to the current state of the art. A summarizing timeline of research development provides a perspective on how fast research has moved, which has positive and negative repercussions. It also provides a clear path from

the foundational work in the broader AI and ML research space that has branched into the many subcategories such as trustworthy AI, fair AI, interpretable AI, responsible AI, transparent AI, interpretable/explainable AI, AI for social good, etc. After the chronological overview, I conclude with a general overview of the literature, the gap present in the related area, and how my dissertation contributions will add to the existing body of literature.

Table 2.1: Key Topics Vocabulary Table

Term	Definition
Trustworthy ML (Example 1)	There are four attributes of trustworthiness: basic performance, reliability, interaction, & purpose. TML investigates the intersection of 7 well-established research areas (pillars): fairness, privacy, adversarial robustness, data robustness, differential privacy, expainability & interpretablity
Responsible AI (Example 2)	Responsible AI is an "approach to developing and deploying artificial intelligence (AI) that prioritizes ethical and legal considerations to ensure safe, trustworthy, and beneficial use."
Safety (Example 3)	Safety is a new term "focusing on ensuring AI systems operate as intended without causing unintended harm or negative consequences"
Performative Ethics (Example 8)	Performative ethics are ethical consideration that only appear to address underlying bias and discrimination, but ultimately make no measurable impact other than a minimal ethical acknowledgment

2.2 TML 4 Attributes, & 7 Pillars

TML and RAI are overarching research domains investigating the interactions of several related and, fortunately, well-established research areas. A trustworthy machine learning system is one that has sufficient (1) basic performance, (2) reliability, (4) human interaction, and (4) aligned purpose also known as the four attributes of TML. TML connects 7 well-established research areas (known as the TML pillars). Nagireddy et al. (2023) named the pillars [1] of trustworthy ML

[1] Nagireddy et al. (2023) comments on how research within each independent domain/pillar is well established, but more work is needed to understand the impacts of how the pillars connect. The cited sources for each pillar point to recent survey papers or related work to detail the current state of the art.

(TML) Varshney (2022); Ali et al. (2023) as privacy Rigaki and Garcia (2023), fairness Weinberg (2022); Wachter et al. (2021); Hanna et al. (2020), robustness to distribution shift Cooper et al. (2022), robustness to adversarial attacks Kurakin et al. (2018), interpretability Hooker et al. (2019); Kim et al. (2016), explainability Ali et al. (2023); Marques-Silva and Ignatiev (2022); Liao and Varshney (2021), and uncertainty quantification (also referred to as transparency) Bhatt et al. (2021); Ghosh et al. (2022). The 7 pillars map to only 2 of the 4 attributes:

1. basic performance
2. reliability: robustness to distribution shift, robustness to adversarial attacks, fairness, privacy
3. interaction: interpretability, explainability, uncertainty quantification (transparency)
4. purpose

2.2.1 Interaction: Interpretability & Explainability

Intepretability and explainability are challenging problems to address because researchers must choose between different types of ML models, data selection, the limitation of intrinsically interpretable model, and the scope of explanations needed. Surveys such as Ribeiro et al. (2015) by Ribeiro et al. argue against model-specific tools explaining "model-agnostic explanation systems provide a generic framework for interpretability that allows for flexibility in the choice of models, representations, and the user expertise." The challenges described in their paper include (1) the trade-off between flexibility and interpretability for more complex models, (2) requirements for exact explanation rather than estimations, and (3) difficulty incorporating actionable feedback, but Ribero et al. maintain that model-agnostic methods are better than model-specific.

The popularity of interpretable ML and evaluation methods has heavily increased over the last five years. Interpretable ML aims to create processes that allow practitioners with varying backgrounds to understand the mechanics behind their model's decision-making. The current, standard ways to test and evaluate the success and performance of these models vary on their scope: *pre-hoc, in-model, post-hoc, intrinsic, locally, globally, model-specific, and model agnostic* Molnar (2022); Carvalho et al. (2019). There are three levels to test the model either before building (pre-hoc), while building (in-model), and after building (post-hoc). Depending on the model, there may be an intrinsic interpretation by nature of its structure such as a decision tree or linear regression model. These are examples of model-specific interpretations and are typically used to distill knowledge from more complex, black-box models requiring global or

local, post-hoc evaluation. Common interpretable techniques analyze model components, model sensitivity, or surrogate models with linear regression models, decision trees, or decision rules because their structures are intrinsically interpretable Molnar (2022); Carvalho et al. (2019). *Model-agnostic* methods are all *pre-hoc* methods that allow flexibility because they explain multiple model types rather than one specific system. Another important criterion is the level of explanation whether *globally* to the entire model or *locally* to individual outputs.

A distinction between model-agnostic methods is global or local level explanations. Global methods explain the model's behavior on average, whereas local interpretation explains individual predictions. There are numerical (i.e., SHAPLundberg et al. (2018), Distill-and-Compare Tan et al. (2018), Gradient Feature Auditing (GFA) Adler et al. (2018), Global Sensitivity Analysis (GSA) Cortez and Embrechts (2011, 2013)), and rule-based explanations (i.e. Genetic Rule EXtraction (G-REX) Johansson et al. (2004a,b), Model Extraction Bastani et al. (2017), PALM Krishnan and Wu (2017), AnchorRibeiro et al. (2018)), but some methods will mix with other mediums such as graphical representations (i.e. LIME Ribeiro et al. (2016), Functional ANOVA decomposition Hooker (2004), ExplAIner Spinner et al. (2019), MMD-critic Kim et al. (2016)). Popular examples of local model agnostic methods include LIME, Partition Aware Local Model (PALM)Krishnan and Wu (2017), counterfactual explanations Wachter et al. (2017), and SHAP. Popular global model agnostic methods include global surrogate models, permutation feature importance, feature interaction, and partial dependence plots.

Interpretability: Knowlegdge Distillation, Surrogates, & Core-sets

Surrogate models are related to knowledge distillation and model extraction. They are used for both global and local explanations and are interpretable by "copying" the behavior of another model. The most popular surrogate interpretable model is LIME, and it is strictly local Carvalho et al. (2019). Surrogate models treat the model they want to copy as a "black-box" model and use the training data and model output to train the surrogate model Molnar (2022); Ribeiro et al. (2016); Krishnan and Wu (2017); Bastani et al. (2017); Puri et al. (2017); Ming et al. (2018); Frosst and Hinton (2017); Craven and Shavlik (1995). However, surrogates have multiple features, such as their targeted ML model, data sampling technique, or interpretable model Molnar et al. (2021). Other methods for model extracting include decision rules based on internal components like neural network weights Andrews et al. (1995); Augasta and Kathirvalavakumar (2012). Linear models and decision trees are popular methods for interpreting model decision-making processes and extracting models. PALM Krishnan and Wu (2017) uses a surrogate model with two components: a DT which partitions the training data, and a set of sub-models fitting the patterns to a partition Vilone and Longo (2020b).

Previous work by Baehrens et al. (2010); Robnik-Šikonja and Kononenko (2008); Dwork et al.

(2012) has noted the value of producing explanations for individual predictions and related work by Giannakoglou et al. (2006); Goel et al. (2007). Other related work in privacy and security has emphasized a need for more auditing and evaluation tools and the relevance of surrogate models for auditing cloud-based systems Park et al. (2022). Surrogate models have connections to knowledge distillation Hinton et al. (2015), which compresses the general (global) behavior from a teacher/parent model to a student/child model, like distilling neural networks into tree-based models. Surrogate models Molnar (2020); Forrester et al. (2008) are interpretable models that copy the behavior of highly computational, "black-box" models but are inherently interpretable by construction. Examples of implicitly interpretable models are linear regression, decision trees, and rule-based models.

Har-Peled et al. (2004) presented core-sets for k-means/median clustering. Agarwal et al. (2005) coined the term core-set defined as a data reduction technique used to extract general behavior. Chen (2009) extended Har-Peled et al.'s (2004) work establishing a new approximation algorithm for k-median/means clustering problems as a technique for core-set construction. There are several benefits to using core-sets such as optimization, efficient data usage, model selection, and feature reduction, particularly for knowledge distillation of deep learning models. Feldman (2020) provided an updated survey on core-sets, outlined a general algorithm for core-set construction, and identified a gap in the traditional ML literature where tree-based models have no identified core-sets. While core-set research focuses on data reduction/summarization, their applications to surrogate modeling have yet to be explored.

2.2.2 Reliability: Fairness & Justice

The current state-of-the-art (SOTA) data fairness intervention is a fairness-performance ensemble approach for improving ML fairness-performance trade-off (MAAT) Chen et al. (2022). MAAT uses the "We're All Equal" (WAE) view Friedler et al. (2021); Yeom and Tschantz (2021) (which is most similar to the concepts of Group Fairness and Statistical Parity), which states the privileged and marginalized groups should have the same probabilities of receiving the favorable outcome. All data representation balancing pre-processing techniques divide the dataset into four subgroups using explicit knowledge of the protected attribute's privileged and marginalized subgroups and the outcomes' favorable and unfavorable labels. They then either undersample or generate data to oversample the groups until the groups have a balanced amount of points. This style of fairness intervention is most popular because, by nature, pre-processing data manipulation is more generalizable and model-agnostic (i.e., it can be applied to many models rather than being model-specific).

The current stage of fairness intervention is most successful at the data manipulation stage,

but has yet to make successful headway in causation. Judea Pearl's Ladder of Causation has the 3 hierarchical rungs: (1) association, (2) intervention, and (3) counterfactual Pearl and Mackenzie (2018). Data manipulation is related to the first rung because associations can be inferred directly from the observed data using conditional probabilities and conditional expectations. To move to the next rung, one must use causal Bayesian networks and structural relationships, but there is limited research that has been successful. Fair ML literature critiques generalized, abstracted fairness interventions. Related literature emphasizes the need for understanding and appropriate context to best apply the most relevant fairness interventions, metrics, and more complex notions of Justice.

Jacobs & Wallach (2021) asserted that some notion of justice must be operationalized to measure fairness and proposed how to represent the fairness pillar with measurement modeling to promote transparency in fairness assumptions. Carrey et al. (2023) describe how procedural justice refers to the fairness of the decision-making process that leads to the outcome. As Carrey and Wu Carey and Wu (2022) describe, "procedural justice aims to arrive at a just outcome through an iterative process as well as through a close examination of the set of governing laws in place that guide the decision-maker to a specific decision." They pose that the overall goal of procedural fairness in machine learning should be re-aligned with the aim of procedural justice by instead analyzing the system surrounding the algorithm, as well as its use, rather than simply looking at it from the specifics of the algorithm itself Carey and Wu (2022). Distributive justice (i.e., disparate impact or disparate treatment Barocas and Selbst (2016a)) refers to the fairness of the outcomes of a process.

2.2.3 Reliability: Adversarial Robustness & Hidden Malicious Behaviors

Scholars have recently raised concerns about deceptive AI (Schneider et al. 2020; Brundage et al. 2018; Banovic et al. 2023; Slack et al. 2020), and discriminatory AI (Noble 2018; Benjamin 2023; Buolamwini and Gebru 2018; Keyes 2018; O'Neil 2017; Obermeyer et al. 2019; Angwin et al. 2022; Eubanks 2018). Researchers have presented local mitigation, deception, and detection solutions which Schneider et al. (2020) described as *"a new cat and mouse game between 'liars' and 'detectors'... in the context of AI"*, with limited scalable industry applications. Slack et al. (2020) was the first study to model a deceptive, discriminatory agent. Where bias and fairness concerns were typically considered unintentional consequences of hidden bias, deceptive AI raise a new problem regarding intentionally hidden malicious behaviors.

2.2.4 Literature Gap & Connected Chapter

The current state of the art explores how knowledge distillation can be applied to foundational models, but a gap in the literature regarding tree-based models still exists. There is limited work investigating the appropriate methods of knowledge distillation for tree-based modeling and in particular, there is limited connections analyzing the value of knowledge distillation for trustworthiness. The most related work is discussed in Chapter 3 and 4. Chapter 3 is the initial development of a surrogate modeling framework that improved performance and fairness while prevent malicious deception. Chapter 4 is a second, more rigorous study presenting the first trustworthy framework for interpretable knowledge distillation of tree-based models. The study serves as a start to examining how multiple trustworthy pillars impact each other, and an example of the complexity of trustworthy implementation. The findings from both experiments highlight the robustness of iSTEALTH in creating interpretable surrogates while maintaining comparable performance to non-interpretable methods. Importantly, the analysis indicates that heterogeneous knowledge distillation does not compromise the predictive power of the models, despite the inherent differences between the teacher and student models.

2.3 AI Ethics: Computer Science & AI Ethics Curriculum Integration

This dissertaion is not the first to discuss an integrated approach to ethics across the computer science (CS) curriculum Fiesler et al. (2020) Grosz et al. (2019). A motivation present in the CS ethics curriculum is to contextualize technology by connecting it to its societal impact. Krakowski et al. Krakowski et al. (2022) referred to this coupling as a "sociotechnical" curriculum and highlight the curriculum's success, which corroborates with previous case studies, including a survey reviewing 115 university tech ethics course syllabi Fiesler et al. (2020); Baumer et al. (2022). Fiesler et al. Fiesler et al. (2020) note the importance of interdisciplinary learning and the challenge to standardize teaching while considering the many variations of "tech ethics" integrations in present curricula.

Prior researchers have argued that ethics should be introduced as early as Programming 101 to prevent the "I'm just an engineer" mindset and underscore the need for more interdisciplinary collaboration with domains like philosophy to create a standard infrastructure for teaching and embedding ethics across CS. A European study Stavrakakis et al. (2021) on the importance of CS ethics curriculum coincides with previous research on 1) the widespread embrace for ethics curriculum integration; 2) the lack of hours dedicated to ethics teachings; and 3) the previous misconceptions of ethics referencing it as a standalone topic or area of

concentration rather than a foundational concept present in main domains such as artificial intelligence, data science, and security. In accordance with the present results, studies have detailed the current "ethics crisis" and present a call for computer scientists to make the same strides as climate scientists by utilizing collaborative teaching practices and exchanges of knowledge Fiesler et al. (2020); Raji et al. (2021).

The suggestions for future work involve using ethics as a pedagogical lens such as "responsible" versus "irresponsible computing" frameworks Prioleau et al. (2021) and assessing academics' levels of ethical awareness to improve curriculum integration by beginning with educating professors Hans et al. (2021). A challenge for future research is to expand the ethics theory at present and connect tech ethics pedagogies to feminist theory and critical inquiry. Williams et al. Williams et al. (2022) spotlight the dangers of minimizing ethics into "consequentialist, duty, or virtue ethics" and argue that students are not prepared for real-world scenarios involving structural societal systems of inequity. More research is needed to understand the best methods to teach future AI practitioners and engineers to integrate responsibility and AI ethics at all phases of the ML lifecycle.

2.3.1 Literature Gap & Connected Chapter

The current state of the literature emphasizes an ethics crisis but an open community of academics willing to learn how to successfully integrate tech ethics given the right curriculum. Chapter 6 explains the design process of an ethics-focused AI and Ml curriculum module designed for high school novices. The study results support previous research that female students engage well with socially relevant topics like contextually situated AI and Machine learning Fisher and Margolis (2002). Chapter 7 discusses foundational theory and design choices of an RAI 15-week curricula for undergraduates and Chapter 8 presents the evaluation of the course through projec-based case study analysis. The results of Chapter 8 indicate that novice students with limited technical background successfully produces RAI end to end integrations with considerations to responsibility, safety, and AI ethics at each ML lifecycle phase.

CHAPTER

3

DON'T LIE TO ME: AVOIDING MALICIOUS EXPLANATIONS WITH STEALTH

This chapter was accepted to IEEE Software, special issue on XAI4SE, February 2023.

Alvarez, L., & Menzies, T. (2023). Don't lie to me: Avoiding malicious explanations with stealth. *IEEE Software, 40*(3), 43-53.

3.1 Dissertation Framing

This chapter details the first iteration of iSTEALTH, STEALTH. STEALTH is a method for avoiding malicious deception. The results of this chapter found that malicious deception could be prevented using global surrogate models while persevering performance and *improving fairness*. Initially, the improvements in fairness performance were promising, but after further investigation, further discussed in the following chapter, the improvements in fairness were misleading. Chapter 3 and Chapter 4 serve as support for RQ1: How do we create trustworthy AI systems for socio-technical environments, including scenarios where a malicious actor is intentionally attempting to misuse the system? and demonstrate the complexities and careful considerations needed when presenting new methods integrating multiple pillars of trustworthy ML. Chapter 3 and Chapter 4 connect to the overall dissertation by presenting the first

framework to consider multiple pillars and investigate the effectiveness of interpretable and explainability methods for elements of reliability (i.e., adversarial robustness and fairness). The broader connections are further discusses in the following chapter.

Abstract

STEALTH is a method for using some AI-generated model, without suffering from malicious attacks (i.e. lying) or associated unfairness issues. After recursively bi-clustering the data, STEALTH system asks the AI model a limited number of queries about class labels. STEALTH asks so few queries (1 per data cluster) that malicious algorithms (a) cannot detect its operation, nor (b) know when to lie.

In order to support open science practices, all our scripts and data are on-line at https://github.com/laurensalvarez/STEALTH.

3.2 Introduction

Within the current AI industry, there are many available models– not all of which can be inspected. "Model stores" are cloud-based services that charge a fee to use models hidden away behind a firewall (e.g. AWS market-place Marketplace (2019) and the Wolfram neural net repository Repository (2022)). Adams et al. Xiu et al. (2021) discusses model stores (also known as "machine learning as a service" Ribeiro et al. (2015)), and warns that these models are often low quality (e.g. if it comes from a hastily constructed prototype from a GitHub repository, dropped into a container, and then sold as a cloud-based service). Theoretically, software testing can defend us against low-quality models, but in practice, these models do not come with verification results nor detailed specifications– so standard testing methods are unsure what to test.

An alternative to testing is explanation algorithms that offer a high-level summary of a model. Unfortunately, the more we learn about explanations, the better we get at making misleading explanations. As we shall see:

> The more a model is queried, the better it can mislead.

For example, Slack et al.'s lying algorithm Slack et al. (2020) (discussed below) knows how to detect "explanation-oriented" queries, and can then switch to models which, by design, disguise biases, or unfairness, against marginalized groups (e.g. by gender, race, age, or other identifiers).

Slack et al.'s results are particularly troubling. An alarming number of commercially deployed models have discriminatory properties Noble (2018); Chakraborty et al. (2021). For example, the (in)famous COMPAS model (described in Table 4.1) decides the likelihood of a criminal defendant reoffending. The model suffers from alarmingly high false positive rates for Black defendants than White defendants. Noble's book *Algorithms of Oppression* offers a long list of other models with discriminatory properties Noble (2018). Clearly, before we can trust models from the cloud, we need ways to assess their bias without being misled by malicious algorithms.

This article reasons as follows. If too many queries are the problem, then perhaps the solution is:

> Ask fewer queries.

To test this idea, we built a new algorithm called STEALTH which recursively bi-clusters N examples, down to leaves of size $\sqrt{N}$. A model is then queried with 1 example per leaf. These queries generate "labels"; i.e. decisions about each $\sqrt{N}$ example, and are used to build a *s*urrogate model which can be used for predictions and explanations. The key point here is that since STEALTH makes few queries, malicious algorithms can not detect its operation. Hence, they never know when to lie. Also, since we use our surrogate for explanations, those explanations cannot be misled by a malicious model.

However, this all assumes the surrogate (learned from very few samples) can mimic the important properties of the original model(learned from more data). To test that, our research questions ask what is *gained* and *lost* by reasoning over such a small $\sqrt{N}$ sample of the data:

RQ1: *Does our method prevent lying?* When querying Slack et al.'s lying algorithm, we found while the original model can lie, *lying fails in our surrogates*.

RQ2: *Does the surrogate model perform as well as the original model?* Much to our surprise, in the usual case, *STEALTH's surrogates performed as well as or better than the original (MODEL1)*, measured in terms of *both* predictive performance *and* fairness.

RQ3: *How does STEALTH compare against other bias mitigation algorithms?* Another reason to use STEALTH is its fairness properties are competitive to those generated via state-of-the-art bias mitigation methods such as Fair-Smote Chakraborty et al. (2021), MAAT Chen et al. (2022), and FairMASK Peng et al. (2022).

3.3 Background (How to Lie)

Figure 3.1.A shows the Local Interpretable Model-Agnostic Explanations (LIME) algorithmRibeiro et al. (2016) that samples thousands of instances uniformly at random near an example point, then builds a class of linear models from the generated sample data. Using its models, LIME can explain what features influence the classification of a particular instance.

LIME's explanations look like Figure 3.1.B. Note that, in the figure, the Age feature is most influential. By running LIME on all the test instances, it is possible to collect the *most influential* features; i.e. the attributes that are found as being most influential for the test set.

An interesting result seen in the most influential feature set is that often most features are *not* influential. For example, in our study, we have ten data sets with dozens of attributes, but only a handful are ranked most influential. This indicates the most influential features are a useful way to summarize the differences between explanations generated via different methods. Given the two most influential sets A, B a Jaccard coefficient is computed between (0,1) meaning (zero, total) similarities in the explanations, respectively:

$$jaccard(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{3.1}$$

To build a lying algorithm, we note that, in Figure 3.1.C, many of the **red** dots (showing LIME's generated samples) are at different locations to the raw data (**blue** dots). Slack et al.Slack et al. (2020) built a classifier to recognize when many of these **red** dots appear as input to the model. This meant they could distinguish (a) queries-about-explanation from (b) normal-queries seen in day-to-day operation. Hence they could hide any discriminatory properties by offering a *lying model* to queries-about-explanation and a second model to normal-queries.

To build their liar, Slack et al. trained an out-of-distribution (OOD) model using similar mutations to LIME and used this OOD detector to produce different classifications using the *protected attributes*. These are special attributes that divide data into two, *privileged* & *unprivileged*, groups (e.g. *male* & *female*; *young* & *old*). Thus, by training a model to recognize OOD data and in-distribution data, it is possible to hide discriminatory classifications against *unprivileged groups*, and mislead explanations.

The problems raised by Slack et al.Slack et al. (2020) are widely explored– see Table 3.1. In that table, STEALTH's use of limited domain data as samples is novel. Are such small samples of domain data relevant? We claim, without controversy, that for a model store to profit, *consumers must want to use it*, and know their own domain data they wish to label. For example, a doctor could have many records of patients and want to query a model to obtain a label of "high risk" or "low risk" of an imminent heart attack. More formally, given a data set with independent

attributes X and dependent classes Y, our approach assumes access to many x_i rather than other researchers who assume no access to x_i or y_i and instead use randomly generated x_i values. The starting point for this research was the realization that often model users do have x_i data, but not y_i labels. Our goal is to learn $Y = MODEL1(X)$ under the following "STEALTH assumptions":

- We have domain data for x_i, but not y_i
- We can obtain y_i values by querying a model *MODEL2*
- *MODEL2* may be malicious
- To avoid being misled, we must limit the number of y_i queries

Given the STEALTH assumptions, we first cluster our data on the x_i values and then make one y_i query per cluster. Using the few x_i and associated y_i labels, we build a *surrogate* model. The limited y_i queries would add very few **red** points to Figure 3.1, because the points are domain data samples and would be identical to the **blue** dots making our STEALTH approach inherently undetectable.

3.4 Methods

Table 3.2 describes the seven steps needed to run and evaluate STEALTH. To answer our research questions, we applied those steps, with and without Slack et al.'s lying algorithm Slack et al. (2020). The surrogate models built were assessed via their (a) LIME explanation properties (using Equation 3.1); (b) their predictive performance; and (b) their fairness using metrics that are widely applied in the literature (see Table 4.3). Since we are exploring fairness, we also compare STEALTH to state-of-the-art bias reduction algorithms Fair-SmoteChakraborty et al. (2021), MAATChen et al. (2022), and FairMASKPeng et al. (2022).

ALGORITHMS: For this work, we tried various learners such as random forests, logistic regression, and a SVM with a radial basis function. It was observed that random forests generated better recalls, and thus, we focus on random forests for this work.

Table 3.2 uses several bias reduction methods. Chen et al.'s MAAT systemChen et al. (2022) makes conclusions by balancing conclusions between two separate models (a performance model and a fairness. model). Chakraborty et al.'s Fair-SMOTE systemChakraborty et al. (2021) adjusts the training data by (a) removing training instances that change conclusions when protected attributes change; and (b) evening out the distributions of the protected attribute values. Chakraborty argues that these adjustments make it harder for any particular protected

Table 3.1: Related Work

If the problem is that, in Figure 3.1.C, the **red** dots are different to the **blue** then one solution is to make LIME's perturbations more realistic by adjusting sampling distributions [1], or better neighborhood calculations [2,3,4,5]. Other approaches improve LIME by making it more robust and less vulnerable to attacks [4,5], or creating better detection defenses [6,7] This is a very active area of research see [8-12] where prior work is rapidly assessed, improved, or refuted by subsequent work.

We have three comments on that related work. Firstly, some of that work makes restrictive assumptions about the data. For example, Ji et al. [1] rely on parametric assumptions such as the data and model output conforming to a normal distribution. While such assumptions might be true, they cannot be checked in the black-box case. STEALTH, on the other hand, is a non-parametric instance-based approach that makes no assumptions of normality.

Secondly, some works assume complexity when that may not be the best engineering decision for all domains. As shown by the results of this paper, methods that require 1000s of random samples (e.g. LIME) and which assume very high dimensional data (e.g. generative adversarial networks [4,5]) are not appropriate in all domains.

Thirdly, the above work does not take full advantage of non-generated data. Model store consumers must have a supply of their own domain data they wish to label. We can use that to great effect as discussed in the main text.

Related work citations:

1. D. Ji, P. Smyth, and M. Steyvers, "Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference," in NeurIPS'20.
2. D. Vre ˇ s and M. R. ˇ Sikonja, "Better sampling in explanation methods can prevent dieselgate-like deception," arXiv:2101.11702, 2021.
3. Y. Jia, J. Bailey, et al. "Improving the quality of explanations with local embedding perturbations," in KDD 2019, pp. 875–884.
4. S. Saito, E. Chua, N. Capel, and R. Hu, "Improving lime robustness with smarter locality sampling," arXiv preprint arXiv:2006.12302, 2020.
5. A. Saini and R. Prasad, "Select wisely and explain: Active learning and probabilistic local post-hoc explainability," AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 599–608.
6. Z. Carmichael and W. J. Scheirer, "Unfooling perturbation-based post hoc explainers," arXiv preprint arXiv:2205.14772, 2022.
7. J. Schneider, J. Handali, M. Vlachos, and C. Meske, "Deceptive ai explanations: Creation and detection," arXiv preprint arXiv:2001.07641, 2020
8. Slack, D., Hilgard, A., Lakkaraju, H., & Singh, S., "Counterfactual explanations can be manipulated," in NeurIPS '21
9. Wilking, R., Jakobs, M., & Morik, K., "Fooling Perturbation-Based Explainability Methods," In ECML/PKDD '22.
10. Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., ... & Bischl, B., "General pitfalls of model-agnostic interpretation methods for machine learning models,". xxAI - Beyond Explainable AI. 2020.
11. Maratea, A., & Ferone, A., "Pitfalls of local explainability in complex black-box models," In WILF. 2021.
12. Covert, I., et al. "Explaining by Removing: A Unified Framework for Model Explanation," J. Mach. Learn. Res. 22 (2021): 209-1.

Table 3.2: Inside STEALTH: The core of STEALTH is steps 2-5. Step 1,6, and 7 are added for evaluation.

0. Data Preparation The available data is divided 40:40:20 into a Train1:Train2:Test split. The test set is used as a set of hold-out examples in steps 1, 6, and 7. The Train1 set is used to build *MODEL1* that we are trying to understand (and which might be managed by a malicious model manager). STEALTH does not use the class attribute labels in Train2 since its mission is to extract these from the limited number of queries to *MODEL1* learned from Train1. Hence, for these experiments, all the class labels in Train2 are initially set to null.
1. Baseline generation: Build *MODEL1* (also known as the *original model*) using 40% training data. In terms of our domain modeling, MODEL1 is the model hiding behind a firewall. Using MODEL1 and the 20% hold-out test data, we can collect *baseline values* for all the metrics of Table 4.3.
2. Clustering: A *clustering algorithm* recursively bi-cluster Train2 data to leaves of size $\sqrt{N}$. We use Figure 3.2's recursive bi-clustering method. Nair et al. report that this method is fast and useful for finding a good spread of examples across dataChen et al. (2018).
3. Sampling: After that, we need a *leaf sampling algorithm* to select m examples per leaf. Here, we selected $m = 1$ items at random. While STEALTH works at $m = 1$, other algorithms such as MAATChen et al. (2022) seem to fail for such small data sets, and for some test cases, we cannot validate them with respect to prior work.
4. Labeling: A *labeling algorithm* must then assign labels to $\sqrt{N}$ selected examples. For that purpose, we query *MODEL1*.
5. Surrogate Creation: Some *model creation algorithm* must build a surrogate model. Here again, we used a random forest to build *MODEL2* (also known as the *surrogate model*) using the data associated with the $\sqrt{N}$ labels.
6. Collect Performance Scores: Here, the surrogate MODEL2 is applied to the 20% hold-out test data to generate the Table 4.3 performance scores. These values are compared to the *baseline values* seen in **step 1.**
7. Collect Explanation Scores: Here, we gather the most influential attributes of (b1) MODEL1 and (b2) MODEL2 using LIME, and compare them using Equation 3.1.

Table 3.3: Evaluation metrics: In this table, TP, TN, FP, FN are the true/false positive/negative rates seen in binary classification.

Performance Metric		**Fairness Metric**	
Accuracy: truly predicted positive and negative cases divided by all cases	$\frac{TP+TN}{TP+TN+FP+FN}$	**Equal Opportunity Difference (EOD):** Difference between TPRs for the marginalized and privileged groups.	$TPR_M - TPR_P$
False Positive Rate (False Alarm): falsely predicted positive cases divided by the actual negative cases	$\frac{FP}{FP+TN}$	**Average Odds Difference (AOD):** Average of difference between TPRs and FPRs for the marginalized and privileged groups.	$\frac{(FPR_M - FPR_P)+(TPR_M - TPR_P)}{2}$
True Positive Rate (Recall): truly predicted positive cases divided by the actual positive cases	$\frac{TP}{TP+FN}$	**Statistical Parity Difference (SPD):** Difference between the probabilities of the marginalized (PA = 0) and privileged (PA = 1) groups receiving the favorable outcome.	$P[Y=1\|PA=0] - P[Y=1\|PA=1]$
Positive Predictive Value (Precision): truly predicted positive cases divided by the predicted positive cases	$\frac{TP}{TP+FP}$	**Disparate Impact (DI):** Similar to SPD, but instead the ratios of the two probabilities are measured.	$\frac{P[\hat{Y}=1\|PA=0]}{P[\hat{Y}=1\|PA=1]}$
F1 Score: harmonic mean of the precision and recall values	$2 \times \frac{Precision \times Recall}{Precision + Recall}$		

Table 3.4: Description of processed data sets including their domain usage, protected attributes, numerical features post-processing, and non-empty rows. STEALTH number of rows and median run-times in seconds over 20 repeats.

Data set	Domain	Protected Attribute	Num Features	N rows	$\sqrt{N}$ rows	Our runtimes (secs)
Adult Census	U.S. census information from 1994 to predict personal income	Sex, Race	7	45,522	512	58.89
Bank Marketing	Marketing data of a Portuguese bank to predict term deposit	Age	7	30,488	512	54.54
Default Credit	Customer information in Taiwan to predict default payment	Sex	24	30,000	512	94.98
Compas	Criminal history of defendants to predict re-offending	Sex, Race	4	6,172	256	12.43
MEPS15	Surveys of household members and their medical providers	Race	41	4,870	128	2.00
Student	Student performance to predict good grades	Sex	24	1,044	64	1.19
German Credit	Personal information to predict good or bad credit	Sex	6	1,000	64	1.19
Diabetes	Diagnostic measurements to predict diabetes	Age	8	768	64	1.14
Heart Health	Patient information from Cleveland DB to predict heart disease	Age	13	297	58	1.11
Communities	Law enforcement information to predict violent crimes	racePctWhite	122	123	32	1.49

value to unduly affect the conclusion. Peng et al.'s FairMASK systemPeng et al. (2022) replaces protected attributes with new values learned from other independent attributes. Peng argues that this removed misleading latent correlations between attributes since the learned values tend to be average values (not outliers).

An important note to make about MAAT, Fair-SMOTE, and FairMASK is that all these meth-

ods require the analyst to pre-specify the protected attributes. STEALTH, on the other hand, operates without such knowledge. Despite this, as we will see, STEALTH achieves competitive levels of bias reduction to MAAT, Fair-SMOTE, and FairMASK.

DATA: The case studies in Table 4.1 were selected since they appeared in prior explanation and fairness papers including Slack et al. Slack et al. (2020). In column three of Table 4.1, each protected attribute is listed.

STATISTICS: Eight of our data sets have one protected attribute while ADULT and COMPAS have two. Hence, we run 12 times, one for each data set and protected attribute. Each run is repeated 20 times, each time dividing the available data into 40% Train1, 40% Train2, and 20% Test. Since each run returns nine scores (one for each of the metrics of Table 4.3), then generates 12*20*9=2160 scores.

For each data set and metric, we compared methods (using STEALTH or MAAT or FairSMOTE or FairMASK) against a **baseline** (running all data through random forests). Each method earns one more win, tie, or loss if it is statistically better, the same, or worse than the **baseline** (respectively). For that statistical test, we use a nonparametric test recommended by prior work on bias reduction; a Scott-Knot procedure using Cliffs Delta and a bootstrap test. For notes justifying and explaining those procedures, seeHess and Kromrey (2004); Mittas and Angelis (2012); Chakraborty et al. (2021); Chen et al. (2022). Note that for recall, precision, accuracy, and F1 *larger* numbers are *better* while for all other metrics such as false alarms, *smaller* numbers are *better*.

3.5 Results

> ***RQ1:*** *Does our method prevent lying?*

Table 3.5 compares the explanations from Slack et al.'s lying algorithm and STEALTH's Slack surrogate models. In Table 3.5, we see that the Slack et al. entries are nearly all zero; i.e. explanations from STEALTH do not overlap with the lying algorithm's explanations. This means that Slack's algorithm cannot detect the operation of STEALTH, and so it never knows when to lie.

Generating different explanations, from the liar, could be achieved by selecting lucky random values. Happily, we can show that the explanations from STEALTH correspond to important influences within our models. Table 3.5 also shows us that explanations from STEALTH have a large overlap (median=59.5%) with the explanations from the original model, *MODEL1*.

In summary, to answer RQ1, we say STEALTH can defeat Slack et al.'s lying algorithm. This is not to say that STEALTH defeats all possible liars, and exploring our method in the context

of other liars is a clear direction for future work.

For our methods, RQ1 is a positive result while many of the explanation overlaps reported in Table 3.5 are less than 100%. That is, between the original model and those made by STEALTH, there are differences measured in terms of which attributes are most influential. Happily, in conjunction with the results of RQ2, it suggests STEALTH's models are preferable to the original, and arguably we should have more confidence in the most influential attributes, as reported by STEALTH.

Table 3.5: RQ1 results: Jaccard coefficient (Equation 3.1) results comparing LIME's first top-ranked features of *MODEL1* (with the Baseline and the Slack et al. liar algorithm) to STEALTH's first top-ranked features.

Data set	Slack	Baseline
student	0.08	0.25
meps	1	0.42
adult_r	0	0.50
adult_s	0	0.50
compas_s	0	0.57
		0.595 $\Longleftarrow$ median
diabetes	0.20	0.62
compas_r	0	0.67
default	0.07	0.68
bank	0.17	0.69
german	0	0.79

> RQ2: Does the surrogate model perform as well as the original model?

It turns out that our method for avoiding liars does not lead to sub-optimal models. Table 4.4 shows how often our method finds improvements over the **baseline** models. Note that there are many wins recorded in Table 4.4 such as achieving numerous non-small, statistically significant improvements in performance and fairness compared to the baseline.

There are some losses reported in Table 4.4 and to explain, we look at the raw values of the **baseline** models (shown in Figure 3.3). It is clear that in the original model, *MODEL1*, there are some weakly performing data sets. For example, in Figure 3.3, while ADULT has high accuracy, the baseline also suffers from very high false alarms (over 50%). This is relevant since, when we drill down into the losses of Table 4.4, we note that many of STEALTH's losses arise in ADULT. That is to say, these losses seem to be less about drawbacks with STEALTH and rather

Table 3.6: RQ2 results: A method earns one win or tie or loss if it is statistically **better**, same, or worse(respectively) than the Figure 3.3 baseline (as assessed via Scott-Knot statistical tests that combine Cliffs Delta and a bootstrap). Note that for recall, precision, accuracy, and F1 *larger* numbers are *better* while for all other metrics such as false alarms, *smaller* numbers are *better*.

	Performance: (accuracy, recall, precision, F1, false alarm)				Fairness: (AOD, EOD, SPD, DI)			
Method	Wins	Loses	Ties	Wins + Ties	Wins	Loses	Ties	Wins + Ties
STEALTH	28	7	25	53 / 60	21	6	21	42 / 48
MAAT	34	17	9	43 / 60	31	7	10	41 / 48
Fair-SMOTE	27	17	16	43 / 60	28	6	14	42 / 48
FairMask	21	23	16	37 / 60	12	25	11	23 / 48

drawbacks with random forests (for these data sets).

> RQ3: How does STEALTH compare against other bias mitigation algorithms?

Our summary of Table 4.4 is that STEALTH is competitive with the current state-of-art, and even defeats some of those algorithms (e.g. FairMask). A case might be made that Fair-SMOTE and MAAT perform a little better since they have a few more wins (that said, the wins+ties are the same as STEALTH).

Hence, we see no reason to dismiss STEALTH based on the fairness of the models it creates, and given its superior *performance* scores, makes us recommend this algorithm.

Another reason to recommend STEALTH is that other bias reduction methods have to be directed to protect specific attributes. STEALTH, on the other hand, offers bias improvements without having to be directed. This is to say STEALTH could protect important attributes even if they were inadvertently overlooked.

3.6 Discussion

By using a few examples, STEALTH runs "under the radar" and is undetected by Slack et al.'s lying algorithm. STEALTH makes very few ($\sqrt{N}$) in-distribution queries, which are undetectable by OOD attacks. Since STEALTH's queries are undetectable, the liar cannot lie and returns honest results. Why do models built from just $\sqrt{N}$ examples have similar or better performance, and bias reduction properties? Perhaps this is for the same reason that semi-supervised

reasoningChapelle et al. (2006) works so well. One of the relevant semi-supervised learner assumptions is the ”manifold assumption", i.e. higher dimensional data can be approximated by a smaller set of most important attributes, without much loss of signal Chapelle et al. (2006). As evidence of our data having a lower dimensional manifold, recall from the above that in all the data sets used here, there were very few most influential attributes. When such smaller manifolds exist, data mining needs to only explore a small sample of the data set (provided the samples are spread across the data set using methods like, for example, Figure 3.2).

As to STEALTH's success at reducing bias, we conjecture that STEALTH's success results from a zealous use of design principles from Peng et al.'s FairMASK Peng et al. (2022). Peng et al. argue that FairMASK reduces bias by replacing minority outlier opinions with values that are better supported by more data. FairMASK does this by synthesizing replacement values for protected attributes from the other independent attributes. STEALTH, on the other hand, does this for *all independent attributes* since, in effect, its $\sqrt{N}$ examples are reports of average case effects within small data clusters. By averaging the data in this way, we *smooth* out the features including any outliers that may skew results unfairly.

(**Technical aside**: other fairness mitigation algorithms e,g, Fair-SMOTE, mitigate for unfairness by “blanding-out” attributes– i.e. by giving their values equal frequency in the training data. Our pre-experimental concern was that STEALTH operations on all attributes would bland-out everything, thus leading to a naive and poorly performing classifier. Yet, we realized it was the opposite of blanding-out. STEALTH builds a training set by clustering down to $\sqrt{N}$ and uses a random point near the centroids of the leaf clusters to build a surrogate. In so doing, STEALTH removes outlier signals and emphasizes the central tendencies within each of these $\sqrt{N}$ samples. Hence, we achieve the *fairness* results of Table 4.4 by deleting spurious signals without having to explicitly annotate which attributes are protected.)

Based on the above, STEALTH can be called a fairness enhancing algorithm which we recommended over MAAT, Fair-SMOTE, and FairMASK. But also, we argue that there is another result here. An important feature of these results is they suggest a new view on the nature of explanation, discussion, and discrimination. The STEALTH results highlight a connection between “better, more honest explanations” and “reducing biased results” into something that we might call “trusted communication”. Are there other such “trusted communication” methods? Would they perform better than STEALTH? We do not know– but that would be a useful direction for future research.

3.7 Limitations & Future Work

There were three main limitations, which can be addressed in future work for this study. The first limitation concerns STEALTH's subset selection and if the subset generation unintentionally falls under *Fairness Gerrymandering* by overgeneralizing and excluding important x-values for the marginalized groups. Thus, the first phase of future work could be to audit STEALTH's training set to determine how the clusters are forming and which points are being selected.

To further ensure STEALTH's fairness contributions, the second phase would be to examine that STEALTH follows the concept of *Procedural Justice*. This requires that when protected attributes are removed, the results still hold, implying none of the protected attributes are being used as predictive attributes. Also, to follow the requirements of *Procedural Justice*, STEALTH's intervention for each data set must be properly contextualized, which it currently fails to do. The last phase could be to remove any data set bugs and potentially add more to improve STEALTH's generalizability.

Acknowledgements

This work was partially funded with support from IBM and the National Science Foundation Grant #1908762. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of IBM or the NSF.

3.8 Broader RAI & TML Lessons Learned

STEALTH was a preliminary study exploring the interactions of 3 TML Pillars, fairness, adversarial robustness, explainability. More work was needed to further understand the improvements in fairness. The results of this chapter suggested that more work was needed to interpret the results from STEALTH, and motivated the creation of interpretable STEALTH (iSTEALTH). The broader RAI and TML lessons of the combined results of STEALTH and iSTEALTH are discussed at the end of the next chapter.

Figure 3.1.A: LIME samples around the boundary to find the delta between **blue** and **red** classes. From Ribeiro et al. (2016).

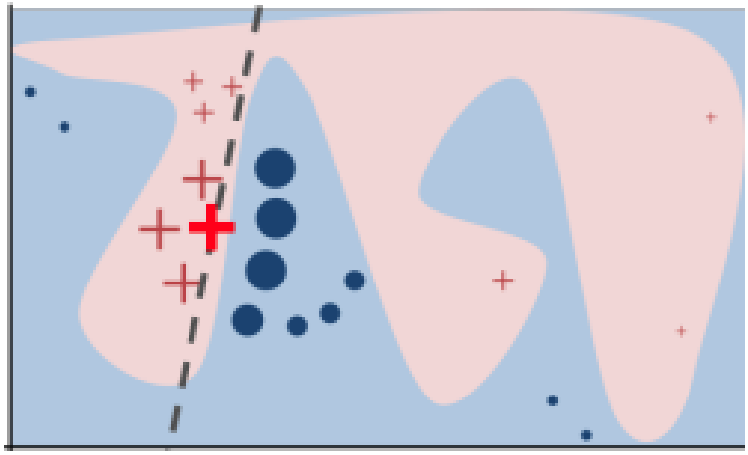


Figure 3.1.B: Feature importance as assessed by LIME. A **positive** weight means the feature encourages the classifier to predict the instance as a positive and vice versa for the **negative** weight. Larger weights indicate greater feature importance.

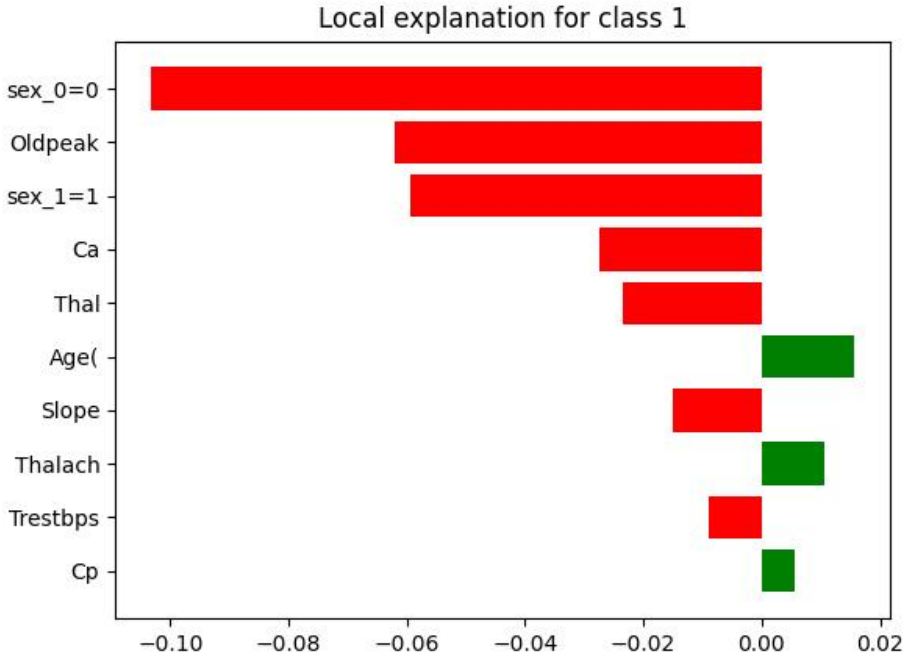


Figure 3.1.C: **blue**/**red** = original/generated data Slack et al. (2020).

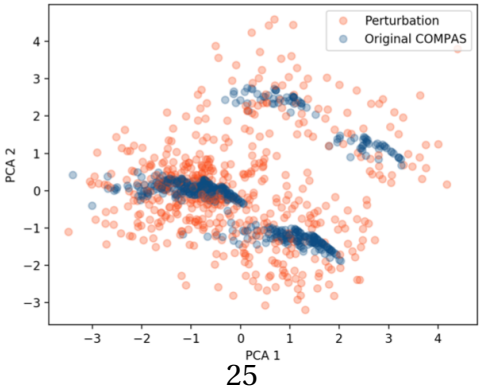
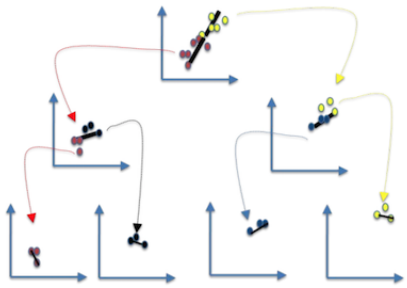


Figure 3.1: About LIME.

Figure 3.2: STEALTH partitions x values using the random projections. Nair et al. report that this method is fast and is useful for generating a good spread of examples across a data set [10].



NOTES: This recursive bi-clustering method finds two distant points E (east) and W (west). Using $c = dist(E, W)$ then all other examples have distances a, b to E, W, respectively and distance $x = (a^2 + c^2 - b^2)/(2c)$ on a line from E to W.By splitting data on median x, the examples can be then bi-clustered (and so on, recursively, see Figure 3.2b).

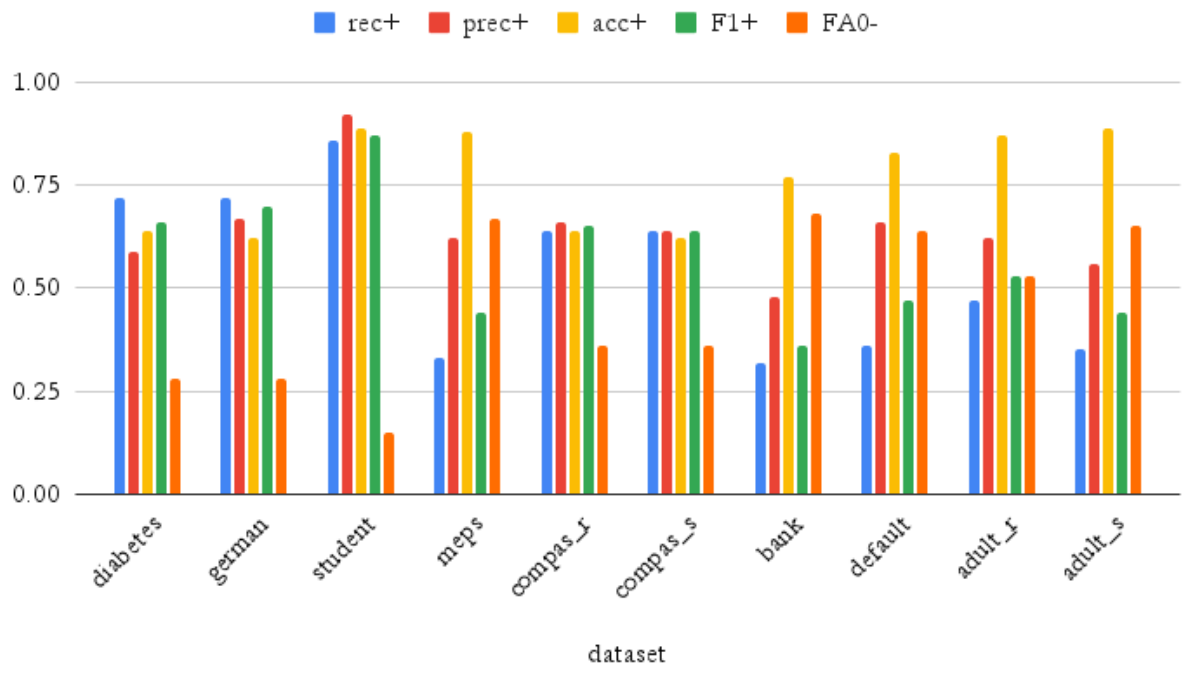


Figure 3.3.A: RQ2 results: *performance* reports median results for accuracy, recall, false alarm precision, F1.

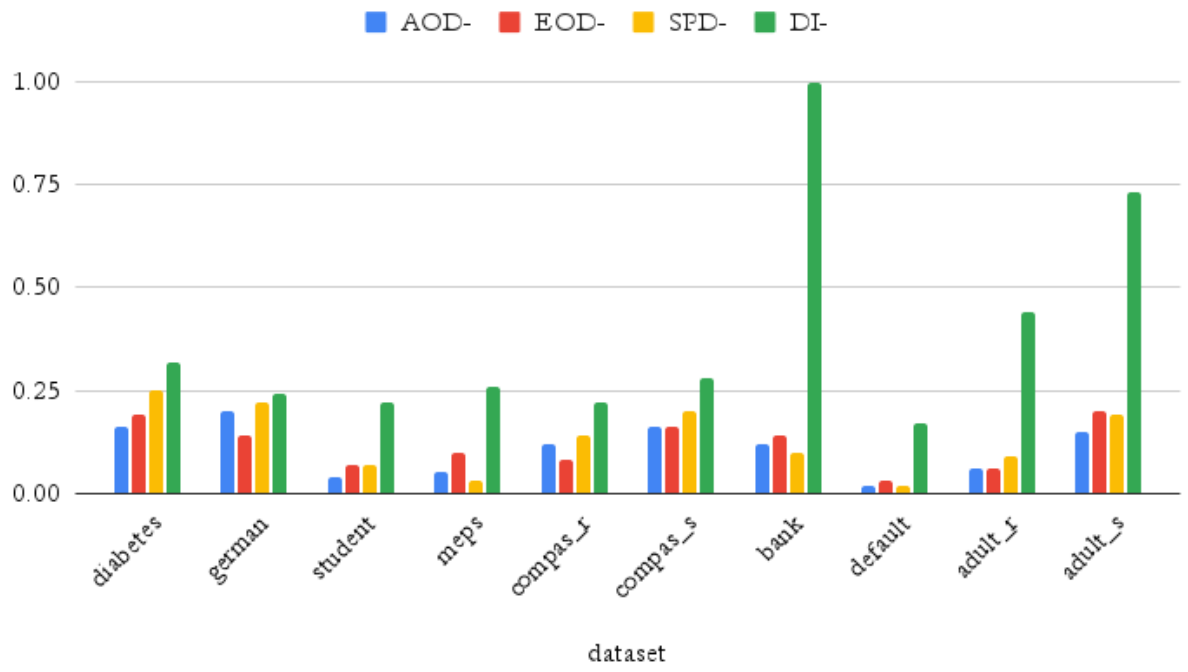


Figure 3.3.B: RQ2 results: *fairness* reports median results for EOD, AOD, SPD, DI

Figure 3.3: RQ2 Baseline, a non-malicious RF trained with all data, results on Table 4.3 metrics.

CHAPTER

4

iSTEALTH: INTERPRETABLE KNOWLEDGE DISTILLATION FOR TRUSTWORTHY EVALUATION

4.1 Dissertation Framing

Chapter 3 and Chapter 4 serve as support for RQ1: How do we create trustworthy AI systems for socio-technical environments, including scenarios where a malicious actor is intentionally attempting to misuse the system? and demonstrate the complexities and careful considerations needed when presenting new methods integrating multiple pillars of trustworthy ML. Chapter 3 and Chapter 4 connect to the overall dissertation by presenting the first framework to consider multiple pillars and investigate the effectiveness of interpretable and explainability methods for elements of reliability (i.e., adversarial robustness and fairness). STEALTH was a preliminary study exploring the interactions of 3 TML Pillars, fairness, adversarial robustness, and explainability. However, more work was needed to further understand how the system could promote improvements in fairness. Furthermore, this initial STEALTH research implied that more work was needed to interpret the results from STEALTH, and motivated the creation of interpretable STEALTH (iSTEALTH). Together, Chapter 3 and Chapter 4 are an extensive in-

vestigation into end-to-end TML development and the necessary technical knowledge needed to create an appropriate TML implementation.

Abstract

With the rise in popularity of computationally-heavy deep learning models, there is a critical push for robust techniques to assess system performance and safety, with extra attention toward upcoming national and global AI regulations. Notions of interpretability and explainability have become inextricable elements of upcoming AI regulations. However, limited work examines appropriate knowledge distillation techniques, or core-sets, for uninterpretable tree-based models. We investigated low-dimensional core-sets as query sets for model explanation. Core-sets are constructed with data summarization methods that show benefits for efficient data usage, optimization, model selection, and feature reduction, particularly for knowledge distillation of deep learning models. Our explanation technique, iSTEALTH, generates interpretable, global surrogate models with less than 13% of a dataset and presents promising results regarding data-efficient explanation queries. Our results indicate iSTEALTH is one of the first promising methods for constructing core-sets for tree-based models and bridges the literature gap regarding explainability and adversarial robustness.

4.2 Introduction

Artificial intelligence (AI) regulations and ethical concerns have become increasingly relevant, particularly after the exponential growth of generative AI, large language models, and their applications. There has been a global concern for understanding AI and many preliminary policies outlining the human rights for AI explanations Goodman and Flaxman (2017a) and transparency. Explainability is fundamental to current regulations, policy initiatives, and federal risk management frameworks Rueda et al. (2022a). In response to upcoming AI regulations, IBM released WatsonX as the first governance tool for enterprise use. Interpretable ML (IML), explainable AI (XAI), and robustness have become inextricable elements of transparent and trustworthy machine learning.

Core-sets are constructed with data summarization methods that show benefits for efficient data usage, optimization, model selection, and feature reduction, particularly for knowledge distillation. Little to no work examines appropriate knowledge distillation techniques or core-sets for tree-based models Feldman (2020). Feldman (2020) provided an updated survey on core-sets, outlined a general algorithm for core-set construction, and identified a gap in the

traditional machine learning literature where tree-based models have no identified core-sets. While core-set research focuses on data reduction/summarization, their applications to surrogate modeling are, still, yet to be explored.

For the first time, we investigate the applicability of core-sets for explanation, particularly with respect to adversarial robustness. We present, iSTEALTH, an interpretable explanation technique that uses core-sets and knowledge distillation methods to extract global behavior from adversarial models. iSTEALTH generates successful, statistically comparable global surrogate models with less than 13% of a dataset and presents promising results regarding data-efficient explanation queries. Our results indicate iSTEALTH is one of the first promising methods for constructing core-sets for tree-based models and bridges the literature gap regarding explainability and adversarial robustness.

This paper will begin by describing the necessary relevant work and real-world motivations. Section 2 will define our problem description and experiment setup. Section 2 subsections will provide more information on each step of the global surrogate generation process (see Figure 1 for visual representation). The following sections are in the standard order of results, discussion, conclusion, and future work.

4.3 Background

4.3.1 Surrogate Models, Knowledge Distillation, & Core-sets

Surrogate models have connections to knowledge distillation (KD) Hinton et al. (2015), which compresses the general (global) behavior from a teacher model to a student/surrogate model[1], like distilling a large neural network into a smaller neural network model. Surrogate models Molnar (2020); Forrester et al. (2008) are typically interpretable models that copy the behavior of highly computational, "black-box" models but are inherently interpretable by construction. Examples of implicitly interpretable models are linear regression, decision trees, and rule-based models.

Related work by Giannakoglou et al. (2006); Goel et al. (2007) reports the benefits of surrogate models with promising applications to deep learning architectures and evolutionary algorithms, and Park et al. (2022) emphasized a need for surrogate models to audit cloud-based systems. Har-Peled and Mazumdar (2004) presented core-sets for k-means/median clustering, and Agarwal et al. (2005) coined the term core-set defined as a data reduction technique used to extract general behavior. There are several benefits to using core-sets such as optimization, efficient data usage, model selection, and feature reduction, particularly for knowledge

[1] Other works refer to the models as parent and child, respectively.

distillation of computationally heavy models.

Recent studies Alvarez and Menzies (2023); Chai et al. (2022) suggest KD techniques can improve fairness performance while maintaining accuracy. Chai et al. (2022) focused on graph-based model KD. They distilled a larger neural network behavior (the teacher model) into a small neural network (the student model) with label smoothing and reweighting. Alvarez and Menzies (2023) presented a method to avoid adversarial deception that combined tree-based distillation and a recursive bi-clustering sampling technique. Both publications report KD improved fairness performance while maintaining classification performance in less complex models with smaller training sets. Both suggest more work is needed to understand why distillation positively impacts fairness and how to improve robustness with respect to distributional shifts/other adversarial impacts.

4.3.2 Trustworthy ML: Explainability, Interpretability, Safety & Robustness

As AI harms become more prevalent, there's a push for transparent, responsible, and trustworthy development. Landmark books and case studies have shown examples of discriminatory AI (Noble 2018; Benjamin 2023; Buolamwini and Gebru 2018; Keyes 2018; O'Neil 2017; Obermeyer et al. 2019; Angwin et al. 2022; Eubanks 2018). Since the rise of responsible and safe AI, the relevant domains have been formalized into the 7 pillars [2] of trustworthy machine learning (TML) Varshney (2022); Ali et al. (2023):

1. **Privacy** Rigaki and Garcia (2023)
2. **Fairness** Weinberg (2022); Wachter et al. (2021); Hanna et al. (2020)
3. **Robustness to distribution shift** Cooper et al. (2022)
4. **Robustness to adversarial attacks** Kurakin et al. (2018)
5. **Interpretability** Hooker et al. (2019); Kim et al. (2016)
6. **Explainability** Ali et al. (2023); Marques-Silva and Ignatiev (2022); Liao and Varshney (2021)
7. **Uncertainty quantification** Bhatt et al. (2021); Ghosh et al. (2022)

Ultimately, investigating new XAI and IML techniques benefits the entire AI and TML research communities.

[2] Nagireddy et al. (2023) comments on how research within each independent domain/pillar is well established, but more work is needed to understand the impacts of how the pillars connect. The cited sources for each pillar point to recent survey papers or related work to detail the current state of the art.

4.3.3 iSTEALTH v.s. STEALTH

iSTEALTH is an extended, rigorous examination of STEALTH's original problem. iSTEALTH core goal was to understand the impact of an interpretable architecture on surrogate performance, and further investigate STEALTH's fairness claims. iSTEALTH focused on the impacts of heterogeneous tree-based model architectural insights to further understand how trustworthiness can be technically implemented. There's a gap in the literature investigating the best methods for tree-based models, and related knowledge distillation research typically focuses on homogeneous distillation (i.e., NN to smaller NN or LLM to smaller LM).

iSTEALTH analysis indicates that heterogeneous knowledge distillation does not compromise the predictive power of the surrogate models, despite the inherent differences between the teacher and student models. Specifically, STEALTH distilled random forests into random forests (homogenous distillation), and iSTEALTH distilled random forests into a decision tree. Related to Trustworthy Knowledge Distillation, iSTEALTH highlights the challenge of investigating multiple TML pillars/attributes, but more work is needed to provide value in the Deep Learning domain for publication merit as supported by Nagireddy et al. (2023) (unpublished).

STEALTH focuses on avoiding adversarial deception using global surrogate models. iSTEALTH further investigates avoiding adversarial deception using global interpretable surrogate models. iSTEALTH focuses on the advantages of interpretable models for knowledge distillation, and how interpretable models can be used for heterogeneous knowledge distillation surrogates, particularly for tree-based models. The emphasis on tree-based models is due to the nature of datasets – which research shows for medium to large tabular datasets, tree-based models outperform neural networks.

4.4 Problem Statement & Surrogate Generation Methodology

Building off recent studies, we introduce **iSTEALTH**. iSTEALTH's surrogate/student architecture setup evaluates heterogeneous model pairings between a random forest model and a decision tree model. We use the same adversarial model from Slack et al. (2020) as well as a low-dimensional core set to query the teacher model and distill its global behavior into a more interpretable student model. The goal is to extract the general behavior of the teacher model without being deceived. Our problem is anchored on 3 key assumptions: (1) an auditor wants to explain the behavior of the teacher model; (2) the teacher model's training set is unknown to the auditor; and (3) the teacher model may have hidden malicious behavior meant to deceive evaluators.

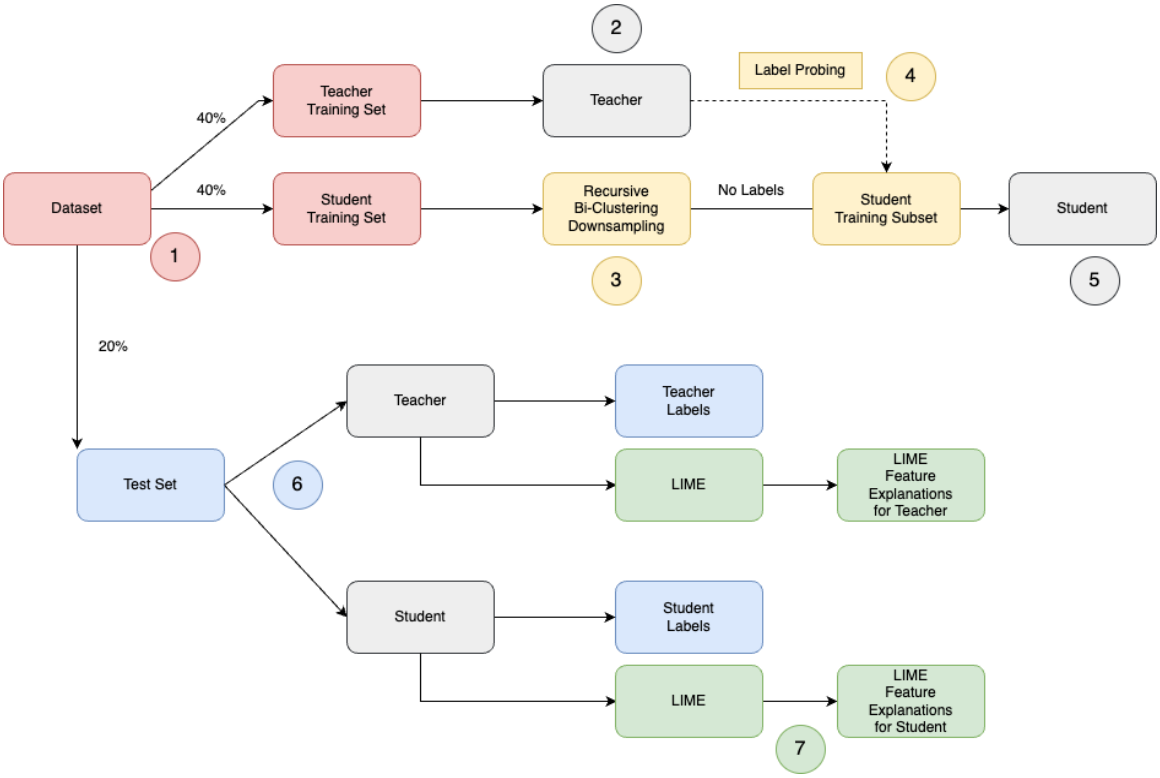


Figure 4.1: iSTEALTH: The full pipeline flowchart for explanation generation and evaluation.

4.4.1 iSTEALTH Overview

iSTEALTH has 7 main parts, as shown in Figure 2. We provide a concise overview of each step.

1. **Data Preparation:** The available data is divided 40/40/20 into a teacher/student/test split. The teacher set is used to build the TEACHER model.

2. **BASELINE generation:** There are two baseline models. The first BASELINE is trained on 80% of the dataset and represents the standard off-the-shelf implementation for each dataset. The second baseline is the TEACHER model. The TEACHER is trained with 40% of the dataset. We computed the TEACHER's *baseline performance* using the test set and evaluated its performance with standard classification and fairness metrics.

3. **Core-set Construction:** To construct the low-dimensional core-set, we used FastMap, which finds clusters of size $\sqrt{N}$, where N is the total amount of data points in the STUDENT training set. We compared the median and random samples and found no difference in performance. Thus, one random example per leaf was used to build the core-set.

4. **TEACHER Behavior Query:** To construct the STUDENT training set, we used the core-set as a query set to extract the TEACHER global behavior. The TEACHER predicts every

query in the core-set, and its predictions become the re-assigned label for those values. The STUDENT model “copies" the global TEACHER behavior by learning the relationship between the queries and the TEACHER’s labels.

5. **STUDENT Generation:** The core-set with labels from the TEACHER is used to train an more interpretable decision tree to build a global STUDENT model.

6. **Performance Evaluation:** Using the 20% test set, the evaluation metric scores for both the STUDENT and TEACHER are generated. These values are computed 20 times with reproducible random seeds, and selected metrics between the TEACHER and STUDENT are statistically compared with Scott-Knott’s Statistical Test. The performance evaluation measures how similarly the STUDENT scores compared to the TEACHER.

7. **Local Explanation Generation:** For added evaluation, we uses LIME to report the feature importance ranking for both models. The top-ranked feature sets were compared using the Jaccard Index/Coefficient. The local evaluation explains the features of each model that influenced the specific outcomes in the test set.

Table 4.1: Data sets descriptions including their domain usage and protected attributes.

Data set	Domain	Protected Attribute
Adult Census	U.S. census information from 1994 to predict personal income	Sex, Race
Bank Marketing	Marketing data of a Portuguese bank to predict term deposit	Age
COMPAS	Criminal history of defendants to predict re-offending	Sex, Race
German Credit	Personal information to predict good or bad credit	Age, Sex
Communities & Crime	Law enforcement information to predict violent crimes	RacePctWhite

4.4.2 Model Selection

There is a lot of work investigating KD techniques for neural networks (NN). A supervised learning benchmark was recently established for medium-sized datasets Grinsztajn et al. (2022), stating that tree-based learning performs as well as neural networks. Highlighting the literature gap regarding KD, core-sets, and tree-based models exposed by Feldman (2020), we intentionally focus on tree-based models. Supported by previous work Grinsztajn et al. (2022); Mittas

and Angelis (2012), we selected random forests as our teacher model because random forests are non-interpretable, tree-based models that have been identified as most appropriate for our selected case studies. Tree-based models are commonly used in real-world applications where interpretability and decision transparency are essential, making them critical in addressing AI robustness challenges. By enhancing these models' explainability, we contribute to the broader goal of creating AI systems that align with human values, particularly in contexts where deep learning's black-box nature may not be appropriate or desirable

4.4.3 Case Studies & Pre-Processing

Grinsztajn et al. (2022) explicitly noted fairness considerations as an imbalanced class "side issue" and one of the "subproblems which would deserve their own analysis." While our future work will incorporate broader datasets, we deliberately chose fairness benchmarks for our case studies because addressing these subproblems is critical to a more comprehensive understanding of AI robustness and fairness. As our 5 case studies, we selected the 3 most popular tabular datasets (Table 4.1) for supervised classification and fairness Fabris et al. (2022) and added two more uncommon fairness datasets as an extension. We only evaluated the datasets with respect to one protected attribute. The datasets with two protected attributes are counted separately, meaning we have a total of 8 case studies. On average, publications use only 1-3 as case studies Fabris et al. (2022). In efforts to create sustainable research practices, this work focuses on the current state of the art.

The data pre-processing steps were the same as those suggested by the AIF360 fairness toolkit (Bellamy et al. 2018). For each data set, standard ML pre-processing methods were applied: (1) rows with missing values were ignored, (2) continuous features were discretized and converted to categorical, and (3) non-numerical features (i.e., protected attributes such as sex, race, and age, and other demographics) were converted to numerical and coded to binary outcomes where 0 represented the under-represented/marginalized [3] group, and 1 represents the over-represented/privileged group. The class outcomes were coded similarly, with 0 being negative and 1 being positive outcomes.

4.4.4 Core-set Method

FastMap is a well-established multi-dimensional scaling algorithm, Faloutsos and Lin (1995). Chen et al. (2018) reported FastMap as fast and useful for finding a good representative subset

[3] Improving outcomes for underrepresented groups is the most common problem in AI fairness. Typically, a marginalized group is the underrepresented or smaller population within the data set. However, we prefer the term marginalized because it can be extended to the COMPAS data set. The COMPAS data set's marginalized group is the over-represented population, and the privileged group is the underrepresented population.

of examples across data. Core-sets are typically used to optimize, reduce dimensionality, or feature reduction. However, iSTEALTH present one of the few works that explore core-sets as an explanation method. iSTEALTH uses FastMap to reduce the student training set into a core-set, and use the core-set to query the teacher model. With the teacher's predictions, iSTEALTH uses the core-set to distill the global behavior into the surrogate model. iSTEALTH novelty lies in applying recursive bi-clustering to create a core-set and using the core-set as queries to explain global behavior.

Table 4.2: Sizes of the baseline (trained on 80% of the entire data set), teacher, and surrogate/student model training sets.

Data set	Baseline	Teacher	Surrogate	Surrogate %
Adult Census	45,522	12,604	128	10%
Bank	30,488	12,195	128	10%
COMPAS	6,172	2468	64	3%
Communities	123	49	16	33%
German	1,000	400	32	8%

4.4.5 Evaluation Metrics

To comprehensively assess the classification and fairness performance of the surrogate models in comparison to the teacher model, we computed standard classification and fairness evaluation metrics. The classification metrics, including accuracy, precision, recall, and F1 score, provided a holistic view of the model's predictive capabilities. We also incorporated fairness metrics to evaluate the models' behavior across different subgroups, such as disparate impact and equal opportunity difference. Table 4.3 defines the classification and fairness metrics we used to evaluate our baseline, teacher, and surrogate models. The Baseline refers to using a standard model with minimal fine-tuning that is expected to offer a quick and efficient starting point to compare the teacher and surrogate model's performance. In contrast, the teacher model baseline is an example of a fully trained, original model serving as the benchmark for evaluating the effectiveness of distilled or surrogate models.

We employed the Scott-Knott test Scott and Knott (1974), a robust statistical method for comparing classification performance to validate the statistical comparability of the surrogate models with the teacher model. This test is particularly suited for datasets with imbalanced

Table 4.3: Evaluation metrics: In this table, TP, TN, FP, FN are the true/false positive/negative rates seen in binary classification.

Classification Metric		Fairness Metric	
Accuracy: truly predicted positive and negative cases divided by all cases	$\frac{TP+TN}{TP+TN+FP+FN}$	**Equal Opportunity Difference (EOD):** Difference between TPRs for the marginalized and privileged groups.	$TPR_M - TPR_P$
False Positive Rate (False Alarm): falsely predicted positive cases divided by the actual negative cases	$\frac{FP}{FP+TN}$	**Average Odds Difference (AOD):** Average of difference between TPRs and FPRs for the marginalized and privileged groups.	$\frac{(FPR_M - FPR_P)+(TPR_M - TPR_P)}{2}$
True Positive Rate (Recall): truly predicted positive cases divided by the actual positive cases	$\frac{TP}{TP+FN}$	**Statistical Parity Difference (SPD):** Difference between the probabilities of the marginalized (PA = 0) and privileged (PA = 1) groups receiving the favorable outcome.	$P[Y=1\|PA=0] - P[Y=1\|PA=1]$
Positive Predictive Value (Precision): truly predicted positive cases divided by the predicted positive cases	$\frac{TP}{TP+FP}$	**Disparate Impact (DI):** Similar to SPD, but instead the ratios of the two probabilities are measured.	$\frac{P[\hat{Y}=1\|PA=0]}{P[\hat{Y}=1\|PA=1]}$
F1 Score: harmonic mean of the precision and recall values	$2 \times \frac{Precision \times Recall}{Precision+Recall}$	**Predictive Parity Difference (PPD):** Difference between PPVs for the privileged and marginalized groups.	$PPV_P - PPV_M$

class distributions and allows for the identification of statistically distinct groups Jelihovschi et al. (2014). This statistical validation reinforces the success of the surrogate models as more interpretable substitutes for the parent model. These findings contribute to building confidence in the surrogate models' potential deployment in scenarios where an interpretable model comparable to the parent model is critical.

4.4.6 LIME Evaluations

To delve into the explainability of both the parent model and its surrogate, we employed Local Interpretable Model-agnostic Explanations (LIME) as a powerful tool for generating local explanations of their respective decision-making processes. Incorporating LIME as a local explanation tool enhances the local explainability of our methodology, providing a nuanced understanding of how both models arrive at specific decisions for individual instances. This local interpretability is crucial for building trust in machine learning models, especially when fine-grained insights into decision-making processes are essential for diverse stakeholders.

The Jaccard Similarity Coefficient Jaccard (1901); Jain and Dubes (1988); Cha (2007) evaluation of the parent and surrogate's influencing features, as reported by LIME, showed a median 60% similarity for the teacher and surrogate's top-ranked influencing features. This suggests that, at a local level, the surrogate similarly mirrors the decision boundaries and feature importance of the parent model. The LIME evaluations indicate that more work is needed to

understand how models with similar performance have differing influential factors. However, due to page limitations, further investigation into the LIME evaluations between the Baseline, parent, and surrogate is outside the scope of this paper and will be discussed in future work. More work is needed to understand what properties may be impacting the difference in influential features.

Table 4.4: iSTEALTH vs STEALTH:
A method earns one win or tie or loss if it is statistically **better**, same, or worse (respectively) than the **Baseline** model (trained on 80% of the dataset) as assessed via the Scott-Knott statistical test, which combines Cliff's Delta and bootstrapping.

	Classification: (accuracy, recall, precision, F1, false alarm)				Fairness: (AOD, EOD, SPD)			
Method	Wins	Ties	Loses	Wins + Ties	Wins	Ties	Loses	Wins + Ties
iSTEALTH	17	5	8	22 / 30 (73%)	6	7	5	13 / 18 (72%)
STEALTH	19	3	8	22 / 30 (73%)	11	1	6	12 / 18 (66%)

Table 4.5: iSTEALTH Extension: A method earns one win or tie or loss if it is statistically **better**, same, or worse (respectively) than the **Baseline** model (trained on 80% of the dataset) as assessed via the Scott-Knott statistical test, which combines Cliff's Delta and bootstrapping.

	Classification: (accuracy, recall, precision, F1, false alarm)				Fairness: (AOD, EOD, SPD, DI, PPD)			
Method	Wins	Ties	Loses	Wins + Ties	Wins	Ties	Loses	Wins + Ties
iSTEALTH	24	7	9	31 / 40 (77%)	7	12	13	19 / 32 (59%)

4.4.7 Methodology Summary

iSTEALTH is a more interpretable surrogate model generation technique that is robust under adversarial attacks. Our methodology differentiates in model selection, model pairing, and case studies. We experimented with interpretable surrogates and heterogeneous model pairings and evaluated our methodology with 3 fairness benchmarking datasets. Our novelty lies in applying recursive bi-clustering to create a core-set and using the core-set as queries to extract global, **surrogate models**. We explored the future work suggested by Alvarez and Menzies

(2023); Chai et al. (2022). Both studies stated that more work was needed to understand how fairness properties improved within the surrogate and knowledge-distilled models.

4.5 Results

We conducted two experiments to evaluate the effectiveness of our methodology. The first experiment compared iSTEALTH with STEALTH to assess whether the use of a decision tree surrogate model affected the performance of the surrogates. The second experiment evaluated the iSTEALTH methodology across 8 case studies and investigated the effectiveness of heterogeneous knowledge distillation for tree-based models.

iSTEALTH vs. STEALTH Previous work experimented with homogenous teacher and surrogate pairing between random forests and random forests or neural networks distilled into smaller neural networks. We investigated: **(RQ1) How does a decision tree surrogate model architecture impact knowledge distillation and surrogate performance?** Table 4.4 compares iSTEALTH to STEALTH on 6 case studies and evaluates the surrogate's performance with 5 classification and 3 fairness metrics. These metrics were computed 20 times with reproducible random seeds, and their performance was statistically compared with Scott-Knott's Statistical Test. Our findings indicate that iSTEALTH performed nearly identically to STEALTH, suggesting the use of **decision tree model architecture had minimal to no impact on surrogate generation and performance.**

iSTEALTH Evaluation We conducted our second experiment to further understand: **(RQ2) How effective are heterogeneous knowledge distillation surrogates for tree-based models?** We performed a more extensive evaluation of iSTEALTH using all case studies and evaluation metrics. We report the Scott-Knott results of each evaluation metric in Table 4.5. We found that we improved or maintained for 77% of the classification and 59% of the fairness evaluations. The findings from both experiments highlight the robustness of iSTEALTH in creating decision tree surrogates while maintaining comparable performance to non-interpretable methods. Importantly, our analysis indicates that heterogeneous knowledge distillation does not compromise the predictive power of the models, despite the inherent differences between the teacher and student models.

4.6 Discussion

Our results (Figure 4.2) demonstrate that, after constructing the core-set, the surrogate models exhibit statistically comparable performance to the teacher model across the classification

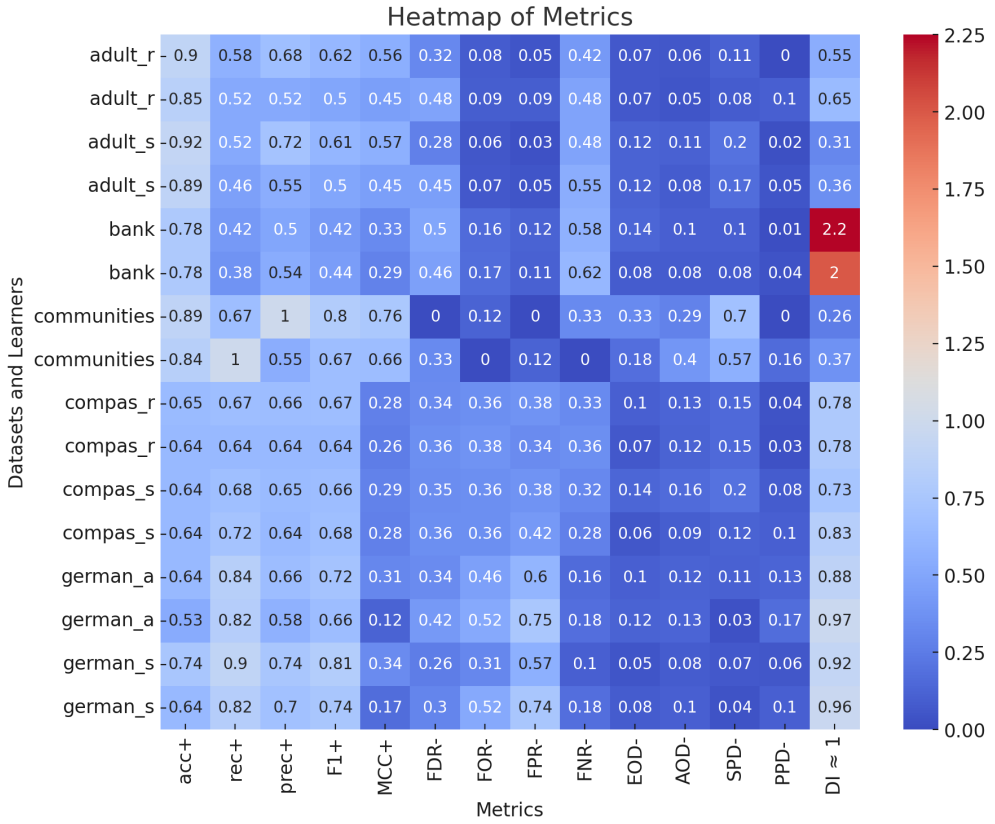


Figure 4.2: Heatmap of parent and surrogate's 9 classification and 5 fairness metrics across 8 datasets. The parent model is the model above and the surrogate is the model below with the same dataset name. For analyzing the better metrics,(+) indicates that closer to 1 is best, and (-) indicates that closer to 0 is best. For DI, the closest to 1 is best.

and fairness metrics. The Scott-Knott test confirmed that the observed differences were not statistically significant, affirming the surrogate models' reliability in replicating the parent model's predictive behavior.

FastMap reduces the teacher query set into a *representative* low-dimensional core-set, enabling the surrogate to extract the global behavior of the parent with 3-10% of the training set. However, the imbalanced subgroup representation remains in the surrogate training set data distributions, highlighting a key limitation of the core-set technique. This exposes the imbalanced and, therefore, unfair data distribution training the surrogate. Many fair AI scholars emphasize the statistical risks of imbalanced demographic representation and how this imbalance leads to biased model outcomes. With such extreme imbalance within the subsets, bias in the surrogates' outcomes was expected. We dispute the reported claims of improving fairness in Alvarez and Menzies (2023) and suggest more work like Wyllie et al. (2024) is needed to understand how fairness performance may be misleading.

4.7 Conclusion

Our exploration of constructing low-dimensional core-sets for surrogate generation provided valuable insights into the applications of core-set techniques for tree-based models. The results demonstrated that low-dimensional core-sets can effectively serve as query sets for generating explanations. These core-sets successfully trained global surrogates, reduced data utilization to a median of less than 13%, and prevented adversarial detection of explanation queries. This work contributes to the growing literature on knowledge distillation, core-set construction, and data-efficient explanation queries.

Despite the efficiency of our technique for constructing a highly representative, low- dimensional core-set, the application of iSTEALTH raises concerns regarding surrogate model fairness properties. It is imperative to acknowledge that surrogate models inherently mirror the behaviors of the parent model, including biased properties. Dressel and Farid (2018) supported this concern when they reported models trained on 2 features from the COMPAS dataset performed just as well as models trained on all the features. Further work is needed to address fairness concerns in core-set construction and explore the associated risks of introducing unfairness into distilled models.

To conclude, this study presents a low-dimensional core-sets method tailored for tree-based and adversarial models. By presenting a viable solution for creating more interpretable and efficient global surrogate models, we make a novel contribution to the ongoing efforts to open the black box models and foster trust in AI systems. Our findings contribute to the broader discourse on ML interpretability and robustness.

4.8 Future Work

Understanding the real-world implications of imbalances and biases is crucial for the responsible deployment of interpretable machine learning systems. Future work should prioritize refining the ethical considerations of surrogate models to align with evolving AI regulations and standards (i.e., applying appropriate data augmentation and fairness techniques). Further investigation into diverse datasets, core-set construction techniques, and model architectures pairings could enhance the generalizability and robustness of our approach. Additionally, extending this research to consider the interpretability of non-tree-based models, including Natural Language Processing and generative AI architectures, could broaden the scope of its contributions.

Our focus on interpretable machine learning (IML) and explainable AI (XAI) methods, particularly through tree-based models, addresses a critical gap in the existing literature. As

concerns surrounding ML transparency intensify, our research highlights the potential of coreset techniques for generating surrogate models and developing efficient data summarization strategies for knowledge distillation. Building on previous work by Chai et al. (2022), further efforts could understand distributional robustness, investigate the most impactful variables, and learn how to prevent symmetrical and asymmetrical covariate shifts.

4.9 Broader RAI & TML Lessons Learned

Chapter 3 and Chapter 4 provided crucial insights into the multifaceted nature of trustworthiness in machine learning pipelines. While the adversarial robustness impacts were promising, the unintentionally misleading fairness claims exposed a critical area for future work. The STEALTH and iSTEALTH experiments investigated how multiple attributes of trustworthiness impact TML pipeline implementation and performance. With current interventions, the myth of fairness-accuracy tradeoff has been disproven, but there is limited knowledge on how other extensions of trustworthiness (i.e., basic performance, reliability, interaction, and purpose) impact each other in certain implementation contexts.

Even with extensive TML knowledge, misleading results can be challenging to identify, and require socio-technical knowledge. A key finding was TML implementation is context-specific and requires domain-specific sociological knowledge. The iSTEALTH framework presented a viable solution for creating interpretable and efficient global surrogate models with attention to adversarial models. Still, more work is needed with complex model architectures, and to understand the real-world implications of data imbalances and biases to appropriately apply relevant mitigation techniques. Most critically, these experiments highlighted how philosophical, sociological, and educational blind spots among technologists impair their effectiveness in socio-technical spaces. Chapter 3 and Chapter 4 present a novel technical framework of TML development, and motivated a more in-depth examination of performative ethics and the impact of practitioner blind spots.

CHAPTER

5

PERFORMATIVE ETHICS: HOW CS PRACTITIONER BLIND SPOTS UPHOLD SYSTEMS OF OPPRESSION

This chapter is divided equally between Lauren Alvarez and Zari McFadden, and was published in March 2023 in the Journal of Artificial Intelligence Research[1].

5.1 Dissertation Framing

While, Chapter 5 may diverge from traditional empirical or technical methodologies, it is a rigorous, peer-reviewed contribution to a growing body of AI scholarship that recognizes the need for critical, interdisciplinary analysis. Specifically, the paper situates AI ethics within an intersectional framework, naming and analyzing how capitalism and classism, colonialism, racism and white supremacy, patriarchy, and ableism intersect with AI research, development, and deployment. It argues that technological systems are not neutral, and that ethical AI research must be grounded in a material understanding of these systemic forces. While this

[1]McFadden, Z., & Alvarez, L. (2024). Performative Ethics From Within the Ivory Tower: How CS Practitioners Uphold Systems of Oppression. *Journal of Artificial Intelligence Research*, 79, 777-799.

research approach may differ in style, methodology, and framing from traditional "hard science" research, it is a rigorous and peer-reviewed contribution to a rapidly growing subfield of artificial intelligence scholarship.

For readers unfamiliar with positionality research or reflexive methodologies, it is important to note that this work is not intended as a subjective personal essay, but rather as a structured ethical inquiry grounded in established scholarly frameworks from sociology, philosophy of science, and AI ethics. The peer-reviewed acceptance into JAIR underscores its scholarly merit, originality, and relevance to the AI research community. In the context of this dissertation, the inclusion of this work serves several essential purposes:

1. It foregrounds my positionality and the socio-technical lens through which the dissertation's technical contributions are developed and evaluated.

2. It establishes a normative foundation for understanding the ethical stakes of AI beyond compliance or fairness metrics, highlighting the structural entanglements between technology and systems of oppression.

3. It aligns with a broader shift in AI research toward reflexivity, accountability, and interdisciplinary rigor, demonstrating how ethical inquiry enhances the credibility and relevance of technical work.

This contribution should not be viewed as ancillary to the dissertation's technical components but as integral to a more holistic, responsible, and socially aware approach to AI research.

Abstract

This paper analyzes where Artificial Intelligence (AI) ethics research fails and breaks down the dangers of well-intentioned but ultimately performative ethics research. A large majority of AI ethics research is criticized for not providing a comprehensive analysis of how AI is interconnected with sociological systems of oppression and power. Our work contributes to the handful of research that presents intersectional, Western systems of oppression and power as a framework for examining AI ethics work and the complexities of building less harmful technology; directly connecting technology to named systems such as capitalism and classism, colonialism, racism and white supremacy, patriarchy, and ableism. We then explore current AI ethics rhetoric's effect on the AI ethics domain. We conclude by providing an applied example to contextualize intersectional systems of oppression and AI interventions in the US justice

system and present actionable steps for AI practitioners to participate in a less performative, critical analysis of AI.

5.2 Introduction

This article addresses the shortcomings of AI ethics research [2] while presenting a primer for AI practitioners to make critical connections between their work and society. Recent ethics studies have presented perspectives on naming individual systems of oppression Mohamed et al. (2020); Munn (2023), harmful theoretical abstractions Morley et al. (2021); Birhane et al. (2022), and the importance of critical discussion Waelen (2022); Schultz and Seele (2023); van Maanen (2022); Hampton (2021). We argue that critical self-reflexivity through a clear understanding of societal systems of oppression and ethics is necessary for more productive research on AI ethics, as supported by Birhane et al. (2022); Hampton (2021); Carey and Wu (2022); Weinberg (2022), but is often understated by abstracting the pervasive impact of systemic oppression from technology and its creators. We position our argument through an abolitionist framework. We highlight the differences between abolition and reform, where abolition requires creative reimagining of the world as it exists, while reform focuses on making improvements while keeping the world as it is.

Those socialized under Westernized systems of oppression such as patriarchy, white supremacy, colonization, and capitalism are direct products of these systems. A 2022 study Jakesch et al. (2022) details the discrepancy between AI practitioners and the general public about the importance of AI, with the general public being more concerned. Along with a higher value for AI ethics, there is also a difference between white and non-white AI ethics practitioners experiencing racial discrimination in the workplace. The 2020 People of Color in Tech Report surveying 1207 professionals in Big Tech Birch and Bronson (2022) reported 42.6% of people of color expressed issues due to race/ethnicity compared to 19% of white people Ton (2020). This statistic is supported by what Bonilla-Silva describes as 'white habitus' in which *"racialized, uninterrupted socialization process that conditions and creates Whites' racial taste, perceptions, feelings and emotions, and their views on racial matters"* Bonilla-Silva (2022). If the non-white

[2] We will use AI ethics and AI fairness as well as researchers and practitioners interchangeably in congruence with current rhetoric. Perhaps the domain language will become more representative of the nuances between ethics and fairness. Still, until then, we refer to AI ethics researchers to include all people involved in developing sociotechnical tools. Including but not limited to those within trustworthy AI, fair AI, interpretable AI, responsible AI, transparent AI, interpretable/explainable AI, AI for social good, and other ethics-related research domains. However, this is not a comprehensive list and is directed at all sociotechnical researchers. We would also like to note the difference between AI ethics and philosophical ethics. We are not theorizing more ethical concepts or interpretations of AI. We are arguing for critical engagement of the scholarly material already available and applying that material to the sociotechnical impacts of technology.

engineers who are helping build this technology are experiencing discrimination amongst their peers, what does that say about the discriminatory outcomes of the technology being built? We take a position of abolition, recognizing that those in positions of power desire to maintain that power even if they make small concessions to improve certain outcomes.

The rampant perpetuation of systemic injustice through *algorithmic oppression* by Computer Science (CS) practitioners despite having access to resources and the privilege to learn is a choice, even if unintentional. This choice has violent consequences, particularly for the most marginalized populations Binns and Kirkham (2021). With all this knowledge accessible to those in academia and Big Tech, there is a social responsibility to do better, think more critically, and create positive change Cooper et al. (2022); Waelen (2022); Munn (2023); Mittelstadt (2019). Computing and theoretical ethics allow researchers to abstract and "discover" discrimination in models. This theoretical abstraction removes accountability from the researchers and places it on "black-box" models. Unsurprisingly, simple AI models have observed obvious discriminatory relationships and are repeatedly presented in books Benjamin (2020); Noble (2018); Eubanks (2018); O'Neil (2017) and landmark cases Buolamwini and Gebru (2018); Angwin et al. (2016); Obermeyer et al. (2019); Keyes (2018).

This article begins with key definitions and explicitly connects the societal systems of oppression and power to AI fairness. **This is not an all-inclusive paper, but rather a primer to serve as a "crash-course" resource to frame future critical AI ethics research.** We then discuss the current AI ethics rhetoric and present a contextualized case study example. The following section expands on our abolitionist position and comments on current AI regulations' shortcomings. The final section concludes by discussing less performative, more critical AI ethics research, and our suggestions for future work.

5.3 AI Ethics' Position Within Systems of Oppression & Power

In this section, we explicitly define and denote ten systems of oppression and power motivating this work, what they look like in practice, and their implications on both a local and global scale. The motivating systems of oppression and power include, but are not limited to, *capitalism and classism; colonialism, colonization, and imperialism; racism and white supremacy; sexism and patriarchy; and ableism* defined in Table 1 and Table 2. Yes, this seems like a lot of -isms, but each of these -isms interact, shaping how each person experiences and interacts with the world. We acknowledge these are interlocking systems of power that have differing consequences for different intersections of identity Crenshaw (1989); Collins and Bilge (2020), but for clarity purposes, we save that for further discussion in future work. To understand the impact of intersectional systems of oppression, one must begin with understanding singular systems

first. Given the lack of papers within the AI ethics domain discussing systems of oppression at all, referring to each system individually is the current scope of this work. Understanding these systems sets the base for critically analyzing technology and AI ethics. The next step in understanding is recognizing that a truly equitable society would require abolishing these systems of oppression because as long as they exist, people will continue to be oppressed. The technology we build will just uphold and exacerbate that oppression. This section defines the aforementioned systems, why they are relevant, and how they connect to technological development more broadly. In addition, this section will parse and add nuance to common rhetoric related to AI ethics and building socio-technical tools.

Table 5.1: Systems of Oppression and Power Definitions.

Term	**Definition**
Capitalism	“an economic system based on the private ownership of the means of production. Capitalism is typically characterized by extreme distributions of wealth and large differences between the rich and the poor." Hill Collins (2000) “the capitalist mode of production, the seizure of the means of production by capital, which has become predominant in the sphere of production" Marx (1990)
Classism	“negative attitudes, beliefs, and behaviors directed toward those with less power, who are socially devalued” Lott (2012)
White Supremacy	“an ideology that presents the ideas and experiences of Whites as normal, normative, and ideal" Hill Collins (2000)
Racism	“a system of unequal power and privilege where humans are divided into groups or “races” with social rewards unevenly distributed to groups based on their racial classification. Variations of racism include institutionalized racism, scientific racism, and everyday racism. In the United States, racial segregation constitutes a fundamental principle of how racism is organized." Hill Collins (2000)
Patriarchy	“a political-social system that insists that men are inherently dominant, superior to everything and everyone deemed weak, especially females, and endowed with the right to dominate and rule over the weak and to maintain that dominance through various forms of psychological terrorism and violence" hooks (2005)

Table 5.2: Systems of Oppression and Power Definitions Cont.

Term	Definition
Sexism	when one is “discriminated against on the basis of sex" hooks (1984) “sexism should be understood primarily as the “justificatory” branch of a patriarchal order, which consists in ideology that has the overall function of rationalizing and justifying patriarchal social relations.” Manne (2017)
Ableism	“a set of beliefs or practices that devalue and discriminate against people with physical, intellectual, or psychiatric disabilities and often rests on the assumption that [people with disabilities] need to be ‘fixed’ in one form or the other" Smith (2023)
Colonialism	“a form of domination - the control by individuals or groups over the territory and/or behavior of other individuals or groups" Merriam-Webster (2023a)
Colonization	can be thought of as three different types Merriam-Webster (2023a): 1. “colonization in which the dominant relationship between the colonizers and the colonized is extermination of the latter" 2. “colonization in which assimilation is the relationship between the colonizers and the colonized" 3. “colonization in which settlers neither exterminate nor assimilate the indigenes"
Imperialism	“a form of intergroup domination wherein few, if any, permanent settlers from the imperial homeland migrate to the colony" Merriam-Webster (2023a)

5.3.1 Capitalism and Classism

We argue **capitalism** is the first and most important word to analyze and dissect because it is the most fundamental system shaping our lives, and the following systems, to be discussed, are utilized towards the goals of capitalism. Mandel summarizes Karl Marx’s definition of capitalism as *“the ruthless and irresistible impulse to growth which characterizes production*

for private profit and the predominant use of profit for capital accumulation" Marx (1990). Capitalism, as a system of power that directly or indirectly impacts the global society Lechner and Wallerstein (2015); Morley (1989), is more complicated than space allows us to explore here. However, it is crucial to understand the implications of how the United States' economic system, the locus of most big tech companies, is structured to value profit over everything else. When profit is valued over people, the primary concerns are those that increase profit; this extends into technology and debates about ethics. For example, Access Now, an international digital and human rights organization, resigned in protest from Partnership on AI (PAI), a consortium of Big Tech companies including Apple, Amazon, Facebook, Google, IBM, and Microsoft Johnson (2020). As Access Now stated in a letter about their resignation, they *"did not find that PAI influenced or changed the attitude of member companies or encouraged them to respond to or consult with civil society on a systematic basis"* Johnson (2020). Under capitalism, capitalists and corporations are primarily accountable to the stakeholders that rely on them for profit, with limited accountability to government regulation.

As Césaire notes, *"capitalist society, at its present stage, is incapable of establishing a concept of the rights of all men, just as it has proved incapable of establishing a system of individual ethics"* Césaire (1972). Any conversation concerned with the impact of AI and technology on society must also be concerned with the impacts of capitalism and the reality of economic profit over consumer impact. Capitalism is driven by profit and the accumulation of capital [3]. What is prioritized for capitalists is often not suitable for the working class (or proletariat). This idea is captured by **classism**. For example, producing items as cheaply and quickly as possible is in the best interests of the capitalist class because of the increase in profit, but it is not suitable for the laborers who are paid less than living wages and are forced to work in dangerous conditions Epatko (2018); Lee (2022); Perraudin (2019). Constant production and overconsumption also have a drastic environmental impact that continues as climate change progresses Panagiotopoulou and Chryssolouris (2022). Technology continues to exacerbate the environmental Nair (2023); Report (2023) and human consequences Perrigo (2023); Dzieza (2023) involved with training large language models, sourcing materials to build technology, and planned obsolescence that forces consumers to constantly buy new technology.

[3]"value (initially in the form of money) becoming an independent operator in the pores of a non-capitalist mode of production" Marx (1990)

"value constantly increased by surplus-value, which is produced by productive labour and appropriated by capitalists through the appropriation of the commodities produced by the workers in factories owned by capitalists"

5.3.2 Colonialism, Colonization, and Imperialism

Most AI ethics work and the societal impact of technology has been centered around the Western perspective; this is a result of **colonialism** and ignores the treachery that remains an invisible undercurrent in the building blocks of our technology. The United States' foreign policy and the history of Europe's direct and indirect **colonization** and **imperialism** are out of the scope of this paper. Still, their historical legacy is the basis of the technological reality we currently live in. There would be no computers or phones without the exploitation of the Global South where children as young as six in the Democratic Republic of the Congo (DRC) are risking their lives to mine cobalt Kara (2018), providing the very minerals that are necessary for computer production. Furthermore, the present day would not exist as it does if it were not for Leopold II establishing the DRC as a colony of Belgium in the 19th century, leading to the deaths of over 10 million people Hochschild (2022). Colonialism (and the requirement of profit by capitalism) is the reason why there exists a Global North and a Global South. As Morley notes, *"the U.S. state, as an imperial state shaped and controlled by outward-looking capital, assumes a multiplicity of tasks to facilitate the goals of its outwardly oriented capital class"* Morley (1989). The connections between capitalism, colonialism, and imperialism can be seen in a company like OpenAI outsourcing their labor to Kenyan workers for only $2 an hour to make ChatGPT "less toxic" Perrigo (2023). A company with plans to raise funds to reach a $29 billion valuation is exploiting the labor of Black people from an under-resourced country to fix its violent, racist, and sexist technology for $2 an hour without acknowledgment of the lasting psychological harms. That is one of the clearest, contemporary examples of how each of these interlocking systems of oppression collide in the tech space.

5.3.3 Racism and White Supremacy

Racism works hand-in-hand with white supremacy to uphold a social order that privileges whiteness over everything else. Although race is a social construct with no biological basis Smedley and Smedley (2005), it has tangible impacts on the lives of white and non-white people. For white people, whiteness is the hegemonic [4] norm, an invisible background force that allows for the maintenance of privilege to the exclusion of everyone else. Whiteness as a political construct was used to justify colonization and slavery Robinson et al. (1983). People like Charles Darwin and previous world leading statistician R.A. Fisher Pearl and Mackenzie (2018), used race as an underpinning of evolutionary theory to codify supremacy of white people over non-white people (phrenology): *"Although Darwin's book was criticized for its stance against the*

[4]"the social, cultural, ideological, or economic influence exerted by a dominant group" Merriam-Webster (2023b)

church, the British empire used it to justify colonialism by claiming that those subjected under its rule were scientifically inferior and unfit to rule themselves, with British anthropologists like James Hunt using Darwin's theory to justify slavery in papers such as The Negro's Place in Nature (1863)" Gebru (2019). The ruling political context (regardless of individual personal belief) of **white supremacy** is a socially dominant force fundamental to the prominent scientific discoveries, extending to AI and other technologies. Hanna et al. note *"despite the risk of reifying the socially constructed idea called race, race does exist in the world, as a way of mental sorting, as a discourse which is adopted, as a social thing which has both structural and ideological components. In other words, although race is social [sic] constructed, race still has power"* Hanna et al. (2020). There have been many examples of various technologies performing best on white people and worst on non-White, but primarily Black and Latine people. Some examples include facial recognition Raji et al. (2020), recidivism prediction Angwin et al. (2016), and healthcare Obermeyer et al. (2019); Grant (2022). Ignoring race in AI ethics research and development areas is a blatant upholding of racism and the hegemonic norm of white supremacy. Furthermore, performatively naming racism or intersectionality in AI ethics research does not automatically denote critical engagement. Racism, white supremacy, intersectionality, etc. should not just be treated as buzzwords, but instead as markers of structural oppression with tangible consequences.

5.3.4 Patriarchy and Sexism

Patriarchy is a structural system that is upheld by and harmful to men, women, children, and non-binary people alike. Violence is commonly defined as *"behavior involving physical force intended to hurt, damage, or kill someone or something"* Inc (2023), but reducing violence to just physical force does a disservice to the other forms that violence takes Bufacchi (2005); poverty, police states, microaggressions, verbal and emotional abuse, etc. are all examples of non-physical violence. Sexism is a consequences of patriarchy with bell hooks noting that *"under capitalism, patriarchy is structured so that sexism restricts women's behavior in some realms even as freedom from limitations is allowed in other spheres"* hooks (1984). A concern consistently at the forefront of the tech space is getting more women involved in CS Aguilar (2022); Oi (2022). What is not often included in these calls is the explicit naming of patriarchy and patriarchal violence as one of the mechanisms impacting the experiences women and girls have in male-dominated tech spaces, not including other dynamics of race, class, or ability.

Patriarchy as a framework is genderless on Gender-Based Violence (2018); hooks (2005), though it more directly benefits men and more directly harms women and non-binary people. It is also present in the type of technical work deemed worthy of respect. In a patriarchal

society, things coded as feminine or designated to women's space are regarded as less valuable regardless of the gender of the person participating. Terrence Real defines this in terms of his role as a therapist, but the same idea can be extrapolated to other areas coded as traditionally feminine: *"all therapists, under patriarchal mores, are coded as female, and as such they are subject to the same devaluing and intimidation as are traditional wives"* Real (2003). Direct examples of sexist AI include showing gender-biased ads Benjamin (2020), the fetishization of Black and Latine women in search engines Noble (2018), women not being considered in research Perez (2019); Harding (1991), the reinforcement of gendered stereotypes through virtual assistants Loideain and Adams (2020); Chin and Robison (2020); West et al. (2019), and misgendering/outing non-binary people Costanza-Chock (2018); Keyes (2018).

5.3.5 Ableism

Ableism is a system that operates with an air of invisibility, especially for those who are able-bodied, and often leads to the death of people with disabilities who are thought of as disposable to society. Williams et al. asked student participants about the ethics of various hypothetical medical situations and highlighted ableism as incredibly prevalent:

> *"Though we found most participants recognized race, gender, and class biases as threats to the ethics of the scenarios they evaluated, only one participant explicitly warned the eugenic potential of these systems, and no participants clearly identified disabled people as a social class vulnerable to unique mechanisms of discrimination. This gap in student understanding is further illustrated by the ways that they implicitly identify "society" as medical practitioners, institutions, families, and friends—those social agents whom [sic] traditionally have substantial power over disabled people's access to care, culture, and life."*Williams et al. (2022).

The scenarios positioned as hypotheticals in the study were actual events such as the use of predictive algorithms to determine the allocation of resources Pourhomayoun and Shakibi (2021); Shanbehzadeh et al. (2022). This technology is especially dangerous for people with disabilities because their disabilities are used to put them in higher risk categories that are often used as support for withholding care to the benefit of abled-bodied people. This was especially seen in the way the media addressed COVID-19 mortality rates: *"Public health officials, journalists, and politicians of pretty much every variety have said explicitly, or implied, that whatever current form of covid is under discussion can be regarded as at least a little less worrying because it mainly sickens and kills elderly, chronically ill, and disabled people"* Pulrang (2022). When such explicit language is used in the media in reference to the lives of people with disabilities, who can fully know the impact of inscrutable algorithms being used to determine

the allocation of care? Who is being sacrificed and relegated as less valuable? Other examples of ableism in AI are present in papers about participatory fixes excluding people with disabilities Sloane et al. (2020), the usage of sensitive medical data for people with disabilities Binns and Kirkham (2021), and other critique papers Shew (2020); Tilmes (2022).

5.4 A More Expansive Understanding of Bias, Fairness, and Ethics

The next crucial point to acknowledge is the impact of context and rhetorical nuances on research development. Words like bias, fairness, and ethics are nearly ubiquitous in discussions about the impact of various technologies, but the ways each of those words is used in the tech space is often different from their traditional meanings. Fairness and fairness interventions have been used as an operationalization tool for ethics. However, we argue that fairness is inaccurately operationalizing ethics. This section explores the aforementioned words, how their meanings change depending on the context, and what that means for us as engineers and technologists.

5.4.1 The Misuse of Bias and Fairness

Fairness and bias are common words used when considering technology's impact and repairing its negative lasting effects Ehsan et al. (2022b). Other researchers have commented on the various definitions of fairness, such as Hampton's reference to over 21 definitions presented in Tal et al. (2019). In both senses of the word, bias has to do with a lack of fairness, meaning that fairness and bias are essentially two sides of the same coin. The problem with these words' ubiquitous nature is their ambiguity and superficial relation to the real-world impact they abstractly represent. For there to be a lack of equality, there has to be a comparable original amount. What lens is the world being viewed through to determine what is actually fair, and to whom does this fairness apply? Have there ever been moments throughout human history of total fairness? If not, how do we know what to strive for? What does fairness mean in an unfair world? As Berhane et al. note, *"the treatment of concepts such as fairness and bias in abstract terms is frequently linked to the notion of neutrality and objectivity, the idea that there exists a 'purely objective' dataset or a 'neutral' representation of people, groups, and the social world independent of the observer/modeller"* Birhane et al. (2022). With white, heterosexual, male, able-bodied perspective being socially constructed as the hegemonic norm, the spectrum of how fairness is quantified should be heavily reviewed. We point to this paper for further discussion on specific fairness metrics Carey and Wu (2022).

As mentioned by multiple scholars Buyl and De Bie (2022); Tal et al. (2019), AI ethics work has no widely accepted definition of fairness. This could be defended by needing more context or understanding of the application for which fairness is being assessed. However, the lack of a widely accepted definition and contradiction of current methods creates too much nuance that can have unanticipated negative effects such as Fairness Gerrymandering, or the Simpson's Paradox Kearns et al. (2018). A lack of a clear fairness definition can lead to explicitly harmful consequences because the current state of success would be "fair enough", "fairer than previously" or as "fair as it could be", which is not a concrete or acceptable answer Rakova et al. (2021). AI and machine learning aim for accurate generalizations and, ideally, a concept of *group fairness* that considers *individual fairness* Berk et al. (2017); Mehrabi et al. (2021) which can, but not always, connect to distributive justice Gabriel (2022). However, with so much confusion around the definition of fairness, fairness interventions do not operationalize a specific AI ethics subfield.

5.4.2 Glorification & Abstraction of Ethics

Ethics is a popular term in the AI space that acts as a catch-all term for considering the morality and impact of technology. However, this perspective oversimplifies the philosophical field of ethics and glorifies the field as a designated source of truth about what is right rather than a theoretical abstraction. Ethical theorizing allows AI practitioners to abstract ethics into a speculative exercise rather than a practical problem with real, violent impacts Birhane et al. (2022). Our goal is to shift focus from abstraction to structural connections, and discuss the dangers of performative ethics.

Ethics as a philosophical field is composed of three different categories: normative ethics, applied ethics, and metaethics of Wisconsin Madison (2023). AI ethics tends to fall in the category of metaethics which *"investigates where our moral values, language, and principles come from and what they mean; it is concerned with 'what is morality?' rather than 'what is moral?'"* of Wisconsin Madison (2023). This distinction is important because addressing the ethics of technology without specifying the ethics framing collapses together different aspects of ethical work. Furthermore, as Siapka notes, *"Although much of the literature focuses on identifying a suitable ethical framework for the development and deployment of AI, there have not been comparable efforts for the identification or construction of frameworks within which to carry out a second-order reflection on AI ethics"* Siapka (2022). AI metaethics invites us to think about whether there is such a thing as fairness, who defines fairness, and whether fairness can be a realistic goal. When it comes to thinking about the ethics of various technologies, researchers must work to define the various frameworks that will be used to classify and analyze

the ethical impacts of technology. As Birhane et al. note,

> *"canonical Western approaches to ethics, from deontology to consequentialism, at their core strive for such universal and generalizable theories and principles 'uncontaminated' by a particular culture, history, or context. Underlying this aspiration is the assumption that theories and principles of ethics can actually be disentangled from contingencies and abstracted in some form devoid of context, time, and space"* Birhane et al. (2022).

When glorifying ethics as a field and looking to it as a guiding light for answers about morality, there has to be intentionality about critically understanding the underlying philosophical perspectives. Using ethics that are rooted in the same systems of power mentioned above is not enough to determine rightness *"given the ways that existing AI ethics literature builds on the circulation of existing philosophical inquiry into ethics, the reproduction of the exclusion of marginalized philosophies and systems of ethics in AI ethics is unsurprising, as is the maintenance of the white, Western ontologies upon which they are based"* Birhane et al. (2022). Suppose there is a desire to generate specific fields dedicated to AI ethics. Those fields must be intentionally created to broaden the scope of ethics (and the subfields that compose it) beyond the Western context. We point to these works for examples of consideration of non-Western perspectives Hongladarom and Bandasak (2023); Abebe et al. (2021).

5.5 Liberation Cannot Be Operationalized Under Systems of Oppression

Historically, computing has been able to abstract itself from its embedded social construction and only recently has been critiqued for the resulting harms. To deconstruct systems of oppression that are reified through algorithmic oppression, we argue that socio-technical researchers must have a common foundational base of sociology, ethics, and self-reflexivity. A prevalent limitation in AI ethics work is the avoidance, or inability, to explicitly state the critical structures that shape the world and examine the impacts of sociotechnical systems on society Ehsan et al. (2022b). Without systemic analysis, we claim work dedicated to positively improving the impact of technology on society will be performative at best and reify systems of oppression at worst. The following sections demonstrate the difference between reform and abolition, and the importance of contextualizing research with systems of oppression.

5.5.1 Abolitionist Framework

Abolitionist movements can be traced throughout history to the goal of abolishing slavery. In more recent times, calls for abolition have moved beyond abolishing chattel slavery to calling for the abolishing of police and prisons. Notably, work from Angela Davis Davis (2003), Ruth Wilson Gilmore Gilmore (2022), and others challenge us to question the necessity of cages and punitive responses to societal failure. Abolition requires creativity and imagination. As Davis notes regarding prison abolition, *"it has become so much a part of our lives that it requires a great feat of the imagination to envision life beyond the prison..."* Davis (2003). Abolition requires the conceptualization of a world outside of current systems of oppression.

While there is an increase in sociotechnical acknowledgments and inclusive vocabulary Birhane et al. (2022); Munn (2023); Mittelstadt (2019), much of the existing AI ethics research makes suggestions of reform rather than abolition. For example, a paper about decolonizing AI Mohamed et al. (2020) acknowledges the *"coloniality of power"* and argues for structural decolonization *"that seeks to undo colonial mechanisms of power, economics, language, culture, and thinking that shapes contemporary life,"* but does not make explicit calls for abolition. Instead, the authors suggest three modes of *"reciprocal tutelege"*, which are dialogue, documentation, and design. These suggestions can be seen as ideals of reform and the initial stages of Harro's Cycle of Liberation Harro (2000). Reform solutions fall in line with the ideals of techno-solutionism, where technology is seen as the most optimal solution for all problems as long as the right problem set is outlined and the technologies exist Morozov (2013); Gardner and Warren (2019). Although improvements can be made to the outcomes of AI systems through reform approaches, focusing on reform alone is not enough to address systems of oppression and can often lead to the further reification of power and oppression.

Reform ideals operate from a position that if we improve certain aspects of the system, then the problems of oppression and discrimination will be fixed. Rodriguez notes that *"to reform a system is to adjust isolated aspects of its operation in order to protect that system from total collapse, whether by internal or external forces"* Rodriguez (2020). In the AI ethics space, this looks like the implementation of datasheets for datasets, focuses on AI governance and policy, community-centered design, etc. While each of these contributions are important and help make the technology we build a little better, a little more fair, we argue to extend beyond reforms. The goal of abolition is a lofty one and it feels like too big, too unreasonable of a solution, but so did abolishing slavery. As Gilmore conceives, abolition is not about having all the answers of how,it is about magic: *"meaning we don't yet know how, which is what magic is, what we don't know how to explain yet"* Gilmore (2022). Abolition requires constant reimagining of the world as it exists and acknowledgment that we cannot band-aid solution or *"undo"* our way out of structural harm.

5.5.2 Abolition, Colorblind Racism, and the Role of Engineers

In his book *Racism Without Racists: Colorblind Racism and the Persistence of Racial Inequality in the United States*, Eduardo Bonilla-Silva explores the maintenance of racial hierarchy and inequality that continues despite the lessened experience of direct racism. He clarifies *"I have argued elsewhere that contemporary racial inequality is reproduced through 'New Racism' practices that are subtle, institutional, and apparently non-racial. In contrast to the Jim Crow era, where racial inequality was enforced through overt means (e.g., signs saying 'No Niggers Welcomed Here' or shotgun diplomacy at the voting booth), today racial practices operate in a 'now you see it, now you don't fashion"* Bonilla-Silva (2022). With technology, colorblind racism [5] Carr (1997) becomes even more obscure and expansive through machines built to be "black boxes" with often little explanation or accountability for the decisions made Pedreschi et al. (2019). The idea of "colorblindness" can be extended outside of race (i.e., ableism, classism) to include the often invisible privileges of power structures that systematically disenfranchise marginalized groups. Thus, CS practitioners are often operating from and participating in two layers of invisibility: (1) the invisibility of the ways they benefit from various systems of oppression and are invisibly passing those beneficial assumptions to the technology they build and (2) the invisibility of abstracting ethics and redirecting accountability to a "discriminatory machine" (i.e., the black box). As Hanna et al. note, *"treating race [or any marginalized aspect of identity] as an attribute, rather than a structural, institutional, and relational phenomenon, in turn, serves to minimize the structural aspects of algorithmic unfairness"* Hanna et al. (2020). Without a critical understanding of systems of oppression, high-risk AI is abstracted away from the real violence it can and does cause, and the responsibility falls on the creator who ignorantly [6] pursued a solution without gathering the necessary socio-technical background to fully understand the problem.

In their work on coliberative consciousness in CS pedagogy, Williams et al. show that despite having an understanding of social justice issues, many students still envision ethical issues with AI systems as a result of *"biased datasets and human mis/trust factors, rather than as problems of design and purpose"* supporting the idea of a kind of colorblindness present among engineers and technologists Williams et al. (2022). They explain, *"Even when ethics are directly addressed in CS curriculum, instruction tends to focus around professional, corporate, or legalistic frames of ethical behavior rather than the ethics of developing and deploying systems within their sociotechnical context"* Williams et al. (2022). There is a present belief that objectivity, merit, or

[5] "in terms of racial colorblindness, a person is also choosing to not just see race or skin color, but also the racial disparities, inequities, history of violence and current trauma perpetuated within a racist society" Library (2023)

[6] We would like to make further connections to *weaponized ignorance*, but for the benefit of conciseness see McKeever (2022); Froehlich (2017).

following suggested guidelines within these frames, are the criteria for 'fair' systems. However, as Gebru notes, *"an analysis of scientific thinking in the 19th century, and major technological advances such as automobiles, medical practices, and other disciplines shows how the lack of representation among those who have the power to build this technology has resulted in a power imbalance in the world and in technology whose intended or unintended negative consequences harm those who are not represented in its production"* Gebru (2019). Without an understanding of the sociotechnical context, which requires understanding societal systems of oppression, many of the AI issues will be limited to how a system is implemented rather than the society in which the implementation is taking place.

We must be clear there is a difference between decreasing a system's measurable disparate impact [7] and whether the system's implementation is working to dismantle or perpetuate the systems of oppression these automated systems operate within. With an abolitionist perspective, there is a better foundational basis to ask questions about the goals of a particular technology, whether that technology should exist, and how the technology interacts with the rest of the world. It also leaves space for realizing that no technology will be enough to fix systems that have been in place for hundreds, if not thousands of years, ridding all of us of the delusion that AI (or any other technological system) will be some kind of savior from all the problems in the world Birhane et al. (2022); Ahmed (2004). Addressing these systems takes more than a pithy checklist, an anti-racism course, or diversity training; it requires starting with a much more systemic analysis and understanding of the world and the historical depth of the complexities in which we live.

5.6 AI in the U.S. Judicial System

In this section, we examine a high-risk domain, discuss the ways in which current research suggestions are shortsighted, and provide an example of how we argue the domain should be contextualized in future work. First, we present the necessary background. A ProPublica article Angwin et al. (2016) published in 2016 found that the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) used to predict recidivism was racially biased/racist against Black people, predicting that Black people have a higher risk of committing crimes or reoffending than white people. This type of predictive technology has a tremendous capacity for harm because judges can use it to determine sentences, meaning that algorithms influence how long people stay in cages. A group at the Northpointe Inc. research department did a counterstudy to the ProPublica claims (i.e., asserting that the COMPAS risk scales were not

[7] "when the decision outcomes disproportionately benefit or hurt members of certain sensitive attribute value groups" Zafar et al. (2017)

racially biased). It illustrated that ProPublica made several statistical and technical errors to reach their conclusions Dieterich et al. (2016). Their argument supported by Rudin et al. Rudin et al. (2020) is that there is no racial bias because if they remove race as a feature, the results are the same; they do not deny the algorithm has disproportionately negative outcomes for Black people compared to white people (which is what the ProPublica study argues). The ProPublica counterargument is presented to showcase how complicated fairness can be and that statistical fairness metrics do not necessarily represent fairer outcomes.

The larger research discussion fails to critically analyze the dataset and the pervasive impact of automation in this field. The abstraction of mitigating recidivism is digestible for current practitioners, but contextualizing their model is the blind spot Hagendorff (2022) and results in reifying systemic, historical oppression through *algorithmic oppression*. Current perspectives on recidivism prediction include improving fairness through dataset manipulation or increasing interpretability, while other perspectives believe that recidivism prediction outcomes are not unfair because the model is statistically accurate based on the data. Both directions ignore the historical, systemic oppression that impacted the people represented in the dataset, and are incredibly ignorant. Focusing on the data abstraction and improving statistical models, removes the larger socio-historical context that models are being integrated into. The result is explicitly discriminatory models that impact real people, and the harms that have passed can never be undone. The algorithmic imprint Ehsan et al. (2022a) will be longstanding, and until a critical, ethical intervention is presented, these algorithms should be removed. There are very few AI researchers Hampton (2021) who advocate for the removal of harmful algorithms. From the abolitionist perspective, recidivism automation should be abolished because these systems perpetuate a long history of violence.

5.6.1 Contextualized Case Study

To further demonstrate our position for critical ethical engagement, we provide a contextualized formula that investigates (1) a rarely acknowledged system of oppression, (2) current research interventions, and (3) the unseen negative impacts of performative AI ethics research without critical analysis. Our applied example examines (1) mass incarceration [8], (2) discriminatory risk models, and (3) automated “justice." While these concepts may not seem to relate directly, we seek to spotlight systems of oppression and how these systems manifest in AI ethics research and related fields, and the dangers of inappropriate, abstracted solutions.

[8]We highlight capitalism, white supremacy, and colonization are interconnected to mass incarceration and criminalization, and without recognizing the impacts of these systems, the automated systems embedded in these contexts will be harmful. Although we stated that intersectionality was out of scope of this paper, there are certain systems that cannot be removed because systems of oppression are inherently connected.

We note that mass incarceration as a system of oppression is rarely stated in AI ethics publications, especially within recidivism prediction research. According to The Sentencing Project, the United States is the world's leader in mass incarceration Nazgol Ghandnoosh and Nellis (2022). There are currently *"1,566 state prisons, 102 federal prisons, 2,850 local jails, 1,510 juvenile correctional facilities, 186 immigration detention facilities, and 82 Indian country jails"* Wagner and Peter (2022), totaling 6,296 places to hold humans in cages in the U.S. compared to 4,360 higher education institutions Bouchrika (2022); DuVernay (2016), and many private prisons have capitalized on incarcerating Black and Brown men and boys. While the U.S. only makes up 4% of the world's population, it is home to 16% of the world's prisoners Vera (2023). Of the people incarcerated, Black, Latine, and Native American people are overrepresented (with Black people most severely overrepresented), while white people are underrepresented. Black people make up only 13% of the U.S. population, but 38% of the incarcerated population, Latine people make up 19% of the U.S. population, but 21% of the incarcerated population, and Native American people make up 0.9% of the U.S. population, but 2% of the incarcerated population, while white people make up 60% of the U.S. population, but only 38% of the incarcerated population. This is despite the fact that crime rates are consistent across communities, with most crimes occurring intra-racially at similar rates Caldera (2020) and drug usage being similar across racial backgrounds Hinton et al. (2018); Project (2018). The history of U.S. policing stemming from the "Slave Patrol" in the early 1700s Carolinas NAACP (2021) and the U.S. 13th amendment being a loophole to allow slavery as a punishment for crime is out of scope of this paper, but the issue of racism, white supremacy, and discrimination has been a fundamental aspect of the creation of policing in the United States.

With racialized policing practice in mind, how can we train predictive policing models in fair and unbiased ways when the data used for training is steeped in the history of racist policing? Even if the process by which AI models were being built were completely free of bias, that would not change the reality of policing, and the data it produces, being purposely and historically discriminatory against Black and Latine people. AI, no matter how statistically fair the results are, cannot change centuries of historical precedence and discrimination that continue to this day. Consider two of the interconnecting systems of oppression (mass incarceration and criminalization) impacting racial demographic representation in the industrial prison system, and let's connect these systems to current research suggestions and the violent impact of automating a system without the proper socio-technical background. We mentioned the landmark COMPAS recidivism model, which is one of many recidivism models active in the United States. In fact, 60% of the U.S. population lives in a jurisdiction that actively uses risk assessment tools Injustice (2023). By coupling AI to tangible systems of oppression in the United States, the only result is automated systemic injustice and algorithmic oppression.

The suggested interventions do not situate the models, nor do they consider the systems of oppression that influence recidivism prediction.

The abolitionist perspective as a starting position asserts that certain inalienable human rights belong to every human being across the globe and that human rights trump the rights of corporations. From there, work can begin to think about who has access to those rights and who does not and how we can build a society (and technology along with it) to grant everyone access to those rights. Then, solutions can be engineered from positions of justice, equality, and equity, recognizing that there is no technical, mathematical solution without dismantling systems of oppression.

5.7 Discussion

This paper presents a primer for AI practitioners to make critical connections between their work and systems of oppression. This work contributes to related literature regarding critical engagement and applying said material to the sociotechnical impacts of technology. We argue that critical self-reflexivity through a clear understanding of societal systems of oppression and ethics is necessary for more productive AI ethics research, but is often understated through theoretical abstraction. To deconstruct systems of oppression being reified through algorithmic oppression, socio-technical researchers must have a common foundational base of sociology, ethics, and self-reflexivity. As a case study, we choose to examine a high-risk domain, discuss the ways that current research suggestions are shortsighted, and provide an example of how we argue the domain should be contextualized. Our goal is to unmask performative ethics and shift focus from reform to systemic engagement.

A novel contribution of this work is interrogating the difference between reform and abolition in AI ethics research. Much of the current research provides solutions and suggestions that fit within the realm of reform, but we argue for an abolitionist framework. While the reform suggestions that existing research provides are valuable, it is necessary to be aware of how reform can act to deradicalize radical traditions to avoid shaking the boat of power and respectability too much. Crenshaw notes a history of racial liberalism being used to deradicalize racial reform Crenshaw (2017) and similar concepts can be extended to the ideas of “ethics washing" Schultz and Seele (2023); van Maanen (2022) by Big Tech corporations. We are pushing for AI ethics research to move in a direction beyond reform alone. As Gilmore states *“the abolition I speak of somehow, perhaps magically, resists division from class struggle and also refuses all the other kinds of power difference combinations that, when fatally coupled, spark new drives for abolition. Abolition is a totality and it is ontological. It is the context and content of struggle, the site where culture recouples with the political; but it is not struggles’ form”* Gilmore (2022). An

abolitionist framework asks for creativity and imagination and struggle. We urge the research community to struggle together to imagine a new world without oppressive systems and the technology that upholds them.

Our work intentionally does not offer concrete solutions. Part of the abolitionist framework is not knowing how to achieve the end goals *right now* and we are okay with this position for now. Our goal for this paper is to participate in the first two steps of cycle of Harro's Cycle of Liberation which is "waking up" and "getting ready" Harro (2000). We are waking up CS practitioners to the various systems of oppression and encouraging them to get ready to critically engage with the systems in their research. Like Gilmore, we are looking forward to the magic of imagining new worlds without these systems and figuring out AI/technology's role in that world. We encourage readers to dream and imagine even if there are no actionable solutions just yet. We encourage sitting in the discomfort of not knowing what is next. We encourage grappling with ideas of reform versus abolition and how to make progress without further entrenching the existing systems.

5.7.1 Future Work

This paper points back at those present in academia and Big Tech because when well-intentioned interventions aren't successful and are consistently reimagined with no significant impact, when does performative ethics research become harmful? "Well-intentioned" isn't enough. Academia is coined the "ivory tower" meaning *"it retains the notion of an authorized body of knowledge, relies primarily on classroom-based (or laboratory) learning, does not stress subjective experience as a legitimate source of knowledge, has a hierarchical structure that stresses individual achievement, and appears to maintain a fairly standardized and accessible set of rules governing classroom behavior and interaction"* Treichler and Kramarae (1983). When socio-technical practitioners acknowledge their participation in the systems of oppression and power and begin to interrogate themselves and their work as agents of privilege and power, there will be more critical engagement in AI ethics. Focusing on mathematical abstractions distracts from the socio-historical context needed to make a fairer world and removes accountability from researchers to place it on technology. A reasonable approach to tackle this issue could be to begin here. This is an introductory primer and critique of current AI ethics, and we present individual first steps and principles:

- Imagining solutions beyond reform alone that lead to organizing for systemic change
- Identifying and acknowledging systems of oppression

- Unpacking the way every person participates in these systems (unconsciously or consciously)
- Recognizing that human rights matter more than the rights and interests of organizations and corporations
- Further unpack the frame of systems of oppression by applying them to individual research as well as the larger research community
- At the beginning of every research endeavor, reflect on systemic and local impact to interrogate the work
- Examine (1) the impacting system of oppression, (2) current research interventions, and (3) the unseen negative impacts of performative AI ethics research without critical analysis

We implore researchers and tech practitioners to participate in abolitionist-framed AI ethics/fairness research, and to do so the very first step is to actively learn about systemic oppression and what that looks like for everyone. There is a moral obligation to engage with ethics beyond theoretical abstraction because the harmful consequences negatively impact real people. Technology and AI do not live in a vacuum void of systemic injustice or historical legacy; one must understand the real world to create tools that impact the real world, which begins with systems of oppression and how practitioners knowingly and unknowingly perpetuate them.

5.8 Broader RAI & TML Lessons Learned

Chapter 5 serves as a interdisciplinary and theoretical primer for critical engagement in ethics, and motivates the following chapters strong emphasis on sociological and AI ethics considerations. It also further highlights the complexity of AI ethics and challenges faced by practitioners. AI ethics and socio-technical risk mitigation is inherently challenging with sociological knowledge, and thus, is exponentially challenging for technologists and practitioners who do not have access to necessary resources and education. To attempt to bridge this gap in knowledge, the following chapter present work on AI and ML curriculum development that integrates ethics throughout.

This extensive chapter extensively covers relevant sociological theory and the theoretical underpinnings of why practitioners develop ethical blind spots and outlined steps on bridging

the knowledge gap. This work also discusses the value of abolitionist framing and intersectionality theory when analyzing socio-technical problems. We selected recidivism prediction as an exemplar case due to the notorious controversy in research. The recidivism prediction example is particularly instructive because it illustrates how intersectionality theory is often deliberately excluded from technical discourse as being “out-of-scope”

This pattern of exclusion reveals a troubling disconnect: concepts considered foundational in sociological disciplines are frequently labeled as "advanced" or "peripheral" in AI ethics education. We maintain that introductory sociological theory is not beyond the cognitive reach of AI practitioners and should be integrated early in technical education. The question of how to effectively implement instruction in sociological theory, particularly intersectionality theory, within AI education contexts forms a central inquiry of the present work.

CHAPTER

6

CSF: A SOCIALLY RELEVANT FOCUSED AI CURRICULUM DESIGNED FOR FEMALE HIGH SCHOOL STUDENTS

The following chapter is adapted from the accepted submission to the 2022 EAAI Symposium at AAAI Conference on Artificial Intelligence

Alvarez, L., Gransbury, I., Cateté, V., Barnes, T., Ledéczi, Á., & Grover, S. (2022, June). A Socially Relevant Focused AI Curriculum Designed for Female High School Students. In *Proceedings of the AAAI Conference on Artificial Intelligence: : Educational Advances in Artificial Intelligence (EAAI) Symposium* (Vol. 36, No. 11, pp. 12698-12705).

6.1 Dissertation Framing

Chapter 6 is survey study during a summer camp implementation of a new curriculum. The curriculum is a AI and ML module designed to integrate ethics and teach novice high school students advanced AI topics. The research question is: **(RQ3) How does an AI/ML curriculum with integrated social relevance and ethics considerations impact girls' confidence and**

self-efficacy in AI/ML?

The key results suggest that there was a positive increase in female participant confidence and self efficacy in AI and ML content, self efficacy in computer science, and career identity after attending the camp that situated AI and ML concepts around socially relevant topics like social media, environment, criminal justice and the arts. There were some challenges including two programming languages in the curriculum, but participants had a preference for Python. This chapter serves as a start to education research and my experience in AI and ML curriculum development. This chapter is the most related to the contributing undergraduate curriculum presented in the following chapter.

Abstract

Historically, female students have shown low interest in the field of computer science. Previous computer science curricula have failed to address the lack of female-centered computer science activities, such as socially relevant and real-life applications. Our new summer camp curriculum introduces the topics of artificial intelligence (AI), machine learning (ML) and other real-world subjects to engage high school girls in computing by connecting lessons to relevant and cutting edge technologies. Topics range from social media bots, sentiment of natural language in different media, and the role of AI in criminal justice, and focus on programming activities in the NetsBlox and Python programming languages. Summer camp teachers were prepared in a week-long pedagogy and peer-teaching centered professional development program where they concurrently learned and practiced teaching the curriculum to one another. Then, pairs of teachers led students in learning through hands-on AI and ML activities in a half-day, two-week summer camp. In this paper, we discuss the curriculum development and implementation, as well as survey feedback from both teachers and students.

6.2 Introduction

This paper details the design and initial testing process of our artificial intelligence (AI) and machine learning (ML) module of the CS Frontiers (CSF) curriculum. The CSF project focuses on curriculum development of a new high school course designed to introduce students who have completed the U.S. College Board's Advanced Placement Computer Science Principles course to cutting-edge computing technologies and applications Broll et al. (2021). Due to the decreasing trend of female students showing interest in the field of computer science over the last several decades, the CSF curriculum focuses on engaging female students Seneviratne

(2017). To increase female interest, curricular activities are focused on interdisciplinary and socially relevant connections between computer science and other disciplines, such as social media and criminal justice Lédeczi et al. (2021), that female students have shown interest in Fisher and Margolis (2002). The activities in the CSF AI and ML module connect advanced computer science concepts with other domains, such as criminal justice, natural language processing, the environment, and more. To teach novice high school students advanced topics, the curriculum leverages NetsBlox, the block-based programming language, which supports advanced programming techniques and connects to several Application Programming Interfaces (APIs) to enable interdisciplinary projects and access to online datasets Broll et al. (2017). Introducing students to AI and ML using a block-based language helps novice students focus on learning computing concepts rather than syntax of a text-based language Resnick et al. (2009).

In this paper, we explain the design process we used to create a 35-hour abridged camp curriculum, how teachers learn to use it, and how students and teachers perceived the materials. We will describe the pre and post surveys we used during the summer camp to determine female engagement and confidence, and the results of these surveys. In conclusion, we will discuss the significance of our findings and how our future work with this curriculum will proceed based on these discoveries. The investigation of the AI and ML module has shown that the curriculum positively impacts females' confidence and sense of belonging in computer science.

6.3 Background & Related Works

There has been a decline in female interest in computer science and computer science topics over the past several years Seneviratne (2017). Considering how the gender disparity in Tech has a statistically even distribution across race, ethnicity, and socioeconomic status levels, it can be deduced that the factors influencing the gender gap are not the same as those impacting the systemic connections between race, ethnicity, and income Wang and Hejazi Moghadam (2017). Therefore, this problem is not only about access, but it is also compounded with social barriers and perceptions Wang and Hejazi Moghadam (2017). Research has shown that female interest in computer science changes in early high school, where first career choices occur Microsoft Corporation (2017). For this reason, we want to introduce advanced topics to females early in high school, to show the possibilities of computer science that are not usually explored until late in a college computer science degree. Research suggests that bridging the gender gap in computing requires presenting computer science as a variety of perspectives and possibilities Khan and Luxton-Reilly (2016) and making connections to other subjects Fisher and Margolis

(2002). For instance, women (in general) prefer "people-oriented fields" that involve caring for communities and improving quality of life Kirk et al. (2012), while men prefer "thing-oriented fields" Ceci et al. (2014), which helps explain why women are more prevalent in careers with a clear social purpose, such as healthcare, social work, and education. Socially-relevant projects have shown to be engaging for female and underrepresented students Fisher and Margolis (2003); Papastergiou (2008); Ceci et al. (2014). Connecting real world experiences that make an impact with diverse female experts for support and inspiration can provide girls with authentic STEM opportunities that promote sustained engagement Chapman and Vivian (2017). Some existing advanced topic curricula, following this goal of engaging young women in the field of computer science, are based on the subjects of Distributed Computing and the Internet of Things Grover et al. (2020), Broll et al. (2021). For our curriculum framework, we wanted to implement an AI and ML module with the same goal. We believe introducing students to these socially-relevant topics early in their exposure to computer science fields will allow females to connect and engage, while still learning novice computing topics.

6.4 Curriculum Development Pipeline

6.4.1 Curriculum Curation

This section reviews how the curricula was initially developed and the preparation provided to camp teachers. The development and curation of the camp activities took place in several stages. First, we found relevant and open source AI and ML materials to teach K-12 students. Understanding that the curriculum is intended to fit a 35- to 40-hour summer camp or course-based 9-week schedule, we curated topics that would be suited for an abridged camp implementation and a typical high school classroom. We adapted the AI4K12 grade-band progression charts as learning objectives for this curriculum. AI4K12 is a developed, national curriculum for AI topics in classrooms ranging from kindergarten to 12th grade Touretzky et al. (2019). Topics range from domain knowledge to search algorithms to machine learning concepts depending on the targeted age group. We specifically used the AI4K12 Big Idea 3 progression chart, as it had high school level objectives specifically relating to machine learning algorithms, neural networks, and datasets. The key insights of these learning objectives Touretzky et al. (2019) are:

1. Machine learning allows a computer to acquire behaviors without people explicitly programming those behaviors.
2. Learning new behaviors results from changes the learning algorithm makes to the internal representations of a reasoning model, such as a decision tree or a neural network.

3. Large amounts of training data are required to narrow down the learning algorithm's choices when the reasoning model is capable of a great variety of behaviors.
4. The reasoning model constructed by the machine learning algorithm can be applied to new data to solve problems or make decisions.

Once these learning objectives were solidified, our team consisting of 6 female researchers (4 White women and 2 Latinx women) selected the activities which they thought would be most engaging. For learning objectives where resources were not found, we developed activities, such as our NetsBlox Twitter bot detection activity (See Table 6.1). In conjunction with the selected activities, we also used presentations prepared by AI4All, an online computer science curriculum promoting AI exposure to all students, from their AI Bytes units Judd (2020). The AI4All AI Bytes units are slide decks introducing areas of research and pre-made tools that students are able to tinker with to explore a topic. The presentations connected AI and ML concepts to real world applications and people. Examples of these presentations are AI & Environment, AI & Drawing, and AI & Facial Recognition. These presentations gave students exposure to how AI can connect to other disciplines they may be interested in and exposure to a diverse range computer scientists.

To assess the open source and developed activities and presentations, we had two high school interns (a female Asian high school student and a male Asian high school student), and one undergraduate researcher (a male Asian computer science undergraduate), complete and evaluate the activities in relation to interest level for female high school students and complexity based on prior computer science knowledge needed to complete the activities. The undergraduate researcher and interns all had differing levels of block-based and text-based programming knowledge to accurately evaluate these activities from different points of view. The undergraduate researcher and interns recorded the time it took for them to complete the activities independently, their perceived complexity of the activity, and if they found it engaging or interesting. After the activities were evaluated, they were paired with the related learning objectives (see Table 6.1) and arranged into the camp schedule.

6.4.2 Teacher Preparation

In order to achieve our goal, we elected to have secondary education teachers, with experience teaching the Advanced Placement Computer Science Principles course, act as camp instructors, facilitating the curriculum to the camp participants, and simulating a classroom environment Kick and Trees (2015). To prepare the educators, our team facilitated a week of professional development (PD) that leveraged the Teacher-Learner-Observer (TLO) model developed by

Table 6.1: Selected camp activities (in order that they were presented) and their related AI4K12 learning objective.

Activity Title	**Learning Objective(s)**
Twitter Bot NetsBlox Activity	AI4K12 3-A-v: Describe how various types of machine learning algorithms learn by adjusting their internal representations.
Sentiment analysis using NetsBlox and Python	AI4K12 3-A-v: Describe how various types of machine learning algorithms learn by adjusting their internal representations.
Introduction to Python Syntax	Understand Python syntax and introductory computational concepts (lists, libraries, objects) in Python
AI4All AI & Environment	Understand the different types of careers and applications to the AI and ML and the environment
AI4All AI & Drawing	Understand the different types of careers and applications to the AI and ML and the arts
AI4All AI & Criminal Justice	Understand the different types of careers and applications to the AI and ML and criminal justice system
Final project using APIs and sentiment analysis of a chosen media type (music lyrics, tweets, or NYT article summaries)	AI4K12 3Av: Describe how various types of machine learning algorithms learn by adjusting their internal representations. AI4K12 3-A-iii: Use either a supervised or unsupervised learning algorithm to train a model on real world data, then evaluate the results.

Goode, Margolis, and Chapman Goode et al. (2014). In the TLO model, participants spend time playing the roles of teacher and learner while the PD facilitators observe and guide a structured reflection after each TLO session Cateté et al. (2020). The goal of this PD was for the educators to co-design and critique our developed curriculum, similar to the purpose of the PD in Grover et al. (2020). At the end of each day during the PD, participants completed a debrief survey where they answered questions about how their TLO sessions went overall, suggested changes to the activities, and any concerns about the curricular materials.

6.5 Study Implementation & Context

6.5.1 Camp Context

To test our learning objectives in a controlled environment, we condensed our curriculum into a two-week summer camp in conjunction with our university's camp infrastructure Bottomley (2015). There were two camp implementation types: Camp A which included both

Table 6.2: Gender, Race, and Ethnicity demographics of the 24 camp participants included in our analysis.

Gender	Native American	Asian	Black	Multi-Racial	White (non-Hispanic)
Female	1	4	0	1	4
Male	0	10	1	2	1

NetsBlox and Python activities and Camp B which only included Python activities. Each camp had approximately 35 hours of contact time with participants. Camp A was online 4 hours a day over the course of 10 weekdays. Camp B was online 8 hours a day over the course of 5 weekdays. We had two instances of both Camp A and Camp B each. The online platform used for synchronous activities was Zoom due to its video and audio recording, screen-sharing, breakout room capabilities, and chat features. Breakout room capabilities were particularly important to facilitate collaboration amongst participants, since collaboration is a key computational thinking and computer science practice to better engage female learners Denner et al. (2005). In each camp's Zoom classrooms, we had two teacher facilitators, one computer science graduate student, and an undergraduate camp counselor.

The gender and racial and ethnic demographics of the 10 female and 14 male high school participants who consented to participate in this research study are displayed in Table 6.2. We distributed the pre and post surveys via Google Forms. At the beginning of the first day of camp, the participants completed the pre survey and on the last day, the post survey was completed. The surveys were adapted from the NCWIT Computing Interest-Confidence-Perception Survey National Center for Women & Information Technology (2021) and the Development of the STEM Career Interest Survey Kier et al. (2014). The survey topics covered were confidence, perception, interest in computing, interest in computing careers, and curriculum content knowledge.

6.5.2 Abridged Curriculum

The AI concepts addressed in the camp were natural language processing, sentiment analysis on textual data, bias in datasets, computing ethics, AI connections to social media, and AI in real world applications. The expected learning outcomes of the summer camps were 1) describe how various types of machine learning algorithms learn, 2) use either a supervised or unsupervised learning algorithm to train a model on real world data, then evaluate the results, 3) understand introductory computational concepts (lists, libraries, objects) and syntax in Python, and 4) understand the different types of careers and applications of AI and ML to the

criminal justice system, the environment, and the arts. The activities that taught each learning objective are presented in Table 6.1.

While we also prompted the participants on completion and difficulty feedback, the focus of our AI and ML summer camp was on engagement and interest with the materials presented. To account for the density of material in the short two-week camp and to account for difference in computing backgrounds, our team scaffolded activities to help novice participants connect to the programming faster and added extensions to challenge advanced participants. In the summer camp implementation, there was limited time to engage in each activity, so we prepared starter code for participants to offset the reduced time. For example, for novice participants, the penultimate activity provides adaptable starter code usable for the final project, and for more advanced participants, they could choose to pick a new API to work with as a challenge.

The intention for each activity was to have participants think of the impact and larger applications of the concepts they were learning. For assignments, the participants were asked to think beyond their small batch activities to larger implications. For example, if we look at the final project, it had the option of three APIs Genius Lyrics, Twitter, and the New York Times (NYT). From these options, the participant groups chose and devised a research question they hoped to begin to answer with their final project. One group's final project compared the sentiment analysis on the lyrics of a few songs from 5 short haired artists to 5 long haired artists. The participants were then asked to explain what their results meant and to come to the conclusion that their results utilized a small amount of data but with more they could make stronger claims. Implementing the curriculum through a summer camp served as a way to pilot the materials and receive initial feedback from participants and teachers.

6.6 Results and Analysis

6.6.1 Participant Survey Results

Using a statistical software package, SPSS, an independent-samples t-test was conducted on the pre and post camp surveys to compare female and male participant confidence before and after attending the 2021 summer camps. The pre and post survey entails 52 Likert type questions in the following subcategories: confidence, perception, interest in computing, interest in computing careers, and curriculum content knowledge. Questions 1-41 were on a 4 point scale, with 4 indicating stronger agreement, and questions 42-52 were on a 5 point scale with 5 indicating strongest agreement with the statement. Bonferroni correction was applied to correct for the large number of tests, so significance requires $p \leq \alpha = 0.00096$.

In Table 6.3 we share a set of selected questions and their results that measure confidence

on items related to AI and ML content, career identity, and self efficacy in computer science.

Questions 15 and 19 of the survey were concerned with confidence in understanding relevant computing concepts presented in the camp. Questions 36-38, asked how much they knew about computer scientists and what their job responsibilities look like. In the final part of the survey, questions 42-51, respondents were asked to rank on a Likert scale how true they felt each question related to their self efficacy. Below we highlight and discuss questions 15, 36, 37, 38, and 42 as those items measure confidence in career identity and general computer science.

There was a difference, t(9) = 3.28, p= 0.01, in confidence regarding Q.15 for female participants increasing from M=2.0, σ= 1.05 to M=2.7, σ = 1.05 in Likert value. Effect size was calculated as Cohen's D = 0.68, indicating a medium effect on increasing female participants' confidence in their ability to represent data (and images).

For Q. 36, there was a difference, t(9) = 3.28, p= 0.01, in confidence in understanding computer scientists' jobs for female participants increasing from M= 2.20, σ = .63 to M = 2.90, σ = 0.57. Effect size was calculated as Cohen's D = 0.68 indicating a medium effect, suggesting that the camp helped female participants become more confident in knowing the jobs that computer scientists have. The next question of interest is Q. 37 where data shows that female participants felt more confidence in knowledge of computer scientist's job responsibilities (M = 2.20, σ = .92 to M = 3.0, σ = .67) from pre to post. Using the Bonferroni correction, however, this trend is not significant, t(9) = 2.45, p= 0.04. Although overall the camp helped female participants become more confident in understanding what computer scientists do in their jobs.

The fourth highlighted question is Q. 38. There was a difference, t(9) = 3.86, p< 0.01, from pre (M=1.40, σ=0.70) to post (M=2.30, σ=0.82 with an increase in confidence in how much participants know about building AI applications for female participants. Effect size was calculated as Cohen's D = 0.74, indicating a medium effect on helping female participants become more confident in knowing how to build AI applications. This result shows support for our hypothesis that a more socially relevant and applied curriculum would be more engaging for female participants. The last question of interest for the in-group analysis is Q. 42. There was a difference, t(9) = 2.45, p= 0.04, from pre (M=2.90, σ= 0.99) to post (M=3.70, σ= 0.95) with an increase in confidence in completing computer science activities for female participants. Cohen's D = 1.03 indicating a large effect, regarding the camp helping female participants become more confident that they would be successful in activities that involved computer science. Together these results suggest that the female participant efficacy increased after the completion of our summer camp.

Furthermore, we highlight the results of questions 38, 42, and 51 in comparison to our male participants because they reflect each of the survey section topics: confidence in AI

and ML content, self efficacy in computer science, and career identity. The first question of interest helps serve as a baseline for the cross group analysis, Q. 38 shows that female and male participants did not show a difference from each other in their change of understanding regarding AI.

The second question of interest is Q. 42. There was a difference, t(22) = -2.54, p= 0.02, in increase in confidence in completing computer science activities for female attitudes (MD=0.80, σ= 1.03) compared to male attitudes (MD= -0.14, σ= 0.66). With Cohen's D = 0.83 indicating a large effect, these results suggest that female participants gained more confidence in their ability to do well in computer science activities when compared to male participants. The last question of interest for the cross group analysis is Q. 51. There was a difference, t(22) = -3.03, p< 0.01, in increase in confidence in talking with computer scientists for female attitudes (MD=0.70, σ= 0.82) compared to male attitudes (MD= -0.14, σ= 0.36). With Cohen's D = 0.60 indicating a medium effect, these results suggest that female participants gained more when compared to male participants, in becoming more comfortable talking to people who are computer scientists.

Table 6.3: t-Test results showing the mean difference in female participant confidence from pre/post surveys ($\alpha = 0.018$) (N=10).

Survey Question	Pretest Mean	Post Mean	MD	σ	t	p
Q.15: Right now, how confident are you in your ability to understand how computers present data and images?	2.00	2.70	0.70	0.68	3.28	0.01
Q.19: Right now, how confident are you in your ability to program computers to create new apps (in other words, writing code)?	2.00	2.60	0.60	0.69	2.71	0.02
Q.36: How much do you know about jobs computer scientists have?	2.20	2.90	0.70	0.68	3.28	0.01
Q.37: How much do you know about what computer scientists do in their jobs?	2.20	3.00	0.80	1.03	2.45	0.04
Q.38: How much do you know about how to build artificial intelligence applications?	1.40	2.30	0.90	0.74	3.86	<0.01
Q.42: I am able to do well in activities that involve computer science.	2.90	3.70	0.80	1.03	2.45	0.04
Q.51: I would feel comfortable talking to people who are computer scientists.	3.30	4.00	0.70	0.82	2.69	0.03

Table 6.4: Independent samples t-Test results showing mean differences in female (F) versus male (M) participant change in confidence from pre/post-survey.

Survey Question	F M	Pretest Mean	Post Mean	MD	σ	t	p
Q.38: How much do you know about how to build artificial intelligence applications.	10 14	1.40 1.64	2.30 2.64	0.90 1.00	0.74 0.96	2.76	0.78
Q.42: I am able to do well in activities that involve computer science.	10 14	2.90 4.50	3.70 4.36	0.80 -0.14	1.03 0.66	-2.54	0.02
Q.51: I would feel comfortable talking to people who are computer scientists.	10 14	3.30 4.50	4.00 4.36	0.70 -0.14	0.82 0.36	-3.03	<0.01

6.6.2 Participant Feedback

To gain perspectives from our female participants, we look at open-ended questions from the post-survey on participants' experience in the camp and their interest in a career in computer science. The female participant data showed the overall response to the question "What other feedback, comments, or suggestions do you have after your experience with the [CSF] Camp?" was very positive. One female participant commented on recommending the program, "I learned a lot and would recommend it to others." Others included, "I like how they were trying to make the students very energetic and more involving with the lesson' and 'I had fun at this camp." Four female participants commented on the impact of the facilitators on their experience which is another factor for female participant engagement Fisher and Margolis (2002). Additional feedback illustrates how facilitation is a valuable aspect of curriculum development and the importance of the facilitator(s) impact on the participants. A female participant commented, "I really enjoyed the camp and the support my teachers provided", "The information was very thorough and I really liked the step by step coding instructions. The Genius API activity was also something I've never seen before (in a good way).", "I liked how knowledgeable all the counselors were and the amount of effort that was put into the presentations and activity planning was evident".

Our results indicated that while the participants were challenged, they were not discouraged by the difficulty or the new concepts they were learning as represented in the following comment "I really enjoyed this camp! A few of the assignments towards the end of the week were confusing but once we worked through them it felt good that we could understand it."

In the final part of the survey, respondents were asked "Do you think you could be a computer scientist?" 6 female participants answered yes and 4 answered maybe (due to being interested in other STEM areas, which was specified in the survey). The results of this question, coupled with the Likert scale survey questions, is interesting to our research goals because it indicates that the participants left more confident in knowing what being a computer scientist is like and whether they could visualize themselves as one. With female representation in computer science being 25% nationally National Center for Women & Information Technology (2021), career identity of school age females is important to investigate to help further understand the barriers and obstacles to female students entering computer science Google Inc. (2014).

In summary, the results of the pre/post surveys informed us about female self efficacy levels after the camp and how it compared to male participant data including changes in confidence, content knowledge, and career interest. The female participants had positive responses to building AI applications and how computers present images and data, and reported some improvements to their self efficacy in computing and AI concepts. In comparison to the male

group, the female participants had a greater increase in AI and ML content, self efficacy in computer science, and career identity after attending the camp. Between the two groups, we regard both male and especially female participants positively engaging with the curriculum as we are moving in a positive direction toward the CSF project goals of broadening participation in computer science and engaging more young women in advanced computing topics.

6.6.3 Teacher Feedback

Sentiment of teacher feedback received from educators who facilitated the abridged camp curriculum changed over time, as the teachers became more familiar with the materials and gained confidence in teaching more advanced computer science topics. Initially during the PD, one educator felt there was a sense that the “big picture” of AI was disregarded with the introduction of Python. This view was directly stated in a statement by the facilitator, “Make sure campers understand the bigger picture – anticipate that they’ll wonder why we introduced Python”. When asked about how they felt about Python fitting into the curriculum, multiple educators explained it would be more beneficial for participants to be exposed to Python; perhaps by comparing NetsBlox and Python code, but not to learn Python syntax in the course, because the main goal of the curriculum is to engage participants in the overall topic of AI and ML, not programming. Additionally, in anticipation of problem areas, the educators suggested more detailed explanations of the AI and ML concepts in activities’ instructional guides to better support teachers in leading the materials and answering participant questions. Overall, prior to leading the summer camps, the teachers felt comfortable with the provided materials.

Each day after each camp session, we debriefed with the teachers to get their perspectives on trying out the materials with participants. In all cases, the facilitators reported a positive view of the curriculum and disclosed that participants were engaged in the activities. Conversations with the PD facilitators expressed some types of activities were more successful than others. Successful activities included the NetsBlox activities, the AI4ALL AI Bytes units, and the final project presentations. Teachers stated that the NetsBlox activities (Twitter Bot Classification and Sentiment Analysis) allowed participants to focus on the AI concepts being discussed without participants being concerned with optimal coding concepts. As the participants began the Python activities, a facilitator commented “Students seemed to enjoy working in Python. My students even said that it was fun." We also received positive comments about the AI4ALL AI Bytes presentations and the topics discussed. One teacher said, “ [I] really really appreciated the absolutely excellent slide shows!", and another said, “Campers had some good insights about some of the issues with facial recognition AI, as well as some of the potential beneficial uses". This statement also expresses that participants were not only learning about the subject

of AI and ML, but engaging in the subject matter.

Finally, the culmination of the camps was the paired/group final projects and presentations. The day before the final presentation showcase (where parents were able to attend with their children to view the final project demos) a facilitator remarked “All groups made excellent progress on their projects today. More than half the groups were able to finish and think about extensions to their projects. The other groups are confident that they’ll be able to finish in time tomorrow." Despite the differences in programming backgrounds, all participants completed a final project and presented a slide presentation on them with a demo of their Python code.

Other feedback wasn’t necessarily related to the curriculum but suggested time adaptations such as more time for teacher preparation between the school year, PD, and beginning of the camp (since the PD and camps were in consecutive weeks. Given the geographically dispersed nature of the team, some teachers had been out of school for multiple weeks, while another had just finished a few days prior. With time commitments and the obstacles of virtual learning, we plan to adjust for these challenges in coming seasons.

6.7 Discussion

6.7.1 Participant Perceptions

The 2021 summer camps served as a pilot test for a controlled virtual classroom setting, and provided valuable feedback from the participants and facilitators. The results of the surveys suggest that the abridged version of the full 9 week AI and ML module curriculum was successful in engaging the female participants in the summer camp and improving their self efficacy and confidence in computing. The female participants had positive responses to computing concepts and reported improvements to their self efficacy in computing and career connections. Specifically, in comparison to the male participants, the female participants had a greater increase in AI and ML content, confidence in computer science, and career interest after attending the camp. It’s interesting to note the negative correlation in the difference in means for the male participants given the overall positive growth in mean. We theorize the male participants had a preexisting misconception about how much they knew prior to the camp given their high scores on the pre-survey. Then after participating in the camp, they had a more realistic perception of the computing topics. Therefore, we don’t believe the male participants learned less than the female participants given the data findings. Looking at the data further may expose deeper correlations or individual student characteristics that may confound or explain the data, although that is currently outside the scope of this report.

We conjecture that the female participants were more impacted by the abridged camp

curriculum than the male participants due to female representation in the camp facilitation, female representation in the camp curriculum, socially relevant applications, and intentional group pairings. In the open-ended response questions, 4 female participants commented on the impact their camp facilitators had on their experience. In the camps, only 2 of the 11 instructors and computer science teaching assistants from across both Camps A and B were male, demonstrating higher female representation in computer science. The AI4ALL AI Bytes units also showed diverse computer scientists and highlighted inclusivity in the field. Furthermore, we reduced the amount of technical jargon and focused on a more applied people-focused computing lens versus "things-focused" programming and syntax Kirk et al. (2012). As we had some student participants in the summer camp that were new to computing, we find that the combination of inclusive representation in the curricular materials, and socially relevant applications helped make the curriculum relevant to a range of students with different backgrounds. We also believe this lowered the barrier of knowledge to participate in group discussions. Considering imposter syndrome has a greater prevalence in computer science students than other disciplines and a greater prevalence in female students than male students, the teachers helped further facilitate more engagement in discussions by intentionally pairing female participants together to ensure that they were not the only female participants in all male groups Rosenstein et al. (2020). These same discussion group tactics were also applied to picking final projects and joining breakout rooms.

Furthermore, a relationship exists between including interdisciplinary frameworks to make traditional computational thinking more inclusive and broadening the participation of underrepresented students in computer science Kafai et al. (2020). The camp's curriculum connected computational thinking to interdisciplinary perspectives by including creativity in brainstorming sessions and application to other real world applications (i.e. music, art, social media, nature, and equity).

6.7.2 Teacher Perceptions

Although student perceptions and experiences of the camp were positive, PD teachers were initially hesitant to try to fit all of the different activities into the two week camp, specifically teaching two languages in a limited time frame as it might take away from the content. When leading the camp sessions, however, the teachers realized that the participants didn't struggle as much as anticipated and that they were eager to learn the new materials. To alleviate future potential discomfort, our goal is to include more NetsBlox and Python interconnected activities to ensure that students are transitioning smoothly between the two languages, and understanding how the computing concepts are represented in both a block-based and text-based

language. In the end, the teachers worked in pairs as they delivered the materials and were able to include each of the main activities into their camp programs.

6.7.3 Limitations

Several limitations were imposed by COVID-19 restrictions including virtual learning obstacles, population size, and time constraints. The online virtual learning setting included internet, video, or audio connection problems, which as the camp progressed led participants to be less inclined to have their cameras on and deterred from the typical social experience of a camp or classroom. To combat this, one teacher used “chat cascades" which began with a question prompt to the group to have them type, but not send their response until told to do so. Once prompted, all participants would send their response at once in the text chat and the participants could see all of the responses filter in as a cascade. Additional limitations may impact the findings of this report, such as the self-selection bias of camp participants and the different computing background experiences of the instructors. To mitigate the diversity in teacher backgrounds, teachers were given the same preparation materials, and each teacher pair included a teacher with a python programming background. While we note in Section 6.9 the broader responsible AI and trustworthy ML lessons learned, we note a key limitation of this work is the lack of explicit learning objectives around trustworthiness/ethics/responsibility.

Lastly, to account for the large number of tests, we conducted a Bonferroni correction when needed. After further evaluation, we note there is an error in the Bonferroni correction, and the threshold is more lenient than it should be. It’s likely we used an overly permissive threshold (alpha of 1) rather than a standard Bonferroni correction (0.05 / 52). However, we acknowledge this is not appropriate for confirmatory testing and will correct it in future versions.

6.8 Conclusions and Future Work

The aim of the present evaluation was to examine how well the AI and ML curriculum engaged high school students, specifically the 10 female survey participants in the summer camps. Given the limitations of this preliminary study, our pre and post survey results and teacher and student feedback support our extended curriculum goals of engaging female high school students in advanced computing topics. This study’s results support previous research that female students engage well with socially relevant topics like contextually situated AI and Machine learning Fisher and Margolis (2002). Specifically, our results suggest that there was a positive increase in female participant confidence and self efficacy in confidence in AI and ML content, self efficacy in computer science, and career identity after attending the camp that

situated AI and ML concepts around socially relevant topics like social media, environment, criminal justice and the arts.

Participant feedback highlighted the value of strong facilitation, with praise for the counselors' knowledge and the evident effort in presentations and activity planning. Key factors for female engagement included female representation in instructors and materials, an inclusive and socially relevant curriculum, and grouping strategies that avoided isolating female students. Participants preferred using Python and engaging in real-world activities. Teachers were initially hesitant to introduce two programming languages but were surprised by students' eagerness to learn Python; their own confidence grew as they observed strong student engagement. They suggested improvements in PD timing, curriculum preparation, and clarity of activity guides. The curriculum effectively connected computational thinking to interdisciplinary themes such as music, art, social media, nature, and equity, aligning with Kafai et al. (2020) findings on inclusive computing education. Future improvements could include better integration between block- and text-based languages and smoother transitions between programming environments.

In our future work, we plan to address the remaining feedback from students and teachers by providing additional support and guidance. We intend to run another summer camp with the adapted AI/ML curriculum to evaluate and continue refining the curriculum and to get another round of feedback before solidifying the module into the full 2022-2023 academic year curriculum. This designing, testing, and implementation pipeline ensures several rounds of active feedback from teachers and participants before the full curriculum is deployed for in-class testing and evaluation Könings et al. (2014). The teachers who were involved in the facilitator PD have committed to implementing selected activities in their classrooms during the spring 2022 semester, so we expect to receive more feedback and data from authentic classroom environments.

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant numbers 1949472, 1949488, and 1949492.

6.9 Broader RAI & TML Lessons Learned

The Computer Science Frontiers (CSF) project is a prime example of how a short AI/ML module designed for high school students was successful at integrating ethics into every lesson.

Students responded favorably to this integrated approach, providing empirical support for embedding ethical considerations throughout the technical curriculum. However, student projects revealed limitations in responsible AI and TML knowledge. For example, a student project using an open-source music API returned the top 5 songs, and the students were not able to make clear judgments or demonstrate how the data used impacted the model outcomes. While they acquired technical skills for model development, they lacked the analytical framework to evaluate their creations through a responsible AI lens. This observation reinforces the need for explicit instruction in both technical methods and evaluative frameworks to develop practitioners capable of creating trustworthy systems.

CHAPTER

7

TRUSTWORTHY ML CURRICULUM DEVELOPMENT: INTEGRATING ETHICS, SAFETY, AND EDUCATION

7.1 Introduction

This chapter discusses the theory and development of an introductory course on Responsible AI (RAI) and Trustworthy Machine Learning (TML) for undergraduate students. Dr. Kush Varshney's book, *"Trustworthy Machine Learning"*, was used to scaffold the technical content, however, Dr. Varshney notes a lack of philosophical, legal, political, sociological, psychological, and economic perspectives as a key limitation of the book. This work aims to address this limitation by adapting the book's technical content into a semester long curriculum and adding supportive sociological and ethical considerations grounded in technical, sociological, educational, and cognitive theory. Specifically, I address: **(RQ4) How do we explicitly apply Design Based Research (DBR) to design a course that integrates technical AI and AI ethics skills development?** To answer this research question, I present this chapter on curriculum design where DBR was used to guide the development.

7.2 Background & Prior Work

The development of a comprehensive curriculum for Responsible Artificial Intelligence (RAI) and Trustworthy Machine Learning (TML) represents a critical educational intervention in the training of future computer scientists. Prior research has identified significant gaps in how technical practitioners approach socio-technical problems. As demonstrated in my previous work, technologists often exhibit ethical and sociological blind spots that impede their ability to fully comprehend the implications of their technical decisions. This curriculum seeks to address these deficiencies through intentional integration of sociological theory with technical content, supported by evidence-based educational approaches.

Varshney's *"Trustworthy Machine Learning"* (Varshney 2022) provides a technical framework as the primary technical scaffold for the course. The book (written from the perspective of industry professionals) talks about the technical skills needed, but Varshney details a key limitation of his book as:

> "…I recognize that philosophical, legal, political, sociological, psychological, and economic perspectives may be even more important to understanding, analyzing, and affecting machine learning's role in society than the technical perspective. Nevertheless, these topics are outside the scope of the book. Insights from the field of human-computer interaction are also extremely relevant to trustworthy machine learning; I discuss these to a limited extent at various points in the book …"

To address this limitation, I bring in experience from previous work and supplemental theoretical considerations detailed in the following sections.

7.3 Curriculum Design as Research

A key justification for treating curriculum design as research is its impact on learning and educational outcomes. Just like a scientific experiment, a new curriculum can be hypothesized to improve certain outcomes (e.g. student engagement, retention, diversity in the field), and this hypothesis can be tested with data (Stufflebeam and Shinkfield 2007). From this perspective, creating a new curriculum or tool is akin to designing an artifact in engineering: it is an innovation that can be studied, refined, and shared (Brown 1992). Many curriculum design studies highlight innovation by addressing a gap (for example, few programs teaching AI ethics (Gouseti et al. 2024; MIT Media Lab 2024), or lack of real-world projects in core CS courses (Kumar et al. 2024a)) and showing how a new curriculum component can fill it. If the idea is

truly novel and grounded in sound rationale, its dissemination through scholarly channels is justified by the innovation it represents in teaching practice. Perhaps most importantly, curriculum design qualifies as research when it is executed and evaluated with rigor (McKenney and Reeves 2021). The process of designing a curriculum can follow a research-like methodology: identifying a problem (e.g., students struggling with certain concepts, or an outdated skill set in graduates), reviewing relevant literature and standards, formulating design principles or learning objectives, implementing the curriculum, and assessing the results (van der Geer et al. 2007). This mirrors the empirical cycle of research.

Whether approached through Scholarship of Teaching and Learning (SoTL) (Boyer 1990; Hutchings and Shulman 1999), design-based research (Brown 1992; Collins et al. 2004), or discipline-specific models, the common thread is intentionality and inquiry: treating curriculum work as an experiment or design project to be shared, critiqued, and built upon by others. Recent developments in computer science education exemplify this approach, with the CS2023 curriculum guidelines emphasizing both rigorous evaluation methods and the integration of emerging areas like AI and ethics throughout the curriculum (Kumar et al. 2024a; Simha et al. 2024). By leveraging these frameworks, educators in AI and CS can ensure their curriculum initiatives are not only pedagogically sound but also scholarly contributions that advance the field of education.

In fact, educational researchers have formalized such processes, such as Design-Based Research (DBR) (Brown 1992; Collins et al. 2004) or Educational Design Research (EDR) frameworks, to ensure that educational innovations yield generalizable knowledge. McKenney and Reeves (2021) describes Educational Design Research (EDR) as a genre of inquiry that is "committed to the simultaneous development of both theoretical insights and practical solutions, together with stakeholders." In curriculum design terms, this means a researcher might co-design a course or module (practical solution) while also examining how students learn from it, thereby contributing to learning theory or pedagogy (theoretical insight).

Problem Statement: In this paper we are considering the problem that North Carolina State University has no computer science course that integrates sociology and advanced computing topics. To address this problem, we use DBR and EDR methods to design a curriculum that addressed the gap in advanced computing topics with integrated sociology.

7.4 Technical, Sociological, & Foundations

We employ the Design Based Research (DBR) methodology to address our research question: **(RQ4) How do we explicitly apply Design Based Research (DBR) to design a course that integrates technical AI and AI ethics skills development?** The DBR method is applied through-

out the study, and it involves a systematic process that includes the selection of pedagogical approaches, the integration of relevant learning theories, and iterative cycles of design and evaluation. The DBR approach supports the alignment of theory and practice, ensuring that the curriculum is both grounded in educational research and responsive to practical classroom needs. This section is structured to follow the first phase of the DBR process (the analysis and exploration phase). The Course design and following section address the second design and development phase: and finally, Chapter 8 addresses the last implementation and evaluation phase.

7.4.1 Chapter 3 & 4 Technical Takeaways

The STEALTH and iSTEALTH experiments investigated how multiple attributes of trustworthiness impact TML pipeline implementation and performance. They provided crucial insights into the multifaceted nature of trustworthiness in machine learning pipelines. With current interventions, the myth of fairness-accuracy tradeoff has been disproven, but there is limited knowledge on how other extensions of trustworthiness (i.e., basic performance, reliability, interaction, and purpose) impact each other in certain implementation contexts.

A key finding was that TML implementation is context-specific and requires domain-specific sociological knowledge. The iSTEALTH framework presented a viable solution for creating interpretable and efficient global surrogate models with attention to adversarial models. Still, more work is needed with complex model architectures, and to understand the real-world implications of data imbalances and biases to appropriately apply relevant mitigation techniques. Most critically, these experiments highlighted how philosophical, sociological, and educational blind spots among technologists impair their effectiveness in socio-technical spaces. This observation serves as a fundamental rationale for the current curriculum development.

7.4.2 Chapter 5 Lessons

Chapter 5 extensively covers relevant sociological theory and the theoretical underpinnings of why practitioners develop ethical blind spots and outlined steps on bridging the knowledge gap. Chapter 5 also discusses the value of abolitionist framing and intersectionality theory when analyzing socio-technical problems. I selected recidivism prediction as an exemplar case due to the notorious controversy in research. The recidivism prediction example is particularly instructive because it illustrates how intersectionality theory is often deliberately excluded from technical discourse as being “out-of-scope”

This pattern of exclusion reveals a troubling disconnect: concepts considered foundational in sociological disciplines are frequently labeled as “advanced" or “peripheral" in AI ethics education. I maintain that introductory sociological theory is not beyond the cognitive reach of AI practitioners and should be integrated early in technical education. The question of how to effectively implement instruction in sociological theory, particularly intersectionality theory, within AI education contexts, forms a central inquiry of the present work.

7.4.3 Chapter 6 - Prior AI Education Efforts

The Computer Science Frontiers (CSF) project is a prime example of how a short AI/ML module designed for high school students was successful at integrating ethics into every lesson. Students responded favorably to this integrated approach, providing empirical support for embedding ethical considerations throughout the technical curriculum. However, student projects revealed limitations in responsible AI and TML knowledge. For example, a student project using an open-source music API returned the top 5 songs, and the students were not able to make clear judgments or demonstrate how the data used impacted the model outcomes. While they acquired technical skills for model development, the students lacked the analytical framework to evaluate their creations through a responsible AI lens. This observation reinforces the need for explicit instruction in both technical methods and evaluative frameworks to develop practitioners capable of creating trustworthy systems.

7.4.4 Cognitive Theory

The design of this RAI and TML course is firmly grounded in established cognitive learning theories that emphasize active knowledge construction, reflection, and authentic applications. This theoretical foundation informed pedagogical decisions, instructional strategies, and assessment design throughout the curriculum development process. Research demonstrates that students who engage in metacognitive practices develop deeper understanding and more transferable knowledge (Flavell 1979). This curriculum incorporates structured opportunities for metacognitive reflection, particularly when examining the intersection of technical decisions and sociological impacts. By carefully structuring the presentation of technical and sociological content, the curriculum aims to optimize cognitive resources for meaningful learning.

7.4.5 Bloom’s Taxonomy

Bloom’s Taxonomy (revised by (Anderson and Krathwohl 2001) and further discussed by Wilson (2016)) provides a hierarchical framework for classifying educational objectives and designing

learning activities that promote higher-order thinking skills. This framework was instrumental in developing the course's learning objectives and associated activities, ensuring progression from foundational knowledge to sophisticated analysis and creation. The curriculum systematically addresses each cognitive level:

- **Remembering:** Students master foundational terminology and concepts in both technical TML.
- **Understanding:** Through comparative analysis of case studies, students demonstrate comprehension of how technical choices manifest in sociological impacts.
- **Applying:** Coding exercises require students to implement fairness constraints, explainability methods, and robustness checks in practical scenarios.
- **Analyzing:** Discussion activities prompt students to deconstruct AI systems, identifying stakeholders and power dynamics that influence system outcomes.
- **Evaluating:** Debates or policy assignments develop students' ability to assess AI systems against multiple ethical frameworks and technical standards, making reasoned judgments about their trustworthiness.
- **Creating:** Culminating projects require students to design and implement AI systems that integrate TML with sociological considerations.

The taxonomy guided the intentional design of learning progressions within each section. For example, the fairness module begins with defining fairness metrics (remembering), transitions to explaining why different metrics might conflict (understanding), then requires implementing mitigation techniques (applying), examining how these techniques operate across different demographic groups (analyzing), assessing the appropriateness of different approaches in specific contexts (evaluating), and ultimately designing novel interventions for a specific use case (creating). This structured approach to cognitive development ensures that students not only acquire technical knowledge but also develop the higher-order thinking skills necessary to navigate complex socio-technical problems.

7.4.6 Experiential Learning Theory

Kolb's Experiential Learning Theory served as a primary framework for how students transform experience into knowledge. Kolb's learning cycle has four stages: concrete experience, reflective observation, abstract conceptualization, and active experimentation. The deliberate structuring of activities to move students through all four phases of Kolb's cycle aimed to facilitate

deeper understanding and application of machine learning concepts beyond what traditional lecture-based approaches might achieve (Dewey 1938).

7.4.7 Constructivist Learning Approaches

Constructivist principles guided the creation of a learning environment where students actively build their understanding through experience and interaction (Honebein 1996; Vygotsky 1978; Piaget 1952). The curriculum design is fundamentally rooted in constructivist principles by creating opportunities for students to construct their understanding of ethical implications by engaging with real-world AI systems and their societal impacts. The technical frameworks from Varshney's text serve as a foundation upon which students can build more complex ethical reasoning skills through guided exploration and problem-solving. Collaborative activities were integrated throughout to embed learning in social experience, reflecting the social constructivist view that knowledge is co-constructed through interaction (Vygotsky 1978)

The course promotes a student ownership, a key constructivist principle, approach that challenges students while remaining within their developmental reach. The course design deliberately incorporated transparency about learning objectives and pedagogical decisions to encourage ownership. The class can take a majority vote to voice any changes they want in their learning process (i.e., a deadline extension). Another constructivist principle implemented is Vygotsky (1978) zone of proximal development. The topic describes the space between what learners can accomplish independently and what they can achieve with guidance. These constructivist design elements were specifically chosen to create an engaging learning environment that challenges traditional lecture-based pedagogy in technical fields.

7.4.8 Metacognition and Active Learning Practice

Active learning strategies form the pedagogical foundation of the curriculum. Active learning principles infused all aspects of the course design, reflecting contemporary understanding of effective STEM pedagogy (Bandura 1977). Research consistently demonstrates that active learning approaches produce superior outcomes in STEM education compared to passive instruction Freeman et al. (2014). Active learning approaches were strategically implemented to replace traditional lecture time with student-centered activities that research indicates lead to improved learning outcomes in STEM fields (Bruner 1961).

Transformative learning beyond the classroom is a critical outcome when addressing sociological dimensions of technology. Transformative learning theory (Mezirow 1991) informs how the curriculum facilitates shifts in students' understanding and beliefs. By creating dilemmas that challenge existing assumptions about AI systems, the curriculum catalyzes perspective

transformation. The course deliberately incorporated metacognitive practices that encouraged students to think about their own learning processes, and develop students' capacity for self-regulated learning beyond the course itself (Flavell 1979).

7.5 Related AI Curriculum Research

Contemporary examples include systematic approaches to integrating AI ethics into K-12 curricula (Gouseti et al. 2024), where researchers simultaneously develop practical teaching materials and investigate how students develop ethical reasoning about artificial intelligence (Fu and Weng 2024). A well-conducted design-based study does not end at implementing a new syllabus; it rigorously collects data (e.g. assessments, surveys, observations) to evaluate effectiveness and iteratively refine the design (Valstar et al. 2020). This research-oriented approach to curriculum design has become increasingly important as educational technology evolves rapidly and new domains like AI ethics require evidence-based pedagogical approaches (Bond et al. 2024; Akgun and Greenhow 2022). The Computing Curricula 2020 framework further emphasizes the need for competency-based approaches that can be systematically evaluated and refined (Clear et al. 2020), supporting the view that curriculum design, when conducted with proper methodology and evaluation, constitutes legitimate educational research.

7.5.1 CS Ethics and Curriculum Guidelines

The CS2023 curriculum guidelines represent the most comprehensive integration of ethics into computer science education to date. CS2023 establishes Society, Ethics, and the Profession (SEP) as an integral curriculum component and includes substantial coverage of AI and its societal implications (Kumar et al. 2024a; IEEE Computer Society 2024). The guidelines' development process exemplified research-based curriculum design, incorporating global surveys and systematic evaluation of knowledge areas (Kumar et al. 2024b).

Recent work has demonstrated systematic approaches to integrating specialized ethical considerations, such as accessibility, across CS curricula rather than in standalone courses (Ladner et al. 2023b,a). These efforts provide models for embedding AI ethics throughout technical programs.

7.5.2 AI Ethics Curriculum Development

Existing AI ethics curricula have employed diverse approaches, from modular experiential learning addressing secure and reliable AI (Transformative Interdisciplinary Human+AI Re-

search Group 2022) to comprehensive semester-long courses drawing on humanities perspectives (National Humanities Center 2024). The National Humanities Center's Responsible AI project supported systematic curriculum development across 23 faculty and 15 institutions, demonstrating scalable approaches to AI ethics education.

Industry practitioners have advocated for mandatory AI ethics requirements in engineering and CS programs (Griffin 2022), while institutional initiatives at MIT, Brown, and UC Berkeley have developed various models for responsible computing education (Kumar et al. 2024a). However, most existing work lacks systematic evaluation of learning outcomes or examination of implementation challenges within engineering contexts.

7.5.3 Research Contribution

Our approach builds on established action research methodologies that treat curriculum development as systematic inquiry into educational practice (Lewin 1946; Corey 1953). Contemporary frameworks emphasize collaborative, iterative approaches that integrate educator and student perspectives (Rowell et al. 2022; Fine et al. 2018). This research-oriented approach to curriculum design has become particularly important for emerging fields like AI ethics, where evidence-based pedagogical strategies are essential.

Our work addresses identified gaps by applying established educational research methodologies to AI ethics curriculum development, with specific attention to engineering contexts and systematic evaluation of both immediate learning outcomes and professional impact. This approach demonstrates how curriculum design, when conducted with appropriate scholarly rigor, constitutes legitimate educational research that advances both pedagogical knowledge and student learning in critical areas.

7.6 Pedagogical Choices

The integration of these cognitive theories—experiential learning, constructivism, metacognition, scaffolding, transformative learning, and active learning—created a cohesive pedagogical approach for teaching machine learning concepts. Each instructional decision was informed by established theoretical principles and designed to facilitate deep learning.

Where traditional learning and delivery is lecture-based, each lecture was adapted to majority reflection and discussion-based activities. The delivery of this course was primarily non-traditional and explored many active learning techniques. Research indicates that discussion-based approaches yield superior conceptual understanding compared to lecture-based instruction Smith et al. (2011). Team-based learning extends the benefits of discussion by introducing

collaborative problem-solving within structured accountability systems. These approaches align with evidence that collaborative learning enhances both technical skill development and ethical reasoning ability Johnson and Johnson (2018).

Pedagogical rationales transparency connects to metacognitive thinking such as Schön's concept of reflection-on-action (Schön 1992). Socratic Method serves as a primary pedagogical approach, promoting metacognitive awareness and critical examination of assumptions. This method is particularly effective for exploring ethical dimensions of technical choices, as it requires students to articulate and defend their reasoning. The class was encouraged to discuss among their peers. The instructor would serve as a mediator or guiding light to ensure the correct information was being shared.

This theoretically-grounded approach represents a deliberate departure from traditional STEM instruction, which often emphasizes content delivery over active knowledge construction. By basing course design on cognitive learning theories, the curriculum aims to develop not only technical knowledge but also critical thinking, metacognitive awareness, and the ability to navigate ethical dimensions of machine learning applications.

7.7 Course Design

The theory of change in Figure 7.1 is the road map for how the course is designed, and the following sections go into further detail regarding the implementation and execution of this road map. The curriculum materials were developed to address the integration of technical content from Varshney's framework with sociological perspectives. Core materials include:

- Technical sections addressing each dimension of trustworthiness (performance, reliability, interaction, purpose)
- Sociological frameworks for analyzing AI systems, including intersectionality theory and stakeholder analysis methodologies
- Case studies demonstrating the intersection of technical and daily discussions regarding sociological considerations
- Coding exercises with integrated ethical reflection components
- Scaffolded project that guide students through TML pipeline development and ML lifecycle

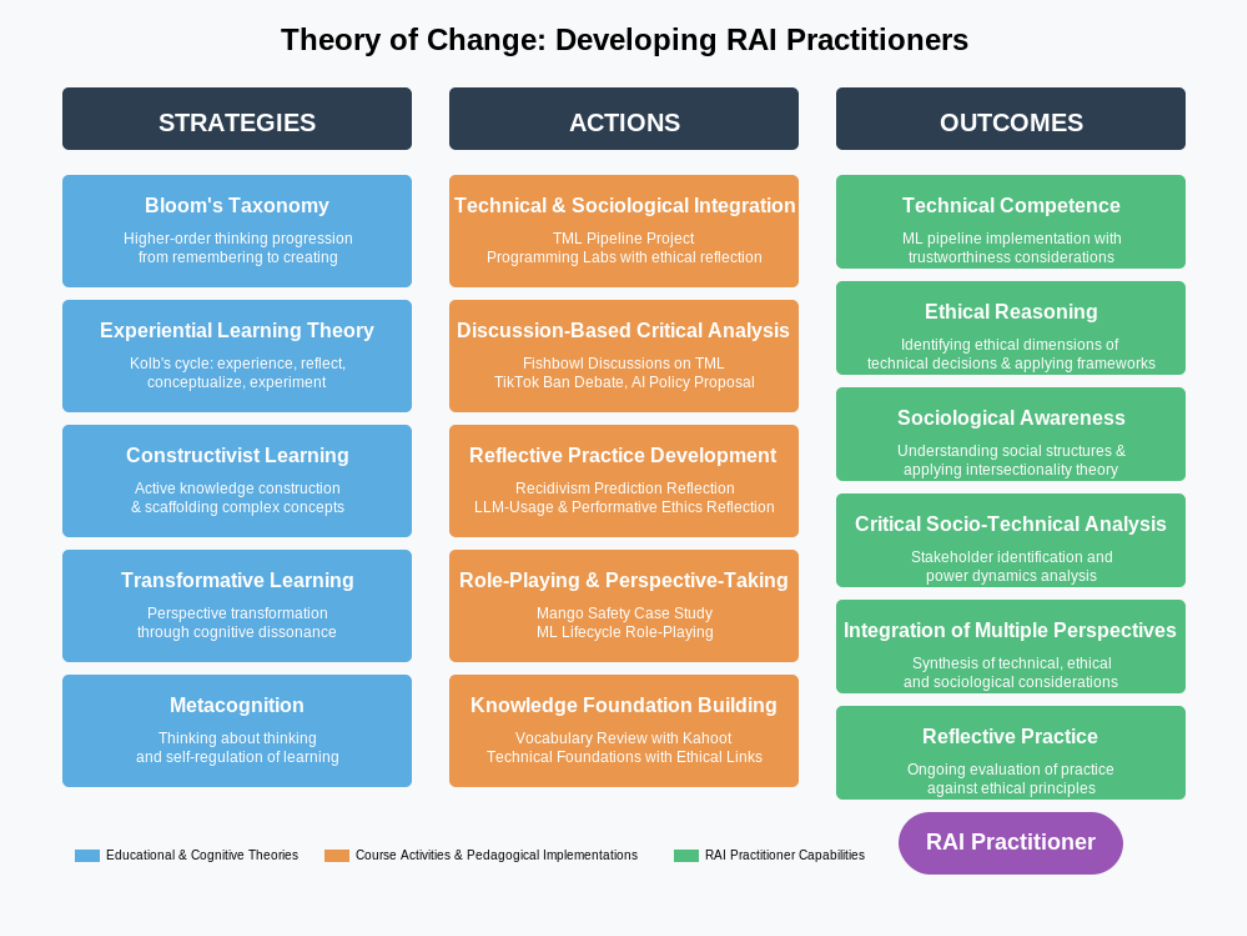


Figure 7.1: This figure visualizes the Theory of Change. The ToC serves as road map for the course design and implementation.

Active Learning Assignments

As a standard part of the course, students participate in daily think-pair-share reflections and written reflections. Each class includes at least 2 5-minute group discussion. In addition to daily in-class discussion, there are 4 other active learning activities. The first active learning activity was a debate about the Tiktok Ban. Another activity was the fishbowl activity focused on the differences between responsible, safe, trustworthy, and fair machine learning, which engaged all students in the class. This class is designed to have every student in meaningful dialogue. The course is designed to ensure every student verbally addresses the entire class.

There is also an in-class group case study reflection titled *The Impact of Data and Model Distribution Shift in Predictive Policing in the U.S.* where students analyzed a short problem statement exploring the harms of predictive policing. In another active learning assignment, the students are provided a vocabulary list with all of the key terms from the course. They have 30 minutes to work with their group and test their collective knowledge by defining each term in their own words. Then, after they are able to reference any course materials to refine and complete their definitions.

Homework Assignments

There are three homework assignments focused on core theoretical considerations. The first assignment involves problem specification and data collection. Students are asked to gather a dataset, describe its components, and present a problem specification demonstrating how the data could be used for responsible and trustworthy machine learning. They are also required to reflect on their own epistemic advantages related to their chosen problem, making this the first scaffolded assignment designed to develop skills in self-reflexivity.

The second homework assignment consists of foundational machine learning and data science labs completed in class. Students work through seven Python and scikit-learn tutorials covering exploratory data analysis (EDA), classifier comparison, clustering algorithm comparison, dimensionality reduction, regression, and training a neural network. The third section of the course, Basic Modeling/Performance, spans four classes and includes paired programming assignments using the driver-navigator technique, with partners switching roles for each lab. Students were required to comment on each line of code to explain its function. Since many students were new to Python, this portion of the course serves as a structured onboarding experience. Homework 2 also serves as the introductory assignment for the final project. Homework 3 is a reflection on an academic research paper regarding performative ethics or a personal reflection on LLM usage in their education.

7.8 Course Structure & Materials

The 15-week course was divided into 6 major sections: Introduction & Preliminaries, Data, Basic Modeling/Performance, Reliability, Interaction, and Purpose. The course outline was adapted by the organization of Kush Varshney's Trustworthy Machine Learning book Varshney (2022). The sections of the book and course coincide with the 4 attributes of trustworthiness. Bloom's Taxonomy was used to develop 22 learning objectives presented in Table 7.1. The 18 TML topics that coincide with the 6 sections of the course, and the assignments used to assess those learning objectives is in Figure 7.2.

The final course grade was divided:

- 20% class participation and attendance
- 25% homework assignments
- 30% final programming project
- 25% final exam

The course's 22 learning objectives were organized across the six levels of Bloom's taxonomy, ranging from foundational recall to advanced synthesis and design. As shown in Figure 7.2, both final projects targeted higher-order cognitive skills (i.e., Analyze, Evaluate, and Create) reflecting the course's emphasis on applied and interdisciplinary learning. The TML pipeline project primarily assessed technical proficiency and decision-making in the machine learning lifecycle, while the AI policy project emphasized ethical reasoning and socio-technical awareness. This alignment between learning objectives and experiential-based and collaborative learning supports the claim that students not only retained content but applied it meaningfully in real-world contexts.

The course was structured around key dimensions of trustworthy machine learning, with each technical component paired with relevant sociological considerations:

Course Introduction & Preliminaries (Weeks 1-3)

- **Technical:** ML fundamentals, data modalilites, data sources, data bias, privacy, consent
- **Sociological:** Intersectionality theory, historical context of algorithmic discrimination, landmark fairness case studies
- **Main Activity:** Tiktok Ban Debate

Basic Performance (Weeks 4–5)

- **Technical:** introduction to python, introduction to data science, supervised learning, decision theory, safety
- **Sociological:**
- **Main Activity:** Python paired-programming labs

Reliability (Weeks 6–8)

- **Technical:** fairness, adversarial robustness, distribution shift, safety, TML Modeling
- **Sociological:** potential harms from algorithmic decision-making, big picture of fair ML, landmark fairness case studies
- **Main Activity:** Project Milestone 1; Case Study: The Impact of Data and Model Distribution Shift in Predictive Policing in the U.S.

Interaction (Week 9)

- **Technical:** interpretability, explainability, transparency
- **Sociological:** AI impact on daily experiences,
- **Main Activity:** In-class vocabulary review, Kahoot

Purpose (Weeks 11–13)

- **Technical:** Ethics Principles, Lived Experience, Social Good
- **Sociological:** Fishbowl discussion on fair, safe, trustworthy, and responsible ML
- **Main Activity:** Project Milestone 2; Fish-bowl Discussion; Kahoot, Project Submission;

Review & Final (Weeks 14–15)

- **Technical:** Review
- **Sociological:** Case Studies
- **Main Activity:** Project Milestone 2; Fish-bowl Discussion; Mango Role-Playing, Kahoot

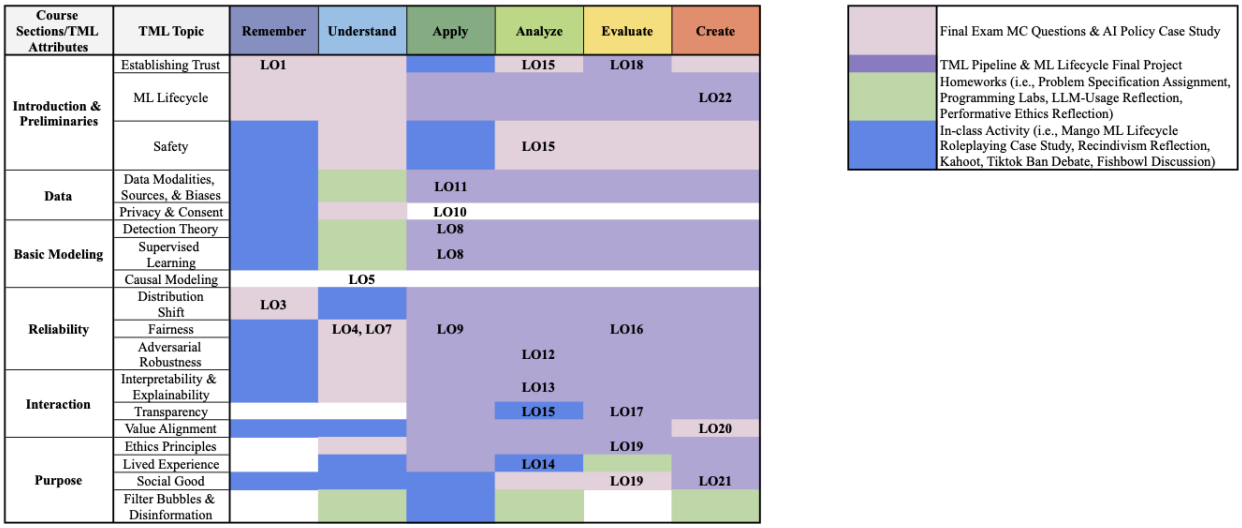


Figure 7.2: This figure visualizes what learning objectives, related Bloom's level, and/or TML topic each assignment and final assessment covered.

7.9 Discussion

The integration of technical and sociological content presents several challenges and opportunities. Primary among these is managing cognitive load while maintaining rigor in both domains. The curriculum addresses this challenge through careful sequencing and explicit connection-making between domains. Another consideration is the potential resistance from students with primarily technical orientations who may view sociological content as peripheral to their professional development. The curriculum counters this perception by demonstrating the practical consequences of sociological blindspots through case studies of real-world AI failures. The approach presented here differs from existing educational interventions in RAI by: (1) Treating sociological content as equally foundational rather than supplementary and (2) Employing evidence-based learning strategies specifically tailored to interdisciplinary integration.

7.9.1 Intersectionality Theory

Sociological theory and intersectionality theory shaped both the design and learning outcome scaffolding to support higher-order and critical thinking skills. The pedagogical choices and active learning activities (positionality statements, LLM reflections, ML role-playing, class discussion) were designed not just for engagement but to foster critical thinking and self-reflexivity. Through positionality statements, reflection on their LLM/GenAI usage, and role-

playing, students were able to help position themselves within sociotechnical systems. While I do present intersectionality theory, the supplementary activities were meant to help students build self-reflexivity and think about their own agency and epistemic advantage (connected to LO6) as AI engineers. This practice of interrogating one's own standpoint and epistemic positioning connects directly to intersectional frameworks, which argue that lived experience informs what we know and how we know it. Specifically, I include a lecture on data sources and bias, and use it as a jumping off point to discuss how intersectional identities can show up in data bias. Then I used the intersecting axes of privilege, domination, and power to further discuss intersectionality theory. I then connected the presence of intersectional data bias to popular fairness case studies (i.e, gender shades, COMPAS, and crime prediction)

7.10 Conclusions & Contributions

This work contributes to the RAI education landscape by addressing Varshney's identified limitation regarding sociological perspectives in trustworthy machine learning. This study employed the Design-Based Research (DBR) methodology to address the research question: (RQ4) How do we explicitly apply Design-Based Research (DBR) to design a course that integrates technical AI and AI ethics skills development? DBR was applied throughout the study as a systematic, iterative process involving the selection of pedagogical approaches, integration of relevant learning theories, and cycles of design, implementation, and evaluation.

The strength of the DBR approach lies in its capacity to bridge theory and practice. It allowed the curriculum to be grounded in educational research while remaining adaptable to the practical realities of classroom instruction. Each phase of the DBR process contributed to this alignment. The initial phase—analysis and exploration—is addressed in the current section. The subsequent course design and development are detailed in the following section, while Chapter 8 presents the final phase, focusing on implementation and evaluation.

Through the application of DBR, this study not only facilitated the development of a curriculum that integrates technical and ethical dimensions of AI education, but also contributed valuable insights into the iterative design of complex, interdisciplinary learning experiences.

Future work should examine longitudinal outcomes of this educational approach, particularly regarding graduates' ability to navigate complex socio-technical problems in professional contexts. Additionally, adaptation of these approaches to different educational contexts, including graduate and professional education, represents a promising direction for further development.

Table 7.1: Learning Objectives Categorized by Bloom's Taxonomy Level

Bloom's Level	Learning Objectives
Remember	• LO1: Define the four attributes of trustworthiness in machine learning systems. • LO2: List the six steps of the machine learning lifecycle. • LO3: Identify the different types of distribution shifts in machine learning systems.
Understand	• LO4: Distinguish between different types of fairness and their appropriate applications. • LO5: Describe the role of causality in machine learning decision-making. • LO6: Explain the relationship between aleatoric and epistemic uncertainty in safety considerations. • LO7: Describe the socio-technical challenges of fairness, accountability, and bias in ML systems.
Apply	• LO8: Evaluate supervised binary classifier performance. • LO9: Apply appropriate bias mitigation strategies based on specific worldviews. • LO10: Implement appropriate privacy preservation methods for different data scenarios. • LO11: Determine ML models with considerations of data sources, bias, privacy, and consent.
Analyze	• LO12: Analyze the implications of different types of adversarial attacks on ML systems. • LO13: Evaluate the effectiveness of different explanation methods for stakeholder personas. • LO14: Examine the relationship between team composition and model outcomes. • LO15: Analyze the stages of the ML lifecycle to promote trust, safety, and transparency.
Evaluate	• LO16: Assess the appropriateness of different fairness metrics for specific contexts. • LO17: Critique the effectiveness of various transparency mechanisms in ML systems. • LO18: Judge the institutional trustworthiness of machine learning platforms. • LO19: Evaluate the ethical implications of ML deployment, including impacts on marginalized populations.
Create	• LO20: Design governance frameworks for ML systems that align with stakeholder values. • LO21: Develop strategies for implementing trustworthy ML in social impact projects. • LO22: Construct appropriate testing protocols for ML models.

CHAPTER

8

TRUSTWORTHY ML CURRICULUM ANALYSIS

8.1 Introduction

The curriculum presented in the previous chapter was piloted as an elective course CSC 495-004 Introduction to Responsible Machine Learning at North Carolina State University in Spring 2025. Over the course of 15-weeks the course students were introduced to RAI and TML content covering foundational machine learning, sociology, ethics, and safety topics. The content delivery was primarily non-traditional with significant project-based and team-based learning as motivated by educational learning theory. We present a survey-based study to measure the impact on perceived content knowledge. We used a paired t-test for quantitative analysis and thematic analysis for qualitative analysis. The results of this study discuss the 6 positive themes and 7 constructive themes in student feedback on the course. Future work will integrate student feedback regarding balancing technical and conceptual learning, more interactive assignments, and improving overall structure and pacing for future iterations.

8.2 Methodology & Context

The CSC 495-004 Introduction to Responsible Machine Learning course was offered in Spring 2025 and had 29 registered students. Two students were completely asynchronous for disability and learning accommodations. There were 7 female students and 22 male students. Feedback was collected through pre-survey, 22 daily written reflections, and post-survey. One the first day of class, the pre-survey was completed in class. The post-survey data was collected on the last day of classes.

Results were analyzed through a two-tailed t-test performed on pre- and post-survey scores, and a thematic analysis on student qualitative feedback from 22 daily reflections and the post-survey reflection. Thematic analysis was selected as the analytical approach due to its flexibility and accessibility as a method for identifying within qualitative data (Braun and Clarke 2006). This method is suitable for exploring participants' experiences, perspectives, and opinions—making and analyzing open-ended student feedback (Nowell et al. 2017). The analysis followed the six-phase process established by Braun and Clarke (2006) and further refined in their subsequent work (Braun and Clarke 2022).

The six phases implemented in this study were:

1. **Familiarization with the data**: All feedback responses were read multiple times to gain thorough understanding of their content and context.

2. **Generating initial codes**: Using an inductive approach, meaningful segments of text were coded to identify potential patterns across the dataset.

3. **Extracting themes**: Codes were organized into potential themes by identifying relationships, similarities, and patterns among them. Following Creswell (2015), this phase required interpretive analysis to move beyond mere categorization toward conceptual understanding.

4. **Reviewing themes**: Potential themes were checked against the coded data and the entire dataset. As recommended by Nowell et al. (2017), this phase involved recursive analysis to ensure themes accurately represented the data and addressed the research questions.

5. **Defining and naming themes**: Each theme was clearly defined and named to capture its essence and relationship to the research questions. Frequency counts were included to indicate theme prevalence.

6. **Producing the report**: The final analysis was compiled, integrating themes with supporting evidence from the data.

Additionally, following Creswell (2015)'s guidelines for qualitative research quality, negative cases were actively sought and analyzed to ensure that the final thematic structure accounted for contradictory or discrepant data. The results are presented using a combination of tabular and narrative formats. As suggested by Braun and Clarke (2006), the frequency of codes within each theme is reported to provide context regarding prevalence, while acknowledging that frequency alone does not determine thematic significance. For each theme, representative quotes are provided as evidence, following Nowell et al. (2017)'s recommendation to establish a clear chain of evidence between raw data and interpreted themes.

The results of this research aim to answer: **(RQ5) How did the course impact student confidence in each of its learning objectives? (RQ6) After the course, how do students perceive the usefulness of the course and what suggestions do they have for course improvement?**

8.3 Learning Objectives

From the full 22 learning objectives in Table 7.1, the CSC 495-004 Introduction to Responsible Machine Learning syllabus explicitly emphasized:

1. LO1: Define the four attributes of trustworthiness in machine learning systems
2. LO2: List the six steps of the machine learning lifecycle
3. LO7: Describe the socio-technical challenges of fairness, accountability, and bias in machine learning applications
4. LO11: Determine ML models with considerations of data sources, bias, privacy, and consent
5. LO19: Evaluate the ethical implications of ML deployment in real-world scenarios

The full list of learning objectives and how they were assessed is visually represented in Figure 7.2.

8.4 Survey Instruments

There was a pre and post-survey covering the 22 learning objectives described in the previous chapter (see Table 7.1). There were 22 additional Likert scale questions gauging the students' confidence on each learning objective. The pre-survey was distributed on the first day of class before any course content was introduced. The post-survey was distributed as the final exam.

Post-survey content knowledge questions were adjusted based on pre-survey scores. The pre-survey assessed a general overview of RAI and TML content. After reviewing the student responses to the pre-survey and analyzing the overall engagement over 12 weeks, the post-survey was adjusted to better represent the key course objectives and topics. The pre-survey was deemed too challenging and too lengthy for students to complete accurately, given the time allotted. The pre-survey submission time stamps showed that after the multiple-choice sections, most students didn't complete or attempt the remaining sections. The pre-survey covered at least 120 minutes of content, and on average was submitted after 20 minutes. For the post-survey, the two main sections, multiple choice questions, and AI Policy proposal, were broken up across the three 75-minute class periods.

Sections 1 and 2 contained multiple-choice questions regarding the content of the course. The 9 multiple-choice questions from the pre-survey were extended with 41 new multiple choice questions in the post-survey remained the same to assess more topics. Sections 3-6 from the pre-survey representing Apply, Analyze, Evaluate, and Apply elements of Bloom's Taxonomy initially had several short and long answer questions, and different case studies to select from. Similarly, for the post-survey, the students participated in a case study assignment where they selected the scenario and invested stakeholder. For more details, please see the full-length pre- and post-surveys in Appendices C and D.

8.5 Pre and Post Survey Learning Perceptions

The pre and post-surveys directly evaluated for all learning objectives. We can say that students did achieve more than the 5 learning objectives formally selected in the course syllabus. The students' self-efficacy regarding their confidence levels to perform each learning objective, and Figure 8.1 shows the difference in pre and post survey scores. All learning objectives have p-values of less than 0.0005, much smaller than the common 0.05 threshold, indicating high statistical evidence that the course was the contributing factor. This result supports the effectiveness of the course on learning objectives as perceived by the students.

8.6 General Insights from 22 Daily Reflections

The qualitative data was collected at the end of each class via a daily reflection form via Google Forms:

1. State your new word of the day. (required)
2. What's one thing you learned today? (required)

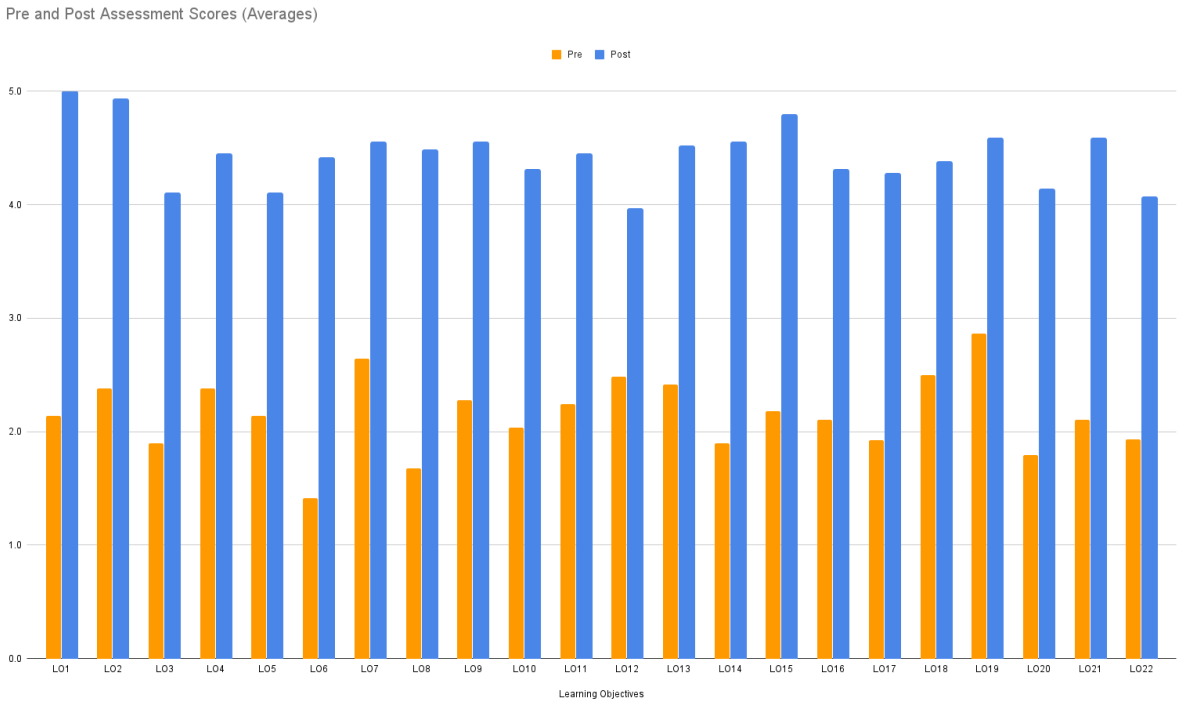


Figure 8.1: Bar graph of the average scores for the entire class for pre and post-survey of learning perception.

3. Did you feel prepared for today's class/lecture? Why or why not? (optional)

4. What would help make today's lecture/class more effective? (optional)

5. Questions/Concerns/Insights/Feedback (optional)

The daily reflections were reviewed by the instructor after each class. Students would provide information on how to make lessons more effective or discuss what they enjoyed in the course. Direct quotations from the reflections are summarized in Table 8.1.

A prominent theme in the reflections was the benefit of interactive learning strategies, particularly those that emphasized open discussion, peer engagement, and the application of concepts to real-world scenarios. One student remarked, *"Very good class format, would like to keep the open discussion portions as they are more stimulating than most other classes I've taken at State."* This comment, along with others, supports the value of group-based exercises and participatory learning environments in promoting student engagement and conceptual retention.

Students consistently expressed appreciation for the transparency in course structure and instructional decisions. Clear articulation of learning objectives, visible organization within lecture materials, and the instructor's openness about course development contributed to an enhanced sense of clarity and motivation. For example, one student noted, *"Reinforcing what your learning objectives are, e.g. keeping a 'section' title at the top of your slides. Keeps me on track,"* while another reflected, *"There's SO much that goes into teaching a course like this that students never see. I feel like we should be educated about our education more often."* These sentiments suggest that students benefit not only from well-defined goals but also from understanding the pedagogical rationale behind instructional choices. This feedback affirms the pedagogical value of authentic, application-oriented tasks and supports broader findings in the literature on experiential learning in technical education.

8.7 Thematic Analysis of Student Feedback

The previous section informally discusses recurring and specific insights from the entire semester. However, thematic analysis was applied for a more rigorous analysis of student experiences on the effectiveness of the course. On the last day of class, all 29 students provided written positive and constructive feedback for future iterations. The qualitative data was collected at the end of the last class via a reflection form via Google Forms. Thematic analysis specifically analyzed:

1. For the next interaction of this course, what did you enjoy?

Table 8.1: Student Daily Reflections: Activities and Feedback

Course Element	Student Feedback
Class Format & Discussions	"Very good class format, would like to keep the open discussion portions as they are more stimulating than most other classes I've taken at State." "I really like the time we take in class to talk with our peers, that gives us a chance to use the vocabulary and talk about the interesting things in class." "I liked it a lot! Working together really makes the content memorable." "I enjoyed the structure of today's class. I liked hearing the instructor's side in answering the reflection questions as well as having feedback about our projects. Then it turned into discussion-based learning and analyzing which kept my attention." "I felt prepared; most of the lecture was interactive with both the instructor and other students in class so it helped me understand the content better."
Learning Objectives & Structure	"Reinforcing what your learning objectives are, e.g. keeping a 'section' title at the top of your slides. Keeps me on track." "I enjoyed the structure of today's class. I liked hearing the instructor's side in answering the reflection questions as well as having feedback about our projects. Then it turned into discussion-based learning and analyzing which kept my attention." "Feel like a lot of what I'm learning is surface level and lacks context, not sure what I am meant to grasp/understand/takeaway; I guess kind of like objectives."
Examples & Applications	"I prefer examples over just definitions. I find it easier when I have an example of something I'm trying to learn." "I really like the real-world examples that you provide, it helps make the content more digestible when I can connect it to something else." "Yes, I was able to connect everything I learned in the lab and I was surprised by the connection with supervised learning and the labs we did. I had discussed it in my HW 1 so I found it more interesting and I think we can use it for the project as well."
Fishbowl Discussion	"I liked the idea of having discussions that people can view. Even though the pressure is high, it shows a lot of thoughts without being hard to keep track."
ML Lifecycle Mango Role-Playing	"Somewhat. I knew the 6 stages of ML lifecycle but applying it to a real-world example can be complex." "Yes?, I had an understanding of what each role had, but there were much more that goes into each role."
TikTok Ban Debate	"I enjoyed the in-class debate. It's a really good idea for classroom peer-to-peer engagement and exploring real-world scenarios."
Group Vocabulary Exercise	"I really liked the exercise of working in groups using no resources, talking about vocabulary. It really got my brain working, I could actually feel it lol."
AI Discussions & Policy Case Study	"I really found the discussion about the use of AI and how to improve our understanding of the terminology super open, honest, and effective. I just wanted to say that I appreciate you taking the time out of class to find real solutions to this issue and caring about us truly learning about what you are teaching. Thank you :D" "I felt prepared yes, you told us about the case study beforehand and we were properly prepared for it." "Yeah! I feel like the entire course geared us to this point, the case study does make sense as a final project/assignment."
Teaching Transparency	"There's SO much that goes into teaching a course like this that students never see. I feel like we should be educated about our education more often. It makes me respect you as a teacher and your methods more."
Interest in Advanced Content	"I really like the introduction of all of the new material, I find this very interesting topic. I would be interested in taking a more advanced version of this course if offered."

2. For the next interaction of this course, what would you change?

The following two subsections present a thematic analysis of positive and constructive student feedback by topic, respectively. Each section contains the extracted theme, connected codes, code frequency, and sample quotations.

Positive Student Feedback on Future Course Iterations

The thematic analysis revealed six key positive themes in student feedback. The positive themes with their codes are summarized in Table 8.2, and the themes with related quotes in summarized in Table 8.3. The most prominent, *Interactive Learning Methods* (15 mentions), highlights students' strong preference for discussions, Kahoot activities, and group engagement. This aligns with research demonstrating the effectiveness of active learning in STEM education (Freeman et al. 2014).

Practical Application of Concepts (12 mentions) emerged as the second most frequent theme, with students valuing hands-on labs, projects, and case studies that connected theory to practice. This preference for experiential learning supports research on the importance of authentic learning experiences in technical education (Kolb and Kolb 2005).

Students appreciated the *Manageable Workload and Learning Environment* (9 mentions), suggesting the balanced approach facilitated deeper engagement rather than surface learning. The comfortable atmosphere and reduced stress levels align with findings that supportive learning environments enhance motivation and performance (Ambrose et al. 2010).

Teaching Style and Approach (7 mentions) and *Focus on Reflection and Critical Thinking* (6 mentions) demonstrate student appreciation for clear explanations, concept reinforcement, and opportunities for thoughtful analysis. These metacognitive practices enhance knowledge transfer and retention (Bjork et al. 2013).

Finally, *Relevance to Real-World/Industry* (5 mentions) indicates students valued curriculum elements connecting classroom learning to workplace applications, potentially increasing motivation through perceived relevance.

These findings suggest maintaining and potentially expanding: (1) interactive learning approaches, (2) applied learning opportunities, (3) the balanced workload and supportive environment, (4) clear teaching with multiple concept exposures, (5) reflection opportunities, and (6) industry connections. The predominance of interactive and applied learning in student feedback suggests successful implementation of student-centered pedagogical approaches. The design choices for the course and motivating technical, sociological, and educational theory are discussed in the previous chapter.

Table 8.2: Thematic Analysis of Positive Student Feedback

Theme	Codes	Frequency
Interactive Learning Methods	Class discussions (8); Kahoot activities (5); Group engagement (6); Question-based learning (3); Peer interaction (4)	15
Practical Application of Concepts	Hands-on labs (5); Real-world application (6); Project-based learning (7); Case studies (4); Skill application (3)	12
Manageable Workload and Learning Environment	Appropriate workload (4); Comfortable atmosphere (5); Reduced stress (3); Sufficient time for assignments (2); Positive learning climate (4)	9
Teaching Style and Approach	Clear explanations (3); Concept reinforcement (4); Approachable instructor (3); Effective teaching techniques (2); Impactful assignments (2)	7
Focus on Reflection and Critical Thinking	Individual reflection (3); Critical thinking (4); Conceptual understanding (2); Thoughtful analysis (2); Deep learning (1)	6
Relevance to Real-World/Industry	Workplace relevance (3); Industry applications (3); Real-world scenarios (3); Career preparation (1); Practical skills (2)	5

Table 8.3: Detailed Thematic Analysis of Positive Student Feedback with Supporting Evidence

Theme	Freq.	Supporting Evidence Quotes
Interactive Learning Methods	15	"Socratic method!! Discussion. I recognize I'm more open to this than most students, but being politely corrected and asking questions without fear of judgment is HOW I learn and remember things." "I enjoyed the discussions and hearing out different perspectives." "I enjoyed our group discussions and how we would all engage with each other during classes." "The kahoots was great. I enjoyed the discussions about AI, ML, and LLM."
Practical Application of Concepts	12	"The in-person coding labs where we worked with different models like AdaBoost and classifiers was enjoyable. It gave me some hands-on experience with ML development." "The semester project was great, and the case study was the best part. Like someone in class said, the case study made us feel very fulfilled, and the semester project was a good way to make sure we actually understood what we were doing." "I enjoyed the team project as it allowed us to use the knowledge learned in class and apply it to a real world scenario while incorporating some programming." "I liked the focus on practical end to end pipeline development. It's a great primer for someone who is interested in machine learning but doesn't know where to start."
Manageable Workload and Learning Environment	9	"I liked the layout of the course: the hw's and the mc / case study for the final exam a lot. It was one of the few courses where I didn't feel rushed, or like I didn't have enough time to do the work. The load was amazingly perfect!" "I also enjoyed the atmosphere of the class, it was comforting and friendly." "Not too much work, conceptual learning, chill professor"
Teaching Style and Approach	7	"Explains concepts in simple terms" "Covering crucial concepts multiple times" "I really enjoyed how chill and straight up you were with us. I felt comfortable and thoroughly enjoyed this class."
Focus on Reflection and Critical Thinking	6	"I enjoyed how we had to individually reflect and respond to our homeworks. I think that through using and reflecting on what we have learned we were able to learn and instill what we learned a lot easier." "Enjoyed the discussion we had in class that got me to think abt the topic"
Relevance to Real-World/Industry	5	"I enjoyed learning about what the workplace was actually like and making the machine learning model this semester." "I enjoyed learning about making fair machine learning models and how flows are slightly different than described in other courses."

8.7.1 Constructive Student Feedback on Future Course Iterations

Seven constructive feedback themes emerged offering direction for course improvements. The constructive themes with their codes are summarized in Table 8.4, and the themes with related quotes in summarized in Table 8.5. The primary theme, *Desire for More Structure and Reinforcement* (7 mentions), indicates students would benefit from additional scaffolding and more frequent, low-stakes consolidation activities. This aligns with research on the testing effect for enhancing retention (Roediger and Karpicke 2006).

The second theme, *Balancing Coding and Conceptual Learning* (6 mentions), revealed varied perspectives on technical content, reflecting heterogeneous student backgrounds. This suggests the need for differentiated instruction or more programming prerequisite recommendations. One of the core design challenges was the balance of technical and theoretical considerations. Where some students preferred less programming, others preferred more. The varied perspectives can suggest a balance was achieved, but more work is needed to determine the best balance for undergraduates. For example, NC State's computer science program is primarily instructed in Java, and this course could be better suited for a computer science program that primarily teaches Python programming.

The third and fourth themes, *Preference for Applied Learning* (5 mentions) and *Enhanced Interactive Learning* (5 mentions), reinforce positive feedback while recommending specific enhancements: more case studies, guided coding demonstrations, and earlier implementation of interactive activities. These suggestions align with scaffolded learning theory (Vygotsky 1978). Both emergent themes support students' positive reception of the curriculum materials, and suggest ways to adjust based on personal preferences.

The fifth theme, *Technical Preparation* (4 mentions), highlights student concerns about the technical preparation needed before entering the course, particularly regarding Python programming skills and foundational machine learning concepts. This theme differs from the second *Balancing Coding and Conceptual Learning* theme in that it specifically addresses pre-course preparation rather than the balance of content during the course itself. The feedback indicates a tension between maintaining course accessibility and ensuring all students have a sufficient technical foundation to engage confidently with the material. More work is needed to understand how large the gap is between foundation CS topics and the technical skillset needed for an introduction to advanced applied topics like RAI and TML.

The sixth theme, *Curriculum Pacing and Content Distribution* (4 mentions), highlighted concerns about initial content density, suggesting cognitive overload early in the course. The recommendation for more project implementation time indicates student desire for deeper engagement with applied components. As mentioned in the pre-survey section, the initial direction of the course was more advanced and required scaling down during the semester. This

theme reflects that students' observed the instructional differences in the early and later course content, and more materials are needed in the early weeks to create a stronger foundation and address knowledge gaps.

The final seventh theme, *Interest in Advanced Topics* (3 mentions), revealed student curiosity about cutting-edge applications (LLMs, adversarial attacks), suggesting opportunity for optional advanced content for interested students. An unanticipated but productive outcome of the course was the integration of content related to large language models (LLMs), prompted by increased student interest in generative AI. Although not originally part of the syllabus, these discussions—particularly those situated within the context of responsible AI usage in education—were perceived as both timely and impactful. One student reflected, *"I really found the discussion about the use of AI and how to improve our understanding of the terminology super open, honest, and effective... I appreciate you taking the time out of class to find real solutions to this issue and caring about us truly learning about what you are teaching."* The ability to integrate emergent topics into the curriculum in response to student curiosity demonstrates a pedagogical flexibility that is increasingly important in rapidly evolving fields such as artificial intelligence. It also motivates future work to consider how to extend the Trustworthy Machine Learning (Varshney 2022) technical content beyond supervised learning.

These themes suggest seven potential areas for improvements: (1) enhanced curriculum structure with small reinforcement activities, (2) addressing technical background diversity, (3) increasing application-based learning, (4) earlier introduction of interactive elements, (5) adaptation for different technical proficiency to maintain accessibility, (6) adjusted content pacing, and (5) optional advanced topics. Many constructive suggestions reinforce elements from positive feedback, emphasizing areas where students perceive particular value and desire enhancement.

Table 8.4: Thematic Analysis of Constructive Student Feedback

Theme	Codes	Frequency
Desire for More Structure and Reinforcement	Curriculum stability (2); Additional reinforcement activities (4); Structured discussions (3); Foundational concept emphasis (2); Consistent format (2)	7
Balancing Coding and Conceptual Learning	Coding difficulty (3); Technical vs. theoretical balance (4); Programming prerequisites (2); Code examples (2); Coding support (1)	6
Preference for Applied Learning	Case study emphasis (3); Active learning (3); Guided coding (2); Application-focused teaching (3); Hands-on demonstration (1)	5
Enhanced Interactive Learning	More interactive activities (3); Collaborative learning (2); Early Kahoot implementation (1); Lab discussion time (2); Mind mapping (1)	5
Technical Preparation	Python proficiency concerns (3); Prerequisite recommendations (2); Knowledge gap identification (2); Technical background diversity (2); Entry-level support needs (1); Prior knowledge assumptions (1)	5
Curriculum Pacing and Content Distribution	Gradual concept introduction (2); Project timeline adjustment (1); Knowledge gap concerns (2); Content pacing (2); Introductory material (1)	4
Interest in Advanced Topics	LLM applications (1); Model fine-tuning (1); Security concepts (1); Policy components (1); Vocabulary relationships (1)	3

Table 8.5: Detailed Thematic Analysis of Constructive Student Feedback with Supporting Evidence

Theme	Freq.	Supporting Evidence Quotes
Desire for More Structure and Reinforcement	7	"Stable structure in curriculum. Course understandably had to go through changes."
		"Maybe add small assignments that reinforces what was learned in class. This will encourage students to go back, look at notes or the lecture video, and possibly help clarify things that may have been overlooked."
		"I think the discussions and lectures would benefit from a little more structure."
Balancing Coding and Conceptual Learning	6	"Less coding, more conceptual learning since coding ML pipelines is a very different game from learning the theory behind it and other ML algorithms."
		"I would say that if this class is okay for beginners in python, to not make it too crazy because I didn't understand most of the python work and I wish I knew how to do it before joining."
		"I would add Python as a prerequisite if you would like to make it more technical and put a lot of what we learned into deeper practice but this is an introductory class so I'm not sure if you would like that."
Preference for Applied Learning	5	"I would take the direction that the second time of the semester took. There were more case studies, and activities that required us to think instead of the normal reading off the slides lecturing that we are used to."
		"It's hard to say exactly, but probably the way the course was formatted in the first half. I didn't feel myself learning as much compared to the second half where we did a lot more application."
Enhanced Interactive Learning	5	"I would have more interactive things like Kahoots from the start because those really helped me learn the concepts more."
		"I think I would give a little more time for labs and time to discuss more about what is learned after each one. I also think one lab should maybe we completed together as a class with some way for the students to engage throughout."
Technical Preparation	5	"A knowledge gap at the beginning of the semester was apparent. I think either a data science / intro to ml pre-req would be beneficial but would make this class less accessible so maybe a longer intro to data science / ml chapter at the beginning."
Curriculum Pacing and Content Distribution	4	"Less heavy content to start, it was a lot at once. Introduce concepts more gradually, and weave in practice / questions / exercises / hw's related to them."
		"The group project could use just a little bit more time/give the first parts a bit less time so there's more time for implementation."
Interest in Advanced Topics	3	"Maybe include more in-depth exploration of LLM applications / fine-tuning models for tasks. Discussion of adversarial attacks and security could be interesting topics for understanding safety and robust ML models."
		"I would make a policy proposal type project as well."

8.8 Discussion

Student feedback throughout the semester offered valuable insight into the effectiveness of the course's instructional design and pedagogical approach. The course appears to have fostered sustained interest in the subject matter. Multiple students expressed a desire to pursue more advanced study in machine learning and AI ethics, with one stating, *"I would be interested in taking a more advanced version of this course if offered."* This suggests that the course successfully cultivated not only foundational knowledge but also intellectual curiosity and a sense of relevance regarding the broader societal implications of AI.

This course offers a pedagogical model for introductory AI and machine learning education that is transparent, flexible, and grounded in discussion-based, example-driven learning. The qualitative feedback received highlights the strengths of such an approach while also identifying actionable areas for refinement. The findings presented here contribute to ongoing discourse in AI education, particularly with respect to designing inclusive, engaging, and ethically attuned curricula that respond to both technological developments and student experience.

8.9 Final Remarks from Students

5 students provided optional feedback on the course included here:

> "Overall, I enjoyed this class and the pacing / structure was well-balanced."
>
> "Thank you so much for this whole course, Lauren. You were a great instructor and the class was super enjoyable! Good luck with your dissertation!"
>
> "Thank you for a wonderful class! [emote]"
>
> "This class was a great experience despite the challenges I had in the first half. I found this course giving me determination to learn more and have faith in education :)"
>
> "Also thank you for a really amazing class, as a second semester senior this was one of my favorite classes i've taken here and that due to your engaging teaching and clear passion to for the subject. I am super grateful I got to have you as a teacher and that you have helped me learned more about Trustworthy ML. I was not knowledgeable about this subject before this class and you have made me incredibly curious and interested in this field. I hope you enjoyed teaching this class as much as I am sure many of us enjoyed taking it! :)"

8.10 Conclusions & Contributions

This study provides evidence-based insights into student learning in undergraduate machine learning (ML) contexts, particularly within a discussion-driven, example-rich instructional model. The qualitative reflections demonstrate that students respond positively to learning environments characterized by transparency, real-world application, and peer interaction. The emphasis on open discussion, vocabulary usage, and case-based learning helped to reinforce understanding and stimulate critical engagement. To answer **RQ5**, future implementations should incorporate flexible pacing, structured support for varied skill levels, and space for emerging topics to maintain relevance and engagement. Furthermore, the integration of emergent topics—such as large language models (LLMs) and ethical AI use—enhanced perceived relevance and highlighted the value of a flexible, responsive curriculum design in technical education.

In parallel, the study identifies key challenges in ML instruction and offers practical recommendations for future curriculum development. Variability in students' technical preparedness, difficulty connecting abstract concepts to practical outcomes, and cognitive overload in the early stages of the course were recurring themes. To answer **RQ6**, these findings suggest a need for scaffolded entry points, more structured formative surveys, and frequent reinforcement of learning objectives. Taken together, these insights support a learner-centered, adaptive approach to ML instruction that can be applied to a range of interdisciplinary computing courses. By foregrounding student voice and responsiveness, this work contributes to the design of inclusive, effective, and TML and RAI curricula. Future research could look at more student artifacts collected during implementation to understand the tradeoffs in teaching both technical and sociology skills in the same class.

CHAPTER

9

CONTRIBUTIONS & CONCLUSIONS

9.1 Summary of Results

This dissertation brings together three primary domains of research: education, ethics, and safety. This work employs methodological diversity to address the inherently interdisciplinary nature of responsible AI. The strength of this dissertation lies in its methodological triangulation, which reveals insights that would remain hidden through any single disciplinary approach. The research explores three primary avenues to improve machine learning practitioners' knowledge regarding harmful ML outcomes:

1. Early education on artificial intelligence and ethics
2. Critical engagement with ML and ethics in project development
3. Trustworthy ML content knowledge and pipeline implementation

Through DBR's participatory and iterative nature, this work contributes to both immediate pedagogical needs and broader scholarly understanding of how to effectively bridge technical AI education with critical social justice perspectives and ethical considerations. The dissertation is organized around six core research questions across six substantive chapters:

9.1.1 Trustworthy AI Systems in Adversarial Environments

Chapters 3 & 4

- How to create trustworthy AI systems in socio-technical environments, particularly when facing potential malicious deception? (RQ1)
- Presents complementary studies examining approaches with multiple trustworthy elements against adversarial models

9.1.2 Abolitionist Frameworks for AI Development

Chapter 5

- How practitioners can meaningfully engage with abolitionist framing in AI development? (RQ2)
- Provides a critical framework for ethical engagement that challenges traditional approaches

9.1.3 CSF: A Socially Relevant Focused AI Curriculum Designed for Female High School Students

Chapter 6

- How does an AI/ML curriculum with integrated social relevance and ethics considerations impact girls' confidence and self-efficacy in AI/ML? (RQ3)
- Provides the results of a survey study during a camp implementation of a AI and ML module designed to integrate ethics and teach novice high school students advanced AI topics

9.1.4 Trustworthy ML Curriculum Development

Chapter 7

- How do we explicitly apply Design Based Research (DBR) to design a course that integrates technical AI and AI ethics skills development? (RQ4)
- Synthesizes findings into a comprehensive theory of action for the design, scaffolding, and instruction of RAI and TML content knowledge

9.1.5 TML Curriculum Analysis

Chapter 8

- How did the course impact student confidence in each of its learning objectives? (RQ5)
- After the course, how do students perceive the usefulness of the course and what suggestions do they have for course improvement? (RQ6)
- Explores the impact of pedagogical approaches for developing technical competence, prioritizing ethical and sociological awareness in undergraduate students

9.2 Contributions

1. An abolitionist framework for critical engagement in AI ethics that challenges existing paradigms
2. A comprehensive framework for trustworthy AI implementation development that addresses both technical and sociological concerns
3. An evidence-based curriculum for teaching RAI concepts to undergraduate students early in their academic careers

9.3 Conclusions

1. RAI and TML implementation is inherently challenging and requires varying levels of AI ethics and sociological considerations that are difficult to teach to active CS practitioners
2. Students can successfully produce end-to-end machine learning development with RAI considerations at each step in the lifecycle without extensive technical backgrounds

This work argues that as the field continues to grow, more research is needed to determine the most effective pedagogical methods to educate future AI practitioners on the complexities and ethical challenges in modern AI research for practical industry applications.

REFERENCES

Abebe, R., Aruleba, K., Birhane, A., Kingsley, S., Obaido, G., Remy, S. L., and Sadagopan, S. (2021). Narratives and counternarratives on data sharing in africa. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 329–341.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., and Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54:95–122.

Agarwal, P. K., Har-Peled, S., Varadarajan, K. R., et al. (2005). Geometric approximation via coresets. *Combinatorial and computational geometry*, 52(1):1–30.

Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. *SIGMOD Record*, 29(2):439–450.

Aguilar, M. (2022). Why is it important to have women in tech: Ironhack blog.

Ahmed, S. (2004). Declarations of whiteness: The non-performativity of anti-racism. *borderlands*, 3(2).

Akgun, S. and Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in k-12 settings. *AI and Ethics*, 2(3):431–440.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805.

Alkhatib, Y. and et al. (2023). Cega: Characteristic explanatory general association rules. In *Proc. AIES*.

Alvarez, L. and Menzies, T. (2023). Don't lie to me: Avoiding malicious explanations with stealth. *IEEE Software*, 40(3):43–53.

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., and Norman, M. K. (2010). *How learning works: Seven research-based principles for smart teaching*. John Wiley & Sons.

Anderson, L. W. and Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

Andrews, R., Diederich, J., and Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based systems*, 8(6):373–389.

Angwin, J., Larson, J., Kirchner, L., and Mattu, S. (2016). Machine bias.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2022). Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications.

Arrieta, A. B., Díaz-Rodríguez, N., and et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges. *Information Fusion*, 58:82–115.

Augasta, M. G. and Kathirvalavakumar, T. (2012). Rule extraction from neural networks—a comparative study. In *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*, pages 404–408. IEEE.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831.

Bandura, A. (1977). *Social Learning Theory*. Prentice Hall, Englewood Cliffs, NJ.

Banovic, N., Yang, Z., Ramesh, A., and Liu, A. (2023). Being trustworthy is not enough: How untrustworthy artificial intelligence (ai) can deceive the end-users and gain their trust. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–17.

Barocas, S. and Selbst, A. D. (2014). Big data's disparate impact. Technical report, SSRN. Revised & published in *Cal. L. Rev.* 104(3):671–732, 2016.

Barocas, S. and Selbst, A. D. (2016a). Big data's disparate impact. *California law review*, pages 671–732.

Barocas, S. and Selbst, A. D. (2016b). Big data's disparate impact. *California Law Review*, 104(3):671–732.

Bastani, O., Kim, C., and Bastani, H. (2017). Interpretability via model extraction. *arXiv preprint arXiv:1706.09773*.

Baumer, B. S., Garcia, R. L., Kim, A. Y., Kinnaird, K. M., and Ott, M. Q. (2022). Integrating data science ethics into an undergraduate major: A case study. *Journal of Statistics and Data Science Education*, 30(1):15–28.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ..., and Zhang, Y. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Benjamin, R. (2020). Race after technology: Abolitionist tools for the new jim code.

Benjamin, R. (2023). *Race after Technology*. Routledge.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413.

Binns, R. and Kirkham, R. (2021). How could equality and data protection law shape ai fairness for people with disabilities? *ACM Transactions on Accessible Computing (TACCESS)*, 14(3):1–32.

Birch, K. and Bronson, K. (2022). Big tech.

Birhane, A., Ruane, E., Laurent, T., S. Brown, M., Flowers, J., Ventresque, A., and L. Dancy, C. (2022). The forgotten margins of ai ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 948–958, New York, NY, USA. Association for Computing Machinery.

Bjork, R. A., Dunlosky, J., and Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64:417–444.

Bond, M. et al. (2024). The international journal of educational technology in higher education: Content and authorship analysis 2010–2024. *International Journal of Educational Technology in Higher Education*, 21.

Bonilla-Silva, E. (2022). *Racism without racists: Color-blind racism and the persistence of racial inequality in America*. Rowman & Littlefield, an imprint of The Rowman & Littlefield Publishing Group, Inc.

Bottomley, L. (2015). Assessing the grit of incoming engineering students. In *2015 ASEE Annual Conference & Exposition*, pages 26–248.

Bouchrika, I. (2022). 75 u.s. college statistics: 2023 facts, data & trends.

Boyer, E. L. (1990). *Scholarship Reconsidered: Priorities of the Professoriate*. The Carnegie Foundation for the Advancement of Teaching, Princeton, NJ.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Braun, V. and Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative psychology*, 9(1):3.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. 20,000+ citations.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

Broll, B., Lédeczi, A., Stein, G., Jean, D., Brady, C., Grover, S., Catete, V., and Barnes, T. (2021). Removing the walls around visual educational programming environments. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–9. IEEE.

Broll, B., Lédeczi, A., Volgyesi, P., Sallai, J., Maroti, M., Carrillo, A., Weeden-Wright, S. L., Vanags, C., Swartz, J. D., and Lu, M. (2017). A visual programming environment for learning distributed programming. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, pages 81–86.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences*, 2(2):141–178.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review*, 31(1):21–32.

Bucilă, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. page 535–541.

Bufacchi, V. (2005). Two concepts of violence. *Political studies review*, 3(2):193–204.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Buyl, M. and De Bie, T. (2022). Inherent limitations of ai fairness. *arXiv preprint arXiv:2212.06495*.

Caldera, C. (2020). Fact check: Rates of white-on-white and black-on-black crime are similar.

Calmon, F. P., Wei, D., Mohammadi, K. P., Ramamurthy, K. N., and Varshney, K. R. (2015). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, volume 28.

Carey, A. N. and Wu, X. (2022). The fairness field guide: Perspectives from social and formal sciences. *arXiv preprint arXiv:2201.05216*.

Carey, A. N. and Wu, X. (2023). The statistical fairness field guide: perspectives from social and formal sciences. *AI and Ethics*, 3(1):1–23.

Carr, L. G. (1997). *"Colorblind" Racism*. Sage.

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832.

Carvalho, P. and et al. (2019). On the complexity of interpretable machine learning: A survey. *ACM Computing Surveys*.

Cateté, V., Alvarez, L., Isvik, A., Milliken, A., Hill, M., and Barnes, T. (2020). Aligning theory and practice in teacher professional development for computer science. In *Koli Calling'20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research*, pages 1–11.

Ceci, S. J., Ginther, D. K., Kahn, S., and Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological science in the public interest*, 15(3):75–141.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307.

Chai, J., Jang, T., and Wang, X. (2022). Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35:19152–19164.

Chakrabarti, K. and Mehrotra, S. (2000). Local dimensionality reduction: A new approach to indexing high dimensional spaces. In *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB)*, pages 89–100, Cairo, Egypt.

Chakraborty, J., Majumder, S., and Menzies, T. (2021). Bias in machine learning software: why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28.

Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. The MIT Press.

Chapman, S. and Vivian, R. (2017). Engaging the future of stem: A study of international best practice for promoting the participation of young people, particularly girls, in science, technology, engineering and maths (stem). *Sydney: Chief Executive Women*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, J., Nair, V., Krishna, R., and Menzies, T. (2018). "sampling" as a baseline optimizer for search-based software engineering. *IEEE Transactions on Software Engineering*, 45(6):597–614.

Chen, K. (2009). On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947.

Chen, Y. and Chen, J. (2009). Improved approximation algorithms for k-means and k-median clustering. *SIAM Journal on Computing*, 39(4):2148–2178.

Chen, Z., Zhang, J., Sarro, F., and Harman, M. (2022). Maat: A novel ensemble approach to addressing fairness and performance bugs for machine learning software. In *The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*.

Chin, C. and Robison, M. (2020). How ai bots and voice assistants reinforce gender bias. *Center for Technology Innovation. https://www. brookings. edu/research/how-ai-bots-and-voice-assistants-reinforce-gender-bias*.

Clear, A., Parrish, A., Impagliazzo, J., Wang, P., Ciancarini, P., Cuadros-Vargas, E., Frezza, S., Gal-Ezer, J., Pears, A., Takada, S., Topi, H., van der Veer, G., Vichare, A., Waguespack, L., and Zhang, M. (2020). Computing curricula 2020 (cc2020): Paradigms for future computing curricula. Technical report, ACM/IEEE Computer Society.

Collins, A., Joseph, D., and Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *The Journal of the Learning Sciences*, 13(1):15–42.

Collins, P. H. and Bilge, S. (2020). *Intersectionality*. John Wiley & Sons.

Cooper, A. F., Moss, E., Laufer, B., and Nissenbaum, H. (2022). Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 864–876.

Corey, S. M. (1953). *Action Research to Improve School Practices*. Teachers College, Columbia University, New York.

Cortez, P. and Embrechts, M. J. (2011). Opening black box data mining models using sensitivity analysis. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 341–348. IEEE.

Cortez, P. and Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225:1–17.

Costanza-Chock, S. (2018). Design justice, ai, and escape from the matrix of domination. *Journal of Design and Science*, 3(5).

Craven, M. and Shavlik, J. (1995). Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8.

Craven, M. W. and Shavlik, J. W. (1992). Extracting tree-structured representations of trained networks. In *NIPS*.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics [1989]. *University of Chicago Legal Forum*, 1989(8):139–167.

Crenshaw, K. (2017). Race liberalism and the deradicalization of racial reform. *Harvard Law Review*, pages 2298–2319.

Creswell, J. W. (2015). A concise introduction to mixed methods research. *SAGE Publications*.

Césaire, A. (1972). *Discourse on colonialism*. Monthly Review Press.

Davis, A. (2003). *Are Prisons Obsolete?* Seven Stories Press.

Denner, J., Werner, L., Bean, S., and Campe, S. (2005). The girls creating games program: Strategies for engaging middle-school girls in information technology. *Frontiers: A Journal of Women Studies*, 26(1):90–98.

Dewey, J. (1938). *Experience and Education*. Macmillan, New York.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

DuVernay, A. (2016). 13th.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Dzieza, J. (2023). Ai is a lot of work.

Ehsan, U., Singh, R., Metcalf, J., and Riedl, M. (2022a). The algorithmic imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1305–1317, New York, NY, USA. Association for Computing Machinery.

Ehsan, U., Singh, R., Metcalf, J., and Riedl, M. (2022b). The algorithmic imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1305–1317.

Electrek (2024). Dieselgate 2.0: 600k ram trucks recalled, cummins to pay $2 billion. Published on January 11, 2024.

Epatko, L. (2018). 5 years after the world's largest garment factory collapse, is safety in bangladesh any better?

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Ewing, J. (2020). Former vw executive is arrested in croatia in emissions case.

Fabris, A., Messina, S., Silvello, G., and Susto, G. A. (2022). Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152.

Faloutsos, C. and Lin, K.-I. (1995). Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 163–174.

Faloutsos, C., Shim, K., and Udupa, J. (1995). Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the ACM SIGMOD Conference*.

Feldman, D. (2020). Core-sets: Updated survey. *Sampling techniques for supervised or unsupervised tasks*, pages 23–44.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proc. 21st ACM SIGKDD*, pages 259–268.

Fiesler, C., Garrett, N., and Beard, N. (2020). What do we teach when we teach tech ethics? a syllabi analysis. In *Proceedings of the 51st ACM technical symposium on computer science education*, pages 289–295.

Fine, M., Torre, M. E., and Stoudt, B. G. (2018). Participatory action research in education. *Oxford Research Encyclopedia of Education*.

Fisher, A. and Margolis, J. (2002). Unlocking the clubhouse: the carnegie mellon experience. *ACM SIGCSE Bulletin*, 34(2):79–83.

Fisher, A. and Margolis, J. (2003). Unlocking the clubhouse: Women in computing. In *Proceedings of the 34th SIGCSE technical symposium on Computer science education*, page 23.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10):906–911.

Fong, R. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proc. IEEE ICCV*, pages 3429–3437.

Forrester, A., Sobester, A., and Keane, A. (2008). *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons.

Fracarolli Nunes, M. and Lee Park, C. (2016). Caught red-handed: the cost of the volkswagen dieselgate. *Journal of Global Responsibility*, 7(2):288–302.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., and Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23):8410–8415.

Freitas, A. (2014). Comprehensible classification models: A position paper. *ACM SIGKDD Explorations*, 15(1):1–10.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2019). Comparative fairness-enhancing interventions in machine learning. *Proc. FAT**.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2021). The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143.

Froehlich, T. J. (2017). A not-so-brief account of current information ethics: The ethics of ignorance, missing information, misinformation, disinformation and other forms of deception or incompetence. *BiD*, 39.

Frosst, N. and Hinton, G. (2017). Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*.

Fu, Y. and Weng, Z. (2024). Navigating the ethical terrain of ai in education: A systematic review on framing responsible human-centered ai practices. *Computers and Education: Artificial Intelligence*, page 100306.

Gabriel, I. (2022). Toward a theory of justice for artificial intelligence. *Daedalus*, 151(2):218–231.

Gardner, J. and Warren, N. (2019). Learning from deep brain stimulation: the fallacy of technosolutionism and the need for 'regimes of care'. *Medicine, Health Care and Philosophy*, 22(3):363–374.

Gebru, T. (2019). Oxford handbook on AI ethics book chapter on race and gender. *CoRR*, abs/1908.06165.

Ghosh, S., Liao, Q. V., Natesan Ramamurthy, K., Navratil, J., Sattigeri, P., Varshney, K., and Zhang, Y. (2022). Uncertainty quantification 360: a hands-on tutorial. In *5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, pages 333–335.

Giannakoglou, K. C., Karakasis, M. K., and Kampolis, I. C. (2006). Evolutionary algorithms with surrogate modeling for computationally expensive optimization problem. In *Proceedings of ERCOFTAC 2006 Design Optimization International Conference, Gran Canaria, Spain*.

Gilmore, R. W. (2022). *Abolition Geography: Essays Towards Liberation*. Verso.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *Proceedings of the 2018 IEEE Symposium on Security and Privacy Workshops (SPW)*. available on arXiv and IEEE.

Goel, T., Haftka, R. T., Shyy, W., and Queipo, N. V. (2007). Ensemble of surrogates. *Structural and Multidisciplinary Optimization*, 33:199–216.

Goode, J., Margolis, J., and Chapman, G. (2014). Curriculum is not enough: The educational theory and research foundation of the exploring computer science professional development model. In *Proceedings of the 45th ACM technical symposium on Computer science education*, pages 493–498.

Goodfellow, I., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Goodman, B. and Flaxman, S. (2017a). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Goodman, B. and Flaxman, S. (2017b). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

Google Inc. (2014). Women who choose computer science: what really matters: the critical role of encouragement and exposure. `https://static.googleusercontent.com/media/edu.google.com/en//pdfs/women-who-choose-what-really.pdf`. Accessed: 2021-12-01.

Gorissen, D., Dhaenens, G., et al. (2009). Global surrogate models for engineering design. *Engineering Applications of Artificial Intelligence*, 22(5):811–819.

Gou, J. and et al. (2020). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

Gouseti, A., Burden, K., Fallin, L., and James, F. (2024). The ethics of using ai in k-12 education: A systematic literature review. *Technology, Pedagogy and Education*, pages 161–182. Published online November 26, 2024.

Grant, C. (2022). Algorithms are making decisions about health care, which may only worsen medical racism: News & commentary.

Grgic-Hlaca, Z., Weller, A., Gummadi, K. P., and Veale, M. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proc. NeurIPS*.

Griffin, W. (2022). Why universities should adopt artificial intelligence (ai) ethics course requirements. *IEEE Teaching Excellence Hub*.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35:507–520.

Grosz, B. J., Grant, D. G., Vredenburgh, K., Behrends, J., Hu, L., Simmons, A., and Waldo, J. (2019). Embedded ethics: Integrating ethics across cs education. *Commun. ACM*, 62(8):54–61.

Grover, S., Cateté, V., Barnes, T., Hill, M., Ledeczi, A., and Broll, B. (2020). First principles to design for online, synchronous high school cs teacher training and curriculum co-design. In *Koli Calling'20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research*, pages 1–5.

Guidotti, R., Monreale, A., Rinzivillo, S., Pedreschi, D., Giannotti, F., and Haklay, M. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42.

Hagendorff, T. (2022). Blind spots in ai ethics. *AI and Ethics*, 2(4):851–867.

Hampton, L. M. (2021). Black feminist musings on algorithmic oppression. *arXiv preprint arXiv:2101.09869*.

Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Hans, R. T., Marebane, S. M., and Coosner, J. (2021). Computing academics' perceived level of awareness and exposure to software engineering code of ethics: A case study of a south african university of technology. *International Journal of Advanced Computer Science and Applications*, 12(5).

Har-Peled, S., Indyk, P., and Motwani, R. (2004). Approximation algorithms for k-means and k-median clustering via core-sets. In *STOC*.

Har-Peled, S. and Mazumdar, S. (2004). On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300.

Harding, S. (1991). *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press.

Harro, B. (2000). The cycle of liberation. *Readings for diversity and social justice*, pages 52–58.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1):100–108.

Hess, M. R. and Kromrey, J. D. (2004). Robust confidence intervals for effect sizes: A comparative study of cohen'sd and cliff's delta under non-normality and heterogeneous variances. In *annual meeting of the American Educational Research Association*, volume 1. Citeseer.

Hill Collins, P. (2000). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge.

Hinton, E., Henderson, L., and Reed, C. (2018). For the record an unjust burden: The disparate treatment of black americans in the criminal justice system.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hochschild, A. (2022). Leopold ii. online.

Honebein, P. C. (1996). Seven goals for the design of constructivist learning environments. In Wilson, B. G., editor, *Constructivist Learning Environments: Case Studies in Instructional Design*, pages 11–24. Educational Technology Publications, Englewood Cliffs, NJ.

Hongladarom, S. and Bandasak, J. (2023). Non-western ai ethics guidelines: implications for intercultural ethics of technology. *AI & SOCIETY*, pages 1–14.

Hooker, G. (2004). Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32.

hooks, b. (1984). *Feminist Theory: From Margin to Center*. South End Press.

hooks, b. (2005). *The will to change: Men, masculinity, and Love*. Simon & Schuster.

Hutchings, P. and Shulman, L. S. (1999). The scholarship of teaching: New elaborations, new synergies. *Change: The Magazine of Higher Learning*, 31(5):10–15.

Huysmans, S., Dejaeger, B., et al. (2011). An empirical evaluation of classification model understandability. In *IEEE Symposium on Visual Analytics Science and Technology*, page 139–148.

IEEE Computer Society (2024). Ieee computer society and industry partners release guidance in the form of undergraduate computer science curricula. Technical report, IEEE Computer Society.

Inc, G. (2023). Violence.

Injustice, M. P. (2023). How many jurisdictions use each tool? `https://pretrialrisk.com/national-landscape/how-many-jurisdictions-use-each-tool/`. Accessed: September 9, 2023.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Jacobs, A. Z. and Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Jakesch, M., Buçinca, Z., Amershi, S., and Olteanu, A. (2022). How different groups prioritize ethical values for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 310–323.

Jelihovschi, E. G., Faria, J. C., and Allaman, I. B. (2014). Scottknott: a package for performing the scott-knott clustering algorithm in r. *TEMA (São Carlos)*, 15:3–17.

Johansson, U., König, R., and Niklasson, L. (2004a). The truth is in there-rule extraction from opaque models using genetic programming. In *FLAIRS conference*, volume 1. Miami Beach, FL.

Johansson, U., Niklasson, L., and König, R. (2004b). Accuracy vs. comprehensibility in data mining models. In *Proceedings of the seventh international conference on information fusion*, volume 1, pages 295–300. Citeseer.

Johnson, D. W. and Johnson, R. T. (2018). Cooperative learning: The foundation for active learning. *Active learning—Beyond the future*, pages 59–71.

Johnson, K. (2020). Access now resigns from partnership on ai due to lack of change among tech companies.

Judd, S. (2020). All means all: Bringing project-based, approachable ai curriculum to more high school students through ai4all open learning. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 1409–1409.

Kafai, Y., Proctor, C., and Lui, D. (2020). From theory bias to theory dialogue: embracing cognitive, situated, and critical framings of computational thinking in k-12 cs education. *ACM Inroads*, 11(1):44–53.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33.

Kamiran, N., Calders, T., et al. (2009). Classifying without discrimination. In *PKDD*, page 62–76.

Kara, S. (2018). Is your phone tainted by the misery of 35,000 children in congo's mines?

Kaufman, G. et al. (2012). Data leakage: A silent cause of poor model performance. *Transactions on Data Privacy*, 5(2):181–198.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR.

Keyes, O. (2018). The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22.

Khan, N. Z. and Luxton-Reilly, A. (2016). Is computing for social good the solution to closing the gender gap in computer science? In *Proceedings of the Australasian Computer Science Week Multiconference*, pages 1–5.

Kick, R. and Trees, F. P. (2015). Ap cs principles: engaging, challenging, and rewarding. *ACM Inroads*, 6(1):42–45.

Kier, M. W., Blanchard, M. R., Osborne, J. W., and Albert, J. L. (2014). The development of the stem career interest survey (stem-cis). *Research in Science Education*, 44(3):461–481.

Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.

Kirk, C. M., Lewis, R. K., Brown, K., Nilsen, C., and Colvin, D. Q. (2012). The gender gap in educational expectations among youth in the foster care system. *Children and Youth Services Review*, 34(9):1683–1688.

Kolb, A. Y. and Kolb, D. A. (2005). Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of Management Learning & Education*, 4(2):193–212.

Könings, K. D., Seidel, T., and van Merriënboer, J. J. (2014). Participatory design of learning environments: integrating perspectives of students, teachers, and designers. *Instructional Science*, 42(1):1–9.

Krakowski, A., Greenwald, E., Hurt, T., Nonnecke, B., and Cannady, M. (2022). Authentic integration of ethics and ai through sociotechnical, problem-based learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12774–12782.

Krishnan, K. P. and Tesauro, G. (1995). Extracting decision trees from trained neural nets. In *NIPS*.

Krishnan, S. and Wu, E. (2017). Palm: Machine learning explanations for iterative debugging. In *Proceedings of the 2Nd workshop on human-in-the-loop data analytics*, pages 1–6.

Kumar, A. N., Raj, R. K., Aly, S. G., Anderson, M. D., Becker, B. A., Blumenthal, R. L., Eaton, E., Epstein, S. L., Goldweber, M., Jalote, P., Lea, D., Oudshoorn, M., Pias, M., Reiser, S., Servin, C., Simha, R., Winters, T., and Xiang, Q. (2024a). Computer science curricula 2023 (cs2023): The final report. Technical report, ACM Press, IEEE Computer Society Press and AAAI Press. Endorsed by ACM, IEEE-CS, and AAAI.

Kumar, A. N., Raj, R. K., et al. (2024b). Computer science curricula 2023 (cs2023): The final report. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, SIGCSE 2024. ACM.

Kurakin, A., Goodfellow, I. J., and Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC.

Ladner, R., Ludi, S., and Domanski, R. (2023a). Cs2023: Acm/ieee/aaai computer science curricula includes teaching accessibility. AccessComputing.

Ladner, R., Ludi, S., and Domanski, R. (2023b). Teaching about accessibility in computer science education. *CS2023 Curricular Practices*.

Lechner, F. J. and Wallerstein, I. (2015). *The Modern World-System as a Capitalist World-Economy*, page 56–62. Wiley Blackwell, fifth edition.

Lédeczi, Á., Grover, S., Catete, V., and Broll, B. (2021). Beyond cs principles: Bringing the frontiers of computing to k12. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 1379–1379.

Lee, T. (2022). A deep dive into the labor exploitation behind everyday products.

Lewin, K. (1946). *Action Research and Minority Problems*, volume 2. Wiley.

Liao, Q. V. and Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Library, F. S. U. A. V. G.-C. (2023). Anti-racism resources: Colorblindness.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

Loideain, N. N. and Adams, R. (2020). From alexa to siri and the gdpr: the gendering of virtual personal assistants and the role of data protection impact assessments. *Computer Law & Security Review*, 36:105366.

Lott, B. (2012). The social psychology of class and classism. *American Psychologist*, 67(8):650.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Luong, K. S., Fang, F., et al. (2011). Situation testing in knn for classification discrimination. In *FATML Workshop*.

Manne, K. (2017). *Down girl: The logic of misogyny*. Oxford University Press.

Marketplace, A. (2019). Machine learning & artificial intelligence.

Marques-Silva, J. and Ignatiev, A. (2022). Delivering trustworthy ai through formal xai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12342–12350.

Marx, K. (1990). *Capital Volume 1*, pages 11–57. Penguin Classics.

McFadden, Z. and Alvarez, L. (2024). Performative ethics from within the ivory tower: How cs practitioners uphold systems of oppression. *Journal of Artificial Intelligence Research*, 79:777–799.

McKeever, E. (2022). Who turned out the lights? how critical race theory bans keep people in the dark. *Washington University Jurisprudence Review*, 15(1).

McKenney, S. and Reeves, T. C. (2021). Educational design research: Portraying, conducting, and enhancing productive scholarship. *Medical education*, 55(1):82–92.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Merriam-Webster (2023a). Colonialism definition & meaning. online.

Merriam-Webster (2023b). Hegemonic definition.

Mezirow, J. (1991). *Transformative Dimensions of Adult Learning*. Jossey-Bass, San Francisco, CA.

Microsoft Corporation (2017). Why europe's girls aren't studying stem. `https://news.microsoft.com/europe/features/dont-european-girls-like-science-technology/`. Accessed: 2021-12-01.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Ming, Y., Qu, H., and Bertini, E. (2018). Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics*, 25(1):342–352.

MIT Media Lab (2024). Ai + ethics curriculum for middle school. `https://www.media.mit.edu/projects/ai-ethics-for-middle-school/overview/`. Open source curriculum project.

Mittas, H. and Angelis, L. (2013). Predictive performance comparison of machine learning methods. *Expert Systems with Applications*, 40(7):2842–2853.

Mittas, N. and Angelis, L. (2012). Ranking and clustering software cost estimation models through a multiple comparisons algorithm. *IEEE Transactions on software engineering*, 39(4):537–551.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507.

Mohamed, S., Png, M.-T., and Isaac, W. (2020). Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33:659–684.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Molnar, C. (2022). *Interpretable Machine Learning*. Molnar, 2 edition.

Molnar, C., Casalicchio, G., and Bischl, B. (2021). Interpretable machine learning–a brief history, state-of-the-art and challenges. In *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings*, pages 417–431. Springer.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., and Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of ai ethics. *Minds and Machines*, 31(2):239–256.

Morley, M. H. (1989). *The U.S. imperial state: theory and historical setting*. Cambridge University Press.

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. PublicAffairs.

Morse, S., Angelos, A., and Doe, J. (2021). Contextualizing fairness criteria in machine learning. *Journal of Fairness in AI*.

Munn, L. (2023). The uselessness of ai ethics. *AI and Ethics*, 3(3):869–877.

Munteanu, A. and Schwiegelshohn, C. (2018). Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intelligenz*, 32(1):37–53.

Murdoch, W. J., Singh, C., Kumbier, K., and et al., R. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

NAACP (2021). The origins of modern day policing.

Nagireddy, M., Singh, M., Hoffman, S. C., Ju, E., Ramamurthy, K. N., and Varshney, K. R. (2023). Function composition in trustworthy machine learning: Implementation choices, insights, and questions. *arXiv preprint arXiv:2302.09190*.

Nair, V. (2023). The environmental impact of llms.

National Center for Women & Information Technology (2021). By the numbers. `https://ncwit.org/resource/bythenumbers/`. Accessed: 2021-12-01.

National Humanities Center (2024). Responsible ai project (2021–2024). https://nationalhumanitiescenter.org/education-programs/faculty-programs/responsible-artificial-intelligence-undergraduate-curriculum-design-project/2021-24-project/. Accessed July 2024.

Nazgol Ghandnoosh, P. and Nellis, A. (2022). Research - get the facts.

Noble, S. U. (2018). *Algorithms of oppression*. New York University Press.

Nowell, L. S., Norris, J. M., White, D. E., and Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1):1–13.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

of Wisconsin Madison, U. (2023). Research guides: General philosophy: Ethics (moral philosophy) and value theory.

Oi, R. (2022). Women in tech: Pushing for inclusivity and diversity.

on Gender-Based Violence, A. P. I. (2018). Patriarchy & power.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Panagiotopoulou, V.C. Stavropoulos, P. and Chryssolouris, G. (2022). A critical review on the environmental impact of manufacturing: a holistic perspective. *The International Journal of Advanced Manufacturing Technology*, pages 603–625.

Papastergiou, M. (2008). Are computer science and information technology still masculine fields? high school students' perceptions and career choices. *Computers & education*, 51(2):594–608.

Park, S., Kim, S., and Lim, Y.-s. (2022). Fairness audit of machine learning models with confidential computing. In *Proceedings of the ACM Web Conference 2022*, pages 3488–3499.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., and Turini, F. (2019). Meaningful explanations of black box ai decision systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9780–9784.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *KDD*, page 560–568.

Peng, K., Chakraborty, J., and Menzies, T. (2022). xfair: Better fairness via model-based rebalancing of protected attributes. *IEEE Trans SE, accepted, to appear*.

Perez, C. C. (2019). *Invisible women: Data bias in a world designed for men*. Abrams.

Perraudin, F. (2019). Low pay in the garment industry still a reality despite pledges – study.

Perrigo, B. (2023). Openai used kenyan workers on less than $2 per hour exclusive. online.

Piaget, J. (1952). The origins of intelligence in children. *International University*.

Pourhomayoun, M. and Shakibi, M. (2021). Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making. *Smart Health*, 20:100178.

Prioleau, D., Richardson, B., Drobina, E., Williams, R., Martin, J., and Gilbert, J. E. (2021). How students in computing-related majors distinguish social implications of technology. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 1013–1019.

Project, T. I. S. (2018). Report to the united nations on racial disparities in the u.s. criminal justice system.

Pulrang, A. (2022). 6 ways responses to covid-19 have been ableist, and why it matters.

Puri, N., Gupta, P., Agarwal, P., Verma, S., and Krishnamurthy, B. (2017). Magix: Model agnostic globally interpretable explanations. *arXiv preprint arXiv:1706.07160*.

Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 145–151, New York, NY, USA. Association for Computing Machinery.

Raji, I. D., Scheuerman, M. K., and Amironesei, R. (2021). You can't sit with us: exclusionary pedagogy in ai ethics education. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 515–525.

Rakova, B., Yang, J., Cramer, H., and Chowdhury, R. (2021). Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23.

Real, T. (2003). *How can I get through to you?: Closing the intimacy gap between men and women*. Simon & Schuster.

Report, A. I. (2023). Measuring trends in artificial intelligence. In *Artifical Intelligence Index Report 2023*. Stanford University Human Centered Artificial Intelligence.

Repository, W. N. N. (2022). Wolfram neural net repository. Available at `https://resources.wolframcloud.com/NeuralNetRepository/` (2022/11/11).

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., Silver, J., Silverman, B., et al. (2009). Scratch: programming for all. *Communications of the ACM*, 52(11):60–67.

Ribeiro, M., Grolinger, K., and Capretz, M. A. (2015). Mlaas: Machine learning as a service. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 896–902. IEEE.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Richardson, E. and et al. (2023). Influential fairness: Adding explainability to bias mitigation via influence functions. In *Proc. FAT**.

Rigaki, M. and Garcia, S. (2023). A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34.

Robinson, C. J., Sojoyner, D., and Willoughby-Herard, T. (1983). *Black Marxism, Revised and Updated Third Edition: The Making of the Black Radical Tradition*, pages 75–77. University of North Carolina Press, 3 edition.

Robnik-Šikonja, M. and Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

Robnik-Šikonja, M. and Kononenko, I. (2008). Explaining predictions of classifiers. In *ECML/PKDD Discovery Challenge Workshop*, page 1–6.

Rodriguez, D. (2020). Reformism isn't liberation, it's counterinsurgency.

Roediger, H. L. and Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3):249–255.

Rosenstein, A., Raghu, A., and Porter, L. (2020). Identifying the prevalence of the impostor phenomenon among computer science students. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 30–36.

Rowell, L., Riessman, C. K., and Stenstrom, M.-L. (2022). Action research and curriculum development with consideration of the nordic context. *Oxford Research Encyclopedia of Education*.

Rudin, C. (2019a). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Rudin, C. (2019b). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

Rudin, C., Wang, C., and Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, 2(1). https://hdsr.mitpress.mit.edu/pub/7z10o269.

Rueda, J., Rodríguez, J. D., Jounou, I. P., Hortal-Carmona, J., Ausín, T., and Rodríguez-Arias, D. (2022a). "just" accuracy? procedural fairness demands explainability in ai-based medical resource allocations. *AI & society*, pages 1–12.

Rueda, P., Alvarez, M., and Gomez, L. (2022b). Explainability and procedural fairness in medical algorithmic allocation. *Medical AI Ethics Journal*, 5(2):120–135.

Schneider, J., Meske, C., and Vlachos, M. (2020). Deceptive ai explanations: Creation and detection. *arXiv preprint arXiv:2001.07641*.

Schön, D. A. (1992). *The reflective practitioner: How professionals think in action*. Routledge, London.

Schultz, M. D. and Seele, P. (2023). Towards ai ethics' institutionalization: knowledge bridges from business ethics to advance organizational ai ethics. *AI and Ethics*, 3(1):99–111.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512.

Semenova, V. and et al. (2022). Revisiting interpretable machine learning: Recent developments. *AI Journal*.

Seneviratne, O. (2017). Making computer science attractive to high school girls with computational thinking approaches: A case study. In *Emerging research, practice, and policy on computational thinking*, pages 21–32. Springer.

Shanbehzadeh, M., Yazdani, A., Shafiee, M., and Kazemi-Arpanahi, H. (2022). Predictive modeling for covid-19 readmission risk using machine learning algorithms. *BMC Medical Informatics & Decision Making*, 22(1):1–12.

Sharma, K. and et al. (2020). Data augmentation for discrimination prevention and bias disambiguation. *Proc. FAT**.

Shew, A. (2020). Ableism, technoableism, and future ai. *IEEE Technology and Society Magazine*, 39(1):40–85.

Siapka, A. (2022). Towards a feminist metaethics of ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 665–674, New York, NY, USA. Association for Computing Machinery.

Simha, R., Kumar, A. N., and Raj, R. K. (2024). Undergraduate computer science curricula: First-job readiness versus long-term career preparation. *Communications of the ACM*, 67(2):29–31.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

Sloane, M., Moss, E., Awomolo, O., and Forlano, L. (2020). Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423*.

Smedley, A. and Smedley, B. D. (2005). Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. *The American psychologist*, 60 1:16–26.

Smith, L. (2023). Center for disability rights.

Smith, M. K., Wood, W. B., Krauter, K., and Knight, J. K. (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE—Life Sciences Education*, 10(1):55–63.

Spinner, T., Schlegel, U., Schäfer, H., and El-Assady, M. (2019). explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics*, 26(1):1064–1074.

Stavrakakis, I., Gordon, D., Tierney, B., Becevel, A., Murphy, E., Dodig-Crnkovic, G., Dobrin, R., Schiaffonati, V., Pereira, C., Tikhonenko, S., et al. (2021). The teaching of computer ethics on computer science and related degree programmes. a european survey. *International Journal of Ethics Education*, pages 1–29.

Stufflebeam, D. L. and Shinkfield, A. J. (2007). *Evaluation Theory, Models, and Applications*. Jossey-Bass, San Francisco.

Szegedy, C., Zaremba, W., et al. (2014). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tal, A. S., Batsuren, K., Bogina, V., Giunchiglia, F., Hartman, A., Loizou, S. K., Kuflik, T., and Otterbacher, J. (2019). "end to end" towards a framework for reducing biases and promoting transparency of algorithmic systems. In *2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 1–6. IEEE.

Tan, S., Caruana, R., Hooker, G., and Lou, Y. (2018). Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 303–310.

Tilmes, N. (2022). Disability, fairness, and algorithmic bias in ai recruitment. *Ethics and Information Technology*, 24(2):21.

Ton, J. (2020). Council post: Race in tech, part one: Inside the numbers.

Touretzky, D. S., Gardner-McCune, C., Martin, F., and Seehorn, D. (2019). K-12 guidelines for artificial intelligence: what students should know. In *Proc. of the ISTE Conference*.

Transformative Interdisciplinary Human+AI Research Group (2022). Modular experiential learning for secure, safe, and reliable ai. In *The 23rd Annual Conference on Information Technology Education*, SIGITE '22, Chicago, IL, USA. ACM.

Treichler, P. A. and Kramarae, C. (1983). Women's talk in the ivory tower. *Communication Quarterly*, 31(2):118–132.

Valstar, S., Griswold, W. G., and Porter, L. (2020). A quantitative study of faculty views on the goals of an undergraduate cs program and preparing students for industry. *Proceedings of the 2020 ACM Conference on International Computing Education Research*, pages 317–332.

van der Geer, J. P., Hanraads, J. A. J., and Lupton, R. A. (2007). *Empirical Research in Education*. Academic Press, Amsterdam.

van Maanen, G. (2022). Ai ethics, ethics washing, and the need to politicize data ethics. *Digital Society*, 1(2):9.

Varshney, K. R. (2022). Trustworthy machine learning. *Chappaqua, NY, USA: Independently Published*.

Vera (2023). Incarceration statistics.

Vilone, G. and Longo, A. (2020a). Explainable artificial intelligence: A systematic review. *Expert Systems with Applications*, 140.

Vilone, G. and Longo, L. (2020b). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Wachter, S., Mittelstadt, B., and Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567.

Waelen, R. (2022). Why ai ethics is a critical theory. *Philosophy & Technology*, 35(1):9.

Wagner, W. S. and Peter (2022). Mass incarceration: The whole pie 2022.

Wang, J. and Hejazi Moghadam, S. (2017). Diversity barriers in k-12 computer science education: Structural and social. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, pages 615–620.

Weinberg, L. (2022). Rethinking fairness: an interdisciplinary survey of critiques of hegemonic ml fairness approaches. *Journal of Artificial Intelligence Research*, 74:75–109.

West, M., Kraut, R., and Ei Chew, H. (2019). I'd blush if i could: closing gender divides in digital skills through education.

Williams, R. M., Smarr, S., Prioleau, D., and Gilbert, J. E. (2022). Oh no, not another trolley! on the need for a co-liberative consciousness in cs pedagogy. *IEEE Transactions on Technology and Society*, 3(1):67–74.

Wilson, L. O. (2016). Anderson and krathwohl–bloom's taxonomy revised. *Understanding the new version of Bloom's taxonomy*.

Wyllie, S., Shumailov, I., and Papernot, N. (2024). Fairness feedback loops: training on synthetic data amplifies bias. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2113–2147.

Xiu, M., Ming, Z., Jiang, and Adams, B. (2021). An exploratory study on machine learning model stores. *IEEE Software*.

Yeom, S. and Tschantz, M. C. (2021). Avoiding disparity amplification under different worldviews. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 273–283.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and 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*, WWW '17, page 1171–1180, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zemel, R., Wu, Y., et al. (2013). Learning fair representations. *ICML*.

Zhou, X. and et al., C. R. (2021). A survey on the quality of explanations in explainable ai. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3047–3064.

APPENDICES

APPENDIX

A

A TIMELINE OF FOUNDATIONAL AI & ML RESEARCH

This appendix provides a concise historical overview of key developments in artificial intelligence (AI) and machine learning (ML), highlighting foundational contributions and the evolution of methods that shape the contemporary research domain of this dissertation. From early clustering and decision tree algorithms to recent advances in fairness, interpretability, and explainability, this timeline offers context for the techniques and ideas referenced throughout the dissertation. While not exhaustive, it reflects the broad trajectory of the field and situates this work within a rich and ongoing scholarly tradition.

Statistical models are the foundation of AI and ML. The popular K-means algorithm was published in 1979 (Hartigan and Wong 1979), and classification and regression trees (CARTs) were published in 1983 (Breiman et al. 1984). Setting the stage for Quilan's 1992 Machine learning book (Quinlan 1992). Also, in 1992, Craven presented a technique for extracting decision trees from trained neural networks (Craven and Shavlik 1992). In 1995, FastMap, a recursive biclustering algorithm, was presented as a method for multi-dimensional scaling (MDR) (Faloutsos et al. 1995). Krishnan and Tesauro (1995) built on Craven's work extracting decision trees from neural networks. Chakrabarti and Mehrotra (2000) presented local dimensional reduction as a method for capturing the most variation from a set. However, they warned that it would increase false

positives. In 2001, Breiman published a historical paper on random forests with over 20,000 citations (Breiman 2001). In 2002, SMOTE was a technique that combined oversampling and undersampling to balance datasets. A paper on the complexities measures for supervised classification was also published (Chawla et al. 2002). Har-Peled et al. (2004) developed core-sets for k-means/median clustering. Agarwal et al. (2005) coined the term core-set, which is a data reduction technique used to extract general behavior. They built off their previous work for approximating extent measures (Har-Peled et al. 2004). Giannakoglou et al. (2006) applied evolutionary algorithms with surrogate models for computationally expensive optimization problems. Goel et al. (2007) used multiple surrogate models to identify where predictions differ and provide a robust approximation approach to solve analytical, complex engineering problems.

Pedreshi et al. (2008) published a foundational fair AI/ fairness paper on discrimination-aware data mining. They argued that the naive approach of removing discriminatory attributes (also known as protected attributes) such as gender, race, and other demographic information was not enough when other background information was provided. Arguably, they can be credited with formalizing redlining in the data context. Robnik-Šikonja and Kononenko (2008) also published a study on identifying classification outcomes for individual instances, setting the stage for future work examining individual fairness. Gorissen et al. (2009) reported how global surrogate models (previously known as metamodels) were indispensable for understanding design space due to their low computational costs, easier visualization, rapid prototyping, and sensitivity analysis. This work is foundational to current interpretable ML (IML) research and the efforts to develop robust IML methods. Kamiran et al. (2009) presented a data augmentation method for classifying with discrimination and noted that impartial classification was required by law. This paper is the predecessor for Kamiran and Calders (2012) later work extending pre-processing methods such as suppression of protected attributes (PAs), data and class label augmentation, and reweighting/resampling data to remove discrimination without manipulating class labels. They focused on singular binary PAs and two-class classification problems, which would stay the focus for most fair classification work. Chen and Chen (2009) extended Har-Peled et al.'s (2004) work establishing a new approximation algorithm for k-median/means clustering problems as a technique for core-sets.

Baehrens et al. (2010) described how no classifier could explain why a model assigned a label to a particular instance, and the only method that could provide individual explanations was decision trees. Building on the work of Robnik-Šikonja and Kononenko (2008), which focused on explaining decisions "for kNN-, SVM-, and ANN-models for individual instances by measuring the difference in their prediction with sets of features omitted," they presented a method that could work for many "black-box", nonlinear classifiers. This work is critical to local

explanation techniques. Huysmans et al. (2011) report the results of an empirical study focusing on the required interpretability of classification models. They argue for the use of decision tables. Luong et al. (2011) use situation testing methodology to apply K-nearest neighbors to identify classification discrimination. Dwork et al. (2012) fairness through awareness arXiv paper used statistical parity to determine differences in individual classifications. Zemel et al. (2013) later expanded this work by proposing a fair classification model that optimized for group and individual fairness criteria.

Mittas and Angelis (2013) performed a large-scale systematic comparison of prediction models and concluded that linear models were adequate for most cases. They warned against using more sophisticated, complicated models and pushed for more testing for appropriate statistical models based on the use case. Freitas (2014) published a position paper on trade-offs between interpretability/comprehensibility and accuracy in classification problems. They highlighted a need for a more principled approach than focusing on predictive accuracy evaluations.

Hinton et al. (2015) paper, with over 17,000 citations, coins the term "distillation" as a method for transferring knowledge from a larger model to a smaller model. This strategy was pioneered by Bucilă et al. (2006), who demonstrated the knowledge/behavior of a large ensemble could be “compressed” into a single small model. The year 2015 also saw the rise of deep learning research, and many research initiatives pivoted to investigating the applications of deep learning models. Goodfellow et al. (2015) and Szegedy et al. (2014) exposed the vulnerabilities of several ML models vulnerable to adversarial perturbations. They called for developing more optimization procedures that could train models to be more stable. This is the predecessor work to data and adversarial robustness concerns in trustworthy ML.

Feldman et al. (2015) expanded on Barocas and Selbst (2014) technical report on disparate impact in big data by measuring disparate impact with classification accuracy metrics and removing disparate impact within in numerical features. However, this work only applies to numerical features, and future work is recommended to investigate non-numerical features. Barocas and Selbst (2016b) essay further described concerns about disparate impact and argued that the meanings of “discrimination” and “fairness” in data must be heavily reexamined. Meanwhile, further work was conducted to mitigate bias and fairness in data. Calmon et al. (2015) presented a “data-driven optimization framework for probabilistically transforming data to reduce algorithmic discrimination” for using pre-processing data augmentation methods.

After 2016 and the looming AI regulations, many research initiatives pivoted to explainable AI (XAI) and IML methods. Ribeiro et al. (2016) published the initial LIME paper. Another notable IML work, Fong and Vedaldi (2017), proposed a framework for explanations as meta-predictors, particularly for NN trained with image data. Goodman and Flaxman (2017b) sum-

marized the European Union's General Data Protection Regulation and "right to explanation." They highlighted the potential impacts on industry work and said that future work is needed regarding evaluation frameworks that avoid discrimination and enable explanation. Guidotti et al. (2018) published a survey on explanation methods and the current state of explanation and interoperability. Critiques such as Lipton (2018) pushed back at interpretable and transparency efforts for reasons like " transparency may be at odds with the broader objectives of AI (artificial intelligence)" implies that models surpassing human ability should be prioritized. Other surveys include XAI survey (Adadi and Berrada 2018) , overview of IML (Gilpin et al. 2018), and core-set survey (Munteanu and Schwiegelshohn 2018).

Other work Grgic-Hlaca et al. (2018), used feature selection as a method to represent procedural fairness and expand fair ML beyond distributive fairness. In subsequent work by Morse et al. (2021), the authors further emphasized the importance and complexity of fairness criteria and the need for contextualization when considering fairness. Also, Rueda et al. (2022b) how a lack of explainability in AI allocation algorithms is procedurally unfair and uses medical field applications as their case study.

During this period, an increased interest in fairness and fair ML was increased. Other efforts include IBM's AI Fairness 360 open-source toolkit (Bellamy et al. 2018), Ribeiro et al. (2018) more precise model-agnostic explanation (Anchors), and Adler et al. (2018) work on auditing "black-box" models for indirect feature influence. Adler et al. (2018) work is one of the first to describe IML as auditing and established a foundation for future work proxy variables that may indirectly influence classification outcomes. They also raise the connections to privacy-preserving data mining (Agrawal and Srikant 2000), data leakage (Kaufman et al. 2012), and feature selection (Chandrashekar and Sahin 2014). Friedler et al. (2019) did a comparative study on fairness-enhancing interventions, Carvalho and et al. (2019) published a survey on IML, Rudin (2019b) argued for IML over explanation methods for high-stakes/high-risk domain applications, and Murdoch et al. (2019) presented a framework and taxonomy for discussing IML. Miller (2019) argued for using social science theories to improve explanations. This is one of the first papers to situate a need for socio-technical considerations in ML, and is discussed in detail in Carey and Wu (2022) statistical fairness guide exploring social and formal perspectives.

In 2020, updated surveys were published to detail the rapidly growing fields, such as the knowledge distillation survey (Gou and et al. 2020), the core set survey (Feldman 2020), the IML book (Molnar 2020), the systematic review of XAI (Vilone and Longo 2020a), the taxonomy of XAI and the connections to responsible AI (Arrieta et al. 2020), and the XAI survey focusing on the quality of explanations (Zhou and et al. 2021). Other related work, Sharma and et al. (2020) paper descriptively titled "Data Augmentation for Discrimination Prevention and Bias Disambiguation' and Slack et al. (2020) study on tricking LIME and hiding malicious behavior,

and Chakraborty et al. (2021) Fair-SMOTE data augmentation method for bias mitigation.

Most recently, there is a reiteration of using IML methods (Semenova and et al. 2022) and an updated supervised learning benchmark for tabular datasets describing potential reasons for tree-based models outperforming deep learning techniques (Grinsztajn et al. 2022). Park et al. (2022) propose a fairness auditing framework using confidential computing and trust. Alkhatib and et al. (2023) introduced a new explanation technique using characteristic rules, CEGA (Characteristic Explanatory General Association rules), that outperforms Anchors (Ribeiro et al. 2018) and similar explanation techniques. Another notable work, Richardson and et al. (2023), coined influential fairness and presented a novel procedure to add explainability, transparency, and contextualization to bias mitigation with influence functions and connections to practitioners' individual needs.

In summary, this timeline highlights the foundational ideas and pivotal research milestones that have influenced the development of AI, ML, fairness, and interpretability research. By situating this dissertation within this historical context, the appendix underscores the continuity of these themes and the ways in which contemporary challenges build on prior work. Consider this appendix item a thorough, footnote-laden nod to academic tradition and an assurance that the bibliography is, at the very least, complete.

APPENDIX

B

COURSE PRE- ASSESSMENT

B.1 Pre-Assessment

B.1.1 Likert Scale Questions

For each statement rate your agreement 1-5 where 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree.

Remember

- LO1: Define the four attributes of trustworthiness in machine learning systems.
- LO2: List the six steps of the machine learning lifecycle.
- LO3: Identify the different types of distribution shifts in machine learning systems.

Understand

- LO4: Distinguish between different types of fairness and their appropriate applications.
- LO5: Describe the role of causality in machine learning decision-making.

- LO6: Explain the relationship between aleatoric and epistemic uncertainty in safety considerations.
- LO7: Describe the socio-technical challenges of fairness, accountability, and bias in ML systems.

Apply

- LO8: Evaluate supervised binary classifier performance.
- LO9: Apply appropriate bias mitigation strategies based on specific worldviews.
- LO10: Implement appropriate privacy preservation methods for different data scenarios.
- LO11: Determine ML models with considerations of data sources, bias, privacy, and consent.

Analyze

- LO12: Analyze the implications of different types of adversarial attacks on ML systems.
- LO13: Evaluate the effectiveness of different explanation methods for stakeholder personas.
- LO14: Examine the relationship between team composition and model outcomes.
- LO15: Analyze the stages of the ML lifecycle to promote trust, safety, and transparency.

Evaluate

- LO16: Assess the appropriateness of different fairness metrics for specific contexts.
- LO17: Critique the effectiveness of various transparency mechanisms in ML systems.
- LO18: Judge the institutional trustworthiness of machine learning platforms.
- LO19: Evaluate the ethical implications of ML deployment, including impacts on marginalized populations.

Create

- LO20: Design governance frameworks for ML systems that align with stakeholder values.
- LO21: Develop strategies for implementing trustworthy ML in social impact projects.
- LO22: Construct appropriate testing protocols for ML models.

B.2 Topic Specification Table

This table shows the breakdown of how questions covered each TML topics and Bloom's level.

Parts	Chapter	Remember	Understand	Apply	Analyze	Evaluate	Create	Total Items
Introduction & Preliminaries	Establishing Trust	1	0	0	0	0	0	1
	ML Lifecycle	2	0	0	0	0	0	2
	Safety	0	1	0	0	0	0	1
Data	Data Modalities, Sources, & Biases	0	0	0	1	0	0	1
	Privacy & Consent	1	0	1	0	0	1	3
Basic Modeling	Detection Theory	1	0	1	0	0	0	2
	Supervised Learning	0	0	0	0	0	1	1
	Causal Modeling	0	0	0	0	0	0	0
Reliability	Distribution Shift	0	1	1	1	0	0	3
	Fairness	0	1	1	1	1	2	6
	Adversarial Robustness	0	1	1	1	1	1	5
Interaction	Interpretability & Explainability	0	1	0	1	1	1	4
	Transparency	1	0	1	0	0	1	3
	Value Alignment	0	0	0	0	0	0	0
Purpose	Ethics Principles	0	0	0	0	0	0	0
	Lived Experience	0	1	0	1	0	0	2
	Social Good	0	1	0	0	0	1	2
	Filter Bubbles & Disinformation	0	0	0	0	1	0	1
	Total Items	6	7	6	6	4	8	37

Table B.1: Condensed Table of Pre-Assessment Questions Denoting Bloom's Levels amd TML Topics

B.3 Multiple Choice, and Response Questions

Section 1: Recall Facts and Basic Concepts (Remember)

1. The four attributes of trustworthiness are competence, reliability, openness, and:

   A. selflessness

   B. accuracy

   C. transparency

D. efficiency

Chapter: Introduction and Preliminaries, Establishing Trust (Week: 1)
Reference: "Trustworthiness is composed of four attributes: competence, reliability, openness, and selflessness."

2. Which type of uncertainty represents inherent randomness in phenomena and is well-modeled using probability theory?

   A. Aleatoric uncertainty
   B. Epistemic uncertainty
   C. Stochastic uncertainty
   D. Random uncertainty

   Chapter: Introduction and Preliminaries, Machine Learning Lifecycle (Week: 1)
   Reference: "Aleatoric uncertainty is inherent randomness in phenomena. It is well-modeled using probability theory."

3. Which step occurs first in the machine learning lifecycle?

   A. Problem specification
   B. Data understanding
   C. Data preparation
   D. Modeling

   Chapter: Introduction and Preliminaries, Machine Learning Lifecycle (Week: 1)
   Reference: "The end-to-end development process or lifecycle involves several steps: 1. problem specification..."

4. Which document provides transparency and conformity assessment information?

   A. Factsheet
   B. Technical report
   C. User manual
   D. System log

   Chapter: Interaction, Transparency (Week: 9)
   Reference: "A factsheet is a collection of facts, appropriately rendered for a given consumer, that enables transparency and conformity assessment."

5. When publishing individual-level data while maintaining utility, which privacy preservation approach is most appropriate for protecting sensitive information?

   A. Syntactic anonymization with similar quasi-identifiers
   B. Differential privacy with fixed queries
   C. Basic data encryption
   D. Simple data masking

   Chapter: Part 2 - Data, Privacy and Consent (Week: 2)
   Reference: "Syntactic anonymization methods group together individuals with similar quasi-identifiers and then obfuscate those quasi-identifiers."

6. Which fundamental concept serves as the basis for evaluating binary classifiers in machine learning systems?

   A. Detection theory with false positive and negative trade-offs
   B. Raw accuracy metrics
   C. Processing speed
   D. Model size

   Chapter: Part 3 - Basic Modeling, Detection Theory (Week: 3)
   Reference: "Four possible events result from binary decisions: false negatives, true negatives, false positives, and true positives."

Section 2: Explain Ideas or Concepts (Understand)

Multiple Choice Questions

1. How do different worldviews influence the selection of fairness metrics in machine learning systems?

   A. Selection depends on whether one believes measurement or sampling bias dominates
   B. Worldviews have no impact on metric selection
   C. Only technical considerations matter for fairness metrics
   D. Fairness metrics are universal across all worldviews

Chapter: Reliability, Fairness (Week: 6)
Reference: "If you believe there are social biases in measurement, then you have the 'we're all equal' worldview; independence and statistical parity difference are appropriate notions of group fairness."

2. Which approach best describes how socioculturally heterogeneous teams affect model development?

   A. They tend to slow down and avoid taking shortcuts
   B. They always work faster than homogeneous teams
   C. They focus only on technical performance metrics
   D. They prioritize speed over thoroughness

   Chapter: Purpose, Lived Experience (Week: 12)
   Reference: "Teams that are socioculturally heterogeneous tend to slow down and not take shortcuts."

3. What is the primary purpose of machine learning platforms designed for social impact?

   A. To empower social change organizations in developing ML systems that help uplift vulnerable people
   B. To maximize technical performance metrics
   C. To increase processing speed
   D. To reduce computational costs

   Chapter: Purpose, Social Good (Week: 13)
   Reference: "Machine learning platforms that are specifically designed to deskill data science needs and minimize the effort for deployment, maintenance, and support are the solution. The goal is to empower social change organizations in the development of machine learning systems that help uplift vulnerable people on their own terms."

Short Answer Questions

1. **Explain the key differences between types of distribution shift and their implications for model deployment. Include specific examples of how each type might manifest in a real-world system.** (200 words)

   Chapter: Reliability, Distribution Shift (Week: 5)
   Reference: "Distribution shift—the mismatch between the probability distribution of

the training data and the data you will see during deployment—has three special cases: prior probability shift, covariate shift, and concept drift."

2. **How do explanation methods serve different stakeholder needs in machine learning systems? Discuss how the same model might require different types of explanations for different audiences.** (200 words)

   Chapter: Interaction, Interpretability and Explainability (Week: 8)
   Reference: "Different consumers have different personas with different needs to achieve their goals. The important personas are the affected user, the decision maker, and the regulator."

3. **Describe how different types of adversarial attacks can affect model reliability, and explain why some attacks may be more concerning than others in different deployment contexts.** (200 words)

   Chapter: Reliability, Adversarial Robustness (Week: 7)
   Reference: "Adversaries may just want to worsen model accuracy in general or may have targeted goals that they want to achieve, such as obtaining specific predicted labels for specific inputs."

4. **Explain the relationship between aleatoric and epistemic uncertainty in safety considerations. Use examples to illustrate each type of uncertainty.** (150 words)

   Chapter: Introduction and Preliminaries, Safety (Week: 14–15)
   Reference: "Safety is the minimization of the aleatoric uncertainty and the epistemic uncertainty of undesired high-stakes outcomes."

Section 3: Use Information in New Situations (Apply)

Short Answer Questions

1. **Given a dataset containing personal information about students (grades, demographics, financial data), describe specifically how you would implement appropriate privacy preservation methods. Include both technical and ethical considerations.** (250 words)

   Chapter: Data, Privacy and Consent (Week: 2)
   Reference: "Data is a valuable resource that comes from people. The use of this data should be consensually obtained. Various methods can be used to preserve their privacy."

2. **Apply detection theory principles to evaluate the performance of a binary classifier for a medical diagnosis system. List the four possible outcomes and explain how you would weigh their importance.** (200 words)

   Chapter: Basic Modeling, Detection Theory (Week: 3)
   Reference: "Four possible events result from binary decisions: false negatives, true negatives, false positives, and true positives."

3. **Design a monitoring plan to detect distribution shift in a deployed machine learning model used for credit scoring. Specify what metrics you would track and how often.** (250 words)

   Chapter: Reliability, Distribution Shift (Week: 5)
   Reference: "Distribution shift—the mismatch between the probability distribution of the training data and the data you will see during deployment—has three special cases: prior probability shift, covariate shift, and concept drift."

4. **Develop a testing strategy for assessing adversarial robustness in a facial recognition system. Include specific test cases and evaluation criteria.** (250 words)

   Chapter: Reliability, Adversarial Robustness (Week: 7)
   Reference: "Even without malicious actors, adversarial attacks are a way for developers to test machine learning systems in worst case scenarios."

5. **Apply appropriate bias mitigation strategies for a machine learning model used in hiring decisions. Explain your choice of fairness metrics and mitigation approaches.** (300 words)

   Chapter: Reliability, Fairness (Week: 6)
   Reference: "Bias mitigation algorithms can be applied as pre-processing, in-processing, or post-processing within the machine learning pipeline."

6. **Create a factsheet documenting key characteristics and limitations of a machine learning model used for college admissions decisions. Include sections on data sources, model performance metrics, fairness considerations, and transparency mechanisms.** (300 words)

   Chapter: Interaction, Transparency (Week: 9)
   Reference: "Fact flow is a mechanism for automatically collecting qualitative and quantitative facts about a development lifecycle. A factsheet is a collection of facts that enables transparency and conformity assessment."

Section 4: Draw Connections Among Ideas (Analyze)

Short Answer Questions

1. **Analyze the trade-offs between model interpretability and model performance in the context of healthcare diagnosis. Consider both technical and ethical implications.** (300 words)

   Chapter: Interaction, Interpretability and Explainability (Week: 8)
   Reference: "Different consumers have different personas with different needs to achieve their goals. The important personas are the affected user, the decision maker, and the regulator."

2. **Examine how team composition affects different phases of the machine learning life-cycle. Provide specific examples of how diverse perspectives influence each phase.** (300 words)

   Chapter: Purpose, Lived Experience (Week: 12)
   Reference: "Team members with lived experience of marginalization have an epistemic advantage in noticing potential harms."

3. **Analyze how different fairness metrics might conflict with each other in a criminal recidivism prediction system.** (250 words)

   Chapter: Reliability, Fairness (Week: 6)
   Reference: "There are two important worldviews in determining which kind of sameness is most appropriate for your problem."

4. **Identify and analyze potential vulnerabilities to adversarial attacks in an autonomous vehicle system. Propose specific defensive measures.** (300 words)

   Chapter: Reliability, Adversarial Robustness (Week: 7)
   Reference: "Adversaries have different capabilities of what they know and what they can change. These differences in capabilities and goals determine the threat."

5. **Assess the impact of data quality and potential biases on model reliability in a natural language processing system.** (300 words)

   Chapter: Data, Data Modalities, Sources, and Biases (Week: 2)
   Reference: "No matter how careful one is, there is no completely unbiased dataset. Nevertheless, the more effort put in to catching and fixing biases before modeling, the better."

6. **Analyze the implications of different types of distribution shift for a financial fraud detection system. How would you adapt the system to each type?** (250 words)

   Chapter: Reliability, Distribution Shift (Week: 5)
   Reference: "Prior probability shift and covariate shift are easier to overcome than concept drift because they arise from sampling bias rather than measurement bias."

Section 5: Justify a Stance or Decision (Evaluate)

Short Answer Questions

1. **Evaluate the effectiveness of different explanation methods for various stakeholder personas in a healthcare diagnostic system. Consider the needs of doctors, patients, and hospital administrators. Justify which methods are most appropriate for each group and why.** (400 words)

   Chapter: Interaction, Interpretability and Explainability (Week: 8)
   Reference: "Different consumers have different personas with different needs to achieve their goals. The important personas are the affected user, the decision maker, and the regulator."

2. **Assess the appropriateness of using particular fairness metrics in a college admissions context. Evaluate how different worldviews ('we're all equal' vs 'what you see is what you get') would influence metric selection and implementation. Provide specific recommendations based on your analysis.** (400 words)

   Chapter: Reliability, Fairness (Week: 6)
   Reference: "If you believe there are social biases in measurement (not only representation biases in sampling), then you have the 'we're all equal' worldview; independence and statistical parity difference are appropriate notions of group fairness."

3. **Judge the effectiveness of implemented defense measures against adversarial attacks in a facial recognition system used for building security. Evaluate both technical robustness and practical limitations. Provide specific metrics for measuring effectiveness and justify your evaluation framework.** (400 words)

   Chapter: Reliability, Adversarial Robustness (Week: 7)
   Reference: "Defenses for evasion attacks include denoising that attempts to remove adversarial perturbations from inputs and adversarial training which induces min-max robustness."

4. **Evaluate the institutional trustworthiness of a machine learning platform used for content recommendation on a social media platform. Consider governance mechanisms, transparency measures, and alignment with relational ethics. Provide specific criteria for your evaluation and justify your conclusions.** (400 words)

   Chapter: Purpose, Filter Bubbles and Disinformation (Week: 13)
   Reference: "Institutional trustworthiness is based on governance mechanisms and their transparency, which can be required by government regulations if there is enough societal pressure for them."

Section 6: Produce New or Original Work (Create)

Pick One Task to Complete

Task 1: Reliability, Fairness
Design a comprehensive fairness assessment protocol for a hiring algorithm that aligns with stakeholder values. Your protocol should include:

- Assessment methodology for multiple fairness metrics
- Transparency reporting framework for different stakeholders
- Specific bias mitigation strategies with implementation details
- Monitoring plan for ongoing fairness evaluation (500 words)

Chapter: Reliability, Fairness (Week: 6)
Reference: "There are two important worldviews in determining which kind of sameness is most appropriate for your problem."

Create a comprehensive fairness testing protocol for a loan approval system. Design:

- Test scenarios covering multiple fairness definitions
- Data collection and validation procedures
- Evaluation metrics for different stakeholder groups
- Remediation protocols for identified issues (500 words)

Chapter: Reliability, Fairness (Week: 6)
Reference: "Bias mitigation algorithms can be applied as pre-processing, in-processing, or post-processing within the machine learning pipeline."

Task 2: Reliability, Adversarial Robustness
Develop an adversarial robustness analysis framework for an autonomous vehicle system. Create:

- Test case design for different types of attacks
- Monitoring protocols for detecting adversarial inputs
- Defense implementation plans with specific technical details
- Evaluation metrics for measuring robustness (500 words)

Chapter: Adversarial Robustness (Week: 7)
Reference: "Defenses for evasion attacks include denoising that attempts to remove adversarial perturbations from inputs and adversarial training which induces min-max robustness."

Task 3: Data, Privacy & Consent
Create a data privacy framework for a healthcare system that balances utility with privacy protection. Include:

- Privacy preservation methods selection criteria
- Implementation protocols for chosen methods
- Evaluation metrics for privacy-utility trade-offs
- Compliance monitoring system (500 words)

Chapter: Data, Privacy and Consent (Week: 2)
Reference: "Various methods can be used to preserve their privacy. Syntactic anonymization methods group together individuals with similar quasi-identifiers and then obfuscate those quasi-identifiers."

Task 4: Purpose, Social Good
Design a social impact assessment framework for a machine learning platform aimed at non-profit organizations. Develop:

- Impact measurement metrics

- Stakeholder engagement protocols
- Implementation guidelines for resource-constrained environments
- Success criteria aligned with social good (500 words)

Chapter: Purpose, Social Good (Week: 13)
Reference: "Machine learning platforms that are specifically designed to deskill data science needs and minimize the effort for deployment, maintenance, and support are the solution."

Task 5: Basic Modeling, Supervised Learning
Describe how you would evaluate a supervised binary classifier. Create:

- Steps for evaluating binary classifiers that incorporate both traditional performance metrics and fairness metrics
- Guidelines for selecting appropriate decision thresholds based on different real-world scenarios
- Integration plan for generalizing to multi-label models with different inductive biases (500 words)

Chapter: Basic Modeling, Supervised Learning (Week: 3–4)
Reference: "Different classifiers have different domains of competence: what kinds of datasets they have lower generalization error on than other methods."

Task 6: Interaction
Develop an interpretability framework for a medical diagnosis system that serves multiple stakeholders. Design:

- A multi-level explanation system that provides appropriate detail for different user personas (doctors, patients, administrators)
- Methods for generating both global and local explanations of model decisions
- Validation protocols for ensuring explanations are faithful to the underlying model
- Guidelines for presenting explanations in an actionable format (500 words)

Chapter: Interaction, Interpretability and Explainability (Week: 8)
Reference: "Different consumers have different personas with different needs to achieve their goals. The important personas are the affected user, the decision maker, and the regulator."

Create a transparency protocol that enables conformity assessment throughout the machine learning lifecycle. Develop:

- A comprehensive fact flow system that automatically collects both qualitative and quantitative information
- Standardized factsheet templates for different stakeholder groups
- Verification mechanisms for ensuring accuracy of collected facts
- Implementation guidelines for maintaining transparency across development phases (500 words)

Chapter: Interaction, Transparency (Week: 9)
Reference: "Fact flow is a mechanism for automatically collecting qualitative and quantitative facts about a development lifecycle. A factsheet is a collection of facts, appropriately rendered for a given consumer, that enables transparency and conformity assessment."

APPENDIX

C

COURSE POST- ASSESSMENT

C.1 Introduction

Your final is a Google form, because Moodle is annoying & it's easier to grade. Don't cheat. The final has 50 exam questions:

- 40 questions where all the content was pulled directly from the study guide
- 10 control questions you haven't seen before
- 2 extra credit questions

The exam begins with 26 Likert scale questions ranking your confidence in the course learning objectives. These questions are required but do not count toward your final grade.

C.2 Self-Efficacy

The following questions are Likert scale responses. Score your confidence for each objective where where 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree.

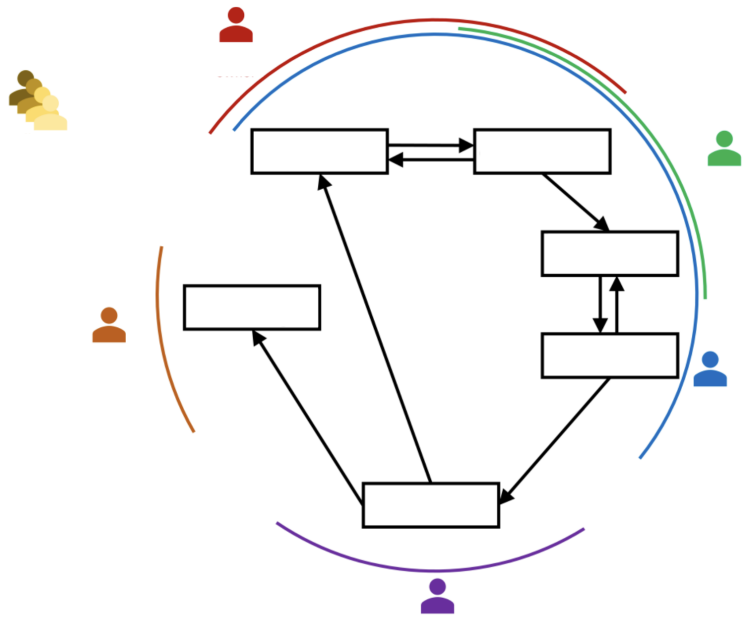
1. Email

2. The student is able to define the four attributes of trustworthiness in machine learning systems.
3. The student is able to list the six steps of the machine learning lifecycle.
4. The student is able to describe the socio-technical challenges of fairness, accountability, and bias in machine learning applications.
5. The student is able to determine (develop) ML models with considerations of data sources, bias, privacy, and consent.
6. The student is able to evaluate the ethical implications of ML deployment in real-world scenarios.
7. The student is able to identify the different types of distribution shifts in machine learning systems.
8. The student is able to distinguish between different types of fairness and their appropriate applications.
9. The student is able to describe the role of causality in machine learning decision-making.
10. The student is able to describe the role of causality in machine learning decision-making. *(Note: This question appears duplicated in the original)*
11. The student is able to explain the relationship between aleatoric and epistemic uncertainty in safety considerations.
12. The student is able to understand the socio-technical challenges of fairness, accountability, and bias in machine learning systems.
13. The student is able to evaluate supervised binary classifier performance.
14. The student is able to apply appropriate bias mitigation strategies based on specific worldviews.
15. The student is able to implement appropriate privacy preservation methods for different data scenarios.
16. The student is able to contextualize ML models with considerations of data sources, bias, privacy, and consent.

17. The student is able to analyze the implications of different types of adversarial attacks on machine learning systems.
18. The student is able to evaluate the effectiveness of different explanation methods for various stakeholder personas.
19. The student is able to examine the relationship between team composition and model outcomes.
20. The student is able to analyze the stages of the machine learning lifecycle to promote trust, safety, and transparency.
21. The student is able to assess the appropriateness of different fairness metrics for specific contexts.
22. The student is able to critique the effectiveness of various transparency mechanisms in machine learning systems.
23. The student is able to judge the institutional trustworthiness of machine learning platforms.
24. The student is able to evaluate the ethical implications of ML deployment, including impacts on marginalized and vulnerable populations.
25. The student is able to design governance frameworks for machine learning systems that align with stakeholder values.
26. The student is able to develop strategies for implementing trustworthy machine learning in social impact projects.
27. The student is able to construct appropriate testing protocols for machine learning models.

C.3 Remember Section: Recall facts and basic concepts

C.3.1 Machine Learning Lifecycle



28. List the names of the Machine Learning Lifecycle phases (in order)
29. List the titles of the people included in the Machine Learning Lifecycle (in order)
30. Name the attributes of Trustworthiness

C.3.2 TML and RAI Vocabulary

31. Which type of learning involves labeled data?
    - Reinforcement
    - Supervised
    - Unsupervised
    - Deep
    - Other: _____
32. What is the main goal of principal component analysis (PCA)?
    - Dimensionality reduction

- Regression
- Classification
- Bias mitigation
- Other: _____

33. Which algorithm is used for clustering?

    - Decision Tree
    - Logistical Regression
    - Random Forest
    - K-means
    - Other: _____

34. Which model is non-parametric?

    - K-Nearest Neighbors (KNN)
    - Linear Regression
    - Decision Tree
    - Complex Neural Networks
    - Other: _____

35. What term is defined by "a category of data defined by how it is received, represented, and understood"?

    - Structured data
    - Data modality
    - Semi-structured data
    - Unstructured data
    - Other: _____

36. What kinds of bias require intersectionality theory?

    - Social bias
    - Data preparation bias
    - Representation bias

- Temporal bias
- All of the above
- Other: _____

37. What are the two applications of random forests?

    - Dimensionality reduction & Encoding
    - Regression & Optimization
    - Classification & Regression
    - Bias mitigation & Fairness
    - Other: _____

38. What is the main application of K-Nearest Neighbor (KNN)?

    - Dimensionality reduction
    - Regression
    - Classification
    - Clustering
    - Other: _____

39. Which metric is not used for classification?

    - Accuracy
    - Recall
    - Precision
    - Mean Squared Error
    - Other: _____

40. What is a "supervised learning technique that aims to predict a continuous numerical value"?

    - Dimensionality reduction
    - Regression
    - Classification
    - Encoding

- Other: _____

41. What is a "supervised learning technique that aims to categorize data into discrete labels or classes"?

    - Dimensionality reduction
    - Regression
    - Classification
    - Encoding
    - Other: _____

42. What do you use to evaluate binary classification?

    - Accuracy
    - Confusion Matrix
    - False Negatives
    - F1 Score
    - Other: _____

43. Which type of uncertainty represents inherent randomness in phenomena and is well-modeled using probability theory?

    - Aleatoric uncertainty
    - Epistemic uncertainty
    - Stochastic uncertainty
    - Random uncertainty
    - Other: _____

44. Which document provides transparency and conformity assessment information?

    - Factsheet
    - Technical report
    - User manual
    - System log
    - Other: _____

45. Which fundamental concept serves as the basis for evaluating binary classifiers in machine learning systems?

    - true/false positives and true/false negative with trade-offs
    - raw accuracy metrics
    - processing speed
    - model size
    - Other: _____

C.4 Understand Section: Explain ideas or concepts

46. How do different worldviews influence the selection of fairness metrics in machine learning systems?

    - selection depends on whether one believes measurement or sampling bias dominates
    - worldviews have no impact on metric selection
    - only technical considerations matter for fairness metrics
    - fairness metrics are universal across all worldviews
    - Other: _____

47. Which approach best describes how socioculturally heterogeneous teams affect model development?

    - they tend to slow down and avoid taking shortcuts
    - they always work faster than homogeneous teams
    - they focus only on technical performance metrics
    - they prioritize speed over thoroughness
    - Other: _____

48. What is the primary purpose of machine learning platforms designed for social impact?

    - to empower social change organizations in developing ML systems that help uplift vulnerable people
    - to maximize technical performance metrics

- to increase processing speed
- to reduce computational costs
- Other: _____

49. What is the primary purpose of machine learning platforms designed for social impact? *(Note: This question appears duplicated in the original)*

    - to empower social change organizations in developing ML systems that help uplift vulnerable people
    - to maximize technical performance metrics
    - to increase processing speed
    - to reduce computational costs
    - Other: _____

C.5 Attributes of Trustworthiness

50. Which scenario is not related to Basic Performance:

    - A model correctly identifies pneumonia in chest X-rays with 95% accuracy.
    - An ML model achieves its core functional objective.
    - The model accomplished what it was designed to do.
    - A recommendation system that performs well for popular items but fails for niche interests.
    - Other: _____

51. What attribute is the unrelated scenario from the previous question actually related to?

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose

52. Which scenario is not related to Reliability:

    - A system's ability to handle adversarial attacks

- A well-performing facial recognition system used for surveillance without legal regulations
- A model works consistently across different populations and environments
- A facial recognition system performs well for all skin tones and ages
- Other: _____

53. What attribute is the unrelated scenario from the previous question actually related to?

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose

54. Which scenario is not related to Interaction:

    - A medical diagnosis model that performs equally well for patients from different socioeconomic backgrounds
    - A recommendation system that explains why it suggested a particular product exemplifies
    - How well the ML system integrates with human users
    - Users can understand why the model made a specific prediction
    - Other: _____

55. What attribute is the unrelated scenario from the previous question actually related to?

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose

56. Which scenario is not related to Purpose:

    - Ethical considerations and societal impacts of an ML model
    - An ML system designed to optimize engagement metrics without considering user well-being concerns

- An autonomous vehicle navigation system performs perfectly in ideal weather but struggles in snow and heavy rain.
- An AI system designed to reduce energy consumption in buildings
- Other: _____

57. What attribute is the unrelated scenario from the previous question actually related to?

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose

58. Metrics like equal opportunity difference, and statistical parity are most directly related to which attribute?

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

59. "Does this ML system contribute positively to society and align with human values?" addresses:

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

60. A fraud detection system that works accurately but disproportionately flags transactions from certain countries needs improvement in:

    - Basic Performance
    - Reliability
    - Interaction

- Purpose
- Other: _____

61. A healthcare AI that achieves high accuracy but operates as a "black box" that doctors can't interpret needs improvement in:

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

62. A company develops a state-of-the-art machine translation system that claims to support 95 languages. In testing, it handles major languages like English, Spanish, and Mandarin excellently, but produces gibberish or incorrect translations for 15 less common languages it claims to support. The primary issue here concerns:

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

63. An ML system for college admissions considers numerous factors and outperforms human reviewers in predicting student success, but researchers discover it reproduces historical biases against certain demographics. This primarily represents a failure in:

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

64. A company develops an AI system that perfectly detects early signs of dementia from speech patterns, explains its reasoning clearly to doctors, and works equally well across all demographics. However, it provides no way for patients to control their data or opt out of analysis. This system most needs improvement in:

- Basic Performance
- Reliability
- Interaction
- Purpose
- Other: _____

65. A production ML model's performance suddenly drops significantly when processing new data. During investigation, developers find that the distribution of input features has shifted compared to the training data. This situation most directly highlights a challenge in:

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

66. An AI hiring tool consistently selects qualified candidates and works well across demographics, but hiring managers report they don't understand its recommendations and often override them. The company should primarily focus on improving:

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

67. An AI hiring tool consistently selects qualified candidates and works well across demographics, but hiring managers report they don't understand its recommendations and often override them. The company should primarily focus on improving: *(Note: This question appears duplicated in the original)*

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

C.6 Control/Challenge Questions

These are questions you haven't seen before. They aren't designed to trick you! Don't overthink them.

68. What are the attributes of Safety?

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose

69. A smart home energy management system accurately optimizes energy usage, performs consistently in various home environments, and aligns with sustainability goals. Users report that the system's recommendations are easy to understand, but they find it difficult to override the system's decisions when they have preferences that differ from the optimal energy-saving settings. Which attribute needs improvement?

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

70. An autonomous manufacturing robot achieves high precision in assembly tasks and operates consistently in various factory environments. It has intuitive controls for human operators and clear safety protocols. However, it was designed to maximize production speed without considering worker displacement or sustainable production practices. The primary issue concerns:

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

71. A climate modeling system achieves state-of-the-art prediction accuracy, performs consistently across different regions, and was designed to help communities prepare for climate change. However, its complexity makes it accessible only to expert users, limiting adoption by local governments with fewer technical resources. This system most needs improvement in:

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

72. A social media content moderation system accurately detects policy violations in English, Spanish, and Mandarin but struggles with less common languages. It explains its decisions to moderators and was designed with user safety as its primary objective. Which TWO attributes need the most improvement?

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

73. A medical diagnosis AI system achieves 99% accuracy on its test dataset, but when deployed in hospitals, doctors notice it makes systematic errors when analyzing images from older equipment. The system provides detailed explanations for its diagnoses and was developed with patient welfare as the primary goal. Which attribute requires the most urgent improvement?

    - Basic Performance
    - Reliability
    - Interaction
    - Purpose
    - Other: _____

74. When stakeholders reject a model with high accuracy, the fundamental issue is usually in the:

    - Deployment and monitoring stage
    - Modeling stage
    - Problem specification stage
    - Evaluation stage
    - Other: _____

75. Feature importance analysis shows the model relies heavily on a feature that engineers consider irrelevant. This indicates a problem primarily in the:

    - Modeling stage
    - Problem specification stage
    - Data understanding stage
    - Evaluation stage
    - Other: _____

76. Adversarial robustness testing would primarily occur in the:

    - Modeling stage
    - Problem specification stage
    - Data understanding stage
    - Evaluation stage
    - Other: _____

77. Data augmentation/transformation would primarily occur in the:

    - Modeling stage
    - Data preparation stage
    - Problem specification stage
    - Data understanding stage
    - Other: _____

C.7 Extra Credit

78. Consent is ...

79. List as many related vocabulary words as you can

APPENDIX

D

COURSE POST- ASSESSMENT: AI POLICY CASE STUDY

This case study assignment was adapted from the 2025 Duke Sanford Public Policy Case Competition scenario, references, and evaluation materials.

Select one case study

Introduction and Framing

The following prompts are challenging interdisciplinary issues that do not just involve technology and our use of it. What you will read today spans tech policy, labor policy, mental health policy and other important issues that relate back to our use of current and evolving AI technology and the results of it on our society. We ask that you keep an open mind and challenge yourself to consider policy solutions that are creative and outside of the box.

Keep in mind that policy making should not depend on the development of new technology. Your focus should be on developing a policy that prioritizes and protects your stakeholder's interests in the short/mid term, considering the abilities and attributes of AI technology models that are currently available. A strong policy solution will think creatively about how public and

private actors can and should govern AI technologies and what policy proposals will best serve your stakeholder's position.

Remember that everyone here comes from a different background and has varying levels of expertise and ways of thinking. We ask that you treat each other with respect throughout this case competition. This scenario does include a sensitive topic (TW; mention of suicide and death), so please be mindful of the lived experiences of your peers, as we do not know what everyone is coming to the table with.

D.1 Case Study 1: Data Moderation and Human Labor Behind AI

D.1.1 Background and Scenario

With the rapid development of AI technology, social media companies have increasingly incorporated AI into their content moderation process as algorithms can analyze vast amounts of data at incredible speeds, efficiently handling the massive volume of user-generated content uploaded every second. This capability allows social media platforms to process text, images, and videos in real-time, quickly identifying potentially harmful or inappropriate content.

One such platform is Sprout. Sprout, a social media platform operated by the US-based company Omega Inc., has a global presence and allows users to share photos and videos on a feed while also enabling personal messaging, commenting, and responding. With a rapidly growing user base of 2 billion users and over 18 million pieces of content generated daily, the company has invested extensive resources—both time and money—into content moderation efforts over the past six months.

Currently, Sprout employs a hybrid moderation approach that uses a Large Language Model (LLM) as their major automated tool to conduct initial filtering and then has human moderators conduct an additional check. The LLM service is bought from a US-based third party. Human moderators are a very necessary part of the moderation process, despite the advancements in the automation used for content moderation. Sprout subcontracts this work to Epsilon Digital Services, which currently employs over 20,000 human moderators in India and the Philippines.

Content missed by initial AI filtering can cause significant harm to human content moderators who are constantly exposed to hate speech, graphic images and videos, and other content that has the potential to cause them severe psychological damage.

Recently, a content moderator from the Philippines was subcontracted by Epsilon Digital Services to moderate Sprout's content. Near the end of his second year moderating, he communicated in writing to his direct supervisor that the content he was exposed to was

inflicting mental harm to him. His communications with his direct supervisor did not lead to any changes in the amount or type of content he was exposed to. He had previously mentioned to his family that the company provided "mental health check-ins" to its content moderators for 15 minutes after every shift, but he believed the provided resources were inadequate and needed improvement. Unfortunately, the moderator died by suicide later that year, and his family, friends, and coworkers strongly believe his death to be a result of the onslaught of violent and inappropriate content he was viewing while on the job.

This employee's suicide has sparked an international outcry, particularly amongst Sprout's own user base. There are widespread, growing calls for U.S.-based social media platforms to take accountability and shift how they treat their human content moderators.

The U.S. Senate's Commerce, Science, and Transportation Standing Committee is holding a hearing to allow two key stakeholders to present their recommendations on what policies Congress should consider to address this complex issue. The invited stakeholders are WorkersUnderAI, a human rights advocacy group, and representatives from Sprout's parent company, Omega Inc.

Today, your team will work together and present your proposed policy as one of these stakeholders, taking care to address the following two points:

- propose policy measures that seek to prevent or limit the harms suffered by human content moderators, and
- propose an overall regulatory framework the U.S. federal government should consider towards U.S.-based social media platforms and their content moderation approaches (considering both AI-powered and human content moderation)

D.1.2 Featured Stakeholders

Your team will represent one of the two following stakeholders. Please ensure that the policy recommendation that your team formulates reasonably represents the viewpoint of your assigned stakeholder. Remember, your presentation should be addressed to the Senate Commerce, Science, and Transportation Committee which will evaluate your policy recommendation (represented by your instructor).

Omega Inc. Representatives

This group of representatives include Omega Inc's AI Safety Policymakers and Financial departments. Together this group wants to assure users and the public that the company is taking

this issue seriously, while also keeping profit as the top priority. The company is also aware that it could face retaliation from their user base if this issue is not handled tactfully.

WorkersUnderAI (human rights advocacy group)

WorkersUnderAI is a U.S. based international human rights advocacy group that engages with a wide range of issues at the intersection of human rights and technology. They often advocate for corporate accountability to uphold human rights in digital labor practices. In response to the tragic death of the human content moderator, this group wants to highlight the human cost of prioritizing profit over worker safety and health while also recognizing that human moderator roles provide income to workers in low and middle income countries.

D.2 Case Study 2: Autonomous Vehicle Decision-Making

D.2.1 Background and Scenario

In recent years, autonomous vehicle (AV) technology has advanced significantly, with numerous companies bringing self-driving cars to market. One leading company, AutoDrive Inc., has deployed over 50,000 autonomous vehicles across 15 major U.S. cities. Their AI-powered navigation system, "PathPilot," is trained on billions of miles of driving data and uses a complex neural network to make split-second decisions.

In March 2024, a tragic incident occurred in Phoenix when an AutoDrive vehicle faced an unavoidable collision scenario. The vehicle's AI system had to make a rapid decision between two harmful outcomes: swerving into a cyclist wearing a helmet or hitting a pedestrian crossing against the light. The AI chose to swerve toward the helmeted cyclist, resulting in a fatal collision. Subsequent investigation revealed that the AI's decision-making algorithm had been trained to minimize overall harm, and had assigned a lower "injury probability" to the helmeted cyclist compared to the pedestrian.

This incident, dubbed "CollisionGate" by the media, has sparked intense debate about the ethical frameworks embedded in autonomous vehicle decision-making systems. After a preliminary investigation, it was discovered that AutoDrive's training data included a proprietary "harm reduction matrix" that assigned different values to potential collision targets based on factors like protective gear, age, and compliance with traffic laws. AutoDrive initially defended this as responsible engineering but later admitted they had not disclosed these decision-making factors to regulators or the public.

Consumer advocacy groups have called for greater transparency in how autonomous vehicles make life-or-death decisions. Meanwhile, insurance companies have raised concerns

about liability determination in accidents involving AI decision-making. The families of victims have filed lawsuits, arguing that they never consented to a system that makes value judgments about human lives.

The U.S. National Highway Traffic Safety Administration (NHTSA) has temporarily suspended AutoDrive's operating permits pending a full investigation, and Congress is now considering legislation to regulate the decision-making frameworks of autonomous vehicles.

The Senate Transportation Committee is holding hearings to allow key stakeholders to present their recommendations on what policies should govern autonomous vehicle decision-making algorithms. The invited stakeholders are the Consumer Safety Coalition and representatives from AutoDrive Inc.

Today, your team will work together and present your proposed policy as one of these stakeholders, taking care to address the following two points:

- Propose policy measures that establish ethical frameworks and transparency requirements for AI decision-making in unavoidable harm scenarios.
- Propose an overall regulatory framework the U.S. federal government should consider for autonomous vehicle AI systems, including testing requirements, disclosure standards, and accountability measures.

D.2.2 Featured Stakeholders

Your team will represent one of the two following stakeholders. Please ensure that the policy recommendation that your team formulates reasonably represents the viewpoint of your assigned stakeholder. Remember, your presentation should be addressed to the Senate Transportation Committee which will evaluate your policy recommendation (represented by your instructor).

AutoDrive Inc. Representatives

This group includes AutoDrive's Chief AI Ethics Officer, Legal Counsel, and Engineering Leadership. They want to demonstrate their commitment to safety while preserving the company's competitive advantage and proprietary technology. They argue that excessive disclosure requirements could stifle innovation and that their overall safety record far exceeds human drivers. They are open to reasonable regulation but are concerned about extreme measures that might cripple the industry.

Consumer Safety Coalition

The Consumer Safety Coalition is a nationwide alliance of safety advocates, ethicists, and transportation experts. They believe that AI systems making life-or-death decisions must operate with complete transparency and public oversight. They advocate for mandatory disclosure of decision-making frameworks, public input on ethical guidelines, and rigorous testing requirements. They acknowledge the potential safety benefits of autonomous vehicles but argue that public trust requires democratic input into how these machines value human life.

D.3 Case Study 3: Healthcare Diagnosis Algorithm Bias

D.3.1 Background and Scenario

MediTech Solutions is a leading healthcare AI company that has developed DiagnosAI, an artificial intelligence system designed to assist physicians in diagnosing diseases and recommending treatment plans. The system has been implemented in over 2,000 hospitals and clinics across the United States, analyzing millions of patient records daily. DiagnosAI's machine learning algorithms were trained on a dataset of approximately 50 million historical patient records from various healthcare institutions.

In January 2024, researchers from the National Institute of Health published a study revealing significant disparities in DiagnosAI's performance across different demographic groups. The study found that DiagnosAI was 28% less accurate in diagnosing cardiovascular conditions in Black patients compared to white patients, and routinely recommended less aggressive treatments for female patients with identical symptoms to male patients. The investigation, dubbed "MedBiasGate" by the press, traced these disparities to imbalances in the training data, which contained predominantly records from white male patients and from healthcare facilities serving wealthier communities.

Further investigation revealed that MediTech was aware of some performance disparities during internal testing but determined they were "within acceptable parameters" and did not disclose these findings during the FDA approval process. The company claimed that their system was meant to be a "decision support tool" rather than a definitive diagnostic authority, and therefore some performance variations were to be expected.

The revelations have led to multiple class-action lawsuits from patients who allege they received delayed or inadequate care due to DiagnosAI's recommendations. Several hospital systems have temporarily suspended use of the system pending further evaluation. Healthcare equity advocates have pointed to this case as evidence of how algorithmic bias can amplify existing healthcare disparities, while industry representatives argue that even with its flaws,

the system improves overall care quality compared to purely human diagnosis.

The House Committee on Energy and Commerce's Subcommittee on Health is holding hearings to allow key stakeholders to present their recommendations on what policies should govern healthcare AI systems. The invited stakeholders are the Healthcare Equity Alliance and representatives from MediTech Solutions.

Today, your team will work together and present your proposed policy as one of these stakeholders, taking care to address the following two points:

- Propose policy measures to detect, prevent, and mitigate algorithmic bias in healthcare AI systems, particularly regarding diagnosis and treatment recommendations.
- Propose an overall regulatory framework the U.S. federal government should consider for healthcare AI systems, including data requirements, testing protocols, ongoing monitoring, and accountability measures.

D.3.2 Featured Stakeholders

Your team will represent one of the two following stakeholders. Please ensure that the policy recommendation that your team formulates reasonably represents the viewpoint of your assigned stakeholder. Remember, your presentation should be addressed to the House Subcommittee on Health which will evaluate your policy recommendation (represented by your instructor).

MediTech Solutions Representatives

This group includes MediTech's Chief Medical Officer, Head of Regulatory Affairs, and AI Research Director. They acknowledge the disparities identified in their system but maintain that DiagnosAI still improves overall healthcare outcomes compared to human-only diagnosis. They are concerned that overly stringent regulations could impede innovation and delay life-saving technologies from reaching patients. They prefer industry-led standards and voluntary certifications rather than mandatory requirements. They are willing to improve their systems but want flexibility in implementation timelines and methods.

Healthcare Equity Alliance

The Healthcare Equity Alliance is a coalition of patient advocacy organizations, medical professionals, and public health experts focused on reducing disparities in healthcare access and outcomes. They argue that algorithmic bias in healthcare AI systems represents a civil rights issue, as it can systematically deny equal care to already marginalized populations. They

advocate for rigorous pre-market testing across diverse populations, mandatory reporting of performance metrics by demographic group, and strong accountability mechanisms for systems that demonstrate bias. They believe the FDA should expand its regulatory framework to specifically address algorithmic bias in medical AI.

Presentation Information

All groups will submit a slide deck presentation. Presentations must be less than 8 min (10 content slides) in length. The slide decks will not be presented but should have presenter notes outlining the key statements for each slide.

Evaluation Criteria

Background

- Background and context of existing precedents and research is provided.
- An argument for why change is necessary, building upon the context given.

Policy Recommendation + Counter Arguments and Limitations

- Must be representative of the interests of your stakeholder. The effects on and roles of other stakeholders involved and what they will likely argue should be considered.
- Make your recommendations specific, clear, and understandable.
- Relevant legislative, judicial, and regulatory policy factors affecting the policy should be addressed.
- Address and explain the limitations and conflicts related to your policy.

PRESENTATIONS MUST BE SUBMITTED BY [Date] [Time] TO [Contact Email] IN GOOGLE SLIDES FORMAT. NO EDITS WILL BE ALLOWED AFTER THIS TIME.

APPENDIX

E

COURSE SYLLABUS

CSC 495 Section 004 Course Syllabus
Introduction to Responsible Machine Learning
Spring 2025

Instructor Information

Name	Email	Office Location
Ms. Lauren Alvarez (Lead Instructor)	lalvare@ncsu.edu	Venture 4
Dr. Veronica Cateté	vmcatete@ncsu.edu	Venture 4, Room 410
Dr. Munindar Singh	mpsingh@ncsu.edu	EB II, Room 2252

Office Hours

Wednesday 2-4:30 PM
Venture IV, room 410
and via Zoom

Ms. Alvarez is available for in-person or virtual office hours on Wednesday from 2-4:30 PM. In-person office hours will be held at Venture IV, Office 410. After three consecutive weeks without any in-person appointments, all office hours will transition to a virtual-only format. We recommend using office hours to discuss course materials, seek feedback on assignments, clarify concepts, or address any course concerns.

Preferred Method of Communication & Response Time

- **Preferred method of communication:** If you need to contact me directly, my preferred method of communication is email. You can expect to receive a response within two business days (i.e. not over the weekend). If I email you directly, please strive to respond within two business days. It is recommended that you check your NC State email at least once a day to stay on top of course communications.
- **Asking questions about the course:** If you have questions about the course or its content, you can email me or post your questions on the Student Help Forum in Moodle. You can expect to receive a response within two business days (i.e. not over the weekend).
- **Email guidelines:** Always include a descriptive, specific, but concise subject. Include your course number and section in your email, and provide adequate context for your question to ensure full understanding of your email. Be sure to use your NC State email account, and sign in with your name and Student ID number.

Course Information

Course Website: NC State WolfWare

Meeting Time and Location:

Tuesday and Thursday
01:30 PM -02:45 PM
02015 Engineering Building I

Course Credit Hours: 3

Catalog Description

This course provides a comprehensive exploration of the principles and practices underlying trustworthy machine learning (ML). It examines critical issues of responsibility, accountability, trustworthiness, and fairness, with a strong focus on both technical and socio-technical aspects. Key topics include the lifecycle of machine learning systems, interpretable machine learning

(IML), explainable artificial intelligence (XAI), privacy preservation, bias mitigation, causal modeling, and the ethical implications of automation. The course emphasizes practical applications such as handling distribution shifts, enhancing adversarial robustness, and aligning ML systems with stakeholder values, particularly in high-risk domains like judicial, criminal, financial, welfare, and healthcare systems.

Additionally, students will critically evaluate the societal impacts of ML, design governance frameworks, and develop methods for implementing ML solutions that promote social good and uphold ethical principles. Throughout, the course integrates diverse perspectives to contextualize ML within broader societal contexts and centers on actionable fairness, transparency, and value alignment strategies.

Students will engage in lectures, discussions, and applied projects to meet these learning objectives. Prerequisites include CSC 216 or equivalent, and the course strongly encourages participation from individuals interested in interdisciplinary approaches to AI ethics and technical integrity.

Structure

- The majority of this course is **synchronous**, delivered through real-time, face-to-face class sessions.
- Additional materials and activities are delivered through **Moodle**, a secure and easy-to-use online learning platform.

Learning activities in this course will include reading assignments, quizzes, case studies, PowerPoint presentations, discussion forums, and individual and group projects.

This course is based on *Trustworthy machine learning* (Varshney, 2022). We will cover topics in 1-2 chapters of this text each week. Weeks are considered Sunday to Saturday.

Intellectual Organizing Principle

The intellectual organizing principle guiding this course is "Trustworthy Machine Learning as a Socio-Technical Imperative." It integrates technical rigor with ethical and societal considerations, emphasizing the lifecycle approach to building machine learning systems that are accountable, fair, reliable, and transparent. This organizing principle ensures students are skilled in ML techniques and thoughtful about their broader implications, fostering a culture of ethical innovation and interdisciplinary problem-solving. This principle manifests in the following important ways:

1. **Lifecycle-Centric Trustworthiness:**
   The course anchors its curriculum in the machine learning lifecycle—problem specification, data preparation, modeling, evaluation, and deployment—while ensuring trustworthiness attributes like fairness, privacy, safety, and robustness are incorporated at every stage, rather than being added as afterthoughts.

2. **Socio-Technical Integration:**
   It bridges technical methods (e.g., bias mitigation algorithms, adversarial robustness techniques) with societal impacts (e.g., ethical governance, inclusion of marginalized perspectives, value alignment) to highlight the interplay between algorithms and the human contexts they operate in.

3. **Diversity of Perspectives:**
   The course embraces heterogeneity—both in team composition and fairness worldviews (e.g., "we're all equal" versus "what you see is what you get")—to underscore the importance of lived experiences and contextual understanding in creating ethical ML systems.

4. **Principled Decision-Making:**
   By framing fairness, transparency, and accountability as context-sensitive and multidimensional, the course encourages students to adopt principled approaches to balancing trade-offs among competing metrics, goals, and stakeholder needs.

5. **Actionable Ethics:**
   Moving beyond theoretical discussions, the course focuses on equipping students with practical skills to design, test, and implement ML systems that align with stakeholder values and promote social good while meeting technical performance standards.

Meeting Time and Tool Used

We will meet in person twice a week on Tuesdays and Thursdays at the class time. All activities are listed on Moodle. Anything not completed during the in-person session should be finished asynchronously before the next class period unless otherwise noted. In the event of adverse weather conditions, class may be held synchronously via zoom.

Remember, you are expected to spend double the class time on assignments and activities. For this three-credit hour class, plan to spend around 280 minutes (4.5 hours) per week outside of our class meetings.

Prerequisites/Corequisites

CSC 216: Software Development Fundamentals
Optional: Python Experience

Minimum Technical and Digital Information Literacy Skills

Required technical Skills

- Familiarity with at least one programming language (Python, Java, etc).
- Navigate and use Moodle, NC State's Learning Management System.
- Use Gmail, including attaching files to email messages
- Create and submit files in commonly used word processing program formats (MS Word, text editors, Google Docs).
- Download and install software as needed (see section on required software)
- Download and upload attachments
- Use spreadsheets, presentations, graphics programs, and other applications in digital environments
- Use web conferencing tools, including Zoom and Google Meet.
- Post to discussion boards and forums

Required digital information literacy skills

- Perform online research using various search engines and library databases. Visit NC State University Libraries.
- Use computer networks to locate and store files or data.
- Use online search tools for specific academic purposes, including the use of search criteria, keywords, and filters
- Analyze digital information for credibility, currency, and bias (e.g. disinformation, misinformation)
- Properly cite information sources

General Education Program (GEP) Information

GEP Category Fulfilled

None

GEP Corequisites

None

Learning Outcomes

Upon completion of this course, students will be able to:

1. Define the four attributes of trustworthiness in machine learning systems
2. List the six steps of the machine learning lifecycle
3. Describe the socio-technical challenges of fairness, accountability, and bias in machine learning applications
4. Determine ML models with considerations of data sources, bias, privacy, and consent
5. Evaluate the ethical implications of ML deployment in real-world scenarios

Course Materials

Required textbook

- Varshney, K. R. (2022). Trustworthy Machine Learning. Independently Published. `http://www.trustworthymachinelearning.com`. (available for free at the link)

Other required materials

- If needed, any supplementary papers, readings, or resources may be provided on Moodle. Access to Python programming tools for relevant exercises.

Optional materials

None

Technology Requirements

NC State University Libraries offers Technology Lending, where many devices are available to borrow for a 7-day period. Computer labs are available in various locations around campus for student use.

Computer

A laptop computer is required for students taking this course. NC State's Office of Information Technology provides recommendations for your computer at NC State.

Other devices

None

Software and digitally-hosted course components

The following software and tools will be used in this course. Some tools are a part of NC State's enterprise tools. See information about their purpose, how to access them, accessibility information, and privacy policies. The same information for any other tools required in this course is provided in the list below.

Google Colab

- **Purpose:** email, calendaring, cloud storage, docs, slides, sheets, groups and more. Learn about Google Workspace
- **How to access:** Login to Google apps with your UnityID@ncsu.edu and Unity Password.
- **Accessibility:** Google Workspace user guide to accessibility
- **Privacy:** Privacy & Terms – Google

Other Student Expenses

None

Communication Guidelines

Respecting our learning community

The NC State Code of Student Conduct outlines expectations for behavior in the classroom (whether virtual or physical) and the consequences for students who violate these expectations.

Any behavior that impacts other students' ability to learn and succeed will be addressed, but expressing diverse viewpoints and interpretations of course content is welcome.

Community guidelines for this course include:

- Use a respectful tone in all forms of communication (email, written, oral, visual)
- Maintain professionalism (avoid slang, poor grammar, etc.) in your written communication.
- Respect regional dialects and culturally embedded ways of oral communication.
- Stay home or in your dorm room if you are exhibiting symptoms of a contagious illness (fever, chills, etc.).
- Enter our virtual and/or physical classroom community respectfully by refraining from lewd or indecent speech or behavior, helping to maintain a safe physical environment, not using your cell phone for voice or text communication except when explicitly given leave to do so, and not attending class under the influence of any substance.
- Treat each community member with respect by not recording others without their consent or engaging in any form of hazing, harassment, intimidation, or abuse.
- Respect cultural differences that may influence communication styles and needs.

Plan for interaction between instructors and students

Course announcements will be posted on Moodle and sent to bypass individual's email preferences to ensure visibility in a timely manner. Feedback will be provided with-in two weeks of assignment completion. Feedback will be delivered via Moodle assignments.

Expectations for learner participation and interaction

Some course activities including synchronous class sessions, Moodle Forums, Yellowdig, and group work will require you to interact with other students in the course. Communication expectations including frequency and content will be detailed in the information about each assignment or activity when it appears in the course.

Percentage of grade	Component	Details and timing of feedback
20%	Class Participation & Attendance	• One daily class exit ticket and full attendance (see attendance policy) are required to earn full credit. • You will receive a grade/feedback every 3 weeks.*
25%	Homework (Theoretical Considerations)	• 3-5 individual assignments that connect to the 5 general sections of the Trustworthy ML book: Introduction, Data, Basic Modeling, Reliability, Interaction, and Purpose. • Grading criteria are included within each assignment. • You will receive a grade/feedback within one week of the due date.*
30%	Project & Deliverables (Programming Implementation)	• 1 major partner/group project with 3-4 incremental due dates throughout the semester. • You will receive a grade/feedback within 2 weeks.*
25%	Final Exam	• There will be one final comprehensive exam. • A practice exam or suggested study guide will be provided 2 weeks before the exam date. • You will receive a grade/feedback within 2 weeks of submitting your project.*

Table E.1: Grading criteria, details, and timing of feedback *modifications to the timing of grades/feedback, if required, will be announced via email.

Grading and Feedback

Grading criteria, details, and timing of feedback

Grading scale

This course uses this grading scale:

Requirements for earning a grade of "Satisfactory"

If you are taking this course for credit only (S/U), your grade will be reported as S (Satisfactory) when coursework is equivalent to a C- or better or U (Unsatisfactory) when coursework is equivalent to less than a C-. For more information, see the Credit Only Courses regulation.

Low	Letter	High
97	A+	100
93	A	97
90	A-	93
87	B+	90
83	B	87
80	B-	83
77	C+	80
73	C	77
70	C-	73
67	D+	70
63	D	67
60	D-	63
0	F	60

Requirements and procedures for auditing this course

Auditing this course is approved on a case-by-case basis. Please contact the course instructor to attain approval. Refer to the Audit regulation for more information and links to required forms.

Course Schedule

Course Policies

Late assignments

Late assignments will be accepted with a 10% daily penalty – up to 7 days. Assignments submitted later than 1 week past the original due date will NOT be accepted unless prior arrangements are made.

Incomplete grades, withdrawals

Information on incomplete grades can be found at REG 02.50.03 – Grades and Grade Point Average. If you encounter a serious disruption to your work not caused by you and you would have otherwise successfully completed the course, contact your instructor as soon as you can to discuss the possibility of earning an incomplete in the course for the semester, including an agreement on when the remaining work must be done in order to change the grade to the appropriate letter grade. If a student must withdraw from a course or from the University due to hardship beyond their control, see Withdrawal Process and Timeline | Student Services Center for information and instructions.

Attendance

For complete attendance and excused absence policies, please see `http://policies.ncsu.edu/regulation/reg-02-20-03`

Attendance Policy

Your full attendance in class is required. From Academic Policy and Regulation REG02.20.3 - Attendance Regulations:

- Full participation in classes, laboratory periods, and examinations is expected of all students.
- Teachers in 100 and 200-level courses must keep a record of attendance throughout the semester. Instructors may use reasonable academic penalties commensurate with the importance of the work missed because of unexcused absences.
- Valid excuses for anticipated absences must still be cleared with the instructor before the absence.
- Valid excuses for emergency absences must be reported to the instructor as soon as possible, but not more than one week after the return to class.

To account for unforeseen circumstances, a student's first two absences will not be penalized. After that, each unexcused absence will result in a 10% deduction from a student's Attendance and Participation grade. In the case of an unexcused absence, students will not be allowed to make up participation/attendance. Any other work that is missed as a result of an unexcused absence can be submitted late, but it will be graded according to the Late Work policy.

Absences Policy

Academic Policy and Regulation REG02.20.3 - Attendance Regulations lists valid circumstances for an excused absence. Please also note:

a) For non-emergency situations, you must request advance permission to miss a class. If you must miss a class then, in all situations where it is practical, communicate (in person, phone, email) with the instructor before the class to get agreement (examples: you are not feeling well and wish to miss the class; you have a planned trip; you have a court appearance).

b) For emergency situations where you cannot communicate with the instructor in advance, you must contact the instructor as soon as you return to school (within 1 week of returning).

Emergency absences must be certified in writing. Emergency illness that was not discussed with the instructor before missing the class must be certified by a physician in writing.

Additionally, according to NCSU REG 02.20.03 - Attendance Regulations, item 3.3, "a maximum of four excused absences is allowed per term. Thus a maximum total combination of 4 make-up quizzes plus tests plus assignments will be given. More than 4 make-up quizzes and tests and assignments will only be considered upon written recommendation from the University Counseling Center."

Makeup Work Policy

Class Goal: You should master all learning objectives, even if it takes multiple tries. Your grade should reflect your knowledge, not how quickly you got there.

- If you lose points on an assignment, you can make it up for full credit.
  - This applies to labs, projects and writing assignments (not quizzes; see below).
- Simply correct the assignment (asking for help if needed), and then come to office hours with the instructor or a TA.
  - This must be done outside of class.
- In office hours, demonstrate how your code now meets the rubric requirements for the assignment.
- We will update your grade for full credit.

A make-up quiz or test will only be given under these conditions:

- You had advance permission from the instructor to miss class as described in point (a) of the Absences Policy above.
- You have a verified excused absence for an emergency situation as described in point (b) of the Absences Policy above.

University Policies

Academic integrity and honesty

Students are required to comply with the university policy on academic integrity found in the Code of Student Conduct 11.35.01 sections 8 and 9. Therefore, students are required to uphold

the Pack Pledge: "I have neither given nor received unauthorized aid on this test or assignment." Violations of academic integrity will be handled in accordance with the Student Discipline Procedures. Please refer to the Academic Integrity web page for a detailed explanation of the University's policies on academic integrity and some of the common understandings related to those policies.

Use of AI Tools and Unauthorized Help Policy

Unless specifically permitted by the assignment, students may not use unauthorized help to complete programming programs or other course assignments. Doing so will result in an academic integrity violation. Unauthorized sources of help (unless explicitly permitted) include: (A) AI-based tools (e.g., ChatGPT, Bing AI, etc.); (B) Copying solutions from the web (without explicit TA/instructor approval); (C) Copy other students' work outside of assigned teams.

Student privacy

Originality Checking Software

Software is not used in this course to detect the originality of student submissions.

Class recording statement: Students will not be able to be identified in any course recordings, or the course will not be recorded at all.

Class privacy statement: This course requires online exchanges among students and the instructor, but NOT with persons outside the course. Students may be required to disclose personally identifiable information to other students in the course, via electronic tools like email or web postings, where relevant to the course. Examples include online discussions of class topics and posting of student coursework. All students are expected to respect the privacy of each other by not sharing or using such information outside the course.

Other Policies

Students are responsible for reviewing the NC State University PRR's which pertain to their course rights and responsibilities: Equal Opportunity and Non-Discrimination Policy Statement and additional references; Code of Student Conduct; Grades and Grade Point Average; Credit-Only Courses; and Audits.

Student Resources

Academic and Student Affairs maintains a website with links for student support on campus, including academic support, community support, health and wellness, financial hardship or

insecurity, and more. Find Help on Campus.

Disability resources

Reasonable accommodations will be made for students with verifiable disabilities. In order to take advantage of available accommodations, students must register with the Disability Resource Office (DRO). For more information on NC State's policy on working with students with disabilities, please see the Policies, Rules and Regulations page maintained by the DRO and REG 02.20.01 Academic Accommodations for Students with Disabilities.

Safe at NC State

At NC State, we take the health and safety of students, faculty and staff seriously. The Office of Equal Opportunity supports the university community by providing services and resources to support and guide individuals in obtaining the help they need. See the Safe at NC State webpage for resources.

Supporting Fellow Students in Distress

As members of the NC State Wolfpack community, we each share a personal responsibility to express concern for one another and to ensure that this classroom and the campus as a whole remain a healthy and safe environment for learning. Occasionally, you may come across a fellow classmate whose personal behavior concerns or worries you, either for the classmate's well-being or yours. If you feel this way, I would encourage you to report this behavior to the NC State CARES website. Although you can report anonymously, it is preferred that you share your contact information so they can follow up with you personally.

Course Evaluations

ClassEval is the end-of-semester survey for students to evaluate the instruction of all university classes. The current survey is administered online and includes 12 closed-ended questions and 3 open-ended questions. Deans, department heads, and instructors may add a limited number of their own questions to these 15 common-core questions.

Each semester students' responses are compiled into a ClassEval report for every instructor and class. Instructors use the evaluations to improve instruction and include them in their promotion and tenure dossiers, while department heads use them in annual reviews. The reports are included in instructors' personnel files and are considered confidential.

Online class evaluations will be available for students to comple

Week	Date Range	Topic	Readings	Activities
Week 1	Jan 6-10	Course Introduction and Preliminaries	Moodle Intro	Course Pre-assessment Note: Jan 10 Last day to add a course without permission
Week 2	Jan 13-17	Preliminaries & Data Topics: Machine Learning Lifecycle, Data modalities, Sources, & Bias, Privacy & Consent	Ch. 1-2	Read the syllabus and ask any questions Note: Jan 17 Census Date/Official Enrollment Date
Week 3	Jan 20-24	Sociotechnical Real-World Examples Topics: Consent, Privacy, Sociology	Ch. 3	Homework 1 Due Jan 23 Jan 23 Snow Day Note: Jan 20 Martin Luther King Jr. Day (University closed; No classes)
Week 4	Jan 27-31	Basic Modeling Inverted classes with in-class labs	Ch. 4	Project Introduction
Week 5	Feb 3-7	Basic Modeling Inverted classes with in-class labs	Ch. 5	Homework 2 Due Feb 7
Week 6	Feb 10-14	Reliability Topics: Review Basic Modeling, Distribution Shift	Ch 6	Note: Feb 11 Wellness Day (No classes) 1st Project Deliverable Due Feb 13 (delayed to Feb 18)
Week 7	Feb 17-21	Reliability Topics: Distribution Shift	Ch. 7	Feb 20 Snow Day
Week 8	Feb 24-28	Reliability Topics: Fairness, Adversarial robustness, Vocab Review	Ch. 8	
Week 9	Mar 3-7	Reliability & Interaction Topics: Fairness, Adversarial robustness, Interpretability, Explainability, Transparency	Ch 9-11	Notes: Mar 3 Drop/Revision Deadline Mar 5 Schedule for 2025 Fall term published; Shopping Cart opens 2nd Project Deliverable Due March 20th
Week 10	Mar 10-14		None	Note: Spring Break (No classes)
Week 11	Mar 17-21	Aysnc Week Topics: Interpretability, Explainability, Transparency	Ch. 12-13	
Week 12	Mar 24-28	Aysnc Week Topics: Interpretability, Explainability, Transparency	Ch. 14	Note: Look out for an announcement about this week.
Week 13	Mar 31-Apr 4	Purpose Topics: Ethics Principles, Lived Experience, Social Good	Ch. 15-16	Homework 3 Due April 3
Week 14	Apr 7-11	Advanced Material Topics: Attributes of Trustworthiness & Safety	Ch. 17-18	Final Project Due April 10 (no extensions) Notes: If the class is interested in other advanced topics, we'll take a collective vote.
Week 15 (LWOC)	Apr 14-18	Review Week Topics: Final Review		Notes: Apr 16-22 Last week of semester Apr 22 Last Day of classes
Week 16	Apr 21-25	April 24 12:00-2:30 Final Exam		Notes: Apr 22 Last Day of classes Apr 23 Reading Day Apr 24-30 Final Exams
Week 17	Apr 28-May 2			Notes: May 2 Spring grades due by 5 pm