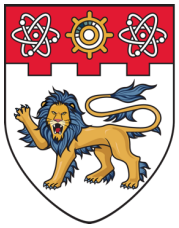
Methodology and Tools For Designing Ethical Artificial Intelligence Systems



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Zhang Jiehuang

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2023

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

16/09/2022

Date

Zhang Jiehuang

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

16/09/2022

Date

Prof Yu Han

Authorship Attribution Statement

This thesis contains material from [5] paper(s) published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 2 is published as J. Zhang, Y. Shu & H. Yu, "Human-Machine Interaction for Autonomous Vehicles: A Review," in Proceedings of the 13th International Conference on Social Computing and Social Media (SCSM'21), pp. 190–201, 2021.

The contributions of the co-authors are as follows:

- Prof Yu Han provided the initial paper direction and edited the manuscript drafts
- I wrote the drafts of the manuscript. The manuscript was revised and proofread together with Shu Ying.

Chapter 3 is published as Y. Shu, J. Zhang & H. Yu, "Fairness in Design: A Tool for Guidance in Artificial Intelligence Design," in Proceedings of the 13th International Conference on Social Computing and Social Media (SCSM'21), pp. 500–510, 2021. and Y. Shu, J. Zhang & H. Yu, "Fairness in Design: A Framework for Facilitating Ethical Artificial Intelligence Designs. International Journal of Crowd Science. 2023 Mar;7(1):32-9.

The contributions of the co-authors are as follows:

- Prof Yu Han provided the initial paper direction and edited the manuscript drafts
- Shu Ying and I prepared the materials, recruited participants and conducted user studies. Data analysis was done together after the study
- I wrote the drafts of the manuscript. The manuscript was revised and proofread together with Shu Ying.

Chapter 4 is published as J. Zhang & H. Yu, "A Methodological Framework for Facilitating Explainable AI Design", in Proceedings of the 14th International Conference on Social Computing and Social Media (SCSM'22), pp. 437–446, 2022. and J. Zhang & H. Yu, "EID: Facilitating explainable AI design discussions in team-based settings." International Journal of Crowd Science (2022).

The contributions of the co-authors are as follows:

- Prof Yu Han provided the initial paper direction and edited the manuscript drafts
- I prepared the materials, recruited participants and conducted user studies. Data analysis was done after the study
- I wrote the drafts of the manuscript. The manuscript was revised and proofread with Prof Yu Han

16/09/2022

Date

Zhang Jiehuang

Acknowledgements

I wish to express my greatest gratitude to my advisor Prof Yu Han for providing his invaluable guidance, feedback and support over the past 4 years. I could only complete my PhD journey with his help, time and effort over countless meetings in and outside the lab. His wisdom in and vision for the future of ethical AI has inspired me to become a better researcher. I cannot imagine a better supervisor with more empathy, kindness and heart for his students.

I am grateful to all the collaborators, lab mates, peers and other seniors and supervisors for the support, time and help they have rendered during the journey of my PhD. They have come to my aid time after time, and given great advice when needed.

Lastly, I would like to thank my spouse and family for supporting me during these 4 years. Their help and words of encouragement helped me to stick to the path and complete my PhD studies.

Abstract

As artificial intelligence (AI) systems become increasingly ubiquitous, there is a need to steer the design and development of such systems in an ethical trajectory. The technological intricacies of advanced AI systems (e.g., self driving cars) is warranting a deeper look into ethical algorithmic decision making. However, existing AI software design and development teams generally lack understanding in the concepts involved in ethical AI, and face a lack of easy-to-use tools to help them incorporate ethical considerations into the AI systems being developed. This thesis aims to address this important gap.

Firstly, even though AI systems produces logical decisions, biases and discrimination are able to creep into the data and model to affect outcomes causing harm. This inspired us to re-evaluate the design metrics for creating such systems and focus more on integrating human values in the system. However, while the awareness of the need for ethical AI systems is high, there are currently limited methodologies for designers and engineers to incorporate human values into their designs. The proposed methodological tool aims to address this gap by assisting product teams to surface fairness concerns, navigate complex ethical choices around fairness, and overcome blind spots and team biases. It can also help them to stimulate perspective thinking from multiple parties and stakeholders. With our tool, we aim to

lower the bar to add fairness to the design discussion so that more design teams can make better and more informed decisions for fairness in their application scenarios.

We then extended the methodology to the field of explainable AI (XAI). The development of AI systems has created many applications that have tremendous current and future value to human society. However as AI systems penetrate more aspects of everyday life, there is a pressing need to explain their decision making processes in order to build trust and familiarity with the end users. In selected fields like healthcare and self driving cars, there are high stakes that require AI to achieve a minimum standard for accuracy and provide well designed explanations for their outputs, especially when it impacts human life. To date, many techniques have been developed to make algorithms more explainable in human terms, however, there are no design methodologies to allow software teams to systematically surface and address explainability related issues during the AI design and conception stage. We proposed the Explainability in Design (EID) methodological framework for addressing explainability problems in AI systems. EID is a step by step guide to the AI design process that has been refined over a series of user studies, and interviews with experts in AI explainability. It is designed to be used by software design teams to uncover and resolve potential issues in their AI products, as well as being able to simply refine and explore the explainability of their products and systems. Through empirical studies involving AI system designers, it has been shown to decrease the barrier of entry, the time and experience needed to effectively make well-informed decisions for integrating explainability into AI solutions.

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction to Ethical Artificial Intelligence

When the scientific field of Artificial Intelligence (AI) matures and is applied to various domains, it is likely to be faced with ethical challenges. As AI technologies enter many areas of our life [1–4], the problem of ethical decision-making, which has long been a grand challenge for AI [5], has caught public attention. As AI approaches and exceeds the performance of human experts, we must consider the desirable traits that it should embody and chart a course towards a paradigm of ethical AI systems. The fields of ethical philosophy and decision-making have been a topic humans mused about for many centuries. Since the time of ancient philosophers such as Socrates and Confucius, humanity has been exploring the many themes that the field of ethics had raised. However, despite the extensive

history of ethics and philosophy, there remain many open questions, due to the complexity, depth and interdisciplinary nature of ethical experiments.

Ethical Artificial Intelligence is a huge and complex field, that compromises many aspects, or values such as privacy, fairness, account and many others [6]. There are various definitions and principles that guide its development, giving rise to subfields where the research community focuses on a different principle. We focus on the two main aspects of ethical AI in this thesis, namely fairness and explainability. We have identified these aspects to be investigated deeply and aim to develop guiding frameworks for each.

A major source of public anxiety about AI, which tends to be overreactions [7], is related to artificial general intelligence (AGI) [8] research aiming to develop AI with capabilities matching and eventually exceeding those of humans. A self-aware AGI [9] with superhuman capabilities is perceived by many as a source of existential risk to humans. Although we are still decades away from AGI, existing autonomous systems such as autonomous vehicles (AVs) already warrant the AI research community to take a serious look into incorporating ethical considerations into such systems. Furthermore, with the advancement of relatively new technology, there will be many new questions raised that we may not have the answers to, as well as novel scenarios that arise due to that advancement. For instance, the notion of revealed rights that are only meaningful in certain technological contexts is a topic that was discussed in [10].

The issue of ethics has been a challenge for humanity since a long time ago and

it has been increasingly so as automation technology becomes ubiquitous. As crowdsourcing is being used to solve problems in fields such as ride-sharing algorithms and multi-agent systems, they are increasingly widely experienced by the public. Frequently appearing in popular cultures such as movies and television shows, public fascination with AI and crowdsourcing tend to lead the discussions into ethical considerations involving automation of decision-making by machines. A topic that has captured public attention is the Artificial General Intelligence (AGI) research, whose objective is to develop AI with capabilities matching and ultimately surpassing human capabilities. However, these fears are unfounded and merely speculation, as AGI is still a distant dream that will take decades or more to materialize. On the other hand, these fears spur us towards taking steps to discuss and define the standards that machine decision-making should uphold. Ethics is a normative practical philosophical discipline of how one should act towards others, and can be separated into 3 distinct dimensions [6]:

- Consequentialist ethics: an agent is ethical if and only if it weighs the consequences of each choice and chooses the option which has the most moral outcomes. It is also known as utilitarian ethics as the resulting decisions often aim to produce the best aggregate consequences.

- Deontological ethics: an agent is ethical if and only if it respects obligations, duties and rights related to given situations. Agents with deontological ethics (also known as duty ethics or obligation ethics) act in accordance to established social norms.

- Virtue ethics: an agent is ethical if and only if it acts and thinks according to some moral values (e.g. bravery, justice, etc.). Agents with virtue ethics should exhibit an inner drive to be perceived favourably by others.

Under the scrutiny of the public eye, there are many concerns raised about the side effects of advancing AI technology. One of the main issues concerns the loss of jobs due to automation[11] , and is a recurring theme in the 2020 United States presidential elections. An interesting concept that could be relevant to the effect of AI on automation is Jevon's paradox [12]. In the industrial revolution, Jevon's paradox describes how technical innovations allowed machines to perform the same output with less coal. Intuitively, this led to thinking that the total amount of coal and other resources required to power the industry would decrease. On the contrary, the decreased cost of operation led to more demand and the absolute coal consumption increased instead [13]. This paradox could be analogous to the rise of AI and its effects on different aspects of human society will play out in the next few decades. These scenarios motivate us to dive deeper into the research of ethical AI systems, in order to fully reap the potential rewards of maturing AI technology while mitigating harmful side effects.

1.2 Motivation

In this section, we describe the historical milestones and highlight the circumstances that led to the rise in the importance of ethical AI systems. In recent years, many

incidents reveal the need for AI to be developed in an ethical way that is beneficial for society as a whole. We indicate the significant events concerning the field of ethical AI in Figure 1.1.

In light of these events, it is evident that algorithms are far from perfect. Furthermore, the issues that plague our algorithms are disturbingly familiar, being analogous to existing problems in human society such as discrimination. Notably, several public figures, such as entrepreneur Elon Musk and renowned physicist Stephen Hawking, have publicly expressed their concerns about the potential consequences of the breakneck pace of the growth of AI technologies. Potential abuse and malicious intent accompany the explosion of data and machine learning techniques [14].

Deep learning pioneer Yoshua Bengio highlighted that the reality that irresponsible use of AI is prevalent in several organizations [15]. This can lead to unforeseen consequences if left unchecked since AI is a new frontier previously unexplored. Subsequently, in the Montreal Declaration for Responsible Development of AI, a set of ethical guidelines is implemented to steer the development of AI in an ethical direction [16]. The magnitude of the impact of AI on humankind is unprecedented and for this reason, these public figures are sounding the alarm.

Another complication within the field of AI is the interconnectivity of ethical values. Ethical AI values are interconnected because they are all related to the principles and guidelines that govern the development and use of artificial intelligence (AI) systems. We include a discussion of these values:

Transparency: The principle of transparency involves making AI systems understandable and explainable to users and stakeholders. It requires providing clear documentation of the system's design, operation, and decision-making processes.

Accountability: Accountability is the principle of ensuring that AI developers and users take responsibility for the actions of AI systems. This includes creating mechanisms for addressing potential harms caused by AI and holding those responsible for those harms accountable.

Fairness: The principle of fairness requires that AI systems treat all individuals and groups equitably and without bias. It includes considerations of bias in data collection and the design of algorithms.

Privacy: Privacy is the principle of protecting individuals' personal information from unauthorized access, use, and disclosure. It includes ensuring that AI systems are designed to protect users' privacy and prevent unauthorized access to their data.

Safety: Safety is the principle of designing AI systems to minimize the risks of harm to individuals and society. This includes ensuring that AI systems are secure, reliable, and resilient.

These ethical values are interdependent, and they influence one another. For example, the principle of fairness is closely related to the principle of transparency because transparency is necessary for detecting and addressing bias in AI systems. The principle of accountability is closely related to the principle of safety because accountability mechanisms are necessary for ensuring that AI systems are safe and do not cause harm. The interconnectivity of ethical AI values highlights the

importance of considering and implementing all of these principles in AI development and use. Failure to consider any one of these values could result in harmful consequences for individuals and society as a whole.

As the stage is set for AI systems to become more ubiquitous, complex and opaque [17], it seems inevitable that it will be intertwined with society. The trajectory of the development of AI will be determined by our current decisions, and it is imperative that we leave our next generation with technologies that leave the world a better place. While it is clear that AGI is not possible in the short term, we have witnessed only the tip of the iceberg of the flaws of these algorithms. When we adopt a long-term perspective of the future of AI, it is indeed wise to commit time and effort to the ethical development of ubiquitous algorithms. To summarize this section, it is critical that we examine in detail the inner workings of our algorithms in order to achieve the vision of ethical AI that contributes significantly more benefits to humankind than harms.

1.2.1 Difficulty of Incorporating Ethical Considerations into AI

Even though the power of technology to transform the world is seemingly limitless, AI is still merely a tool in the hands of humans. A tool that can be used for good or evil, depending on the actions of the wielder. As we embark on the journey to bring to life products that rely on AI, there are potential pitfalls that deserve a second look. AI has the ability to amplify the power of existing threats, while also

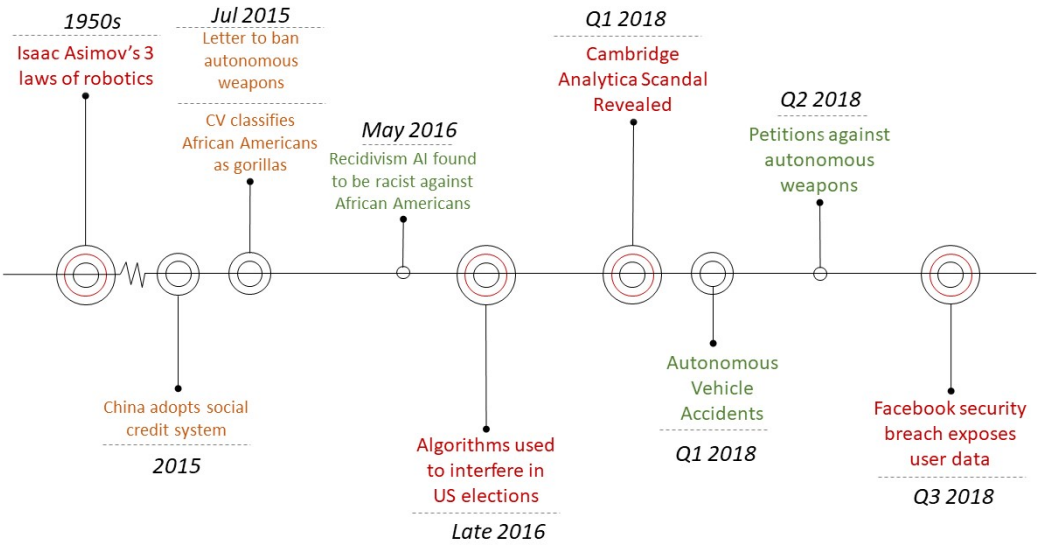


FIGURE 1.1: Timeline of major events leading to the importance of building ethics into AI

creating new attack mechanisms. [14] explores the threats of AI-imbued systems in the digital, physical and political spheres. We will explore these various scenarios in the following sections.

In order to highlight the duality of AI, we use the following example in the study of using algorithms for posting bail in court cases. Shortly after an arrest, a judge has to decide: will the defendant await their legal fate at home? Or must they wait in jail? This is no small question. A typical jail stay is between two and three months. In making this life-changing decision, by law, the judge has to make a prediction: if released, will the defendant return for their court appearance, or will they skip court? And will they potentially commit further crimes?

The authors find that there is considerable room to improve on judges' predictions. The estimates show that if pre-trial release decisions were made using the

algorithm's predictions of risk instead of relying on judge intuition, crimes committed by released defendants can be reduced by up to 25 percent without having to jail any additional people. Or, without increasing the crime rate at all, up to 42 percent fewer people would be jailed. With 12 million people arrested every year in the U.S., this type of tool could let us reduce jail populations by up to several hundred thousand people. Furthermore, this sort of intervention is relatively cheap. Compared to investing millions (or billions) of dollars into more social programs or police, the cost of statistically analyzing administrative datasets that already exist is next-to-nothing. Plus, unlike many other proposals to improve society, machine learning tools are easily scaled. With this simple example, it is understandably complicated when the judge is replaced by an algorithm. Similar stakes and impact concern humans when algorithms drive cars, diagnose illnesses and diseases, as well as many other real-life case studies. However, despite the many claims of algorithms, there also are issues that we have to resolve before they are implemented on a wide scale. These are discussed in the following sections.

There are still many unanswered questions such as the trolley problem [6]. When we allow AVs on the roads, they are essentially given the power to make decisions of life and death during emergencies. Although the process of a passenger takeover exists, we still expect that in some situations the self-driving system has to reach a time-sensitive decision on its own. Researchers from interdisciplinary fields are encouraged to engage with each other in order to collect more data on various ethical dilemmas within the context of diverse cultural and social norms. Besides AVs, other Artificial Intelligence technologies such as autonomous weapons, and

surveillance systems are becoming a reality and have a real impact on society. This calls for a global and unified AI regulatory framework that needs to be established in the short term to address the ethical issues [18].

Other issues that many have highlighted are privacy concerns and security issues [19]. AVs on the road today use sensors, complex algorithms and other tools such as lidar. These vehicles share sensitive data which could include passengers' personal data. Although having interconnected vehicles can optimise traffic conditions, privacy and cybersecurity issues are still a reality. An AV is still a computing device that can be hacked, and the hacker can potentially take over the vehicle's steering, acceleration and other controls. Although the consequences of a hacked AV are dire, paying attention to the security architecture can potentially address this issue. It also requires cooperation between automakers, cybersecurity experts and government agencies to reach a concerted effort. It is necessary for us to build up the defence before we enjoy the benefits of an AV-powered traffic network.

We also need to take into consideration the impact of AI systems under different social, cultural and political settings [20]. Advanced AI systems cannot be analysed or regulated by a simple one size fits all solution, due to the diversity in cultural and social norms. As we advance toward a vision of an AI-enabled society, we must consider the revision of current social contracts. Parker et al. 2019 [21] highlights the notion of a revealed right - a right that is only meaningfully revealed in certain technological contexts. Novel situations may arise out of this new method of mobility in the age of automation, and there is a need to establish regulations and transparency on the responsible party when things malfunction. The AI systems

we use on a regular basis need to be able to explain or show how things go wrong in a manner that humans can understand. However, not all situations and decision-making can be reasonably programmed into AI systems despite best efforts.

1.3 Regulating AI Systems

AI is not just a series of sophisticated mathematical functions, it is unexplored, nascent territory that humans are still researching and making sense of [22]. According to the OECD Principles on AI [23], the following five complementary values-based principles are necessary for the responsible stewardship of AI:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
- AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
- There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
- AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.

- Organizations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

However, in reality, these qualities are not necessarily simultaneously attainable and there usually is some trade-off when designing algorithmic systems. It is critical for the relevant authorities to implement regulations in order to steer the design and deployment of AI toward safe and ethical directions. While there are some efforts taken by governments and other organizations, regulating complex AI systems can quickly become abstract. Due to issues of foreseeability, regulations typically are reactive, and nearly always playing catch-up with technological advances. Compared to other sources of public risk [24], AI self-driving systems may be far more opaque than conventional automobiles, for instance. Critical features underlying an AI system's operation may not be immediately apparent or readily susceptible to reverse engineering. Thus, defects or biases in their design might be undetectable not only to consumers but also to downstream manufacturers and distributors.

Many organizations such as government bodies and technology companies are already formulating guidelines and principles on achieving ethical AI. The research community also recognize the need for such efforts and has committed funds and manpower to this great endeavour. For example, Google, and Microsoft already set up their internal ethics board, while OpenAI was started with the vision for AGI to benefit all of humanity as a whole.

While progress in AI research is accelerating, inappropriate regulatory measures risk suppressing the desirable aspects of applied AI. To highlight the potential of AI for good, we point to an application area in disaster relief. Satellite images are analyzed using computer vision techniques for coordination and prioritization of humanitarian relief efforts [25]. The other examples of how AI can be harnessed for social good are well documented in [17], ranging from substituting humans in the raising of children [26], to shaping romantic matches for online dating services [27].

In the 100 years of AI report [28] and [29], the authors draw an analogy of regulating AI to the privacy regulations in European countries and the United States. Some countries have strict and detailed regulations, which led to a "compliance mentality" in the regulated organizations. This mentality in turn produced the effect of discouraging both innovation and robust privacy protections. The organizations viewed privacy simply as a compliance activity and focused on avoiding penalties, rather than proactively designing technology and adapting practices to protect privacy. By contrast, regulations in the United States and Germany combined ambiguous goals with tough transparency requirements and meaningful enforcement. They were more successful in incentivizing companies to proactively pursue and invest in privacy. When regulating AI, the relevant authorities can also strengthen a virtuous cycle of activity involving internal and external accountability, and transparency rather than narrow compliance. Policies should be evaluated carefully in order to foster the development and equitable sharing of AI's benefits.

Another significant challenge of effectively regulating AI stems from the difficulty

of defining what AI is [24]. The integration of AI into our daily lives is accelerating so quickly that the line between dumb and smart machines is inevitably becoming fuzzy. The concept of "harm" and "reward" can be ambiguous to an algorithm. For example, an AI told not to stab people may get confused about vaccinations, which is technically a form of stabbing. This is summed up by Murphy's Law of AI, that when we give it a goal, the AI will attempt to achieve that goal regardless of whether we like the implications [30]. Against the backdrop of these challenges, the authors then proposed the following approaches to regulating AI: firstly, AI systems should be subject to the full gamut of laws that apply to their human operators. Second, An autonomous agent should clearly disclose that it is not human. Third, the AI should not retain or disclose confidential information without explicit prior approval from the source. Lastly, the AI must not increase any biases that already exist in our systems.

Much akin to parents raising and inculcating the right values in their children, researchers and other stakeholders have a collective responsibility to supervise the development of ethical AI. Especially when AGI is being achieved in the coming decades, parenting AI would inevitably become a focal point when building ethics into AI systems. [31] presents an interesting approach by adding the lens of parenting to weaving ethics in AI research. The authors specifically raised the radical, queer theories of parenting that actively nurture agents whose experiences, objectives and understanding of the world will be necessarily distinct from their parents. By proposing a spectrum of principles which might undermine efforts, it may encourage new schools of thought about the development, design, training

and release into the era of increasingly autonomous agents.

1.4 Ethics in Human-AI Interactions

In AI applications which attempt to influence people's behaviours, the principles established by the Belmont Report [32] for behavioural sciences have been suggested to be a starting point for ensuring ethics [33, 34]. The principles include three key requirements: 1) people's personal autonomy should not be violated (they should be able to maintain their free will when interacting with the technology); 2) benefits brought about by the technology should outweigh risks; and 3) the benefits and risks should be distributed fairly among the users (people should not be discriminated based on their personal backgrounds such as race, gender and religion). The challenge of measuring benefits and risks remains open for application designers albeit the IEEE Ethically Aligned Design guidelines can be a useful starting point [35]. Human-centric values have been incorporated into the objective functions of recent AI-powered algorithmic crowdsourcing approaches [36–38].

A recent example of persuasion is Collaborative Optimization and Planning for Transportation Energy Reduction (COPTER) [39]. The authors created COPTER to be an intelligent travel assistant that evaluates multi-modal travel alternatives to find a plan that is acceptable to a person tailored to their context and preferences. The introduction of COPTER resulted in a 4 percent energy reduction in a deployment scenario conducted in Los Angeles, California, USA. Another example of the application areas in which AI attempts to influence people's behaviours is

persuasion agents [40, 41]. In [42], the authors conducted a large-scale study to investigate human perceptions on the ethics of persuasion by an AI agent. The ethical dilemma used is the trolley scenario which involves making a participant actively cause harm to an innocent bystander by pushing him onto the train track in order to save the lives of five people. It is a consequentialist ethical outcome which requires the decision-maker to violate a sacred value (i.e. one shall not kill). The authors tested three persuasive strategies: 1) appealing to the participants emotionally; 2) presenting the participants with utilitarian arguments; and 3) lying. The three strategies are delivered to some participants by a person playing the role of authority (the station master of the train station) and by a persuasion agent. The results suggested that participants hold a strong preconceived negative attitude towards the persuasion agent, and argumentation-based and lying-based persuasion strategies work better than emotional persuasion strategies. The findings did not show significant variation across genders or cultures. The study suggests that the adoption of persuasion strategies should take into account differences in individual personality, ethical attitude and expertise in the given domain.

Although emotional appeals may not be an effective persuasive strategy under ethical dilemmas, ethically appropriate emotional responses from agents can enhance human-AI interaction. In [43], an approach based on the *Coping Theory* [44] to allow agents to deal with strong negative emotions by changing the appraisal of the given situation was proposed. The agent assesses the ethical effects of its own actions and other agents' actions. If its own action violates a given moral value, the *shame* emotion is triggered which serves to lower the priority of continuing

with the given action. If another agent's action violates a given moral value, the *reproach* emotion is triggered in the observing agent which serves to increase social distance with the given agent (e.g., by reducing trust [45]). The ethical decision-making process is similar to existing individual ethical decision frameworks. The triggering of emotions serves as an implicit reward for the agent and facilitates communications with humans in the loop.

1.5 Foci of this Thesis

In view of the many considerations surrounding the ethical applications of AI, we propose to zoom into several scenarios to better understand the problem and arrive at potential solutions. Some of the scenarios include fields like algorithmic fairness, data privacy measures and decision-making in fully autonomous vehicles where ethics are an inherent part of the discussion. In the next few chapters, we conduct questionnaires and user studies in order to stimulate brainstorming and deep analysis of ethical AI issues. In the process, we aim to achieve a unified methodology that stakeholders can use to surface and address potential ethical issues surrounding the operations of their AI system, as well as for facilitating the incorporation of ethical considerations in the design phase. In order to build ethics into AI, we must first be able to understand and guide its use towards beneficial applications for the good of humanity. For the purpose of this thesis, we focus on the ethical AI aspects of fairness and explainability.

Our research questions are listed here:

- How can system designers narrow down and select the most relevant ethical principle that matches their application domain?
- What can system designers do to identify and resolve ethical issues in their application domain?
- How should system designers stimulate perspective taking from the viewpoints of different stakeholders in their application domains?

In order to achieve our vision of a unified methodology for building ethics into AI and to answer our research questions, we are planning for a series of user studies, online questionnaires and expert interviews. These studies will allow us to study in depth the ethical considerations in the context of advanced AI systems such as self-driving vehicles, and algorithmically driven decision support systems, as well as privacy concerns for personal data.

1.6 Organization of this Thesis

In this section, we briefly discuss the content in the following chapters. As autonomous vehicles are a major context for the discussion of ethics in AI, we devote a significant portion of the end of chapter 2 to it. We explore the various ways AVs build trust with us, both people within and outside the vehicle. The concept of man-machine interfaces is explored, as well as the driver monitoring systems that enhance synergy between the two. We identified AVs as an important context in the study of ethical AI, and will investigate the ethical qualities in this scenario.

In Chapter 2, we highlight the literature review, recent advancements in the field of ethical AI, as well as the key terms being used among the community. We then list the key qualities of ethical AI such as Fairness, Accountability, Transparency, Privacy, these qualities will be the focal point of this thesis. Afterwards, we state and evaluate the existing methodologies available in this field, and explain why our proposed solutions are a suitable fit for this research gap.

In Chapter 3, we introduce the Fairness in AI methodology and the user study implementation and results. There are 2 types of our tool available, a physical user study and an online tool that is more suitable for the new normal caused by the covid 19 pandemic. We have conducted a comprehensive evaluation of our methodology and will discuss the results of our user studies. Next, in Chapter 4, we extend the methodology to the ethical AI aspect of explainability, then provide a similar comprehensive evaluation and discuss the results of a series of user studies.

In Chapter 5, a brief summary of the work done and future plans is given. We illustrate our propose further work to extend the Fairness in AI methodology to cover other ethical qualities like privacy and explainability. In addition to the extension, an iteration of why and how we wish to achieve our vision of an ethical AI methodology is given. Given the state of the field, we then close the thesis by explaining how our work fits into the extended literature as well as the value it creates for the research community and beyond.

Chapter 2

Literature Review

2.1 Overview of Ethical AI

In this chapter, we aim to improve the clarity of the myriad concepts and ideas in the field of ethical AI. We classify the sections into the following areas: Firstly, we discuss the many terms in this field and their definitions for clarity in the following chapters. Following this, we discuss the many aspects of ethical AI such as fairness, accountability, transparency and privacy. Finally, we explore the ways that governing bodies and the authorities can use to regulate current and future AI systems.

2.2 Definition of Key Terms

According to Cointe et al. [46], *ethics* is a normative practical philosophical discipline of how one should act towards others. Ethics is also seen as a reflection of morality. A distinction can be made between fundamental ethics, which is concerned with abstract moral principles, and applied ethics [47]. Applied ethics includes ethics of technology, which in turn contains AI ethics as a subcategory and is the focus of this section. It encompasses three dimensions:

- *Consequentialist ethics*: an agent is ethical if and only if it weighs the consequences of each choice and chooses the option which has the most moral outcomes. It is also known as utilitarian ethics as the resulting decisions often aim to produce the best aggregate consequences.

- *Deontological ethics*: an agent is ethical if and only if it respects obligations, duties and rights related to given situations. Agents with deontological ethics (also known as duty ethics or obligation ethics) act in accordance to established social norms.

- *Virtue ethics*: an agent is ethical if and only if it acts and thinks according to some moral values (e.g. bravery, justice, etc.). Agents with virtue ethics should exhibit an inner drive to be perceived favourably by others.

Over the past decade, there are multiple fields that surface due to a changing technological climate and with it comes vocabulary and technical terms to define.

We would like to disambiguate and extend the definition beyond philosophical ethical dimensions to the following terms:

- *Responsible AI*: Responsible AI frameworks aim to introduce a customizable framework, tools and processes designed to help harness the power of AI in an ethical and responsible manner, from strategy through execution.

- *Ethically Bounded Systems*: designing autonomous systems that constrain or restrict their decision-making according to subjective preferences and ethical principles. This can be difficult due to the subjective variation of what is ethical in different situations.

- *Ethically Aligned Systems*: Establishment of societal and policy guidelines in order for autonomous systems that remain human-centric, serving humanity's values and major ethical principles. These systems are to be designed in such a way that it is beneficial to people beyond reaching functional goals and addressing technical problems.

- *Ethically Aligned Design*: A set of standards and guidelines for the design of AI systems initiated by the Institute of Electrical and Electronic Engineers (IEEE). The standard envisions a future where autonomous and intelligent systems prioritize human well-being.

- *AI for Good*: An initiative by Microsoft to provide technology, resources and expertise to empower those working to solve humanitarian issues and create a more sustainable and accessible world.

- *Trustworthy AI*: The European Commission set up an expert group to steer AI systems toward a set of guidelines that facilitates human trust in AI systems that they interact with regularly.

- *AI Safety*: An area of AI ethical AI that aims to address risks involved with increasingly autonomous AI, as we progress towards AGI. It also implements countermeasures against the misuse or abuse of AI.

- *Human-Agent Collective*: An emerging class of systems that reflect the close partnership and flexible social distance between the system and humans and the computers.

  These 7 terms list the different aspects of ethical AI as a whole. Each term points a to sub-field that aims to achieve ethical AI by driving design, research and implementations in a desirable direction. We desire that the reader may gain a thorough understanding of ethical AI, and is able to identify the many sub-fields, sometimes fragmented, of ethical AI.

  We continue with the terms that we will regularly use in the following sections of our manuscript:

- *AI Governance Techniques*: Methods to set up or improve internal governance structure and measures to incorporate values, risk management and responsibilities related to algorithmic decision making.

- *Ethical Dilemmas* refer to situations in which any available choice leads to infringing some accepted ethical principle and yet a decision has to be made [48].

- *Decision Making Frameworks* is a set of rules or guidelines that allows an individual or organisation to decide how to behave, especially in the context of ethical dilemmas.

We bring attention to several concepts relevant to the emergence of ethical AI, namely Human-Agent Collective (HAC), Social Construction of Technology Theory (SCoT) and Value Sensitive Design (VSD). These concepts are coined and developed in parallel with ethical AI and may have a direct impact on its adoption. Beyond traditional Human-Computer Interaction (HCI), a relatively new field named HAC has emerged, due to the profound effect of AI on the way we work with computers. We are no longer just issuing instructions to passive machines but instead work together with interconnected autonomous entities in a larger network. This is characterized by a dynamic and flexible partnership between humans and AI agents in order to achieve individual and collective goals [49]. HAC involves the flexible autonomy of agents taking actions without reference to their owners, however, this actually causes the diminishing of the internal locus of control valued in HCI. Agile teaming allows multiple humans and agents to group and disband dynamically based on the demand, while incentive engineering rewards them to encourage socially desirable outcomes. SCoT posits that relevant social groups naturally attach a technological frame to a technological artefact. While the artefact has interpretive flexibility, many different frames spring up thus leading to conflicts brought about by inter-group differences. Over time, the period of conflict is reduced and the technological artefact stabilizes and provides singular meaning to all users. VSD is a theoretically grounded approach to the design of technology that

accounts for human values in a principled and comprehensive manner throughout the design process. It employs an integrative and iterative tripartite methodology, consisting of conceptual, empirical, and technical investigations.

Lastly, we list the different classifications of ethical agents. Moor et al. [50] discusses the classification of agents, including artificial agents into the following types

- *Ethical impact agents* are the first level of ethical agents whose actions have ethical consequences whether they are intentional or not. Most robots can be considered to be ethical impact agents since their actions cause harm or benefit to humans.
- *Implicit impact agents* are agents that have ethical considerations built into their design. Usually, these are safety or security concerns put in place to prevent undesired outcomes. They have designed reflexes for situations which require human supervision to ensure security.
- *Explicit impact agents* are more central to ethics in AI and are agents that can process ethical information about diverse situations and make sensitive determinations about the subsequent courses of action. When ethical principles are in conflict, these agents can analyze available information and output reasonable resolutions.
- *Full ethical agents* encompasses central metaphysical features that we usually attribute to human ethical agents. These features include consciousness, intentionality and sentience. Normal adult humans are examples of full ethical agents.

2.3 Algorithmic Fairness

Ensuring that AI systems are fair and adhere to desired qualities and values is vital in our vision of building ethical systems. It is empirically clear that machines are not perfect and can exhibit signs of algorithmic bias. Bias was first defined as any basis for choosing one generalization over another, other than strict consistency with the observed training instances [51]. In modern times, algorithmic bias is understood as systemic errors that lead to unfair outcomes, such as privileging or discrimination against a group or type. As AI systems are trained to recognize and leverage statistical patterns in the data, imperfections and biases in the data are likely to make a difference in the predictions downstream. Quantitative definitions of what is fair or not have been introduced in the fields of education, hiring and machine learning well over 50 years ago [52]. [53] provided a definition of fairness, and later this work was refined by [54] using adFversarial learning to "de-bias" latent representations.

To gain a better understanding we refer to a recent review paper [17], where Rahwan et al. discusses the empirical signs of algorithmic fairness by mentioning the following areas: in computer vision, multiple commercial gender classification are shown to be problematic when classifying dark-skinned females [55]. In Bolukbasi et al. [56] suggests that gender bias and stereotypes are discussed in word embeddings, a popular framework to represent text data as vectors. The authors presented their findings as "disturbing" due to the widespread use of word embedding amplifying these biases. Another investigation into Correctional Offender

Management Profiling for Alternative Sanctions (COMPAS), a risk assessment tool used to predict the risk of recidivism, concludes that it is biased against African American defendants [57]. Although met with much criticism, the investigation does shed some insight into how machines can be tainted by human prejudice.

We list the following types of biases that can affect AI systems, this list is not exhaustive:

- *Sample Bias*: occurs when the training data is not an accurate representation of the environment the model is operating in, this can occur due to non-random sampling of subgroups.
- *Historical Bias*: occurs when training data is influenced by cultural, societal or other stereotypes. The pre-existing bias in the world can seep into the model from the data gathering process and distort results.
- *Measurement Bias*: is a systemic distortion that results when a method or device for collecting data is affecting the values collected. This bias happens from the way we choose, utilize and measure a particular feature.
- *Algorithm Bias*: is a property of algorithms and is contrasted with variance. Models with high bias are more rigid, less sensitive to variations in data and noise, and prone to missing complexities. In contrast to variance, where the model easily fits into training data and welcomes complexities but is more sensitive to noise. Ultimately, a balance has to be found between variance and bias.

In the following paragraphs, we explore key techniques used to define and achieve algorithmic fairness. [58] provides an overview of the key definitions of the notion of fairness.

- *Equalized Odds*: This definition states that the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected and unprotected (male and female) group members [53]. In other words, the equalized odds definition states that the protected and unprotected groups should have equal rates for true positives and false positives.

- *Equal Opportunity*: This definition means that the probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected (female and male) group members[59]. In other words, the equal opportunity definition states that the protected and unprotected groups should have equal true positive rates.

- *Demographic Parity*: This definite states that the likelihood of a positive outcome should be the same regardless of whether the person is in the protected (e.g., female) group [59].

- *Fairness Through Awareness*: An algorithm is fair if it gives similar predictions to similar individuals" [60]. In other words, any two individuals who are similar with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome.

- *Fairness Through Unawareness*: An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process [61].
- *Treatment Equality*: Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories [62]

We facilitate the discussion using the context of a resume filtering algorithm, and a protected attribute that is used to guarantee fairness [63]. The simplest method to ensure fairness in this context is to not give the algorithm access to the protected attribute. [64] indicates that this may ameliorate prejudice biases in an imperfect dataset. However, it is difficult as many attributes may be correlated to the protected one. For example, Israeli men serve 3 years in the armed forces compared to 2 years for women, hence the algorithm is able to identify the gender of the applicant using the length of military service.

Another method, statistical parity, or group fairness, is the property that the demographics of those receiving positive or negative classifications are identical to that of the population as a whole [60].

$$\Pr\left[h(x) = 1 | x \in P^C\right] = \Pr[h(x) = 1 | x \in P]$$

However, statistical parity, or group fairness, may also lead to unfairness. For example, the resume screening algorithm may deliberately select unqualified members of the protected group to be interviewed in the expectation that they fail. Hence it

is desirable to include individual fairness concurrently. Individual fairness ensures that any 2 individuals who are similar with respect to a particular task should be classified similarly.

In more recent work, [65] explores various other means of fairness such as demographic parity, equality of odds and equality of opportunity. The authors then incorporated these concepts into a loss function in order to apply them to supervised deep learning tasks. Their main objective is to maximize the predictor's performance while minimizing the adversary's ability to predict the protected variable, based on the architecture of generative adversarial networks [66].

In [67], the authors incorporated algorithmic fairness into a crowdsourcing system built to address the societal challenge of large populations seeking ad-hoc work in China. The system, Algocrowd, employs Jain's Fairness Index [68] in order to allocate tasks to workers fairly. Algorithmic fairness currently may result in trade-offs with performance in some scenarios, however, it is imperative that we continue to drive research efforts and achieve ethical AI through fairness.

2.3.1 Fairness Concepts

There have been considerable efforts in the field of AI to determine how to operationalize fairness in a way algorithms can understand [69][52][59]. As the philosophical theories of fairness are discussed in words, the attempt is to construct similar concepts at a mathematical level. So far, two different categories have been distinguished for fairness definitions: individual and group fairness [60][70].

Individual fairness requires similarly situated individuals to be treated similarly [60]. It is fulfilled when the algorithm gives similar outputs to individuals who have similar attributes that determine the results. There are three statistical definitions under individual fairness which are fairness through awareness[60], fairness through unawareness[61], and counterfactual fairness[70]. Group fairness on the other hand requires different groups to be treated equally [60]. It is fulfilled when the distribution of outcomes is the same for each group. There are seven statistical definitions under group fairness which include demographic/statistical parity[59], conditional statistical parity [71], equalized odds [72], equal opportunity[72], test fairness[73], treatment equality[62], and fairness in relational domain[74]. We will be using these ten main fairness definitions as the focus of the tool.

2.4 Accountability

A primary consideration for the governance of AI is accountability, which deals with the clear acknowledgement of responsibility for actions, decisions, products and policies. 2018 was a year where huge scandals were exposed, as well as significant events that have cast AI in a negative light [75].

The response from the public was swift and harsh, demanding greater accountability from the large technology companies and their products. The challenge to implement governance and accountability for AI systems is one of the many important themes when building ethics into AI. In recent international conferences, such G7 Conference on responsible adoption of AI and the AI Now 2018 symposium,

AI accountability was a central theme that dominated the conversations. In [76], accountability literature was split into 3 areas:

- Accountability is a feature of the AI system. By building explainability into the AI, it would partially address the AI's accountability.
- Second area focuses on determining which individuals or groups are accountable for the impact of AI.
- Finally, accountability is seen as a feature of the broader socio-technical system that develops, procures, deploys and uses AI.

Each area is an active research field that points towards actionable guidelines in the design of ethical AI systems. In order to provide some context for discussion, we point to the scenario when AI systems make mistakes or harm in AV accidents. This is because AVs are already a reality and our society is actively headed towards the widespread use of fully autonomous self-driving cars.

For conventional vehicle accidents, the responsibility usually can be attributed to 3 areas: 1) driver error, 2) vehicle malfunction or defects and 3) external conditions. In comparison, the unavoidable AV accidents raise the question of who should be liable and how casualties should be compensated when a vehicle controlled by an algorithm rather than a human causes injury [77]. For the first time in the age of autonomous cars, an Uber self-driving car malfunctioned and crashed into a woman and killed her. The answers to this question have implications way beyond the resolution of individual AV accident cases. Our legal system's ability to handle

these new cases impacts the adoption rate of AVs, and in turn, the reduction of road accident casualties.

[78] suggests categorizing the accident scenarios into 4 levels of driver attentiveness, from distracted to fully attentive. The authors commented that the responsibilities should be borne by the AV manufacturer first and foremost, then shifting back to the driver depending on his ability to intervene and stop the accident from occurring. They also suggested that courts and legislatures need to address tort liability for accidents caused in autonomous mode to ensure that the correct party bears responsibility for accidents.

For a better understanding of how to deal with algorithms causing casualties, we refer to aviation, where planes flying on autopilot mode can be considered to be analogous to AVs. In cases of aircraft collision, the responsible party is directed at the pilot instead of the auto-pilot system, because the presence of automation does not reduce a pilot's responsibility to keep a constant lookout [79]. However, whether these cases have any impact on the judgment of AV accidents is hard to foresee. If AVs are not taken up by the public quickly, there is an unnecessary risk of people getting killed or injured by misjudgments of human drivers. In order to address these issues in the short term, [80] proposes a specially designed, no-fault, quasi-judical victim compensation fund that is used to assist victims of AV accidents. This fund can protect AV designers as well as manufacturers from a plethora of uncertainties about liability exposure while assuring the general public that they will be compensated fairly and quickly in the event of an accident. It can

also provide legal systems with the adapt product liability law to the new paradigm of AVs in our society.

From this overview, it is clear that there is much work needed to be done in order to strengthen the attribute of accountability in building AI systems. This needs to be a concerted effort among all the stakeholders if we are going to achieve AI that is ethical and accountable.

2.5 Transparency

An ethical AI system must strive to achieve transparency. Without transparency, it is difficult to trust that the system outputs fair, objective and sound results, falling short of the standards of an ethically aligned system. Explainability and interpretability are essential to the building and evaluation of ethical AI systems. As algorithms become tremendously complex, the gap between the system's inner workings and the designer's understanding of it becomes wider [81]. In [82], the authors expound on the risks of deploying "unaudited black box" systems, arguing that we will lose touch with how decisions are made, and in turn difficult for developers to identify or respond to bias, errors and other problems.

We initiate this discussion by highlighting the difference between explainability and interpretability [83]. While they are often used interchangeably, it is worth distinguishing the two. Interpretability pertains to the extent to which a cause and effect can be observed within a system, in other words how much we can predict

what is going to happen given a change in input or algorithmic parameters. It is being able to look at an algorithm and being able to see what is happening within.

Explainability on the other hand is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms. It's easy to miss the subtle difference with interpretability, but explainability is the ability to explain the algorithm to a human being. Interpretability and explainability, both contribute to the transparency of AI systems, overlapping in many areas, but do also have subtle differences.

The scope of this discussion is limited to the following areas that are deemed to be the most important areas where transparency value adds to. In healthcare, where diagnosis made by an algorithm needs to be explainable so that patients can trust the decision. In some cases, there is urgency for a decision made by the algorithm to be explainable so that lives can be saved and suffering minimized. AI has the potential to revolutionize the medical field and bring about a new paradigm of health and fitness. In the field of autonomous vehicles, we need transparency in how algorithms make decisions to control the vehicle. In the event of mishaps or accidents, the inner workings need to be examined in order to decide if the AV has malfunctioned. In the previously discussed area of criminal recidivism, when algorithms make a wrong decision it can cost people many years of their lives. As AI is being applied to datasets involving humans instead of just numbers or pictures, the importance of transparency grows as inappropriate decision-making can inflict real, lasting damage on the lives of innocent people.

The European Union (EU) adopted a set of comprehensive regulations for the collection, storage and use of personal information, the General Data Protection Regulation(GDPR) [84]. This includes transparency in the form of a right to explanation for algorithmic decisions made by AI systems and reflects the increasing importance of human explainability in algorithm design. Explainable artificial intelligence (XAI) is defined as the techniques used in AI systems that can be trusted and easily understood by humans, as contrasted with the "black box" approach where even the system developers were unable to explain how the system arrives at a specific decision. According to DARPA (Defense Advanced Research Projects Agency), XAI aims to produce more explainable models while maintaining a high level of learning performance and enable humans to understand, trust and manage AI partners [85]. Explainability and interpretability measures are proposed as intermediate solutions that alleviate the problems associated with the rising importance and ubiquity of algorithms in our daily lives. With the use of XAI systems, the public or relevant stakeholders are then positioned to make meaningful responses or disputes to the use of algorithmic systems in the event of suspected flaws in the system.

[86] explains that for highly consequential and ethical situations, we require a deep understanding of the decision-making process of the algorithm in order to trust and evaluate them. The authors attempt to develop XAI systems into the following categories:

- *Opaque Systems*: The mechanisms mapping inputs to outputs are invisible to the user.

- *Interpretable Systems*: The user is able to see and understand how inputs are mapped to outputs. However, this implies model transparency and requires a level of understanding of the technical details of the mapping.

- *Comprehensible Systems*: This system emits symbols along with its output. These symbols, such as words or visualizations, allow the user to relate the properties of the input to their output. This assumes that the user has some implicit "cognitive intuition" about how the input, symbols and output relate to each other.

The authors concluded by suggesting a final desired category: truly explainable systems, where automated reasoning is central to output crafted explanations without requiring human post-processing as the final step of the generative process. There are several surveys providing a comprehensive discussion of the XAI research contributions in recent years. [87] describes recent advances along four main axes: explainability strategies, evaluating explanations, XAI perception and explain or predict. For a holistic survey of existing XAI techniques and developments, we refer the reader to the Adadi et al paper and the following survey papers.

[88] provides a comprehensive overview of the existing techniques for visual interpretability in Convolutional Neural Networks (CNN). These contributions are significant as they can be applied in the field of healthcare, where AI is highly

required to explain its decisions. The authors classified progress in this field into 5 different directions:

- *Visualization of intermediate network layers*: Synthesize the image that maximizes the score of a given unit in a pre-trained CNN or invert feature maps of a conv-layer back to the input image.
- *Diagnosis of CNN representations*: Diagnose a CNN's feature space for different object categories or discover potential representation flaws in conv-layers.
- *Disentanglement of patterns encoded in each filter*: Disentangle complex representations in conv-layers and transform network representations into interpretable graphs.
- *Building explainable models*: Representations in the middle layers are designed to have clear semantic meanings.
- *Semantic-level middle to end learning via HCI*: Methods to learn new models via HCI and active question-answering with limited human supervision.

[89] classifies approaches to explainability in artificial neural networks (ANN) into 2 types: integrated (transparency-based) and post-hoc. Integrated explainability seeks to make use of the model to explain itself. For example, [90] proposed a unified framework for interpreting predictions: SHAP (SHapley Additive exPlanations) assigns each feature an importance value for a particular prediction. However, this approach is limited to algorithms with relatively lower complexity and thus cannot be applied to ANNs. Furthermore, there is a tradeoff between explainability and

predictive performance [91] [92], this might mean that not all systems are required to be explainable depending on the goal. In contrast, post-hoc interpretability extracts information from already learned models and does not precisely depend on how the model works. The advantage over integrated approaches is that it does not impact the performance of the model and thus can be used for a diverse range of systems.

We highlight several significant techniques used in promoting the transparency of algorithms:

Surrogate Models: Interpretable model trained on predictions of the original black-box model. Examples such as Local Interpretable Model-Agnostic Explanation (LIME), an explanation technique that explains the predictions of any classifier in an interpretable model locally around the prediction [93]. Leave-One-Covariate-Out operates in a similar manner, generating local explanation models that offer local variable importance measures.

Partial Dependence Plot(PDP): graphical representations that help to visualize the average partial relationship between one or more input variables and predictions of a black-box model [94]. Later PDPs are extended by Individual Conditional Expectation (ICE), where plots reveal interactions and individual differences by disaggregating the PDP output.

Layer-wise Relevance Propagation(LRP): LRP redistributes the prediction function backwards, starting from the output layer of the network and backpropagating up to the input layer. This method identifies important pixels after running a

backward pass in the neural network, the backward pass is a conservative relevance redistribution procedure, where neurons that contribute the most to the higher layer receive the most relevance [95].

Due to the tradeoff between explainability and other attributes of AI systems, we cannot achieve full explainability all the time. While explainable systems are clearly desired, it is not always a necessity. Requiring every AI system to explain itself can result in reduced efficiency, forced design choices and a bias towards explainable but less capable and versatile systems. It is also expensive and requires resources when using both integrated and post-hoc transparency techniques. Hence there needs to be a balance between performance and transparency.

2.6 Application Domain Example: Autonomous Vehicles

A direct result of the exponential growth in technology is that autonomous systems are becoming more common in our daily lives. From industrial machines to autopilot functions, computing systems assist humanity in performing repetitive, dull and tedious tasks [96]. Specifically, AI (Artificial Intelligence) has the potential to accelerate breakthroughs in many fields, especially in the area of AVs (Autonomous Vehicles). Consequently, AVs are one of the most widely investigated technologies within the automotive field [97] as intelligent features such as adaptive cruise control and lane keeping aid are introduced.

As these technological advances improve at a breakneck pace, the importance of trust in the human-machine partnership takes centre stage. This synergy between man and machine is highlighted in [98] [99] [100] and demonstrated in more recent work [101], where an end-to-end learning strategy can be harnessed to improve the combined performance of human-automation teams by considering the distinct abilities of both people and automation respectively. The authors show that the human-machine teams outperform both the individual performances of the human and machine.

The benefits of autonomous vehicles have been widely recognized (e.g, reducing traffic fatalities, improving navigation efficiency and reducing human error). Before these visions can be realised, a certain level of trust must be built between users and AVs such that societies are comfortable with adoption on a large scale. Yu et al. 2018 [6] briefly discuss the role of trust in the ethics of human-AI interaction, highlighting the importance of facilitating the cooperation of humans when using AI systems. The authors also explored the ethical dilemmas involving AVs, such as the moral machine project [102]. Based on feedback from the public, it is clear that different regions prioritize different sets of values due to cultural and societal norms.

In order to design autonomous vehicles that are effective at communication, we need to investigate how trust is built between users and machines. Trust is a term that is differently defined in many different contexts, such as psychology, human-computer interaction (HCI), economics, and computer science. Most research works of trust in HCI aim to establish a quantifiable model of trust so that the level of trust can

be monitored. According to [103], there is still noticeable scepticism in society regarding AVs and their use. Furthermore, one of the most challenging barriers to entry is the average consumer's significant distrust of fully autonomous vehicles. Public trust is essential to the success and adoption of new technology such as automated driving, and it should be a key focus during the design process.

Uggirala et al. 2004 [104] postulate that uncertainty is more easily quantifiable compared to trust, and used this concept to compare the participant's judgement of the performance of autonomous vehicles. The results conclude that trust in the system scale with the understanding of how it works. This finding has implications on the required transparency of AVs, that some basic understanding of the inner workings of self-driving might be conducive for human-machine cooperation.

Kyriakidis et al. 2015 [105] conducted a crowd-sourced online study on the public opinions of partial, highly and fully automated vehicles. The authors discovered that despite significant support of the benefits of AVs, participants were most concerned about software hacking or misuse as well as legal and safety issues. These concerns may extend to data security and privacy issues that are part of the backdrop of the conversations among the wider public. Nevertheless, these findings enable AV manufacturers deeper insight as to how to address doubts in order to streamline adoption.

Wagner et al. 2015 [106] postulate that advancing autonomous driving technology has rendered traditional software safety techniques inadequate, and a new software safety philosophy is required. This new philosophy is vital in the process of deciding

whether the software in control of an autonomous vehicle is safe and trustworthy or not. The process should include testing for the potential of the AV to be trustworthy enough to be on the road driving itself and to develop its trust with the passenger over time. The authors also highlighted that there are significant expectations that the algorithms behind self-driving technology achieve close to 100 percent accuracy, however, this is beyond the performance of most machine learning and image processing methods. As the accuracy and demands of AI improve, we may see a paradigm shift whereby the self-driving AI techniques outperform humans by a large enough margin that it no longer makes sense for humans to drive vehicles, considering a large number of accidents due to human errors.

Akash et al. 2017 [107] presents a model of the dynamic levels of trust between AVs and their passengers. The authors recommend that this data can then be integrated into a feedback control system for improving the AV's response to human trust. The online crowd-sourced study involves 581 human participants to judge whether they trust a computer vision-based obstacle detection system. As expected, trust decreased significantly during faulty scenarios, while the rebuilding of trust slightly takes place after 8 to 10 performing trials by the system. This study also considered the effects of nationality, culture and gender on trust. Americans usually are less trusting of AVs compared to Mexicans and Indians, while women tend to have a larger rate of change of trust compared to men.

Shahrdar et al. 2019 [108] extends the literature by studying how trust can be eroded, rebuilt or enhanced in the context of a passenger riding in a fully autonomous vehicle. The authors argue that in most situations, trust can be rebuilt

after a reasonable time frame despite prior faulty behaviour. In the case of pedestrians, they are unlikely to encounter the same AVs multiple times in a short duration. However, prolonged exposure to negative experiences with AVs tends to erode trust and amplify negative emotions associated with these AVs.

Alvarez et al. 2019 [109] studies pedestrian behaviour when a slow-moving AV with a "eye contact" eHMI communicates with them. The findings of the study conclude that the implementation of visual cues in eHMIs was not specifically necessary in a shared space where informal traffic rules are used. They are more likely to help when the ORUs and AV have the potential that causes danger. This can be inferred that for commonplace, low-risk interactions, pedestrians have high trust in AVs which may cause them to act normally regardless of whether a vehicle has a driver or not.

Olaverri et al. 2020 [110] suggests that it is paramount to understand the actions of different road users and their reactions to AVs. Initially, the interaction of the public and AVs will lead to some unexpected situations that can affect the level of trust. In order to develop confidence in the technology, manufacturers of AVs can consider the expectations and user experience of the average road user. Promoting trust between AVs and ORUs can seem like a daunting task, but AVs have the potential to change transport as we know it. Hence it is worthy of our effort to help facilitate society to widely embrace AVs.

Human transport has come a long way since the steam engine, and we are now just around the corner of the age of fully autonomous vehicles. While the breakneck

pace of scientific innovation and technological advancement continues, researchers and other groups need to look at the human side of the equation. AVs should be positioned as an extension of the basic human need for mobility, ensuring that our commuting experience is elevated while also reducing our impact on the environment. It is key that we put the focus on people when designing AVs, because technology should be pursued for the betterment of humanity. We envision that future research can help improve drivers' trust and user experience in AVs, thereby enabling the harmonious integration of future autonomous vehicles into our societies.

2.7 Existing Design Methodologies

In this section, we evaluate the various tools that are already available in the literature and expound on their strengths and limitations. Generally, there are options in the literature for systems focused on speed, and profitability, however, in contrast, there is a lack of tools that support projects with the goal of enabling ethical measures in modern AI systems. We list a few existing tools available and briefly discuss their strengths and limitations.

2.7.1 Agile Software Design

Agile is a software development process created in 2001 with the following 4 major principles [111]:

- Individuals and interactions over processes and tools
- Working software over comprehensive documentation
- Customer collaboration over contract negotiation
- Responding to change over following a plan

The tool focuses on the end user in mind, to create a system that satisfies the user's requirements and demands. The software development team comprises multi-disciplinary people with the necessary skills to complete the system timely and effectively. These teams can include quality assurance engineers, databases, front-tend and backend engineers as well as analysts depending on the type of software required.

We discuss a popular type of Agile, diving deeper into the methodology called Scrum. According to definitions [112], Scrum is a framework utilizing an agile mindset for developing, delivering and sustaining complex products. This framework is especially applicable for software development, but also has applications in other fields such as research, sales, marketing and other advanced technologies. The speciality of Scrum is in the time-boxed iteration, aptly named sprints, that take no longer than a month. At the end of the sprint, the team demonstrates the work done to stakeholders to elicit feedback, in order to continue refining their work.

2.7.2 Value Sensitive Design

Value Sensitive Design (VSD) is the most famous theoretical approach to the value design of technology that focuses on integrating human values in a principled and methodical way throughout the design process [113]. It encompasses a large literature of targeted methods to engage with values differently. Qualities of VSD methods include being committed to the theoretical constructs of VSD, being in the descriptive form to help designers gain more insights, and staying open to integration with other methods, adaptation and changes [114].

One method is direct and indirect stakeholder analysis. This analysis allows the identification and legitimisation of stakeholders, as well as explores how they might be impacted [115]. This is useful as designers tend to focus on direct stakeholders, who are those that interact directly with the technology and often miss out on indirect stakeholders, those that do not interact directly with the technology but are still affected.

There are also two prominent methodology tools from VSD: Envisioning Cards and Judgment Call. Envisioning cards consist of 32 physical cards and are built on four criteria: stakeholders, time, values, and pervasiveness. The cards allow designers to think about the long-term and systemic issues in technology design [116]. Similarly, Judgement Call is a game that allows the design team to identify ethical concerns in a specific system via self-generated feedback [117]. It also consists of physical cards with four categories: the ethical value, the stakeholder, the number of stars, and wild cards. Our tool draws inspiration from both of them as they provide

an open, engaging, and creative process to think about the impact of values in technology.

2.7.3 Evaluation

Although they promote the discussion of AI ethics, The limitations of tools originating from VSD are that they do not provide user friendly guidelines on thinking about various ethical AI facets suitable for non specialist design team members. We cannot assume that everyone in the team understands everything on ethical AI issues, much less how to address them. With our methodology, we aim to lower the barriers of entry for more lay people to participate in this discussion, people who are direct stakeholders of the new systems that AI promises to bring to reality. This is because most tools available are developed with the purpose of maximising either accuracy, profitability or speed. However With the increased scrutiny that governments and regulatory bodies bring, companies are gradually realising the importance of ethical qualities. While these metrics are important, we need to steer the conversation towards the growing social needs that many have highlighted. Modern AI comes with its own set of challenges that most developers have either overlooked or do not realise at all. If we are going to co-existing harmoniously with AI, there is a need to shift our focus from money or performance towards humanity of our AI. While there is some effort attempting to bring ethics to our enterprises, most of the tools in the market only consider ethics as a side note. In our new methodology, we bring ethics to the forefront and encourage developers to think deeply and critically about how to incorporate them into their systems. Especially

for the team members who may not be well versed in ethics, this tool aims to lower the barrier of entry to allow them to have a voice at the table.

2.8 Summary

In this chapter, we have discussed the key terms and definitions of the main concepts in the field of ethical AI, as well as explore the literature for related work in the design of methodological frameworks:

Transparency: The principle of transparency involves making AI systems understandable and explainable to users and stakeholders. It requires providing clear documentation of the system's design, operation, and decision-making processes.

Accountability: Accountability is the principle of ensuring that AI developers and users take responsibility for the actions of AI systems. This includes creating mechanisms for addressing potential harms caused by AI and holding those responsible for those harms accountable.

Fairness: The principle of fairness requires that AI systems treat all individuals and groups equitably and without bias. It includes considerations of bias in data collection and the design of algorithms.

Privacy: Privacy is the principle of protecting individuals' personal information from unauthorized access, use, and disclosure. It includes ensuring that AI systems are designed to protect users' privacy and prevent unauthorized access to their data.

Safety: Safety is the principle of designing AI systems to minimize the risks of harm to individuals and society. This includes ensuring that AI systems are secure, reliable, and resilient.

However, the state of ethical AI as a field is its fragmented definitions and organisation, which increases the difficulty of creating a unified comprehensive framework to understand ethics in algorithmic systems. In the next chapters, we will discuss our unique approach to tackling this gap in the literature.

Chapter 3

Fairness in Design Methodology

3.1 Introduction

As the pace of advancement of artificial intelligence (AI) technologies increases, it has become intertwined with our everyday life [118]. Alongside the fourth industrial revolution, automation enabled by digital transformation such as AI systems can potentially initiate a new paradigm in the way we work and live [119]. Algorithmic tools are gradually replacing human decision-making. This is of especially important implications in life critical fields like medicine and heavy industries. As such, we must consider the desirable traits that AI should embody and chart a course towards a paradigm of ethical AI.

Currently, the key performance metrics for AI systems are about their effectiveness and efficiency. Most AI development teams are focused on these metrics. However, in recent times, we are increasingly being exposed to evidence that AI systems

can be vulnerable to bias and fairness issues. In 2019, a large scale risk-prediction algorithm in healthcare was found to be less likely to identify African Americans for intensive care management due to a faulty metrics [120]. A high profile investigation in 2016 surrounding the recidivism software Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) found that the algorithm falsely predicted a high rate of recidivism for African Americans offenders [121]. We now understand that biases in training data can negatively affect AI solutions, especially when underrepresented minority groups are involved [122, 123].

As a result of these negative instances of fairness issues in AI systems, it is increasingly important that software development teams to identify and address fairness related issues in their designs. As a community, we need to consider integrating ethical values early in the design and conception stage [115, 124].

The field of digital transformation (DT) has contributed to the increased attention of integrating ethical values to the design and conception of AI products and services. [125] reviewed a large body of works to build a framework that foregrounds DT as a process where technologies create disruptions. This is a process that helps organisations to tweak their value creation paths, improve structural changes and remove barriers. This closely relates to the ethical AI life cycle where software teams aim to reduce the barrier to entry, discover and resolve potential undesirable issues or outcomes before they happen. Lee et al. [126] presented a novel machine learning system for topic modeling to review and analyze advanced DT technologies. The existing literature was automatically classified into several topics for further investigation [127]. Lee et al. [128] created an AI model using data

science and reinforcement techniques to forecast pricing and raw material procurement in the petrochemical industry, thereby enabling business process automation. Methodologies to enable DT such as these frameworks are critical to the continuous improvement and enhancing business core competitiveness.

For so long, key design metrics of AI systems were reliability, efficiency, and accuracy. We were mesmerized by the speed and capacity of these machines and we aim to train them to be faster and better. However, the focus on these metrics above all else overlooked that these AI systems are not as impartial and reliable as we thought them to be. Even though they are machines that follow logic, biases and discrimination can creep into the data and models to affect outcomes to the point of causing harm [122] [123]. The most notable example would be the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software used in state courts systems in the U.S. whereby the model predicted twice as many false positives for recidivism for African-Americans offenders than white Americans [121]. In 2019, a healthcare risk-prediction algorithm that was used on more than 200 million people in the U.S. was found to show racial bias as it was less likely to identify eligible black people for high-risk care management due to a faulty metric [120].

Due to such serious negative consequences of algorithm bias, it is of utmost priority to develop responsible and ethical AI. This pushes us to re-evaluate the design process for creating such systems. One way is to start integrating human values early in the design process, especially when AI systems are capable of embodying political, social, and ethical values. Since product teams train AI with the data

they choose, they have the power to ensure that machines hold up end-user values with a human-centric focus [124]. Therefore, methodologies for product teams to incorporate such values, as well as to avoid and minimize the negative impact of overlooking these values become crucial [115]. Mere awareness from product teams about the importance of human values in AI design is not enough. There needs to be tried and tested tools to help them identify the human values critical to the system, engage such values in their consideration, and discover any potential fatal problems, especially at the earliest stages of the design process when the specifications and prototypes are still fluid [129].

Our methodological tool aims to enable product teams to not only be more aware of fairness criteria for systems but also allow them to surface potential fairness concerns for their application scenario by stimulating perspective-taking from different stakeholders. In this thesis, we will first elaborate on fairness in machine learning and value sensitive design as the foundation for the methodology tool. The instructions on using the tool as well as components of the tool will be described in detail subsequently. We conclude by discussing the potential benefits as well as outlining our future works. We also hope to inspire researchers to construct more methodological tools that enable the integration of human values into the design process.

3.2 Related Work

The AI research community has spent significant effect to formulate notions of fairness mathematically to support algorithmic fairness research [52, 59, 69]. Dwork et al. [60, 70] have divided the notion of fairness into two main categories: individual fairness and group fairness. Individual fairness revolves around one person, and requires that people with similarly attributes receive similar outcomes from AI decision-making [60]. Here, we briefly list the statistical definitions under individual fairness, which are be used in our user studies. Individual Fairness notions include: 1) Fairness Through Awareness [60], 2) Fairness Through Unawareness [61], 3) Counterfactual Fairness [70], and 4) Fairness in Relational Domain. Group fairness, on the other hand, requires that different groups to be treated equally [60]. There are 6 statistical definitions under group fairness: 1) Conditional Statistical Parity [71], 2) Demographic Parity [59], 3) Equal Opportunity [72], 4) Equalised Odds [72], 5) Test Fairness [73] and 6) Treatment Equality [62].

Madaio et al. [130] proposed a co-design checklist to leverage industry practitioners' experience for designing fairness-aware AI. The authors highlighted that despite organisations publishing high level principles to guide the ethical development of AI products, there has been a disconnect between intention and execution. The complexity and abstract nature of AI ethics increase the difficulty of operationalizing it for practitioners. Thus, the proposed ethics checklists in these organisations enable co-designing of ethical AI with active participation from the practitioners working on AI products. Nevertheless, this checklist does not provide an actionable

framework of guidance to help an AI solution design team to organize their brainstorming activities to uncover fairness related issues specific to their application scenarios. Another toolkit - Fairlearn - allows developers to improve the fairness of their AI systems [131]. It aims to mitigate fairness-related harms by framing the fairness issue as a socio-technical problem. However, Fairlearn is limited to addressing only unfairness in classification and regression models. In addition, it is not designed for the team brainstorming stage, but more as a mitigation tool when fairness issues emerge during later stages of development.

Existing ethical AI design methodologies are mainly derived from the Value Sensitive Design (VSD) methodology [113]. It is a theoretical framework that focuses on integrating human values into the design of technologies in a principled and methodical way throughout the design process. SD can help users envision different situations from the perspectives of direct and indirect stakeholders. This analysis helps the users create a list of stakeholders to emulate, as well as explores how they are impacted by specific technology designs [115]. Indirect stakeholders, people who do not use the target technological artefact directly but are affected by its use, tend to be overlooked by system designers. This is despite them being as important as the direct stakeholders, sometimes even more.

The two prominent VSD-inspired technology design methodological tools are: 1) the Envisioning Cards [116], and 2) the Judgment Call game [117]. Envisioning cards use specially designed cards to help stimulate critical thinking from the technology designers. They cover four main criteria: stakeholders, time, values, and pervasiveness. These cards allow designers to consider the long-term and potential

systemic issues in technology design. As it is not specifically designed for AI ethics related issues, the directions of discussion prompted by the envisioning cards do not provide adequate guidance for a design team to uncover AI ethics related issues in their proposed AI solutions. To address this shortcoming, the Judgement Call game provides a more specialized set of cards to cover major AI ethics dimensions (e.g., fairness, privacy, explainability, security and robustness). In addition, it provides stakeholder cards and rating cards to further guide the brainstorming activities by individual team members around specific ethical values. Nevertheless, the methodology assumes that the participants are well-versed in the nuanced notions of the various ethical AI dimensions involved.

The proposed FID methodology addresses the limitations of the envisioning cards the Judgement Call game. Compared to these methodologies, FID provides more nuanced and actionable guide on how to organize team discussions on specific notions of fairness in order to prioritize their project development resources.

3.3 The Proposed Fairness In Design (FID) Methodology

Fairness in Design consists of two forms, the online tool and physical card game. We first developed the physical card game for in depth in person user studies but had to complete the online tool in order to scale and reach more people in the COVID-19 pandemic. The cards are meant to be used by diverse users that

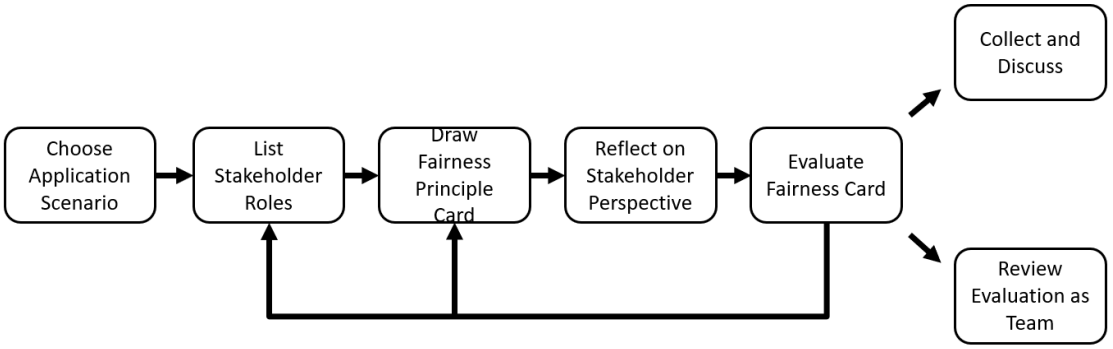


FIGURE 3.1: The FID workflow.

require a methodology to uncover potential ethical problems in their AI products. The methodology works for many types of application domains, from E-commerce to computer vision based systems. While some background experience can enhance the effectiveness of FID, the users are not required to have any prior knowledge to use the methodology, as it is designed to be inclusive to laypeople. The workflow for using FID to facilitate AI design team discussion is shown in Figure 3.1. An overview of the general AI product life cycle is shown in figure 3.2. The stage of interest is in the design and conception life stage, however the methodology can be applied to all stages.

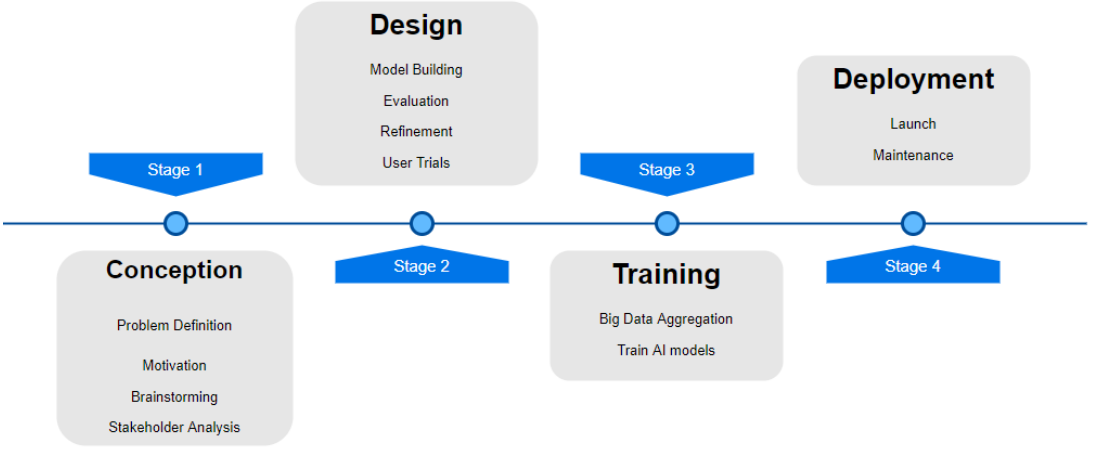


FIGURE 3.2: The FID workflow.

We have consolidated the fairness definitions and notions from the literature into 10 principles, further categorised into group fairness and individual fairness. These principles are represented as cards[1], to be used during the reflection of stakeholders' perspectives in the FID workflow.

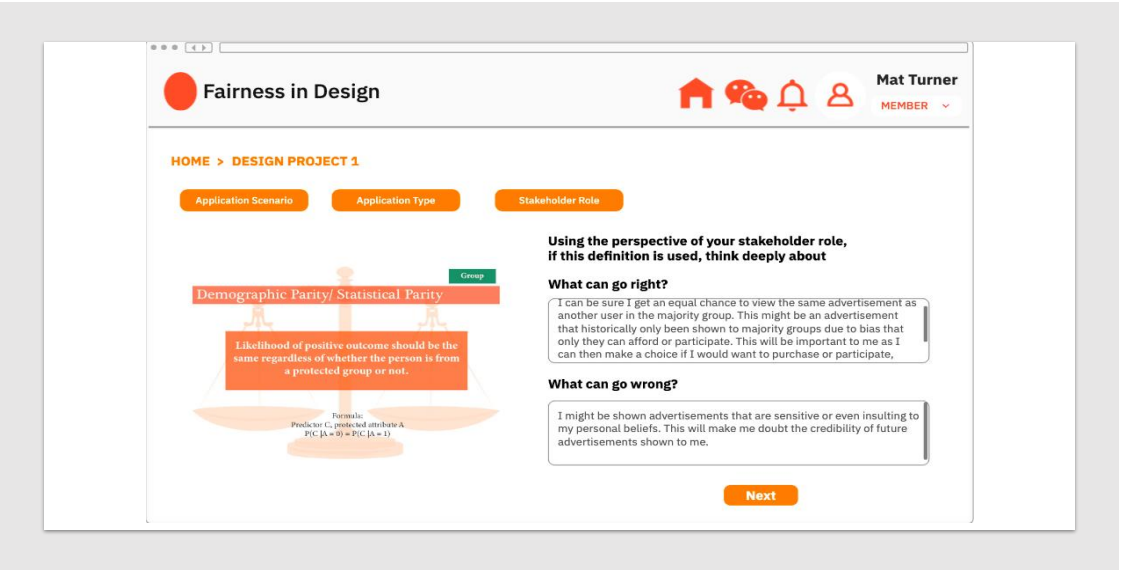


FIGURE 3.3: An example user interface of FID (writing an envisioned stakeholder review for the given fairness notion - demographic parity)

- Initially, the team members need to decide on the application domain that sets the environment for the rest of the session. This can be a fictional or a real world product, depending on the needs of the team. As many details of the AI system or product should be included as possible for FID to achieve a meaningful output, this is because there different considerations and trade offs in different domains which can in turn affect the decision making process.

- Next, in step 2, the team members have to pick a type of application card that best describes their application domain. We have adapted 5 categories

[1]The complete set of cards can be found at: `https://sites.google.com/view/zhang-jiehuang/fairness-in-design`

of application domains from Shneiderman's classification for usability motivation in the Human Computer Interaction literature. They are namely: 1) life-critical systems, 2) industrial and commercial uses, 3) office, home and entertainment, 4) exploratory, creative, 5) collaborative applications, and 6) socio-technical applications.

- Next, in step 3, the application domain will have a set of stakeholders that are key to the analysis. We differentiate the two types of stakeholders, namely direct and indirect stakeholders [115] in this step. Direct stakeholders are people who use the AI system in the application domain directly, while in contrast, indirect stakeholders are people who do not use the system directly but are impacted by its use. Each team member will identify one stakeholder group, and brainstorm issues that they may face from the perspective of that stakeholder.

- For step 4, each team member draws a fairness principle card and applies it to the application domain. If the fairness principle card is not applicable to the application domain, he can choose to discard it and draw another. These cards illustrate the fairness metric, according to the 10 fairness principles we discussed earlier and how it is applied to AI systems.

- In step 5, the team member then applies the fairness metric to the application domain and stimulates the potential problems or solutions that that stakeholder will face. We ask the question of what could go right or wrong

for that stakeholder to elicit critical thinking from the team member. He will then write his thought process on the card.

- For the final step 6, the team compiles all the responses of the members and randomize them before reading it out loud. In this way, the responses will be anonymous and this can encourage team members to be more truthful. The team can discuss and evaluate the responses to decide if they are valid and worth addressing. The team members then rate fairness principles on a Likert scale to evaluate their importance to the application domain.

- Once the process is completed, the team can repeat the process by going back to step 3 and conducting a new stakeholder analysis, or simply conclude the session.

Figure 3.3 shows an example user interface of FID. In this case, the user is at Step 3 of writing an envisioning review from his adopted stakeholder perspective. The given fairness notion in this case is demographic parity for which there is a card-based quick reference guide for the user to refer to. He is asked to think about what can go right and what can go wrong for the current design of his AI solution under the given context. In addition, to supporting ideation, the design input and decisions recorded by FID can be a useful source of information for tracing the origin of design issues. A demonstration video can be found at `https://youtu.be/nnowNLss_wQ`.

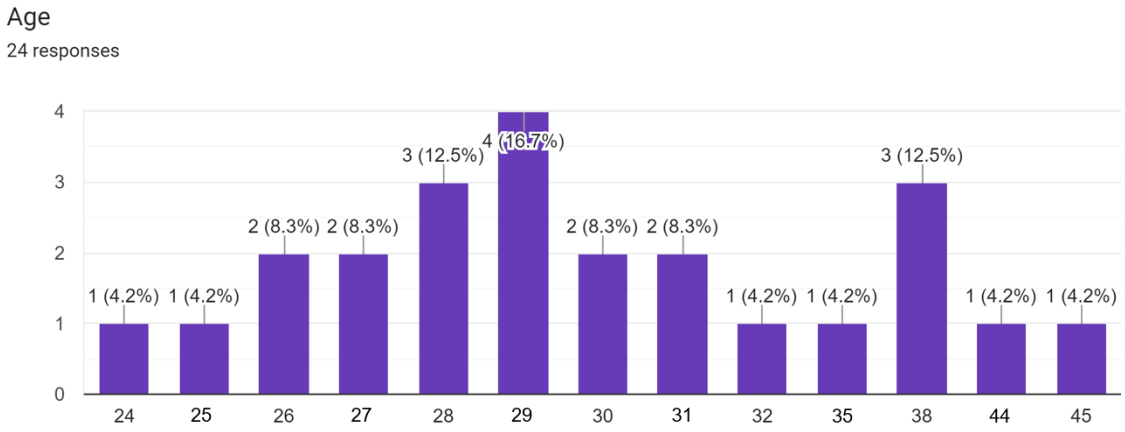


FIGURE 3.4: Frequency (y-axis) and age (x-axis) of the Participants.

3.4 Empirical Evaluation

We conduct user studies to empirically evaluate the proposed FID framework.

3.4.1 Study Design

A total of 24 participants (18 male, 6 female) were recruited for the user study. All of the participants are experienced researchers or engineers who are currently or have previously worked on software systems involving AI. Additional recruitment criteria include the ability to understand the basics of fairness concepts, and consenting to be recorded. We recruited participants with a diverse range of ages for this user study in order to investigate how the methodology can impact usage based by users with different levels of seniority. Most of the participants fall into the 30s age group (Figure 3.4), which is representative of the typical target users of FID.

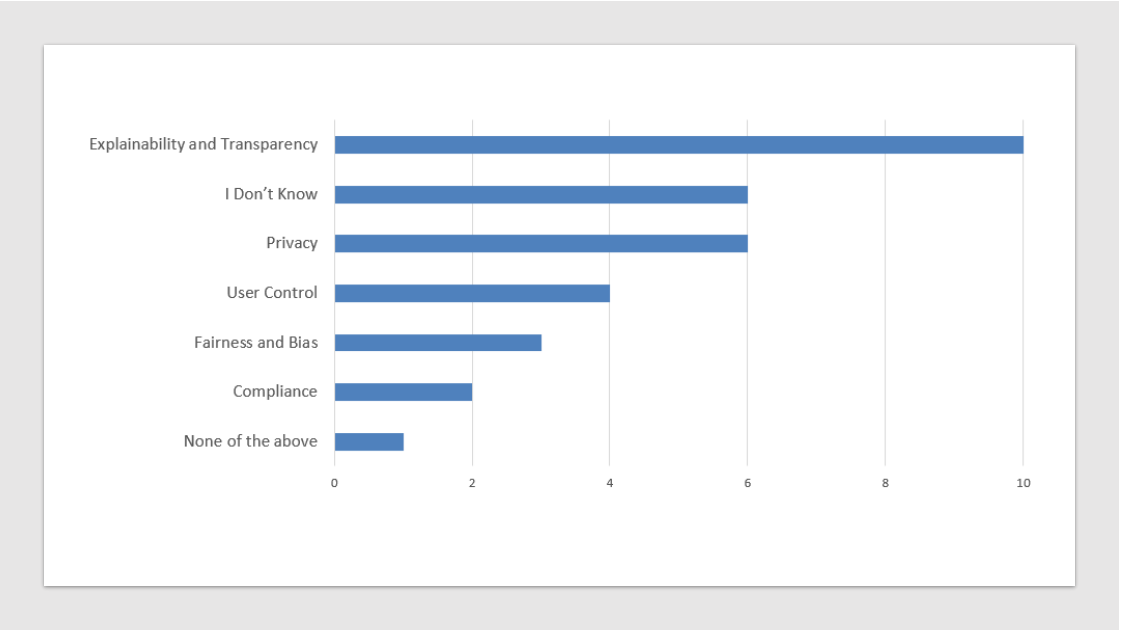


FIGURE 3.5: Participants' ethical AI prioritization.

Before the start of the actual user study, we instructed participants to complete a pre-study questionnaire through Google forms in order to understand how they prioritize ethical considerations in their AI solution development experience (if at all). As shown in Figure 3.5, most of the participants indicated explainability and transparency as the most important criteria in this question. This is understandable as there has been increased attention in this area of machine learning in the research community. While fairness and bias ranked third from the last, the participants do not mean that fairness is any less important. Rather, it is primarily because of the lack of support for the complex multi-faceted concept of fairness to be considered during the design stage.

As shown in Figure 3.6, most of our participants are working on AI solutions in the healthcare application domain. We included a redundancy test in the questionnaires, by asking the same question twice, once in a positive way and once

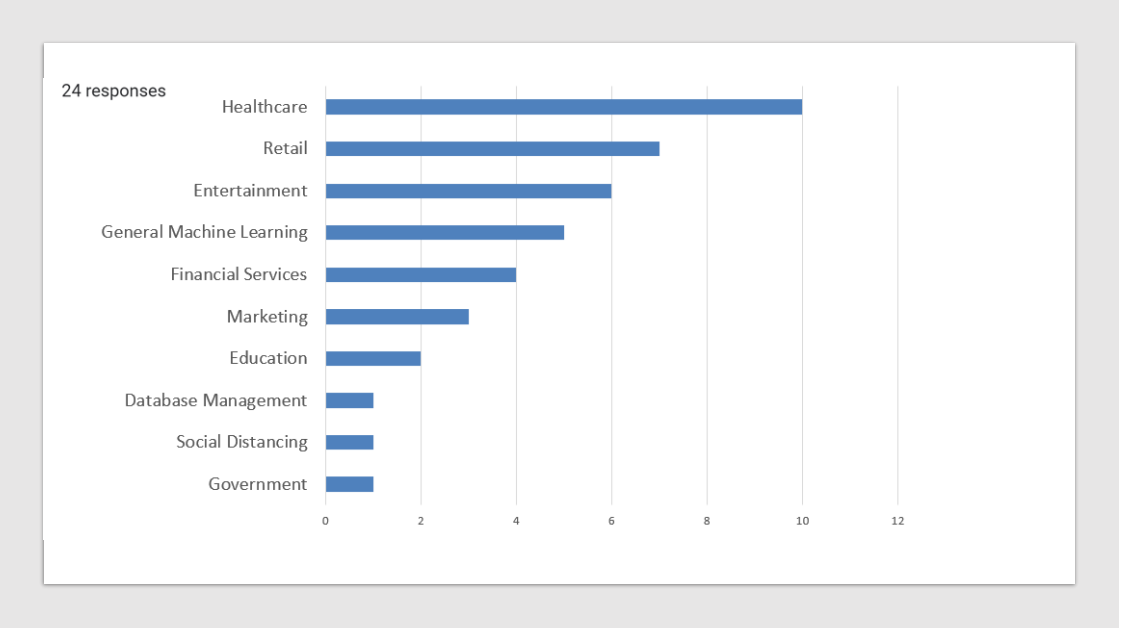


FIGURE 3.6: Participants' experience developing AI applications.

in a negative way. For example, we asked the following positive question "I am able to compare across different possible strategies for addressing a fairness issue in my application area", and subsequently the negative question "I do not know how to evaluate different fairness solutions". By using the redundancy check, we are able to detect responses that were not valid. Additionally, we ensure that the post-study questionnaire was completed within 3 days of the user study.

The three main hypotheses for this study are as follows:

- The FID methodology helps participants determine the fairness criteria that are the most relevant to their AI applications.
- The FID methodology helps participants surface fairness concerns in their AI applications.

- The FID methodology helps participants envision the perspectives from different stakeholders.

We designed our pre- and post-study questionnaires for the participants to conduct self assessment of their fairness related techniques. Each hypothesis is intended to assess the individual ability of the participants to choose an applicable fairness solution, brainstorm and surface fairness concerns and approach problems from the perspective of stakeholders respectively. They are required to rate their understanding of fairness problems from a Likert scale of 1 to 5. Based on the results of the questionnaire, we conduct statistical analysis to evaluate the effectiveness of the three proposed hypotheses[132] [133].

3.4.2 Results and Analysis

In this section, we analyse the results from the empirical studies by presenting the findings related to each hypothesis.

3.4.3 Hypothesis 1

Hypothesis 1: The FID methodology helps participants determine the fairness criteria that are the most relevant to their AI applications.

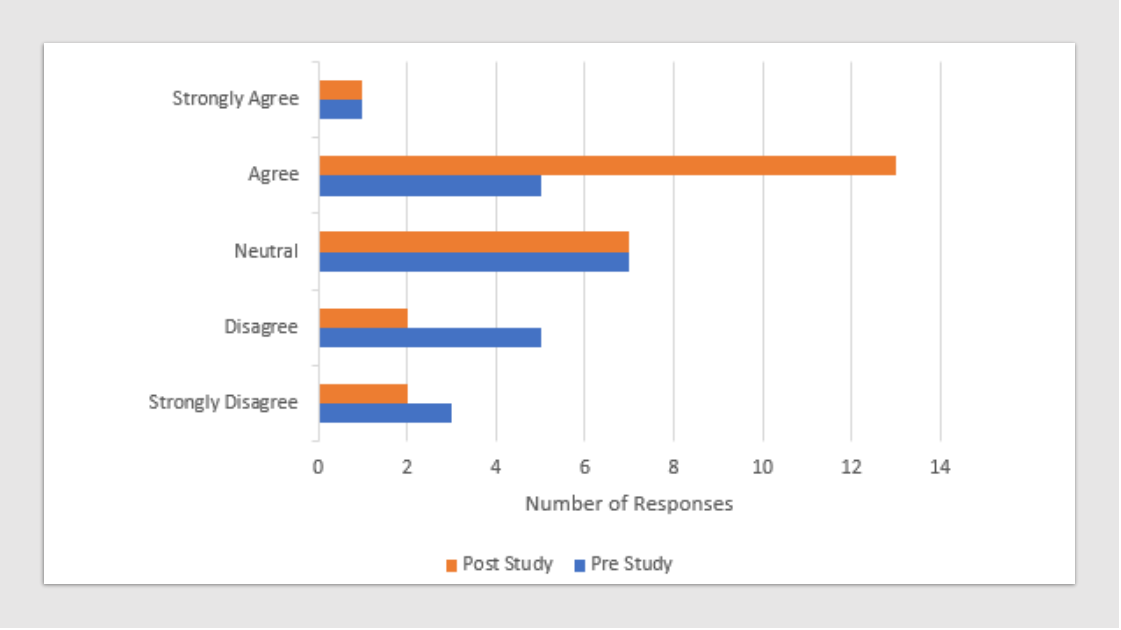


FIGURE 3.7: Participants' self-reported capability of making design decisions related to fairness before and after using FID.

Figure 3.7 illustrates the results from participants' responses on questions related to H1. It can be observed that in the pre-study, participant responses follow a normal distribution centred on "Neural". This shows that the distribution of their capabilities to make design decisions related to the fairness aspect of AI was typical of a population of AI solution designers. After using FID in the empirical study sessions, there is a significant increase in the number of participants who responded with "Agree", while the number of "Disagree" and "Strongly Disagree" responses decreased. The number of "Strongly Agree" and "Neural" responses remains unchanged. The results show that participants found FID to be useful for helping them think about design decisions related to incorporating fairness into AI solutions. As shown in Figure 3.8, the participants' average response scores in the post studies are significantly higher than in the pre-studies.

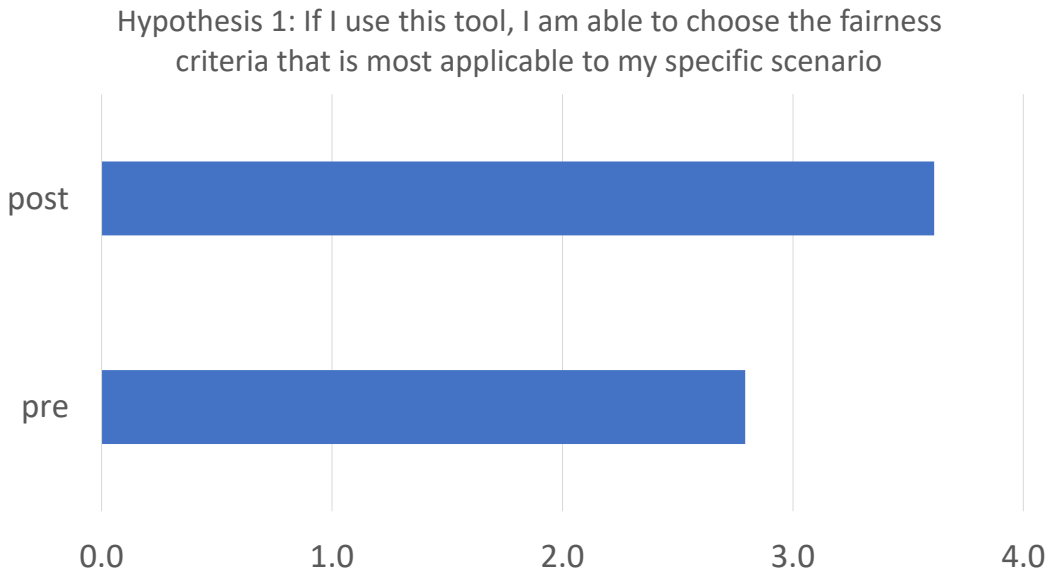


FIGURE 3.8: Participants' average scoring for the pre- and post-studies for hypothesis 1

After conducting a students' t-test analysis of questionnaire results from H1, we conclude that the null hypothesis can only be rejected at the 90% confidence level.

3.4.4 Hypothesis 2

Hypothesis 2: The FID methodology helps participants surface fairness concerns in their AI applications. This hypothesis pertains to the participants' self assessment of their competency in discovering fairness issues ahead of time.

In Figure 3.9, we illustrate the results from the participants' responses about surfacing fairness concerns, focusing on Hypothesis 2. This question pinpoints the self-assessed ability of participants to identify ahead of time what type of fairness issues can arise in their specific application domain. As before participants responses are

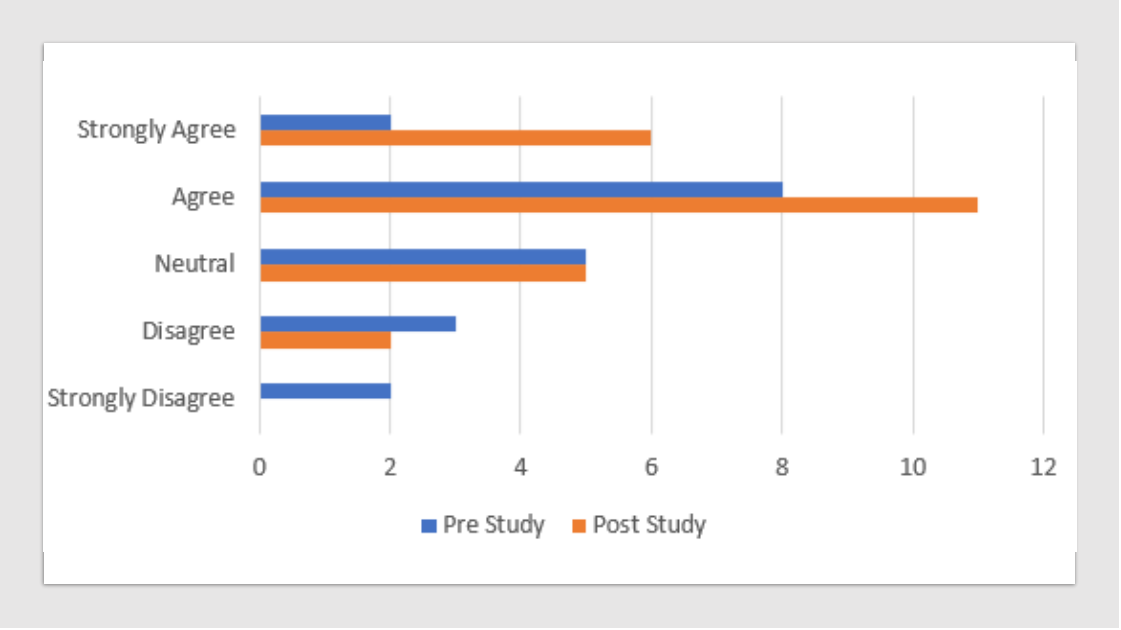


FIGURE 3.9: Participants' self-reported capability of surfacing fairness concerns before and after using FID.

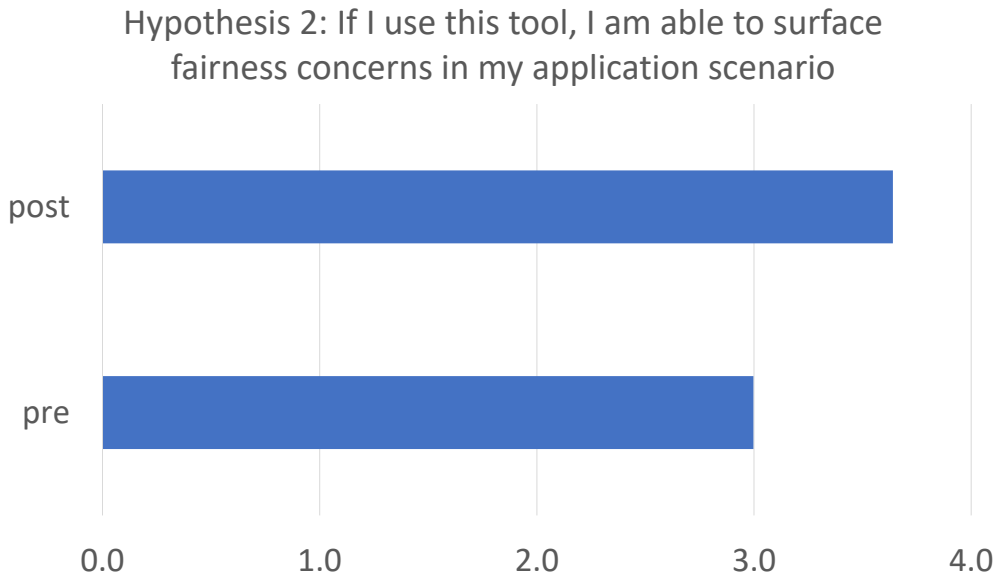


FIGURE 3.10: Participants' average scoring for the pre- and post-studies for hypothesis 2.

roughly normally distributed, centred on neutral. After using FID, there is a significant increase in the number of responses "Agree" and "Strongly Agree", as well as a corresponding decrease in the number of responses for "Disagree" and "Strongly

Disagree". The results indicated that FID is effective in assisting participants to surface potential fairness issues in their application domains.

For H2, we find that the questionnaire response averages increased by more than 0.5 in the post-studies compared to the pre-studies (Figure 3.10). After conducting a students' t-test, we are able to reject the null hypothesis at 95% confidence level.

3.4.5 Hypothesis 3

Hypothesis 3: The FID methodology helps participants envision the perspectives from different stakeholders. We conceptualised this hypothesis to assess the ability of participants to stimulate thinking in the perspective of other relevant groups of people.

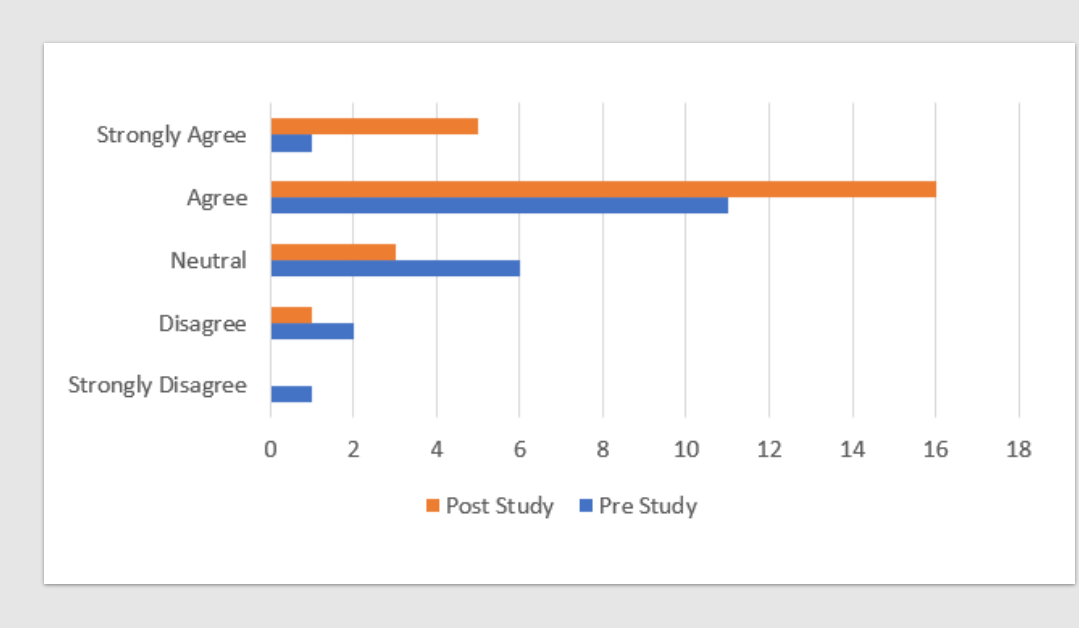


FIGURE 3.11: Participants' self-reported capability of thinking from stakeholders' perspective before and after using FID.

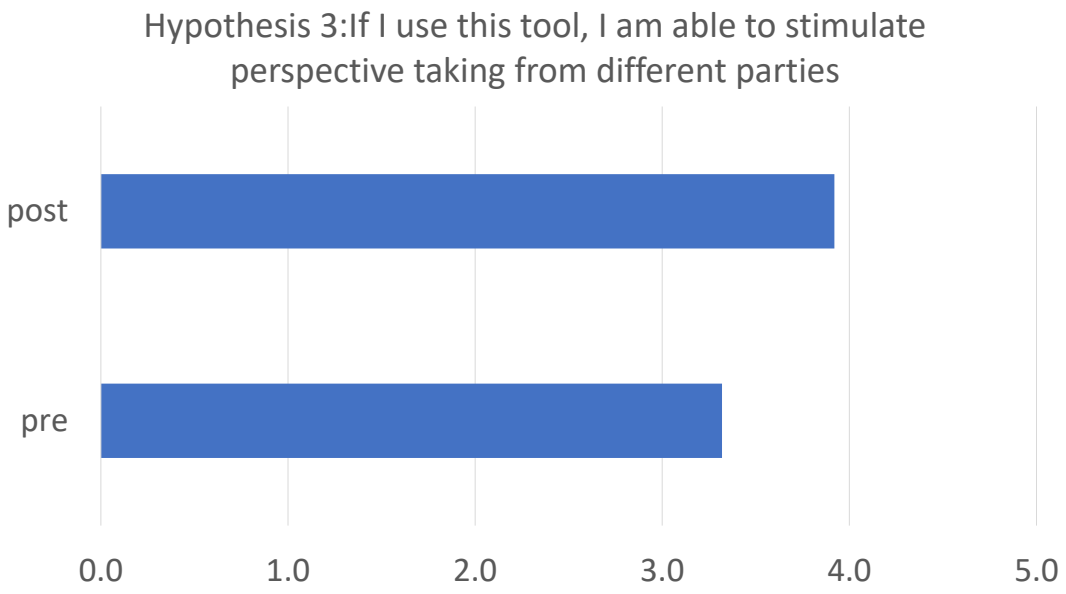


FIGURE 3.12: Participants' average scoring for the pre- and post-studies for hypothesis 3.

In Figure 3.11, we highlight the results from the participants' responses on thinking from the perspective of stakeholders. This question focuses on hypothesis 3 and challenges the participants to visualise and think from the shoes of 2 types of stakeholders, such as the end-user, legal and marketing staff, as well as indirect stakeholders such as families of end users. The proportion of "Disagree" and "Strongly Disagree" greatly decreased and the participants changed their answers to "Agree" and "Strongly Agree". In particular, the frequency of "Strongly Agree" increased from 1 pre-study to 5 post-study. The methodology was able to greatly improve the self-assessed ability of participants to stimulate stakeholder perspective, which is a valuable skill set in AI development teams.

For H3, we find that the questionnaire response averages increased by more than 0.5 in the post-studies compared to the pre-studies (Figure 3.12). conducting a students' t-test analysis of questionnaire results from hypothesis 3, the null hypothesis

can only be rejected at 90% confidence level.

3.4.6 Discussions

We found FID to be effective in promoting conversations about surfacing fairness concerns of AI projects. However, it is only suitable for projects that are in the design stage, due to the reason that fairness metrics need to be built into the system early before other tasks. We highlight one of the important questions in our questionnaire regarding making decisions about fairness in their projects. We found that a significant number of participants are more confident after the study, 14 agreed or strongly agreed post-study compared to 6 pre-study.

Additionally, since fairness is a complex topic in AI development, most system designers are unfamiliar with it, or see it as an unnecessary trade-off for performance or efficiency. In other words, most AI developers are not willing to forgo the reduction in their algorithm's performance, unless there is a specific requirement. This response has been expected, however we believe the direction of the field is that eventually teams can deliver a system that minimises the compromise on both values.

Over the course of multiple user studies, we discovered that the target systems are to specifically collect data on protected and unprotected individuals/groups, then subsequently making decisions based on these data. We also tweak the process numerous times in response to the feedback obtained. For example, we realise the need to provide the option of discarding irrelevant fairness principle cards to the

participants since some principles were more important to the application scenario than others.

3.4.7 Limitations

Our user study currently only includes 24 recruited participants, we are planning for a larger-scale study with only online participants to conduct a more in-depth evaluation of the FID methodology. Additionally, it has been explored that self-reported preferences often do not align well with user's actual behaviours [134]. It remains to be seen if the findings from existing and past work contributes to significant improvements in our methodology. As the fairness notions elaborated to the users in FID are adapted from the artificial intelligence literature, the FID tool is useful for algorithms that collect data to train a machine learning model. We also take into consideration the minority groups or communities that are at high risk of being intentionally or unintentionally overlooked.

3.5 Conclusions and Future Work

In this chapter, we identified the research gaps present in the ethical AI literature and highlighted the need for a methodological tool that allows for deep analysis of potential ethical issues. After exploring the fairness literature as well as VSD, we identified the significant fairness principles and created a methodology to assist software designers to understand fairness issues, then create strategies to address

them and overcome biases. Our conversations with product teams revealed that they usually viewed fairness as an after thought and there are many barriers to even consider the fairness issues in their teams. We designed this FID methodology to have a low barrier to entry and is easy for laypeople to use effectively. Our target audience of AI product teams find the methodology effective and useful to explore and make fairness decisions in their application domains. With this methodology, we hope to inspire others in the ethical AI research community to construct more methodological frameworks that assist AI product teams to consider ethics in their AI systems. To the best of our knowledge, FID is the first technical tool to facilitate AI solution development teams to incorporate fairness into their designs. Empirical results from our user studies involving 24 AI solution developers show that FID can improve design teams' understanding of fairness concepts and is perceived to be useful for their projects.

In subsequent research, we will be looking to scale the usage of our FID methodology to a larger base of users such that more improvements and tweaks can be made. By leveraging on the synchronization of circumstances, stakeholders and technology in the digital transformation, we aim to optimise complex concepts and processes for the benefit of laypeople. At the same time we aim to extend the methodology to other ethical values such as privacy and explainability, and create a unified methodological framework that is the go-to tool to consider ethics in AI.

Chapter 4

Explainability in Design Methodology

4.1 Introduction

In the age of digitization and the fourth industrial revolution [118], several enabling technologies includes artificial intelligence (AI). AI systems are the key to new breakthroughs in important fields and numerous fields, such as medicine [135], algorithmic crowdsourcing [136] and self driving vehicles [137]. AI systems, especially machine learning and neural network-based or deep learning systems, have allowed us to perform many task with greatly increased scale and finesse, technological breakthroughs that were believed to be unattainable without AI. While these advancements facilitated by AI has resulted in a great changes in the way we live and work [119], there are complex and multi-faceted morality issues present in

these systems that warrants attention. This is significant as more decision support systems otherwise solely handled by humans are being transferred to the responsibility of AI. Due to this shift of autonomy and responsibility to algorithms, the chances of mistakes or improper decision making must be addressed in a timely manner before more side effects emerge. As part of a concerted approach, societies heavily reliant on AI systems must consider the repercussions of such a shift and regulate the advancement of AI towards a future direction with proper oversight where the benefits outweigh the potential harms [138].

At the current stage of the AI community, most of the practitioners in the AI circle assess the performance of AI systems based on their accuracy scores and impact on computing resources. Despite the usefulness of these metrics, evidence also exists to prove that they may not give a complete representation on the inner workings of the decision making process. While SOTA (state-of-the-art) AI systems can assist or replace many process in the workplace and peoples' personal lives, they generally lack explainability and even the system designers are unable to fully explain how they work [139]. Despite being trained on factual, logical datasets, these algorithms are not invulnerable to mistakes of misjudgements and various other issues that can be difficult to detect [140]. Therefore, it is possible that on top of being unable to understand how algorithms reach their decision output, there might be problems that even go undetected for a long period of time. For example, we refer to spectral heat-maps, Lapuschkin et al. discovered that standard performance evaluation metrics can possibly be unaware of certain types of issues in the decision making [141]. Ultimately, the consensus in the community is that black

box AI used in modern times do have problems relating to the transparency and explainability of their inner workings, especially in specific fields where the users and other stakeholders need to know how the output is being reached. Some examples of these fields are medical diagnosis and self driving cars. Hence it is urgent and necessary in these fields that a satisfactory explanation can be generated, or given when it is prompted.

Explainability is a complicated endeavour in the context of AI as it is multifaceted and its definition can be fluid depending on the context and type of requirements. The field of explainable artificial intelligence (XAI) [89] aims to create an arsenal of machine learning tools that enable users to understand, trust and constructively regulate the advancing generation of AI systems [142]. These goals are within reach when designers intentionally create AI algorithms to have features that enhance explainability, and are comprehensible from the perspective of human users. While achieving XAI has been increasingly difficult, progress has been accelerating in this field [143] (e.g., Layer-wise Relevance Propagation [144]). These steady improvements in explainability are laying the foundation for the vision of XAI as we gain greater clarity into how complicated AI models such as deep learning neural networks function. On top of creating tools and methods to enhance explainability in algorithms after they have created an output, or in other words, post hoc manner, there is a need to tackle the problems early in the design and conception (DC) stage. There is currently a lack of methodological frameworks for AI software development teams and research groups to integrate explainability measures in their AI products and/or services.

This chapter proposes the Explainability in Design(EID) methodology, which is a step by step framework that guides software design teams and research groups to systematically consider, explore, surface and resolve any explainability related issues and problems in their AI systems. The EID methodology is designed simultaneously to elicit critical thinking during all stages of the AI life cycle, and reduce the barrier of entry to allow lay people to participate in the conversation of ethical AI. Explainability is one of the major pillars of which ethical AI is built upon, and in many ways allow for improvements in other pillars such as privacy and fairness. However due to the complicated nature of AI explainability, lay people often find the concepts and technicalities hard to understand, let alone contribute to this endeavour. To further increase the difficulty, the desire to enhance explainability may lead to tradeoffs in the accuracy or efficiency as human or computing resources might be removed from the main objectives in the AI life cycle. EID aims to address these issues, by introducing teams to the systematic process of brainstorming and discussing the likelihood of explainability related problems for their AI products. This allows for software teams and research groups to identify, or create explainability requirements and objectives specific to the AI system and context, while stimulating perspective thinking from other groups of stakeholders, whether direct or indirect.

Through empirical user studies involving 35 AI designers, EID is shown to significantly enhance the ability of software design teams to identify, explore and resolve ethical issues and problems surrounding explainability. In additional, EID also helps to reduce the barrier of entry for team members to effectively participate in

the design and conception process, allowing a greater pool of participants to improve AI software products and services. EID can assist software teams in diverse application domains, from e-commerce, to facial recognition systems and beyond. By designing it to be application agnostic, we enable a large number of AI engineers and researchers to benefit from using EID in their work.

4.2 Related Work

The field of XAI can be quite rather large and complex, making the difficulty of creating a taxonomy high. For the purpose of this chapter, we broadly classify the techniques into two main categories, Post Hoc and Integrated approaches [89]. This classification differentiates the two categories by the stage of the life cycle where the techniques are applied. Integrated XAI refers to the in-building of explainability features during the design and construction of the algorithm, while post hoc XAI means that the explainability of an algorithm is only investigated after the output has been produced. Between the two categories, there are advantages and flaws associated that we will explore further in the chapter and subsequent user studies. For now, the obvious advantage of post hoc explainability is that there is a low chance that this approach will interfere with the performance of the AI system. The research community in XAI is active and many new improved, novel methods of enchancing explainability in AI systems have emerged in recent times, some examples include Shapley values [145] and LIME [93].

To the best of our knowledge, Value Sensitive Design (VSD) is currently the prominent toolkit in the field of ethical AI methodological frameworks [113]. VSD is closely related to the field of human computer interaction (HCI) and information systems design, the toolkit aims to resolve design issues by centering the analysis around the ethical values such as privacy and fairness. These values are used in the workflows, allowing system designers to gain deeper insights and integrate with other methodological tools. The brainstorming sessions also considers the roles, values and goals of both direct and indirect stakeholders, by stimulating perspective taking in the process. The main difference between direct and indirect stakeholders are that direct stakeholders use the AI product or service directly, while indirect stakeholders are impacted by the use, but do not directly use the AI. Since the effect of AI use is less apparent on indirect stakeholders and as a result has a higher probability of being overlooked, it might be better to allocate more time and attention to the analysis of indirect stakeholders. VSD has been the basis of two methodological card games, Envisioning Cards [116] and Judgement Call [117]. Envisioning Cards, as its name suggests, help to stimulate critical thinking and emphasizes players focus on the timelines, stakeholder interests and values, as well as pervasiveness. They prompt participants to consider the long term and likely systemic problems in system design. In contrast, Judgement Call is a card game that AI developer groups can use to surface ethical problems in a AI product. It consists of cards that primarily focuses on scoring reviews and using wild cards that facilitate critical thinking.

Liao et al. [146] brings to light that while AI systems need explainability features,

there is a consensus to address on the ground, real world user requirements before we understand algorithms. The authors invented a question bank in which user needs for explainability are portrayed as prototypical questions users might ask about the AI and use it as a study probe. Then, they consulted Usability Experience (UX) and design experts on the current gaps between XAI algorithmic work and practices. [147] showed that there is a lack of a principled framework that can provide the basis for the development of a XAI framework. Then the authors came up with four foundational components that can assist to create a simple methodological framework to facilitate the design of XAI systems.

Nevertheless, there is no existing software engineering design methodology to guide an AI solution development team to brainstorm and determine what XAI principles should be incorporated into a given AI system design. The proposed EID methodological framework is designed to fill this gap.

4.3 Preliminaries

In this chapter, we have classified the principles of XAI into the three main types as shown in Figure 4.1: 1) transparency, 2) interpretability and 3) explainability [89]. This allows lay people to focus on the most relevant or important principles of XAI, and brainstorm how to apply to their application scenarios.

- Transparency: an algorithm is transparent if it is transparent when viewed

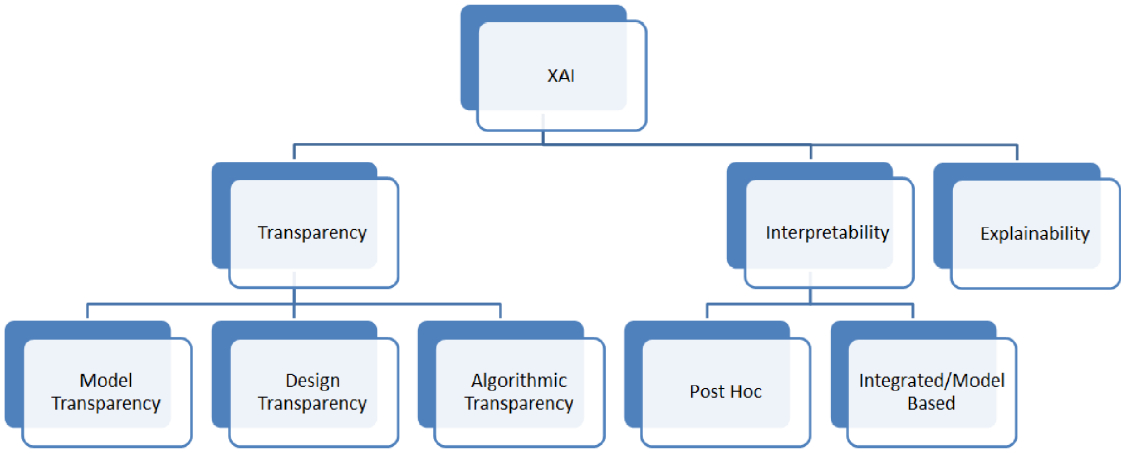


FIGURE 4.1: Metrics of Explainability in AI Categorised into 6 types

in a stand-alone manner. An algorithm can feature different degrees of understandability

1. Model Transparency: Degree of human understandability of how the components (e.g. filters used, layers of a NN) of the trained model contribute to an output.

2. Design Transparency: Degree of human understandability of design decisions made to create the machine learning model.

3. Algorithmic Transparency: Degree of human understandability of the training process that resulted in the trained machine learning model.

The model and design transparency can be distinguished by the specific section of the transparency analysis: model transparency focuses on the way individual components are sequenced, while design transparency refers to the design choices made by the engineer to enhance the transparency of the algorithm. For example, for an algorithm to be more design transparent, a simpler algorithm such as decision forest can be used. While in order to

enhance model transparency, the components of the algorithm need to be easily explainable to a lay person. To the untrained lay person, the two types of transparency can be viewed as largely indistinguishable, the goal of this classification system is to inform them of the minute but significant differences.

- Interpretability: the ability to explain or to provide the meaning in understandable terms to a human

    1. Integrated Interpretability: Design of a ML model that involves specific design choices for better understandability.
    2. Post Hoc Interpretability: Ability to analyse information pertaining to how the output of a trained ML model is obtained from the input.

- Explainability: associated with the notion of explanation as an interface between humans and a decision maker, the focus is on the human and how the human can understand the mechanics of an algorithm.

There is also a need to differentiate between interpretability and explainability. The main difference is the human factor, the measure of explainability includes a human being understanding the output explanation. In contrast, interpretability is the degree of understandability of the algorithmic decision-making process, not based on human factors but relative to other algorithms. In this methodology, the humans concerned are the AI system designers, as well as the direct and indirect stakeholders. The subtle differences in these principles allow for interesting trade-offs and interplays when the specifics are given due attention and analysis.

FIGURE 4.2: Comparison between direct and indirect stakeholders

4.4 The Explainability in Design Methodology

The Explainability in Design (EID) Methodology is integrated into the gameplay processes of a card game. The methodology is designed to be relevant to a wide range of application scenarios, in other words, model agnostic, allowing for the team to adjust fine details to tailor the methodology to their AI product or service. By lowering the barrier to entry to use EID, both lay participants and more experienced people are enabled to join the process and collectively brainstorm towards realistic improvements.

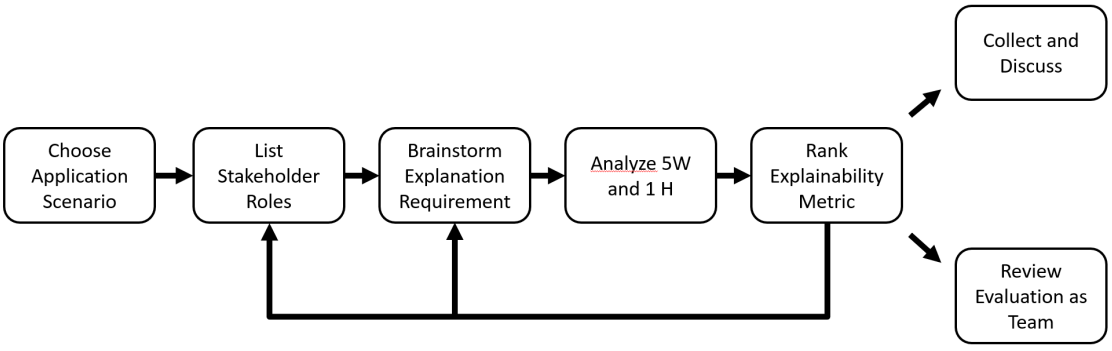


FIGURE 4.3: The Explainability in Design Workflow

The workflow for a team to use EID to facilitate team discussions around XAI

issues is shown in Figure 4.3. We discuss the details of the step by step guide to using EID workflow:

- Initially, each team member selects a scenario that establishes the context for the entire user study. The scenario can be real or fictitious, but very importantly it is preferably in a field where explainability is of high priority and better if the dataset of the scenario partially or fully contains dataset generated by people. If possible, most of the team should be acquainted with the scenario such that as many fine details are considered in the flow of the user study. This is due to the reason is that sometimes in selected scenarios, there are considerations and trade offs that can impact the decision-making process.

- In Step 2, participants must select the type of application card that accurately reflects their scenario. We have chosen some considerations from the work from Shneiderman's classification for usability motivation in the Human Computer Interaction literature [148]. They are 1) life-critical systems, 2) industrial and commercial uses, 3) office, home and entertainment, 4) exploratory, creative, collaborative applications, and 5) socio-technical applications.

- In the next step 3, the team must find and list the types of stakeholders that are central to the process. Armed with the list of direct and indirect stakeholders, each individual must stimulate the perspective of a stakeholder and perform an analysis of the principles pertaining to that stakeholder. To

facilitate this process, the EID methodology includes a list of exploratory questions in this step. Some examples include:

1. Other than the stakeholders, who can the output explanations be targeted at?
2. Does the timing of the explanation generated matter to the user?
3. How can the frequency of use impact the trust levels of the user?
4. Does the emotional state of the user impact other factors?
5. How can the algorithm perform this operation but in a more transparent way?
6. Determine the scope and depth of the explanations needed in this scenario.

As a result of this step, each individual participant should grasp the perspective of their stakeholder deeply, which in turns facilitates the XAI principles to be explored. These requirements are even more pronounced when the analysis of indirect stakeholders occur and the impact on them may not be so apparent initially.

- In the next step 4, the team members must select the highest rated explainability principles relevant to the scenario. They have to justify the selection and why the other principles were not chosen. When a specific principle is chosen repeatedly, we can identify it as the default principle that works well for the scenario.

- In the final step 5, the leader of the group collects the responses of the team and randomly shuffles them before reading them to the team. With the element of anonymity, the members are spurred to be more forthright in their decision-making processes. The leader will assess the responses to decide if any further action or step is to be taken.

At the end of the workflow, the leader of the group can consider if there is a need to return to step 3 to initiate a new stakeholder analysis. The members can also shift their focus to another part of the study to ensure it is well covered. If the results of the workflow are satisfactory, he can end the process.

The methodology aims to provide the following deliverables:

- Selecting the explainability metrics most suitable for the scenario.
- Determine priorities of specialised requirements for the format, scope and type of explanations.
- Measure the differences in the knowledge and experience levels of team members.
- Facilitate the discovery of explainability issues and where in the AI system it is located.

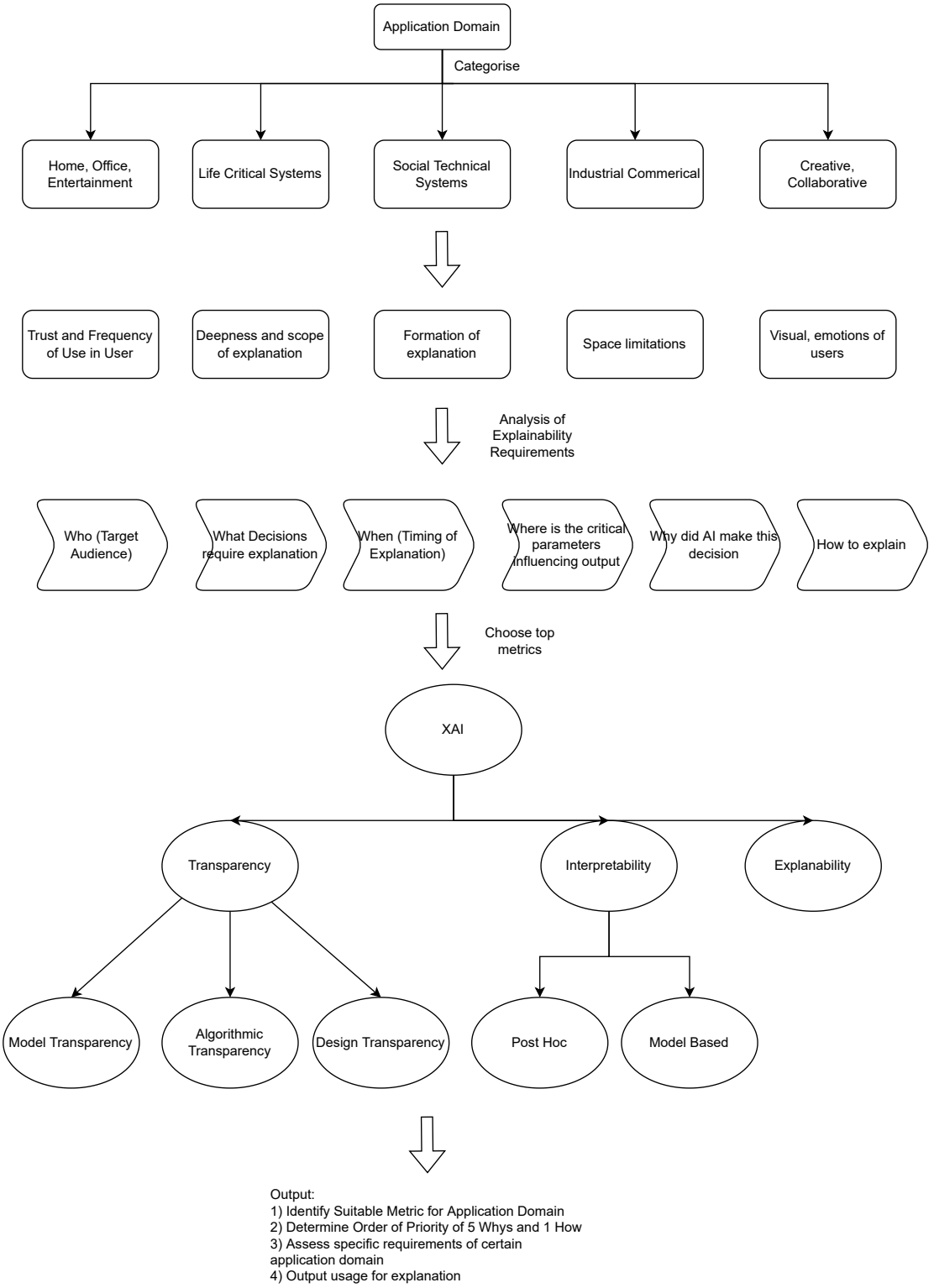


FIGURE 4.4: Overview of the entire EID methodological workflow

4.5 Empirical Evaluation

We conducted user studies to empirically evaluate the proposed explainability in design framework, and our hypotheses.

4.5.1 Study Design

A total of 35 participants were recruited for the user study. All of the participants are experienced researchers or engineers who are currently or have previously worked on software systems design involving AI technologies. We also considered additional criteria, for example the ability to understand basic explainability concepts surrounding the machine learning literature, as well as consenting to being recorded. We recruited participants with a diverse range of ages for this study, intending to also investigate how the methodology can impact usage based by users with different levels of seniority. Most participants fall in the 20 - 30s age group, which is representative of the typical target users of our framework.

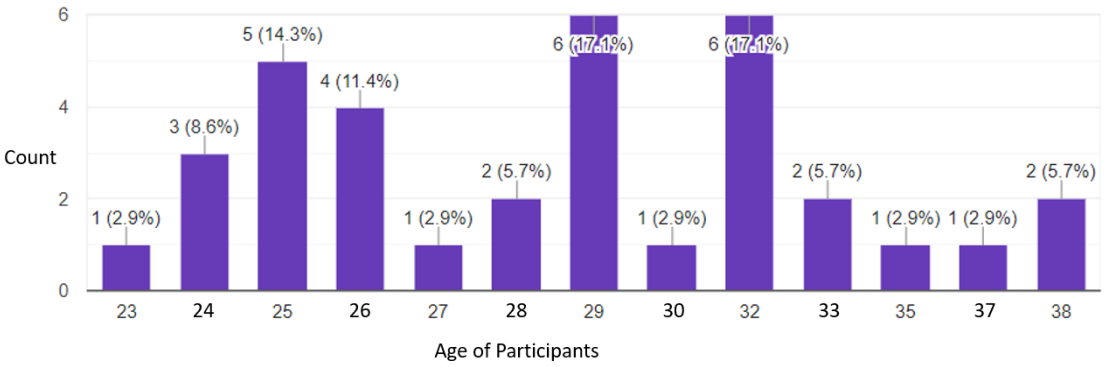


FIGURE 4.5: Frequency (y-axis) and Age (x-axis) of the Participants.

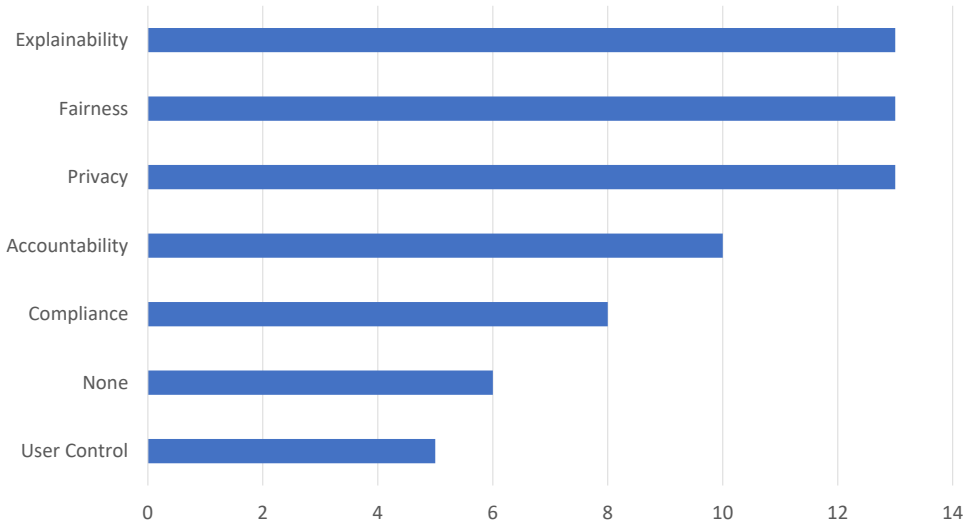


FIGURE 4.6: Participants' ethical AI prioritisation.

In the pre study questionnaire, we asked participants to report how they prioritise ethical considerations in their AI solution development experience. As shown in figure 4.6, most participants chose explainability, fairness and privacy as the top 3 ethical considerations, followed by accountability and compliance. This corresponds with the explainability aspect of ethical AI that we have chosen to focus our user study on.

We also asked participants to identify the application domain where they have worked on their AI products and services. According to 4.7, most of the participants are working in the healthcare sector, general purpose machine learning (ML) applications as well as government related projects. In order to improve consistency in our questionnaire, we included a redundancy test by asking the same question twice, once in a positive way and once in a contrasting negative way. For example, we asked the follow positive question "I can navigate complex ethical

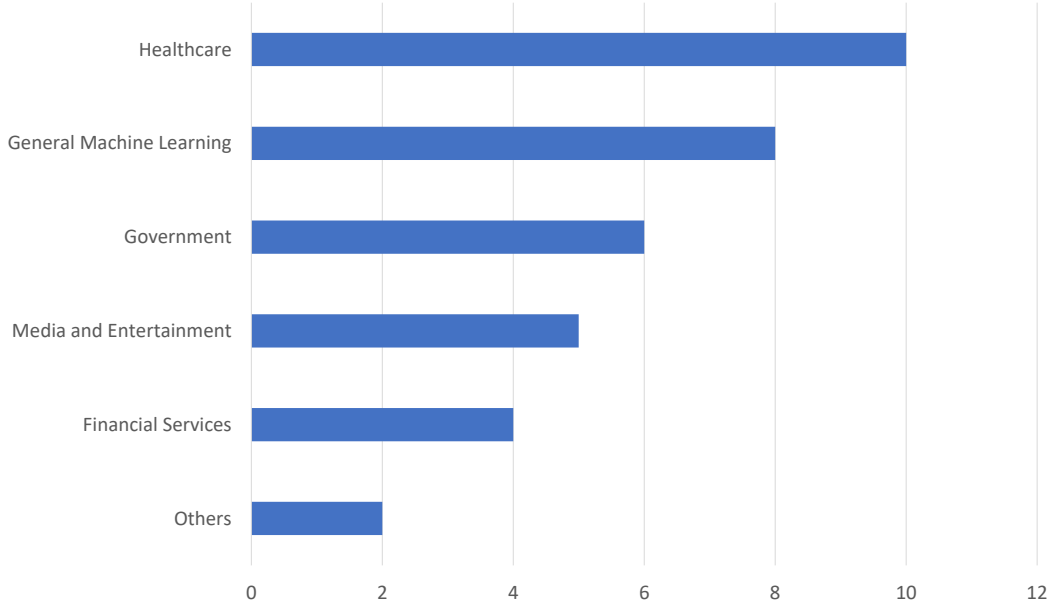


FIGURE 4.7: Participants' Application Domains.

choices around AI/ML explainability", and subsequently the negative question "I don't know how to make decisions regarding explainability in AI/ML". By using this redundancy check, we can detect and discard responses that were not valid. Furthermore, we informed the participants to complete the post study questionnaire as soon as possible after the study, and all participants completed it within 1 day of the study.

The 3 main hypotheses for this user study are as follows:

- The EID methodology helps participants determine the explainability criteria that are the most relevant to their AI applications.
- The EID methodology helps participants surface explainability concerns in their AI applications.

- The EID methodology helps participants envision the perspectives from different stakeholders.

We created the pre and post study questionnaires for the participants to self assess their understanding of the explainability concepts and how to apply it to their AI products and services. Each hypothesis is designed to assess the individual ability of the participants to choose an applicable explainability solution, brainstorm and surface explainability concerns, and stimulate the thinking process and perspectives of stakeholders respectively. They are asked to rate their understanding of explainability problems from a likert scale of 1 to 5, 1 being strongly disagree (SA) and 5 being strongly agree (SA). Based on the results of the questionnaire, we conduct statistical data analysis to evaluate the 3 hypotheses.

4.6 Results and Analysis

In this section, we analyse the results from the empirical studies by presenting the findings for each hypothesis.

4.6.1 Hypothesis 1

Hypothesis 1: The EID methodology helps participants determine the explainability criteria that are the most relevant to their AI applications.

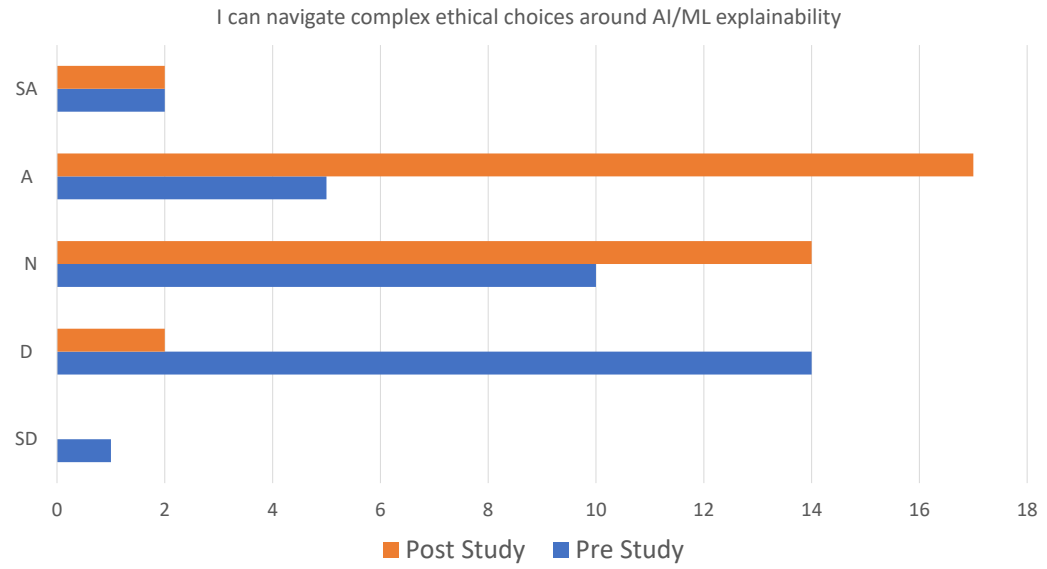


FIGURE 4.8: Participants' self-reported capability of making design decisions related to explainability before and after using EID.

Figure 4.8 illustrates the results from the participants' responses on questions related to H1. We have used "Strongly Agree" as SA, "Agree" as A, "Neutral" as N, "Disagree" as D and "Strongly Disagree" as SD.

The responses can be observed to be largely negative and following a distribution roughly centred on "Disagree", signaling a general lack of confidence in the self assessment by the participants. This result was expected as most of the participants have not actively worked on explainability issues in the AI domain, therefore were unlikely be competent or experienced in this area. This also signifies that the distribution of the participants' capabilities to make design decisions related to the explainability aspect of AI was typical of a population of AI solution designers. Subsequently, we observed that after using EID in the empirical study sessions, there is a significant increase in the number of participants who responded with

"Agree", while the number of "Disagree" and "Strongly Disagree" responses decreased significantly. The results indicate that the participants found the methodology to be useful in helping them think about the explainability principle in their application domains.

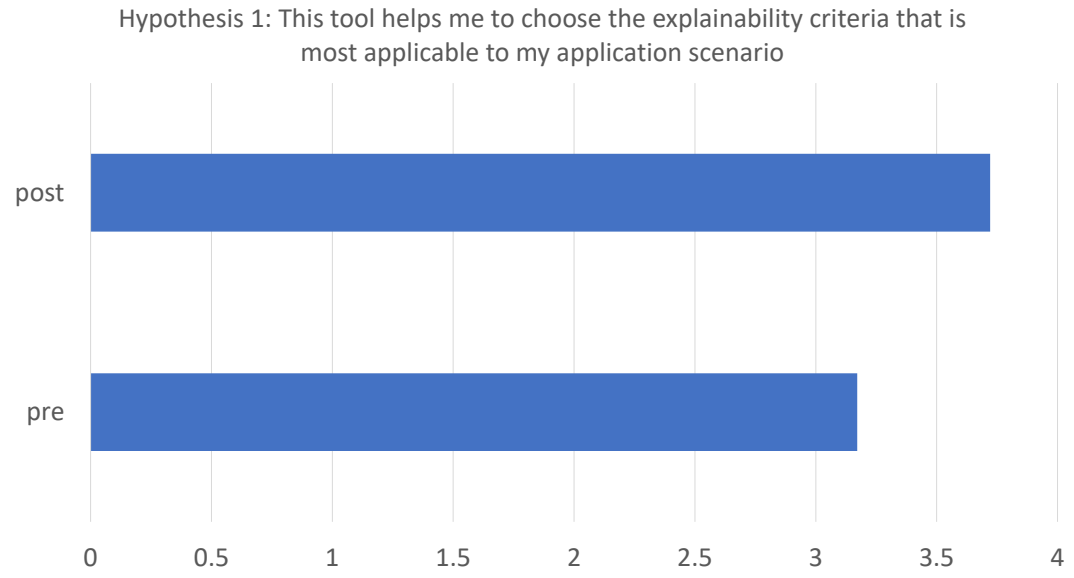


FIGURE 4.9: Participants' average scoring for the pre- and post-studies for hypothesis 1

As shown in 4.9, the participants average response scores in the post study is significantly higher than in the pre study. After conducting a students' t test of questionnaire results from H1, we conclude that the null hypothesis can be rejected at 95 percent, with cronbach alpha at 0.7256.

4.6.2 Hypothesis 2

Hypothesis 2: The EID methodology helps participants surface explainability concerns in their AI applications. This hypothesis pertains to the participants' self

assessment of their competency in discovering issues or concerns about the explainability aspect.

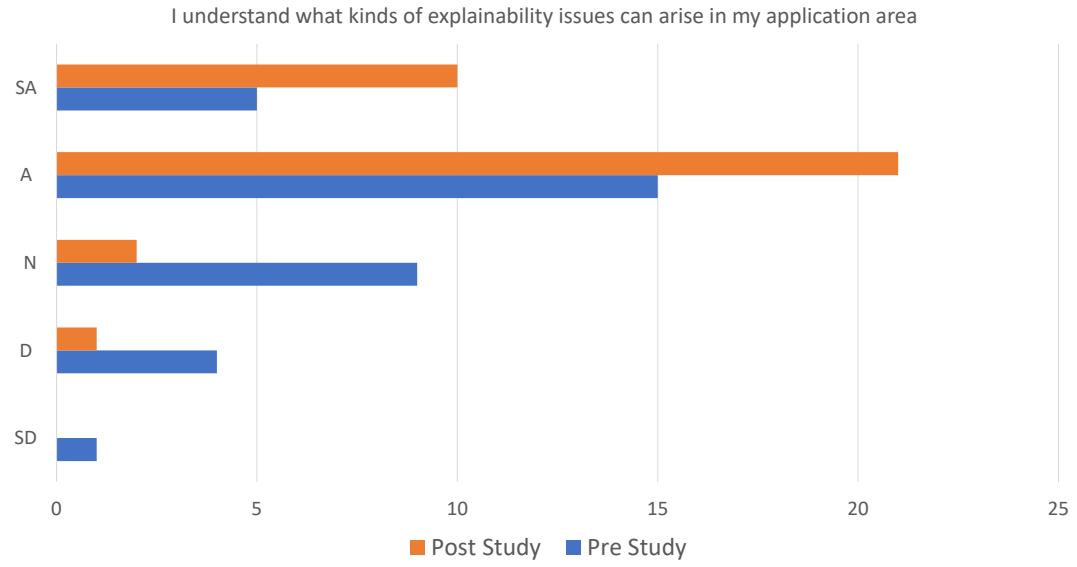


FIGURE 4.10: Participants' self-reported capability of surfacing explainability concerns before and after using EID.

In 4.10, we illustrate the results from the participants' responses about surfacing explainability concerns, focusing on hypothesis 2. This question indicates the self assessed competency of participants to identify in advance what kind of explainability issues can happen in their application domains. Similar to before, the responses are roughly centred around "Agree". In contrast after using the EID methodology, there is a significant increase in the number of responses for "Strongly Agree" and "Agree", as well as a corresponding decrease in the number of responses for "Strongly Disagree" and "Disagree". The results indicate that EID is effective in identifying potential issues or concerns about explainability in advance.

For hypothesis 2, we find that the questionnaire response averages increased by more than 0.5 in the post study compare to the pre study. After conducting a

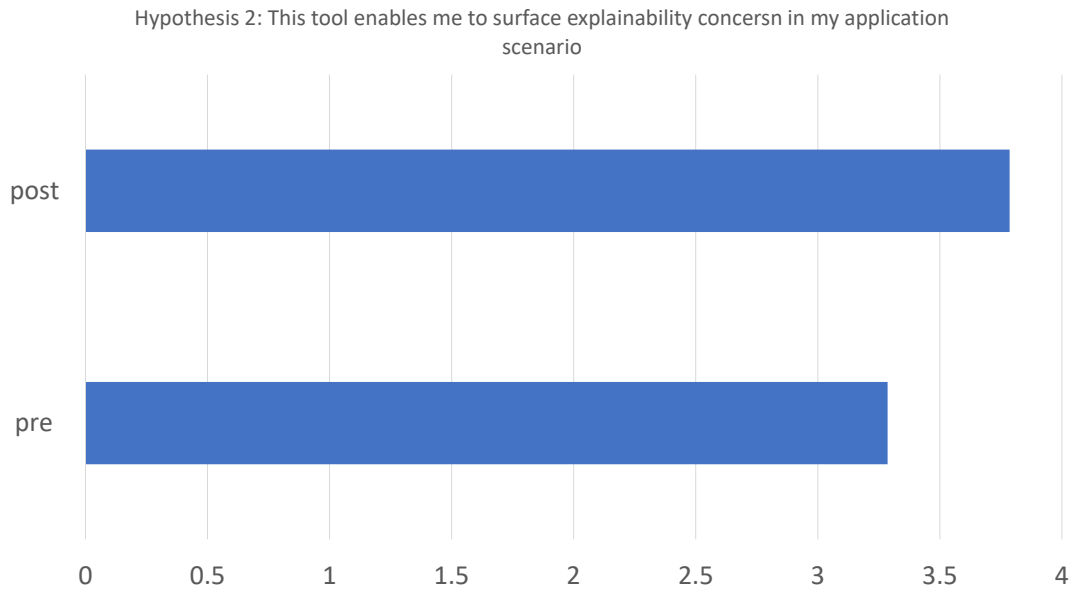


FIGURE 4.11: Participants' average scoring for the pre- and post-studies for hypothesis 2.

students' t-test, we are able to reject the null hypothesis at 95 percent confidence level, with the cronbach alpha at 0.7456.

4.6.3 Hypothesis 3

Hypothesis 3: The EID methodology helps participants envision the perspectives from different stakeholders. We conceptualised this hypothesis to assess the competence of participants to stimulate thinking in the perspective of other relevant group of stakeholders.

In figure 4.12, we highlight the results from the participants responses on thinking from the perspective of stakeholders. The question focuses on hypothesis 3 and challenges the participants to visualising the perspective of 2 types of stakeholders, namely the direct and indirect stakeholders. For example, direct stakeholders

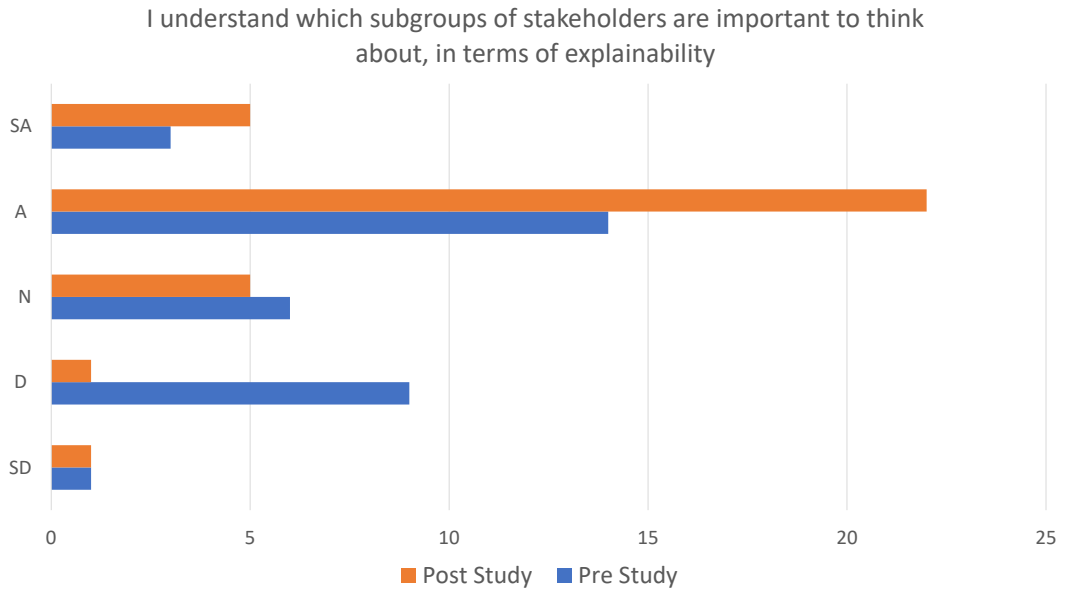


FIGURE 4.12: Participants' self-reported capability of thinking from stakeholders' perspective before and after using FID.

can be the end user of the AI system, the system designer and engineers, while indirect stakeholders include the family members of the end users. The proportion of "Strongly Disagree" and "Disagree" greatly decreased and many participants changed their answers to "Agree" and "Strongly Agree". Hence the methodology was able to greatly improve the self assessed ability of participants to stimulate stakeholder thinking perspective, which is a valuable skill set in AI development teams.

According to figure 4.13, we find that the questionnaire response averages increased slightly in the post study compared to the pre study. However, when we conduct a students' t test analysis of the questionnaire results from hypothesis 3, the null hypothesis cannot be rejected.

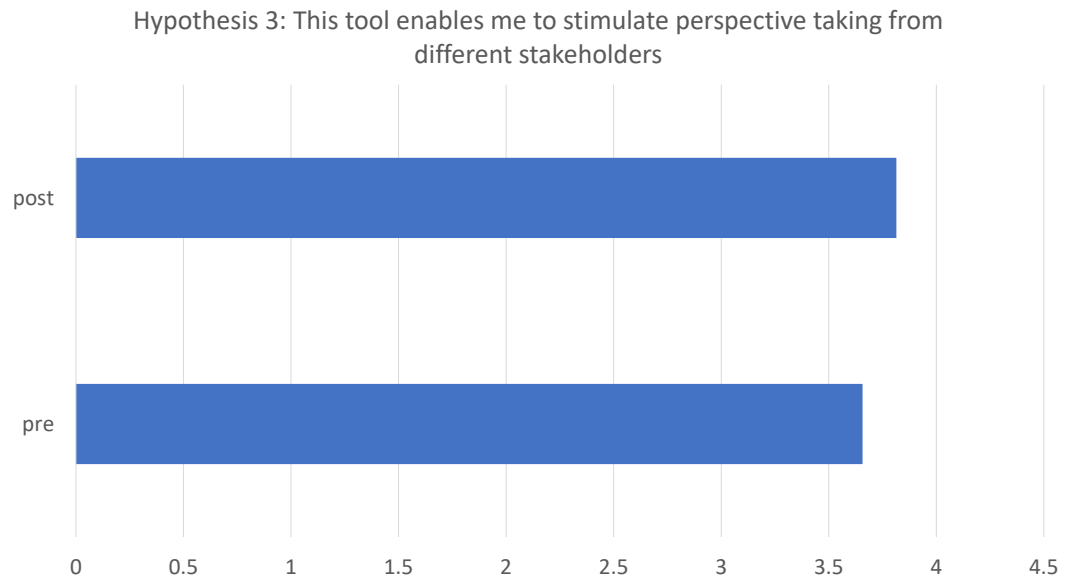


FIGURE 4.13: Participants' average scoring for the pre- and post-studies for hypothesis 3.

4.7 Discussions and Limitations

We found EID to be an effection framework in promoting conversations and eliciting critical thinking about explainability in AI. Over the course of a series of user studies, we found that EID not only introduced the participants about explainability, the framework also provided them with deeper insights on the complexities of explainability in AI. In the post study, a total of 19 participants expressed their confidence to make complex decisions around AI explainability, compared to just 7 participants before the study. However, the EID methodology should be used early in the design and conception stage of the AI life cycle as the metrics need to be integrated early into the product or service, and these design decisions will also impact the rest of the AI pipeline.

While the EID framework is a good starting point for software teams with no experience in explainability, the field of AI explainability can become quite complex and fragmented especially when going deeper into the technical details. We made the assumption to simplify the explainability principles by narrowing them down into 6 main categories, in order to facilitate our user studies without going too in depth into the complexities of AI explainability. However, with more time and resources, it is possible to provide a more balanced view of many aspects of explainability in the ethical AI field. For this reason, most AI system designers are unfamiliar with explainability and/or see it as an unnecessary trade off for performance or efficiency. Most AI developers are not willing to forgo the reduction in their algorithm's performance, unless there is a specific requirement to include explainability. This response has been duly noted, however we believe the trajectory of this field is such that eventually teams can deliver a system that minimises the compromise on both performance and explainability.

Over the course of multiple user studies, we made tweaks to the EID framework or user study procedures to improve it. For example, we realised that in some application domains, some explainability principles were irrelevant to the participants, and we decided to give the participants the option to discard irrelevant explainability principles.

While our use study consists of 35 participants, it might be better to recruit more people to conduct a larger scale online only study to evaluate the EID framework. Additionally, it has been studied that slef reported preferences often do not align well with participants' actual behaviours [134]. It is still an open question if the

findings from existing and past work contributes to significant improvements in our methodology. As explainability in AI is a large field, it may also make sense to divide the investigation into the different sub fields of AI.

4.8 Conclusions and Future Work

An overview of the state of the field in ethical AI design was discussed in this chapter, along with the gaps in the research literature and proposed solutions. Using existing toolkits such as the theory of VSD and various recent studies, we proposed and tested a methodological framework, Explainability in Design. This framework assists software design teams to facilitate complicated ethical choices around explainability. By designing EID to be efficient on time and effort, as well as having a low entry knowledge barrier, it effectively allows team members to improve the decision making process for explainability in their AI products and services.

In proposed future work, we aim to conduct larger scale online only user studies to evaluate the effectiveness our EID, and assess if the project goals are attainable. The format of our users studies is currently conducted in teams of more than 1 member that are working or have worked on AI products previously. As the new normal of working from home has been in effect for the past years, the online form of EID will be taking priority in our propose future research goals. Then the team members can collaborate and use the EID methodology online accordingly. We aim to include project management functions for all team members to manage

work allocation and timeline management. In addition to the above, we plan to identify specific application domains such as autonomous vehicles [149] and medical healthcare diagnosis [135] to apply our methodology and investigate deeper on how the different complexities interact.

Chapter 5

Conclusion, Discussion and Future Work

5.1 Conclusions

As technological advances improve the lives of many and create new opportunities in the fourth industrial revolution, our society is increasingly reliant on the AI systems that drive the revolution. Unfortunately, there usually are groups of people that get left behind or neglected when this happens. In the previous chapters, we explored the background, motivation and objectives for this research. We discussed the need for centring human values in the development of AI systems and why methodological frameworks are needed to assist AI practitioners and researchers to navigate complex ethical choices. Current tools are insufficient in providing user friendly guidelines to account for people without the prerequisite knowledge or

experience. If AI is to truly be built around the needs and requirements of human society, we need to ensure that there is a seat at the table for all people, especially the minority groups that tend to be overlooked. One of the overarching themes of this thesis is the effort to reduce the barriers of entry to lay people, in fact the indirect stakeholders who may not appear in the life cycle of AI products and services. However in contrast, they are the main groups of people who urgently require more thought to their situation. Additionally, we highlight the inherent nature of ethics in AI, its interconnected qualities and discuss our plan for the extension of our existing fairness and explainability in AI methodology.

As mentioned in the previous chapters, works in the ethical AI field are still in the nascent stage and tend to be fragmented as researchers have yet to come together to decide on a common focus. Besides researchers in the scientific and academic communities, we expound on the importance of engaging with parties outside of these communities, such as AI manufacturers, social scientists, government statutory boards, and many others. In view of the situation, we hope that our work will be considered a milestone to kick-start the conversation and bring as many groups to the table as possible. As difficult as we may endeavour to bring our vision to reality, a real concerted effort is required to initiate the next step, which is simply to align the various groups of people to a common vision.

In this thesis, we highlighted the gaps in ethical AI design methodologies and expressed the need for a tool that deciphers an ethical value deeper. We engaged the existing fairness literature in machine learning and the theory of VSD to create a tool that aims to help users to surface fairness concerns, navigate complex ethical

choices around fairness, and overcome blind spots and team biases. The instruction for using the tool has also been described in detail. With this methodological tool, product teams that view fairness as an important value in their AI system can now be focused on it. We understand that product teams often face barriers to improving fairness in their products even when they are motivated to do so. Since our tool takes a short amount of time and is easy to follow, we hope that the bar to include fairness into the design discussion will be lowered and more product teams can make better and more informed decisions for fairness in their application scenarios. We also wish to inspire researchers to construct more methodological tools that enable the integration of human values into the design process.

We have identified several application domains of AI that is of particular interest to the field of ethical AI. Firstly, fully self-driving vehicles may actually become a reality in the coming decades. When these vehicles are given the responsibility of split-second decision-making, it warrants a comprehensive look into the ethics of autonomous vehicles. Building on the topic of autonomous vehicles, we demonstrate the importance of exploring and discussing ethical issues in Artificial Intelligence. Other fields involving ethics include healthcare and medical diagnosis, finance and banking algorithms, facial recognition and surveillance systems, robotics and many others. We have explored various of these application domains in our series of user studies.

5.2 Discussion

Based on recent discoveries in the field of AI governance systems, a variety of techniques in ethical decision making frameworks, rule based and example based methodologies are required to tackle problems in ethical AI. While there are some datasets available in the research community, more data regarding the various ethical issues are needed, especially in different cultural and social contexts. It is clear that AI needs to be designed centred around human beings, and how AI systems react to input and tweaks by human beings should be studied deeply. Besides ethical AI researchers, AI engineers and data teams are the core group of people that needs to collaborate on the pursuit of building ethics into AI. These groups of people can leverage their real world experience to provide wider and deeper insights into how to improve algorithm techniques. Since such AI technologies as autonomous vehicles, autonomous weapons, and cryptocurrencies are becoming a reality and affecting societies, a global and unified AI regulatory framework needs to be established as soon as possible to address the ethical issues by drawing on interdisciplinary expertise [150].

Even though we found our methodological framework, FID and EID, to be effective tools to promote engaging conversation and critical thinking, it is evident that the tool is effective only at the early stage of design and conception. This is due to the fact that metrics need to be integrated early into the AI product or service before other considerations. This however can be rather limiting as sometimes, there is a need to trade off performance or efficiency at the expense of building

fairness or explainability metrics in the system. Currently, when faced with this decision, usually most AI engineers will prioritise performance or efficiency. Over time, we expect the direction of the field to be headed towards a system where both performance and ethical metrics can be implemented without compromise.

With AI becoming increasingly ubiquitous in our daily life, we should start the conversation on revising our current social contracts. Research in this field will help us establish regulations about who is accountable when mistakes happen and how to monitor and enforce these regulations. The aftermath of a mistake made by an algorithm can have unforeseen impact and these should be also explored, especially when human individuals or groups are affected. This research direction is inherently dynamic and interdisciplinary in nature as it must be updated with changing cultural, social, legal, philosophical and technological realities.

5.3 Future Work

Our future work will be focused on conducting user studies to evaluate if the tool is effective in achieving the aims. We will be testing the tool with actual product teams that are working on different areas of AI systems. At the same time, due to Covid 19 and the increased uptake in working-from-home arrangements, we realize that most product teams will be working together online instead. This makes physical cards impractical since most employees will not be in the office together. As a result of this new norm, we are also extending efforts to digitalize our physical tool into a web application. Team members can collaborate and go

through the methodology online. There will also be functions for facilitators to manage their teams and control the process. Having an online tool will help to reduce the need to arrange meetings and provide the flexibility to work on fairness at their convenience.

As AVs are equipped with sensors such as cameras, LIDAR and other monitoring devices, they have generated huge amounts of data. As with any system that uses a large dataset, there inadvertently will be privacy concerns on the way that the way is being handled and stored. If AVs are to become widely adopted, we need to ensure that trust is built effectively between man and machine. We plan to investigate the trust matters in the context of a robotaxi, by creating a virtual reality experience where participants will ride in a mock up wizard of oz vehicle. We will measure the degree of trust and the impact of the privacy, explainability measures the vehicles will have on the passengers. Besides that, we can also test the effectiveness of our fairness in AI methodology when applied to the context of robotaxis, personal transport vehicles that are fully autonomous and self-driving when the passengers are within. These 2 studies will be conducted concurrently, and we envision that multiple valuable findings will be achieved at the end of the proposed work.

We have plans to investigate AI qualities beyond the fairness and explainability aspect, extending to privacy concerns in highly advanced systems such as AVs, as well as accountability considerations in systems that require clarity. Due to the interrelated qualities of ethical AI, and the nascent nature of the field, it is challenging to work on each quality in isolation. In order to achieve a unified methodology,

there must be an investigation of the interconnected nature of ethical qualities. For example, fairness measures need to be complemented by explainability, if we are unable to explain in human terms how an algorithm works or how fairness in the algorithm works, then fairness simply becomes unfeasible.

The number of participants in our user studies are currently 24 and 35 for FID and EID respectively. While these numbers are satisfactory at this stage of research work, we envision a larger-scale study with the number of participants in the hundreds. This can be made possible with several crowdsourcing tools such as Amazon Mechanical Turk. With the larger scale empirical evaluation, we can then understand how effective the methodological tools are when used in these diverse contexts. Furthermore, it has been well explored that self-reported preferences usually do not align well with peoples' actual behaviours [134], hence it would fit in our research goals to investigate this discrepancy.

Ethical AI is incomplete without an emphasis on data privacy and security, we consider it a vital part of making AI safe and beneficial for all. There is increasing public awareness about large companies compromising on data security and user privacy. There has been much backlash in response to these scandals, and many countries are improving their laws to address data privacy and security [151]. For example, the European Union instituted the General Data Protection Regulation (GDPR) in order to protect user's personal privacy and data security [152]. As AI continues to be refined and developed, privacy is becoming a critical attribute that AI researchers cannot ignore. In addition to improving data privacy, the problem of isolated and fragmented data is a worthwhile research problem to address. In

most industries the data exists in the form of isolated islands due to the reality of competition, privacy, security and complicated administrative procedures. This has led to significant developments in the field of federated learning [153]. We aim to extend our methodology into the aspect of privacy, as well as to investigate the interactions and tradeoffs with other aspects of ethical AI.

Appendix A

Fairness Principles for Chapter 3

A.1 Principles for Fairness

This section lists the fairness principles that we have used for the FID methodology

Appendix A

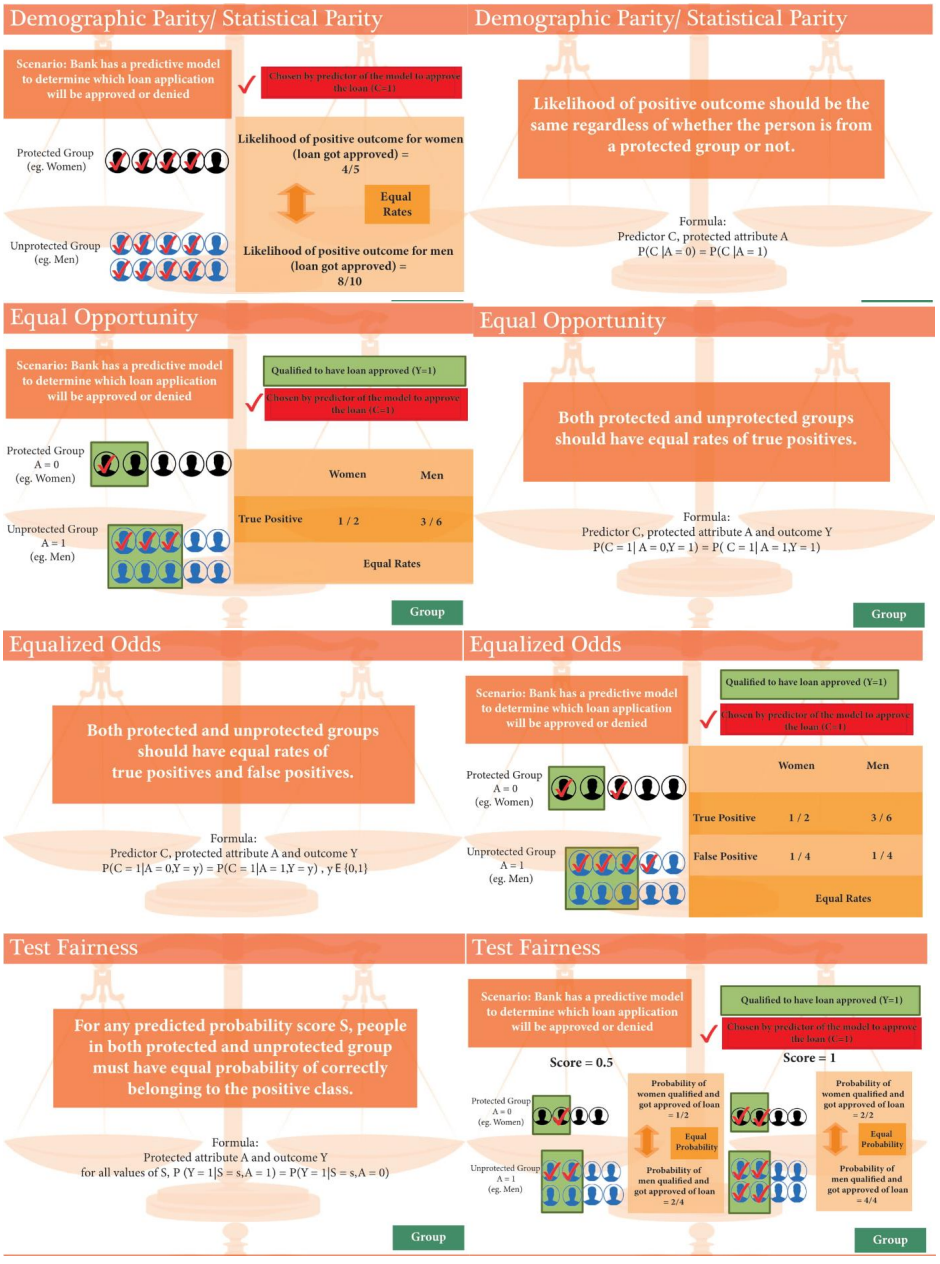


FIGURE A.1: The FID Principles

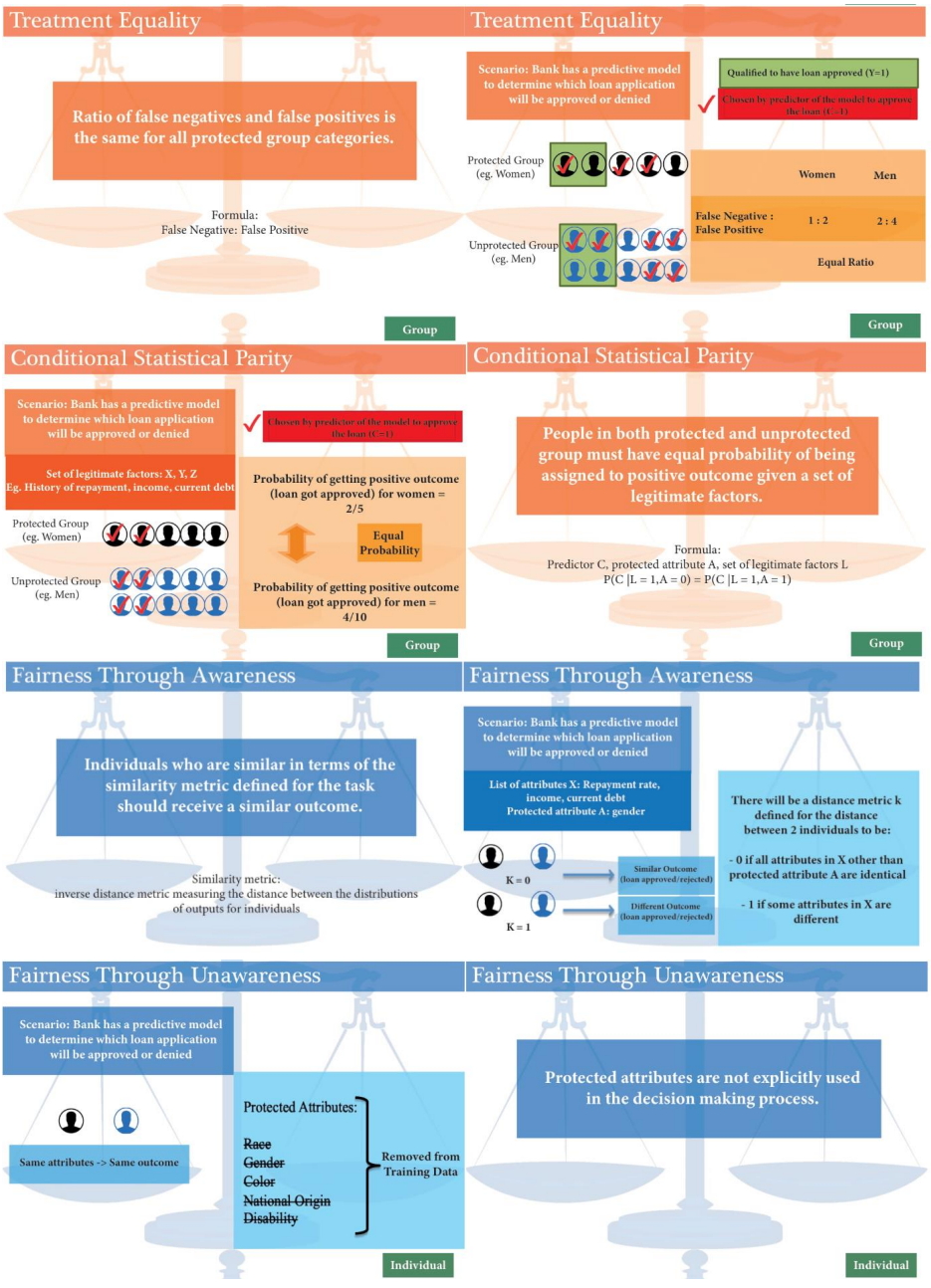


FIGURE A.2: The FID Principles

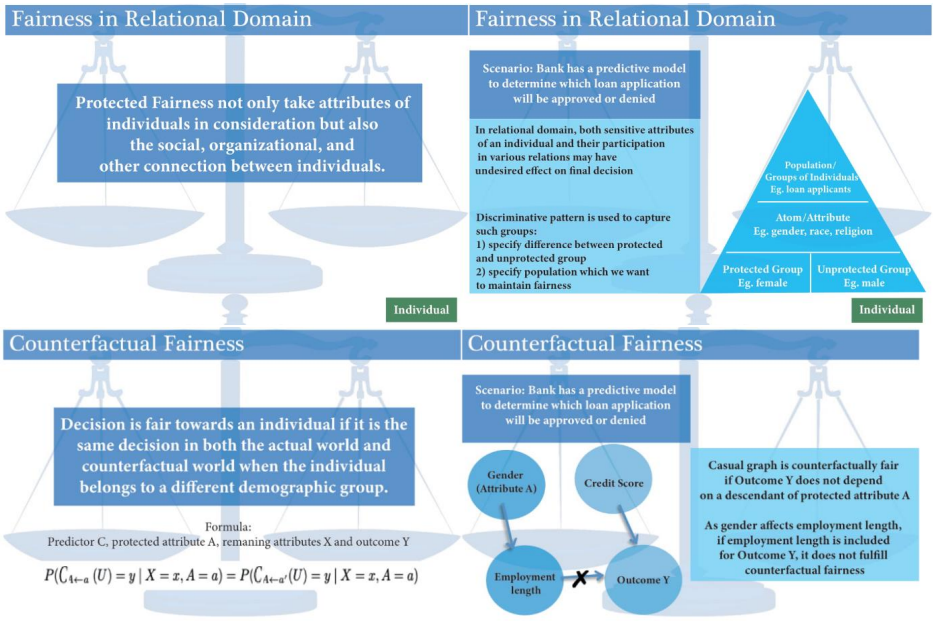


FIGURE A.3: The FID Principles

Appendix B

Explainability Principles for Chapter 4

B.1 Principles for Fairness

This section lists the fairness principles that we have used for the EID methodology

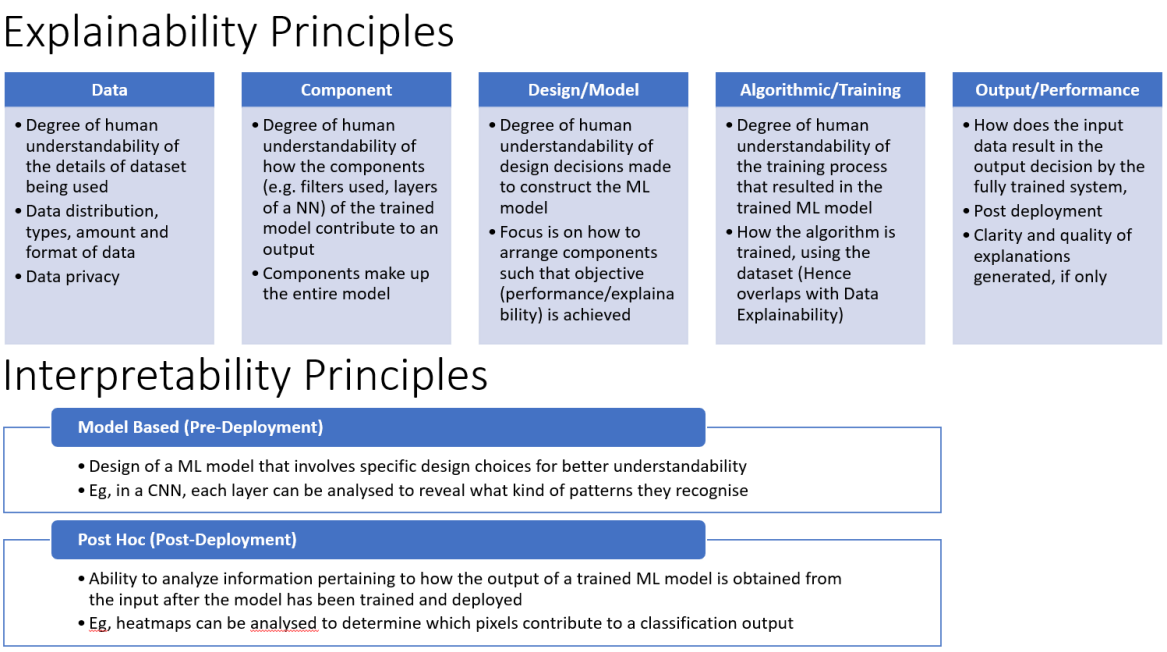


FIGURE B.1: The EID Principles

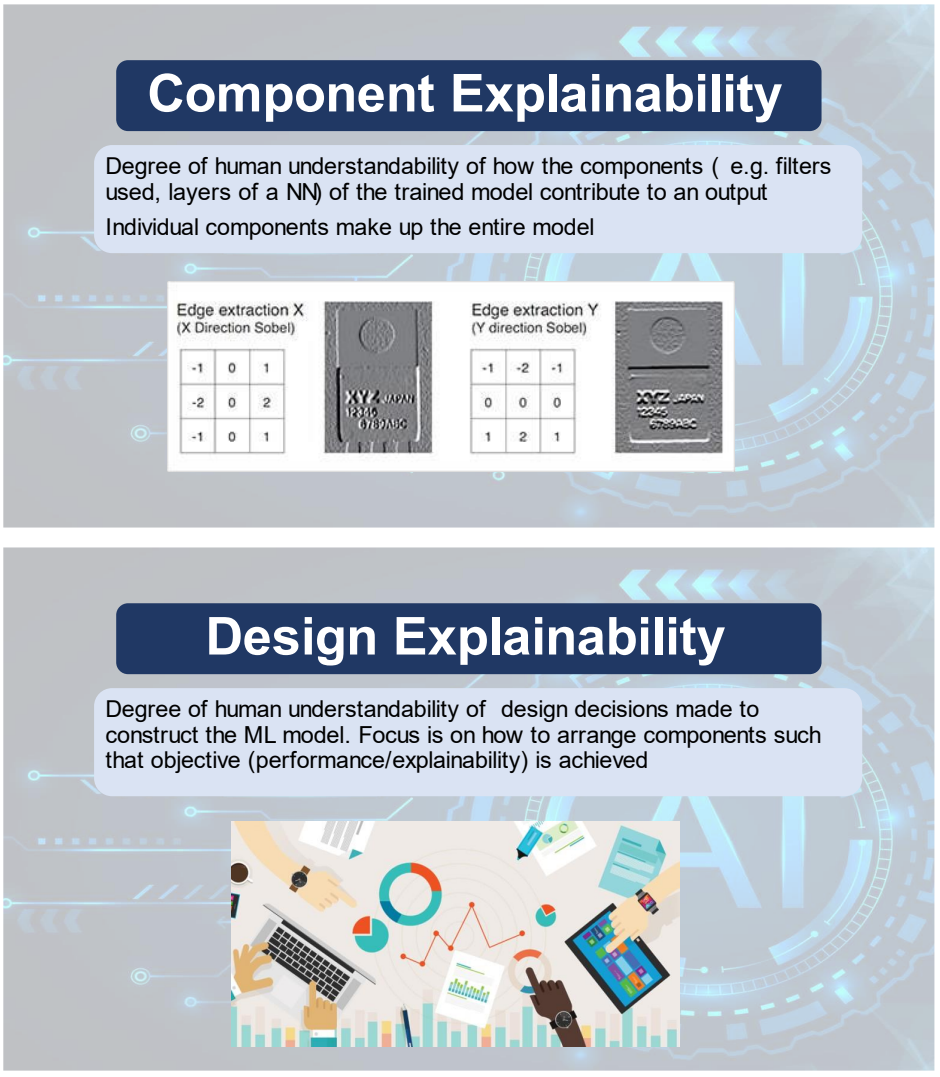
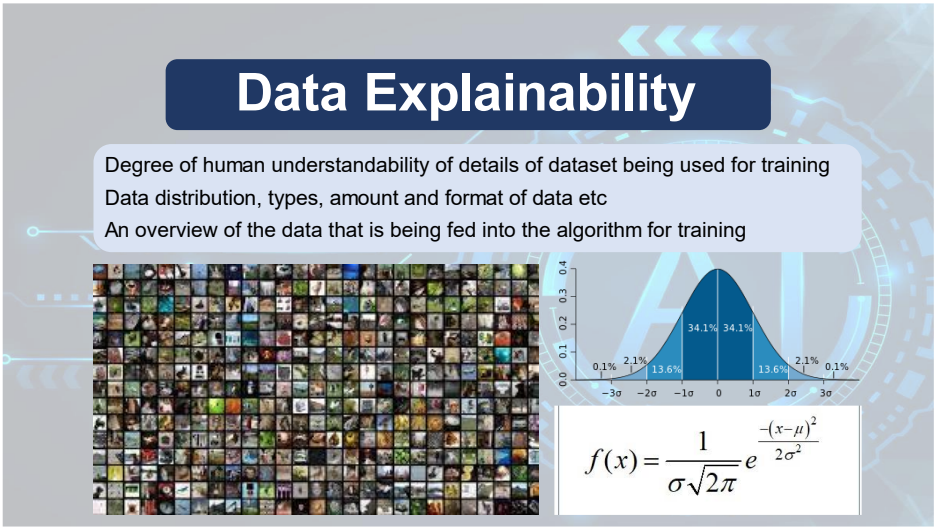


FIGURE B.2: The EID Principles



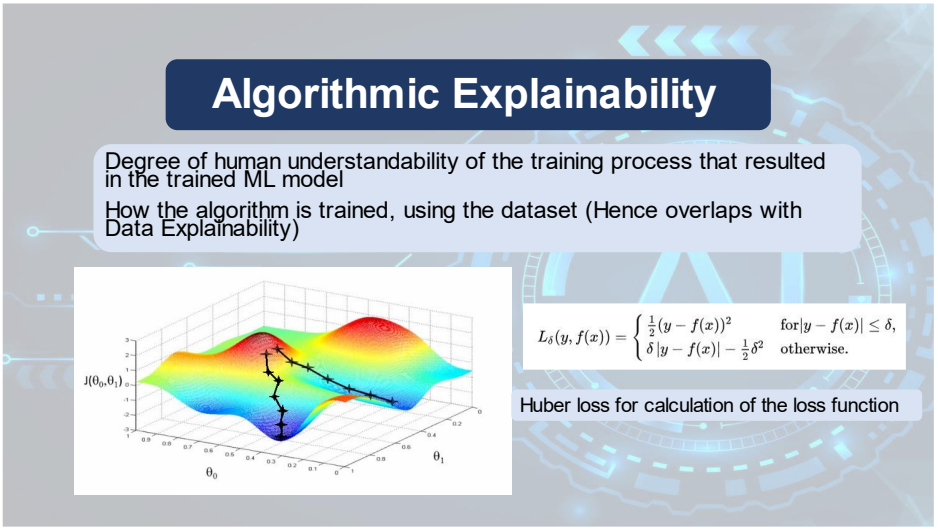
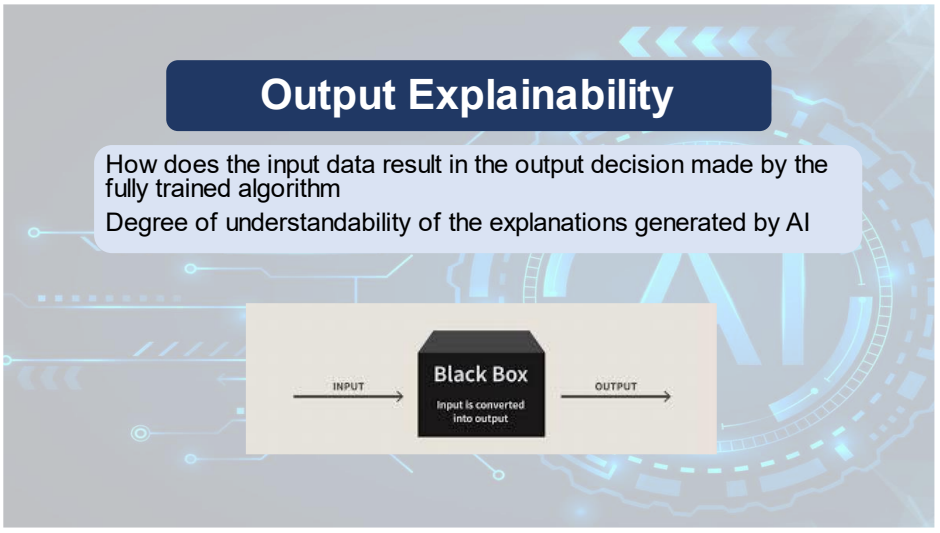


FIGURE B.3: The EID Principles



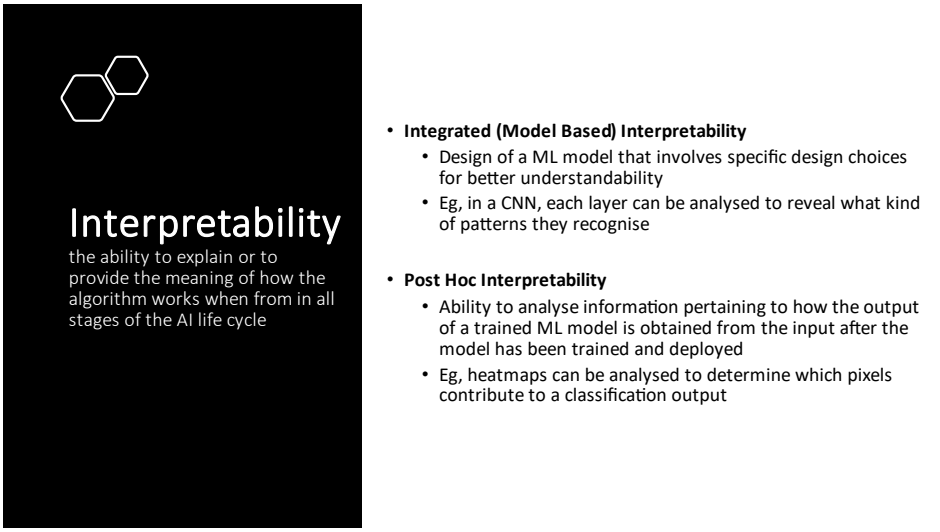


FIGURE B.4: The EID Principles

List of Author's Awards, Patents, and Publications[1]

Journal Articles

1. **Jiehuang Zhang**, Ying Shu and Han Yu, "Fairness in Design: A Tool for Guidance in Artificial Intelligence Design" *International Journal of Crowd Science* (Accepted and Published).

2. **Jiehuang Zhang** and Han Yu, "A Methodological Framework for Facilitating Explainable AI Design" *International Journal of Crowd Science* (Accepted).

Conference Proceedings

1. **Jiehuang Zhang** and Han Yu, "A Methodological Framework for Facilitating Explainable AI Design", in *Proceedings of the 14th International Conference on Social Computing and Social Media (SCSM'22)*, pp. 437–446, 2022.

[1]The superscript * indicates joint first authors

2. **Jiehuang Zhang**, Ying Shu and Han Yu, "Human-Machine Interaction for Autonomous Vehicles: A Review", in *Proceedings of the 13th International Conference on Social Computing and Social Media (SCSM'21)*, pp. 190–201, 2021.

3. Ying Shu, **Jiehuang Zhang** and Han Yu, "Fairness in Design: A Tool for Guidance in Artificial Intelligence Design", in *Proceedings of the 13th International Conference on Social Computing and Social Media (SCSM'21)*, pp. 500–510, 2021.

Bibliography

[1] Yundong Cai, Zhiqi Shen, Siyuan Liu, Han Yu, Xiaogang Han, Jun Ji, Martin J. McKeown, Cyril Leung, and Chunyan Miao. An agent-based game for the predictive diagnosis of parkinson's disease. In *AAMAS*, pages 1663–1664, 2014. 1

[2] Yuliang Shi, Chenfei Sun, Qingzhong Li, Lizhen Cui, Han Yu, and Chunyan Miao. A fraud resilient medical insurance claim system. In *AAAI*, pages 4393–4394, 2016.

[3] Zhengxiang Pan, Han Yu, Chunyan Miao, and Cyril Leung. Crowdsensing air quality with camera-enabled mobile devices. In *IAAI*, pages 4728–4733, 2017.

[4] Yongqing Zheng, Han Yu, Lizhen Cui, Chunyan Miao, Cyril Leung, and Qiang Yang. SmartHS: An AI platform for improving government service provision. In *IAAI*, pages 7704–7711, 2018. 1

[5] Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, Oxford, UK, 2008. 1

[6] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R Lesser, and Qiang Yang. Building ethics into artificial intelligence. *arXiv preprint arXiv:1812.02953*, 2018. 2, 3, 9, 42

[7] Joanna J. Bryson and Philip P. Kime. Just an artifact: Why machines are perceived as moral agents. In *IJCAI*, pages 1641–1646, 2011. 2

[8] Ben Goertzel and Cassio Pennachin. *Artificial General Intelligence*. Springer, Berlin, Heidelberg, Berlin, Heidelberg, 2007. 2

[9] Stanislas Dehaene, Hakwan Lau, and Sid Kouider. What is consciousness, and could machines have it? *Science*, 358(6362):486–492, 2017. 2

[10] Jack Parker and David Danks. How technological advances can reveal rights. 2019. 2

[11] Aaron Smith and Janna Anderson. Ai, robotics, and the future of jobs. *Pew Research Center*, 6, 2014. 4

[12] Blake Alcott. Jevons' paradox. *Ecological economics*, 54(1):9–21, 2005. 4

[13] Institute for ethical AI. Jevon's paradox. https://ethical.institute/principles.html#commitment-5, 2019. 4

[14] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018. 5, 8

[15] Yoshua Bengio. Ai pioneer: 'the dangers of abuse are very real'. https://www.nature.com/articles/d41586-019-00505-2, 2019. 5

[16] Declaration Montreal. Ai pioneer: 'the dangers of abuse are very real'. MontrealDeclarationforaResponsibleDevelopmentofArtificialIntelligence, 2019. 5

[17] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477, 2019. 7, 13, 27

[18] Olivia J Erdélyi and Judy Goldsmith. Regulating artificial intelligence: Proposal for a global solution. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 95–101, 2018. 10

[19] Yeongkwun Kim and Injoo Kim. Security issues in vehicular networks. In *The International Conference on Information Networking 2013 (ICOIN)*, pages 468–472. IEEE, 2013. 10

[20] Kate Crawford and Ryan Calo. There is a blind spot in ai research. *Nature News*, 538(7625):311, 2016. 10

[21] Jack Parker and David Danks. How technological advances can reveal rights. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 201–201, 2019. 10

[22] Michael Spencer. Artificial intelligence regulation may be impossible, 2019. URL https://www.forbes.com/sites/cognitiveworld/2019/03/02/artificial-intelligence-regulation-will-be-impossible. 11

[23] OECD. Oecd ai princples. https://www.oecd.org/going-digital/ai/principles/, 2019. 11

[24] Matthew U Scherer. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. JL & Tech.*, 29:353, 2015. 12, 14

[25] Jigar Doshi, Saikat Basu, and Guan Pang. From satellite imagery to disaster insights. *CoRR*, abs/1812.07033, 2018. URL http://arxiv.org/abs/1812.07033. 13

[26] Jacqueline M Kory Westlund, Hae Won Park, Randi Williams, and Cynthia Breazeal. Measuring young children's long-term relationships with social robots. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*, pages 207–218. ACM, 2018. 13

[27] Gunter J Hitsch, Ali Hortaçsu, and Dan Ariely. Matching and sorting in online dating. *American Economic Review*, 100(1):130–63, 2010. 13

[28] Peter Stone, R Brooks, E Brynjolfsson, R Calo, O Etzioni, G Hager, and A Teller. One hundred year study on artificial intelligence. *Artificial Intelligence and Life in*, 2030, 2016. 13

[29] Kenneth A Bamberger and Deirdre K Mulligan. Privacy on the ground: Driving corporate behavior in the united states and europe (chapter 1). *Privacy on the Ground: Driving Corporate Behavior in the US and Europe (MIT 2015)*, 2015. 13

[30] Oren Etzioni. Point: Should ai technology be regulated?: yes, and here's how. *Communications of the ACM*, 61(12):30–32, 2018. 14

[31] Peter Eckersley. Theories of Parenting and Their Application to Artificial Intelligence Lack of Perspective Contributes to AI risk. 2018. 14

[32] The Belmont report. Technical report, 1978. 15

[33] Rose Luckin. Towards artificial intelligence-based assessment systems. *Nat. Hum. Behav.*, 1(0028):doi:10.1038/s41562–016–0028, 2017. 15

[34] Han Yu, Chunyan Miao, Cyril Leung, and Timothy John White. Towards AI-powered personalization in MOOC learning. *npj Sci. Learn.*, 2(15): doi:10.1038/s41539–017–0016–3, 2017. 15

[35] IEEE. Ethically aligned design. Technical report, 2018. 15

[36] Han Yu, Chunyan Miao, Cyril Leung, Yiqiang Chen, Simon Fauvel, Victor R. Lesser, and Qiang Yang. Mitigating herding in hierarchical crowdsourcing networks. *Sci. Rep.*, 6(4):doi:10.1038/s41598–016–0011–6, 2016. 15

[37] Han Yu, Chunyan Miao, Yiqiang Chen, Simon Fauvel, Xiaoming Li, and Victor R. Lesser. Algorithmic management for improving collective productivity in crowdsourcing. *Sci. Rep.*, 7(12541):doi:10.1038/s41598–017–12757–x, 2017.

[38] Han Yu, Zhiqi Shen, Simon Fauvel, and Lizhen Cui. Efficient scheduling in crowdsourcing based on workers' mood. In *ICA*, pages 121–126, 2017. 15

[39] Shiwali Mohan, Frances Yan, Victoria Bellotti, Ahmed Elbery, Hesham Rakha, and Matthew Klenk. On influencing individual behavior for reducing transportation energy expenditure in a large population. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, New Orleans, LA, USA*, pages 2–3, 2018. 15

[40] Yilin Kang, Ah-Hwee Tan, and Chunyan Miao. An adaptive computational model for personalized persuasion. In *IJCAI*, pages 61–67, 2015. 16

[41] Ariel Rosenfeld and Sarit Kraus. Strategical argumentative agent for human persuasion. In *ECAI*, pages 320–328, 2016. 16

[42] Oliviero Stock, Marco Guerini, and Fabio Pianesi. Ethical dilemmas for adaptive persuasion systems. In *AAAI*, pages 4157–4161, 2016. 16

[43] Cristina Battaglino and Rossana Damiano. Coping with moral emotions. In *AAMAS*, pages 1669–1670, 2015. 16

[44] Stacy Marsella and Jonathan Gratch. Modeling coping behavior in virtual humans: Don't worry, be happy. In *AAMAS*, pages 313–320, 2003. 16

[45] Siyuan Liu, Han Yu, Chunyan Miao, and Alex C. Kot. A fuzzy logic based reputation model against unfair ratings. In *AAMAS*, pages 821–828, 2013. 17

[46] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Ethical judgment of agents' behaviors in multi-agent systems. In *AAMAS*, pages 1106–1114, 2016. 22

[47] Angela Daly, Thilo Hagendorff, Hui Li, Monique Mann, Vidushi Marda, Ben Wagner, Wei Wang, and Saskia Witteborn. Artificial intelligence, governance and ethics: Global perspectives. *The Chinese University of Hong Kong Faculty of Law Research Paper*, (2019-15), 2019. 22

[48] Keith Kirkpatrick. The moral challenges of driverless cars. *Commun. ACM*, 58(8):19–20, 2015. 24

[49] Nicholas R Jennings, Luc Moreau, David Nicholson, Sarvapali D Ramchurn, Stephen Roberts, Tom Rodden, and Alex Rogers. Human-agent collectives. *Communications of the ACM*, 57(12):80–88, 2014. 25

[50] James H Moor. Taking the intentional stance toward robot ethics. *APA Newsletter*, 6:14–17, 2007. 26

[51] Tom M Mitchell. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research ..., 1980. 27

[52] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58. ACM, 2019. 27, 31, 57

[53] Moritz Hardt, Eric Price, ecprice, and Nati Srebro. Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in*

Neural Information Processing Systems 29, pages 3315–3323. Curran Associates, Inc., 2016. URL `http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf`. 27, 29

[54] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017. 27

[55] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018. 27

[56] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016. 27

[57] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016. 28

[58] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019. 29

[59] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018. 29, 31, 32, 57

[60] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012. 29, 30, 31, 32, 57

[61] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016. 30, 32, 57

[62] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018. 30, 32, 57

[63] Gal Yona. A gentle introduction to the discussion on algorithmic fairness. *Towards Data Science*, 2017. 30

[64] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013. 30

[65] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018. 31

[66] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 31

[67] Han Yu, Yang Liu, Xiguang Wei, Chuyu Zheng, Qiang Yang Tianjian Chen, and Xiong Peng. Fair and explainable dynamic engagement of crowd workers. 2019. 31

[68] Raj Jain, Arjan Durresi, and Gojko Babic. Throughput fairness index: An explanation. In *ATM Forum contribution*, volume 99, 1999. 31

[69] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR, 2018. 31, 57

[70] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017. 31, 32, 57

[71] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017. 32, 57

[72] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016. 32, 57

[73] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017. 32, 57

[74] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. Fairness in relational domains. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 108–114, 2018. 32

[75] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. AI Now Report 2018. Technical Report December, 2018. URL `https://ainowinstitute.org/AI{_}Now{_}2018{_}Report.pdf`. 32

[76] Jason Millar, Brent Barron, Koichi Hori, Rebecca Finlay, Kentaro Kotsuki, and Ian Kerr. ACCOUNTABILITY IN AI Promoting Greater

Social Trust Acknowledgements and Authors Note. Technical report, 2018. URL `https://g7.gc.ca/wp-content/uploads/2018/06/FutureArtificialIntelligence.pdf`. 33

[77] Carrie Schroll. Splitting the bill: creating a national car insurance fund to pay for accidents in autonomous vehicles. *Nw. UL Rev.*, 109:803, 2014. 33

[78] Jeffrey K Gurney. Sue my car not me: Products liability and accidents involving autonomous vehicles. *U. Ill. JL Tech. & Pol'y*, page 247, 2013. 34

[79] Gary E Marchant and Rachel A Lindor. The coming collision between autonomous vehicles and the liability system. *Santa Clara L. Rev.*, 52:1321, 2012. 34

[80] Tracy Hresko Pearl. Compensation at the crossroads: Avs and alternative victim compensation schemes. *AIES*, 2019. 34

[81] Davide Castelvecchi. Can we open the black box of ai? *Nature News*, 538 (7623):20, 2016. 35

[82] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic impact assessments: a practical framework for public agency accountability. *AI Now*, 2018. 35

[83] Richard Gall. The difference between machine learning explainability and interpretability. `https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html`, 2018. 35

[84] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3): 50–57, 2017. 37

[85] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017. 37

[86] D Doran, S Schulz, and TR Besold. What does explainable ai really mean. *A new conceptualization of perspectives. CoRR, arXiv: abs/1710.00794*, 2017. 37

[87] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018. 38

[88] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018. 38

[89] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018. 39, 79, 81, 83

[90] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017. 39

[91] Yaochu Jin and Bernhard Sendhoff. Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(3):397–415, 2008. 40

[92] Alex A Freitas. A critical review of multi-objective optimization in data mining: a position paper. *ACM SIGKDD Explorations Newsletter*, 6(2): 77–86, 2004. 40

[93] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016. 40, 81

[94] Donald P Green and Holger L Kern. Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees. In *The annual summer meeting of the society of political methodology*, 2010. 40

[95] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 41

[96] Shervin Shahrdar, Luiza Menezes, and Mehrdad Nojoumian. A survey on trust in autonomous systems. In *Science and Information Conference*, pages 368–386. Springer, 2018. 41

[97] Sven A Beiker. Legal aspects of autonomous driving. *Santa Clara L. Rev.*, 52:1145, 2012. 41

[98] Jean-Michel Hoc. From human–machine interaction to human–machine cooperation. *Ergonomics*, 43(7):833–843, 2000. 42

[99] Jean-Michel Hoc. Towards a cognitive approach to human–machine cooperation in dynamic situations. *International journal of human-computer studies*, 54(4):509–540, 2001. 42

[100] Patrick Millot. Toward human-machine cooperation. In *Informatics in Control, Automation and Robotics*, pages 3–20. Springer, 2009. 42

[101] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020. 42

[102] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018. 42

[103] Monika Hengstler, Ellen Enkel, and Selina Duelli. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, 105:105–120, 2016. 43

[104] Ananth Uggirala, Anand K Gramopadhye, Brain J Melloy, and Joe E Toler. Measurement of trust in complex and dynamic systems using a quantitative approach. *International Journal of Industrial Ergonomics*, 34(3):175–186, 2004. 43

[105] Miltos Kyriakidis, Riender Happee, and Joost CF de Winter. Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation research part F: traffic psychology and behaviour*, 32:127–140, 2015. 43

[106] Michael Wagner and Philip Koopman. A philosophy for developing trust in self-driving cars. In *Road Vehicle Automation 2*, pages 163–171. Springer, 2015. 43

[107] Kumar Akash, Wan-Lin Hu, Tahira Reid, and Neera Jain. Dynamic modeling of trust in human-machine interactions. In *2017 American Control Conference (ACC)*, pages 1542–1548. IEEE, 2017. 44

[108] Shervin Shahrdar, Corey Park, and Mehrdad Nojoumian. Human trust measurement using an immersive virtual reality autonomous vehicle simulator. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 515–520, 2019. 44

[109] Walter Morales Alvarez, Miguel Ángel de Miguel, Fernando García, and Cristina Olaverri-Monreal. Response of vulnerable road users to visual information from autonomous vehicles in shared spaces. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3714–3719. IEEE, 2019. 45

[110] Cristina Olaverri-Monreal. Promoting trust in self-driving vehicles. *Nature Electronics*, 3(6):292–294, 2020. 45

[111] Agile. URL `https://www.infoworld.com/article/3237508/what-is-agile-methodology-modern-software-development-explained.html`. 46

[112] what is scrum. URL `https://www.scrum.org/resources/what-is-scrum`. 47

[113] Batya Friedman. Value-sensitive design. *interactions*, 3(6):16–23, 1996. 48, 58, 82

[114] Batya Friedman, Peter Kahn, and Alan Borning. Value sensitive design: Theory and methods. *University of Washington technical report*, (2-12), 2002. 48

[115] Batya Friedman, David G Hendry, and Alan Borning. A survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction*, 11(2):63–125, 2017. 48, 54, 56, 58, 62

[116] Batya Friedman and David Hendry. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1145–1148, 2012. 48, 58, 82

[117] Stephanie Ballard, Karen M Chappell, and Kristen Kennedy. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 421–433, 2019. 48, 58, 82

[118] Klaus Schwab. *The fourth industrial revolution*. Currency, 2017. 53, 77

[119] Spyros Makridakis. The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures*, 90:46–60, 2017. 53, 77

[120] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. 54, 55

[121] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016. 54, 55

[122] Kate Crawford. Artificial intelligence's white guy problem. *The New York Times*, 25(06), 2016. 54, 55

[123] Adrienne Yapo and Joseph Weiss. Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018. 54, 55

[124] John Havens. *Heartificial intelligence: embracing our humanity to maximize machines*. Jeremy P. Tarcher/Penguin, 2016. 54, 56

[125] Gregory Vial. Understanding digital transformation: A review and a research agenda. *Managing Digital Transformation*, pages 13–66, 2021. 54

[126] Ching-Hung Lee, Chien-Liang Liu, Amy JC Trappey, John PT Mo, and Kevin C Desouza. Understanding digital transformation in advanced manufacturing and engineering: A bibliometric analysis, topic modeling and research trend discovery. *Advanced Engineering Informatics*, 50:101428, 2021. 54

[127] Ching-Hung Lee, Amy JC Trappey, Chien-Liang Liu, John PT Mo, and Kevin C Desouza. Design and management of digital transformations for value creation, 2022. 54

[128] Chia-Yen Lee, Bai-Jian Chou, and Chen-Feng Huang. Data science and reinforcement learning for price forecasting and raw material procurement in petrochemical industry. *Advanced Engineering Informatics*, 51:101443, 2022. 54

[129] Mary Flanagan, Daniel C Howe, and Helen Nissenbaum. Values at play: Design tradeoffs in socially-oriented game design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 751–760, 2005. 56

[130] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*, pages 1–14, 2020. 57

[131] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020. 58

[132] Banda Gerald. A brief review of independent, dependent and one sample t-test. *International journal of applied mathematics and theoretical physics*, 4(2):50–54, 2018. 67

[133] Joost CF De Winter. Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, 18(1):10, 2013. 67

[134] Ethan Zell and Zlatan Krizan. Do people have insight into their abilities? a metasynthesis. *Perspectives on Psychological Science*, 9(2):111–125, 2014. 74, 101, 111

[135] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 77, 103

[136] Han Yu, Chunyan Miao, Yiqiang Chen, Simon Fauvel, Xiaoming Li, and Victor R Lesser. Algorithmic management for improving collective productivity in crowdsourcing. *Scientific reports*, 7(1):1–11, 2017. 77

[137] Jiehuang Zhang, Ying Shu, and Han Yu. Human-machine interaction for autonomous vehicles: A review. In *International Conference on Human-Computer Interaction*, pages 190–201. Springer, 2021. 77

[138] Sky Croeser and Peter Eckersley. Theories of parenting and their application to artificial intelligence. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 423–428, 2019. 78

[139] Michael Heinert. Artificial neural networks–how to open the black boxes. *Application of Artificial Intelligence in Engineering Geodesy (AIEG 2008), S*, pages 42–62, 2008. 78

[140] Ying Shu, Jiehuang Zhang, and Han Yu. Fairness in design: A tool for guidance in ethical artificial intelligence design. In *International Conference on Human-Computer Interaction*, pages 500–510. Springer, 2021. 78

[141] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019. 78

[142] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, 2019. 79

[143] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, pages 563–574. Springer, 2019. 79

[144] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019. 79

[145] Marco Ancona, Cengiz Oztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *International Conference on Machine Learning*, pages 272–281. PMLR, 2019. 81

[146] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020. 82

[147] Mi-Young Kim, Shahin Atakishiyev, Housam Khalifa Bashier Babiker, Nawshad Farruque, Randy Goebel, Osmar R Zaïane, Mohammad-Hossein Motallebi, Juliano Rabelo, Talat Syed, Hengshuai Yao, et al. A multi-component framework for the analysis and design of explainable artificial intelligence. *Machine Learning and Knowledge Extraction*, 3(4):900–921, 2021. 83

[148] Ben Shneiderman and Harry Hochheiser. Universal usability as a stimulus to advanced interface design. *Behaviour & Information Technology*, 20(5): 367–376, 2001. 87

[149] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561*, 2021. 103

[150] Olivia J. Erdélyi and Judy Goldsmith. Regulating artificial intelligence: Proposal for a global solution. In *AIES*, 2018. 108

[151] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10 (2):12:1–12:19, January 2019. ISSN 2157-6904. doi: 10.1145/3298981. URL `http://doi.acm.org/10.1145/3298981`. 111

[152] EuropeanUnion. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). 2016. URL `https://eur-lex.europa.eu/eli/reg/2016/679/oj`. 111

[153] H Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Aguera y Arcas. Federated learning of deep networks using model averaging. 2016. 112