UNIVERSITY OF GOTHENBURG

DEPARTMENT OF APPLIED IT

ASSESSING PUBLIC OPINION ON ALGORITHMIC FAIRNESS

Reviewing practical challenges and the role of contextual factors

Author: Veronica Kecki

Thesis:	30 hp
Program:	Digital Leadership
Level:	Second Cycle
Year:	2022
Supervisor:	Alan Said
Examiner:	Jonas Ivarsson
Report nr:	2022:058

Abstract

AI ethicists often claim that where algorithmic decision-making is impacting human lives, it is crucial to strive for transparency and explainability. As one form of achieving these, some authors have argued for socio-technical design of AI systems that involves the user in the design process. And while there is no shortage of cases where this step is absent due to blatant disregard for users' interests, one can say that even where that is not the case, this is no easy task due to a mounting knowledge gap among the general public on the subject of AI.

This Master thesis aims to demonstrate the above issue in concrete terms by attempting to collect public opinion on algorithmic fairness. The survey conducted for this thesis asks participants to pick among four different algorithmic models that they think achieves the best fairness in the presented scenarios. Results indicate that (1) contextual factors do play a role, and (2) that attempting to collect public opinion on the subject is challenging as there is insufficient knowledge on the topic and, therefore, poor understanding of the presented options.

As urgent as it is to conduct public consultations on algorithmic decision-making where human lives are increasingly impacted, it is even more urgent to improve public knowledge on the subject so that people could actually make informed choices. Understanding the complexity of contextual factors offers substantial support in that endeavor.

Keywords

AI, artificial intelligence, ML, machine learning, algorithms, algorithmic decision-making, fairness, socio-technical design, AI ethics

Foreword

I would like to thank Jonathan Rebane who has first introduced me to the subject of algorithmic fairness. I do not think I would consider this topic on my own, and even if I did, I would question how far I could actually take it without the guidance I received from Jon.

I want to thank my supervisor Alan Said who was very responsive and accommodating to the unexpected last moment adjustments that had to be made throughout my work.

And finally, I want to thank all of the teaching staff at the Department of Applied IT who have been continuously challenging us and expanding our horizons for the past two years, and for simply being there for us, always eager to lend a helping hand. It was a true pleasure and privilege to study at this program.

Abbreviations

AI – Artificial Intelligence

ML – Machine learning

PG – Protected group

NLP – Natural language processing

HRO – High reliability organization

POC – Point of contact

Table of contents

1 Introduction

Artificial Intelligence (AI) not only enables the ongoing process of digital transformation (Wischmeyer and Rademacher 2020) but also transforms the economy and society at large (Coeckelbergh 2017). The advancements of AI have become quite pervasive and often invisibly embedded into our daily lives (Boddington 2017). Today, anyone who uses modern technology is most likely enjoying features that, in one way or another, can find their origin in AI research (Frankish & Ramsey 2014). AI is already occupying a major spot across numerous domains, and, as many authors and experts believe, is going to play an increasingly pertinent role in our lives in the coming decades (*ibid*.).

Autonomous systems not only cut costs and enhance efficiency but also enter entirely new domains previously inaccessible to technological advancements. The numerous capabilities developed by AI due to the exponential growth of computing power, big data and fast mobile networks have enabled algorithms to take over such functions as planning, face recognition, natural language processing (NLP), and, finally, decision-making. Today, we see such novel AI applications in multiple spheres such as transport, finance, insurance, marketing, science, military or healthcare (Coeckelbergh 2017).

While AI opens a whole new array of opportunities, with them come many risks and threats. One question that often comes up, *"how many decisions and how much of those decisions do we want to delegate to AI? And who is responsible when something goes wrong?"* (Coeckelbergh 2017). AI ethics is the branch that typically deals with tackling such issues. It concerns technological change and its impact on individual lives, as well as transformations in society and the economy more broadly (*ibid*.).

Some authors posit that the vast majority of ethical problems generated by the use of AI can be described as unintended consequences of technology (*ibid*.;

Wischmeyer *et.al.* 2020). These technological solutions are typically created to address a specific problem and make improvements to a certain area, but then the question often arises, *"...improvement for whom? The government or the citizens? ... The retailer or the customer? The judges or the accused?"* (*ibid*.). The unintended negative consequences may arise even where the best of intentions are vested into the interest of all parties involved.

As difficult as the resolution of the concomitant issues may be in the above scenario, what makes matters even worse is where the negative consequences arise due to a blatant disregard for the interests of some of the parties. The development of AI brings up questions of power play, such as where a particular technology is being shaped by a few megacorporations (Nemitz 2018). This, in turn, begs the question, who shapes the future of AI, and what significance does this give to AI in social and political issues? (Coeckelbergh 2017).

As a result, a multitude of AI solutions are not only used in commercial applications at the discretion of the user but oftentimes also by public institutions or for-profit organizations without the consent (or even knowledge) of the end user. For instance, employers can monitor employee activities and performance (Tona *et.al.* 2021), users' personal data can be collected by advertising or social media websites and sold to third parties, and entire new markets of behavioral prediction and modification are being created by large search engine corporations (Zuboff 2015). In addition to the obvious breach of ownership, breach of trust and privacy are becoming ever more common: surveillance cameras on the streets can not only perform facial recognition but also read moods, while algorithms can spread hate or false information on social media (Coeckelbergh 2017).

It is becoming more and more evident that urgent measures need to be undertaken to address such issues although there are many questions surrounding this such as who is ultimately responsible, and who ought to take action? On the one hand, some authors stress that it is not sufficient to trust that organizations making use of AI will actually stick to the ethical principles

voluntarily, and argue in favor of legal regulation (Wischmeyer *et.al.* 2020). On the other hand, even for the organizations that are in fact willing to do the right thing, it may not always be so obvious what that "right thing" precisely is.

In algorithmic decision-making, for instance, there are many models each of which will alleviate at least one condition but then will also deteriorate at least one other condition at the same time (e.g. Hiller 2021), what is known as competing notions of fairness (Samuel 2022). It becomes impossible to state definitively that one model is more fair than the other generally speaking; rather, one model could be more fair in one particular aspect and less fair in another.

One notable case that demonstrates this conflict is the COMPAS algorithm used in the US judicial system to predict criminal recidivism (Abu-Elyounes 2020). It surfaced in the media in 2016, and became widely known as the "racist algorithm". COMPAS is a proprietary actuarial risk and needs assessment tool developed by Northpointe that is used by criminal justice agencies in many jurisdictions across the United States.

Upon looking into the case, the investigative journalism non-profit organization ProPublica has demonstrated that the above-mentioned algorithmic system made discriminatory predictions since black defendants who do not recidivate were nearly twice as likely to be classified by COMPAS as higher risk compared to their white counterparts (45 percent vs. 23 percent). In response, Northpointe commented that in their view, the algorithm was fair because the same percentage of black and white defendants did recidivate. In academia, opinions split: while some supported ProPublica's findings (e.g. Hao 2019), others attributed the results to external factors such as a different base rate among black and white defendants (e.g. Corbett-Davies 2016).

While there have been numerous calls for transparency and explainability of AI by AI ethicists or academics when it comes to algorithmic decision-making that impacts human lives (e.g. Kizilcec 2016), the example above is somewhat of a

different kind than, say, AI used in recommender systems or algorithmic grading of student papers. These could relatively speaking be described as low-stake scenarios. Verdicts passed in a criminal court, however, is a very different scenario because here, people's freedom is at stake.

It appears that very different approaches are required depending on the type of a scenario that is in question. This logic is somewhat comparable to the functioning of High Reliability Organizations (HROs) such as hospitals or airports that aim at 0% errors because any error will have a very high cost such as substantial monetary losses or even people's lives. Despite the general trend for organizations to attempt to leverage AI for efficiency to a maximum possible degree, in the case of HROs, achieving accuracy is much more critical even if it comes at the cost of efficiency (Salovaara *et.al.* 2019). This article concludes that it is critical to complement the autonomous (mindless) component with a human (mindful) component to achieve the most reliable results (*ibid*.).

Similarly, it appears necessary to involve the public in the very design of algorithmic decision-making systems that are going to impact areas of life where the stakes are very high before employing such systems in practice. Any algorithm in itself can be said to be a "mindless" component making human input necessary, but the issue becomes especially pressing in view of the competing notions of algorithmic fairness and high-stake contexts.

The aim of this thesis is to examine public opinion on the subject of algorithmic fairness more broadly, and to gain insight into a few more specific issues such as preference for a particular algorithmic model in a given context, and any difficulties that may arise among participants when reflecting on this matter. To achieve this, the following research questions will be asked:

RQ1. Does the public opinion on algorithmic fairness depend on contextual factors, and if so, what are some of these factors?

RQ2. What are some of the challenges of collecting public opinion on the subject of algorithmic fairness?

To answer the RQ1, a survey and a focus group will be conducted with the healthcare setting used as an example. Here, participants will be asked to pick an algorithmic model which they consider to be the most fair in a given scenario. Two major conditions will be contrasted: a high-stake condition (i.e. where people's lives are at stake), and a low-stake condition (i.e. where the decision made by an algorithmic model can either improve or deteriorate a person's well-being to a certain degree but is not life threatening). The results are expected to detect the difference in people's preferences depending on the scenario.

To answer RQ2, other responses of the survey and feedback collected throughout the focus group will be analyzed. One of the resulting contributions here is the five-point context model which could be utilized by practitioners or in future research.

2 Related Work

Previous studies directly related to the subject of this thesis can be broken down into the following categories based on individuals' preference for fair resource allocation:

- Fairness in general socio-economic contexts;
- Fairness in healthcare contexts;
- Algorithmic fairness in general socio-economic contexts and in the healthcare context.

2.1 Overview

2.1.1 Fairness in general socio-economic contexts

Literature on people's fairness preferences in general socio-economic contexts is in fact quite extensive. First, it is important to mention here that any studies on fairness are very much based on the conceptions of equity and equality: in short, the former could be described as equality of opportunity and the latter—as equality of outcome. Because the difference between the two concepts is based on individual contributions, when such contributions are the same, equality and equity coincide.

Another important concept is the trade-off between equity and efficiency as first discussed by Arthur M. Okun (1975). This trade-off appears at the center of numerous socio-economic relations, and is a major source of difficulty in policy formulation (Gordon-Hecker, Choshen-Hillel, Shalvi & Yoella Bereby-Meyer 2017). Understanding the challenges of this trade-off lies at the core of understanding fairness.

Below, a few studies touching upon these above-mentioned concepts will be discussed in brief. The main purpose of this discussion is to demonstrate the inherent nature of fairness in human beings in general, as well as the role of contextual factors in forming individuals' preferences for resource allocation.

The state in which equal work results in equal pay and unequal work results in unequal pay (also often described as the state in which the input/output ratio is constant for all members of society), has resulted in the equity theory. This theory posits that people consistently pursue equitable situations across contexts, and it is backed up by quite a few studies. Firstly, one can mention here that individuals have been shown to be naturally drawn towards equal allocations (Halevy & Chou, 2014). Secondly, when placed under cognitive load, participants are more willing to forfeit some of their payoff for the sake of reducing any existing inequities between themselves and other participants (Schulz, Fischbacher, Thöni & Utikal, 2014).

Additionally, a meta-analysis has shown that individuals prefer allocation patterns which result in maintenance of equity between oneself and others over the types of allocations which are merely advantageous for the self but harm the other (Van Lange, 1999). So maintaining equity has been demonstrated to be a default preference in human beings, the state which individuals are willing to support even despite own interest in many contexts (Van Lange, De Bruin, Otten, & Joireman, 1997).

Also, some of the work on the subject has pointed to the fact that the greater the deviation from these default equitable states, the more individuals feel distressed (e.g. Walster, Berscheid, & Walster, 1973). Multiple studies have shown that people are prone to exhibit inequity aversion, meaning that they try to avoid outcomes that deviate from equity, irrespective of whether such inequity is advantageous or disadvantageous for them (e.g. Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999; Loewenstein, Thompson & Bazerman, 1989).

To continue, individuals have also demonstrated sensitivity to the invested effort (as opposed to the perceived inherent right to equity). In view of that, individuals exhibit a tendency to allocate higher rewards to those who put in more effort (Leventhal & Michaels, 1971). Consequently, many sources have shown equity to be perceived as fairer than equality.

Speaking more specifically about economic relations, however, efficiency has been shown to take prevalence in people's preferences over equity (Gordon-Hecker *et.al.* 2017). In the organizational setting, research similarly shows that people cast preference for differential and potentially more efficient allocations in monetary contexts (Conlon, Porter, & Parks, 2004; DeVoe & Iyengar, 2010; Martin & Harder, 1994; Tornblom & Foa, 1983). This indicates a general preference for efficiency over equality in monetary allocations (Li *et.al.* 2019, 697).

2.1.2 Fairness in healthcare contexts

There is rather limited research on individuals' preference for resource allocation specifically in the healthcare context; below, two examples will be provided.

A paper by Sheldon and Smith published in 2000 has addressed some of the issues that arise when health inequalities in a context of limited health care resources are being tackled (Sheldon & Smith 2000). In this paper, the authors discuss the different possible logics behind the design of health allocation schemes. Here, one possible direction could be the focus on efficiency. That could translate, for instance, into maximizing benefits from resource use or saving lives.

The paper, however, makes an example of how prioritizing equity in the systems of such countries as Sweden or the UK has been beneficial. It concludes that equity can (and should) play a major role in resource allocation of the central budget to healthcare providers. It is important to keep in mind here, though, that this paper discusses allocation logic as seen by the allocating structures, not as perceived by individuals impacted by such allocations.

If looking at the issue of resource allocation on the hospital level, one interesting article here is by Li *et.al.* (2019). It looks at the challenge of allocation of transplant organs which are a very rare resource (and, therefore, not everyone who is in need of a transplant will get one). In such cases, hospitals are faced with a decision on how exactly they are going to be making the appropriate allocation.

This example is a perfect representation of the equity-efficiency trade-off: the challenge that hospital management is faced with here, should the total benefit be maximized, or should individual need be prioritized? Some of the possible distribution logics here could be:

- first-come-first-served basis;
- prioritizing younger patients;
- prioritizing healthier patients;
- aiming to maximize the total life-years saved.

As it was mentioned above, literature has pointed to people's preference for efficiency when it comes to economic relations. As for the healthcare context, there does not seem to be such a definite indication. Here, the authors are pointing out that the findings tend to oscillate between equity and efficiency depending on the specific issue at hand. For instance, one study finds more than half of jurors cast their vote in favor of allocating screening tests to all Medicaid recipients (i.e. based on need, or equity), despite this coming at the price of saving fewer lives in total (Ubel, DeKay, Baron, & Asch, 1996). Other studies have shown the results to be subjectable to framing effects (Colby, DeWitt, & Chapman, 2015; Li & DeWitt, 2017; Li, Vietri, Galvani, & Chapman, 2010; Ubel, Baron, & Asch, 2001), and preference for equity does not always represent the majority.

Li *et.al.* point out that even though the literature seems to indicate a general preference for equality when it comes to healthcare allocations involving lives (and, likewise, preference for efficiency in allocations involving money), these findings cannot be considered to be directly comparable because they come from

very differing lines of literature (Li *et.al.* 2019, 698). To gain some clarity, the authors of this paper have conducted 15 studies and did manage to detect a consistent pattern:

> "People demonstrate increased concerns for efficiency when lives are involved in the allocation decision compared to when lives are not involved."
>
> (*ibid*., 704).

In other words, this means that this study has shown preference for saving as many human lives as possible when the latter are at stake. To translate this into the terminology of the topic of this thesis, the above study has demonstrated people's preference for accuracy in high-stake contexts.

2.1.3 Algorithmic fairness in general socio-economic contexts and in the healthcare context

There is a growing body of literature on a multitude of adjacent concepts and issues related to algorithmic decision-making and its impact on people's everyday lives. Some studies have looked at the human perception of human decision-making as opposed to that of an automated system, such as an article titled "What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents" (Schlicker *et.al* 2021). There are studies demonstrating bias and discriminationas a consequence of utilizing algorithmic decision-making (e.g. Köchling 2021). Some studies propose concrete machine learning (ML) solutions to safeguard the interests of a protected group (the term which will be explained in the Methods section) presuming superiority of equity over accuracy by default (see Rajkomar *et.al.* 2019). Some authors have altogether attempted to argue for an alternative approach to understanding algorithmic fairness that no longer focuses on the traditional expectation to maximize the satisfaction of the mathematical criteria (e.g. Holm

2022). There is, however, not much work that directly tests for individuals' preference for a particular algorithmic model in a given socio-economic setting.

The paper by Srivastava *et.al.* (2019) presents a study in which the authors have asked participants to choose a specific algorithmic model which they believe achieves the best fairness in a presented scenario. By this, the authors attempted to capture lay people's perception of the (un)fair outcome of using different algorithmic models across different social contexts.

The paper begins by criticizing the attempts to arrive at a universal mathematical model for algorithmic decision-making that would maximize the resulting benefit, even if not fully satisfying all of the involved conditions (which has long been proven to be impossible). The authors posit that because fairness is highly contextual, preference for a particular model needs to be tested for a very narrow context. Similarly to Holm (2022), the authors argue that with this study, they are proposing an alternative approach to evaluating the social impacts of algorithmic decision-making than the traditional mathematical approach, namely, a descriptive ethics approach.

This study looks at three contexts: (1) recidivism risk assessment; (2) context of medical predictions, and (3) contexts where the decision-making stakes are high. Without going too much into the details of the methodology and study design, the overall results can be summarized by saying that demographic parity best captures people's perception of fairness (Srivastava *et.al.* 2021, 11). At the same time, the experiments dealing with high-stake conditions have shown that participants consider accuracy more important than equality when stakes are high.

Even though it will be explained in more detail in the Methods section, it will be briefly noted here for the sake of clarity that demographic parity, also referred to as statistical parity, represents an algorithmic model which corresponds to equity (as opposed to efficiency, if viewed through the lens of equity-efficiency trade-off),

and is often viewed as conflicting algorithmic models which prioritize accuracy. Also, it should be noted here that the second main finding of this paper is perfectly in line with the study by Li *et.al.* discussed above (2019).

2.2 Research Gap

Because both the literature on fairness in general socio-economic contexts and on algorithmic fairness pinpoint contextuality of fairness, it is critical to test for individuals' preferences for a specific algorithmic model in as many contexts as possible in the following ways:

- Different socio-economic domains can be examined individually (e.g. banking, education, criminal justice, healthcare etc.);
- Such domains can be compared to each other; additional variables can be studied within each of these domains (e.g. high-stake and low-stake conditions in healthcare; differences in preferences based on participants' demographics or other individual attributes etc.);
- Different combinations of algorithmic models in specific contexts can be tested (i.e. there may be stronger preference for a particular model only as opposed to other specific models, not necessarily in general regardless of the available options);
- And, finally, different forms of information delivery need to be tested. This could include online vs. in-person studies; interactive vs. non-interactive form; verbal vs. graphical vs combinatorial information delivery; delivering minimal vs. maximum. vs. balanced amount of detail (minimal risks leaving out essential information while maximum threatens information overload, and it is difficult to determine where exactly to strike the balance); etc.

This thesis examines individual preference for a specific algorithmic model out of four (Demographic Parity, Equal Accuracy, Equal Odds and Positive Predictive Value) in a high-stake and low-stake healthcare context, testing both interactive and non-interactive forms of information delivery, as well as both verbal and graphical representations, with balanced amount of detail.

3 Conceptual Framework

This section is going to lay out the main concepts pertinent to the discussion of algorithmic decision-making and fairness, namely, artificial intelligence, algorithms, machine learning, algorithmic decision-making itself and, then, algorithmic fairness. The latter will be expanded into the exploration of a more general understanding of fairness from a social perspective, legal understanding and different aspects of this notion as understood in computer science.

3.1 AI, machine learning and algorithms

The Cambridge Handbook of Artificial Intelligence opens up by stating that "very generally, artificial intelligence (AI) is a cross-disciplinary approach to understanding, modeling, and replicating intelligence and cognitive processes by invoking various computational, mathematical, logical, mechanical, and even biological principles and devices" (Frankish & Ramsey 2014, 1). AI can be both abstract in that it contributes to our understanding of natural cognition, and pragmatic in that it provides the foundation to the engineering of smart machines and applications (*ibid*.). Encyclopedia Britannica defines AI as the ability of a digital computer or computer-controlled_robot to perform tasks commonly associated with intelligent beings. Other sources describe artificial intelligence as "the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions": here, the author notes that "the term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving" (Frankenfield 2021).

Application of AI is based on utilization of large data sets (Big Data), suitable computing power, as well as specific analytical and decision-making capabilities to enable computers to perform tasks that resemble human capabilities—in some instances even exceeding them (Wischemeyer & Rademacher 2020, 2). Much of

the output produced by AI systems is delivered via machine learning (ML). The term was first proposed by Arthur Samuel who defined it as a field of study that provides learning capability to computers without being explicitly programmed for that (Samuel 1959). ML has multidisciplinary origins combining ideas from neuroscience, biology, statistics, mathematics, physics, psychology and philosophy (Marsland 2014; Alzubi, Nayyar & Kumar 2018).

Generally speaking, learning is a process that involves acquiring new, or modifying existing behaviors, values, knowledge, skills, or preferences (Alzubi, Nayyar and Kumar 2018). In humans and animals, learning is what allows us to adapt to the environment and is therefore regarded as one of the most fundamental components of intelligence (Marsland 2014). Some of its key aspects include remembering, adapting and generalizing. Learning in computer science implies producing an output from reiterated processing of data that would be as close to the correct one as possible; the gist of ML is, then, in making computers modify or adapt their actions so that such actions become more accurate (*ibid*.).

IBM describes ML as "a branch of AI and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy" (IBM 2020). ML is a term that is often used to broadly denote computer programs that are capable of learning from records of past conduct (Wischmeyer *et.al.* 2020). Past conduct, or experience, refers to the past information available to the learner which most commonly assumes a form of electronic data collected and made available for analysis, with both quality and size of such data being crucial to the success of the predictions (Mohri, Rostamizadeh & Talwalkar 2018).

ML enjoys a vast array of applications today. ML-based techniques have already been applied in such fields as pattern recognition, computer vision, finance, entertainment, computational biology, biomedical or medical applications (El Naqa 2015). It is highly likely that the range of such applications is only going to see expansion moving into the future.

As far as algorithms are concerned, there is currently no agreement in the literature regarding a universal definition for the term. There are quite a few academic articles dedicated just to the discussion of defining an algorithm. One such article opens by referencing a definition provided in Introduction to Algorithms by Corman *et.al.*:

> "Informally, an algorithm is any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output".
>
> (Corman *et.al.* 2002)

The author of the article is instantly subjecting this definition to sarcasm:

> "'Informally'? Can such a comprehensive and highly technical book of 1180 pages not have a 'formal' definition of an algorithm?"
>
> (Yanofsky 2011).

Jokes aside, the term "algorithm" is most commonly used to refer to some sort of a mathematical application or address a specific problem or a set of instructions (Hiller 2021). In data science, IBM defines an algorithm simply as "a sequence of statistical processing steps" (IBM 2020). Other sources define an algorithm as a set of instructions for solving a problem or accomplishing a task (Downey 2022) or a step-by-step procedure which allows to define a set of instructions that must be carried out in a specific order to produce the desired outcome (Upadhyay 2022).

To briefly sum up, different sources may use the terms "learning algorithm", "machine learning" and "artificial intelligence" interchangeably when referring to the outcomes of data analytics fairness or algorithmic decision-making. Nonetheless, it would be more precise to say that artificial intelligence is a broader type of intelligence than algorithms which has the capability to recognize new patterns across data sets. Machine learning is then, rather, a subset of artificial

intelligence that can recognize new patterns in data and then adapt to the identified changes (Hiller 2021).

Algorithms broadly speaking, and especially machine learning algorithms, typically perform well at tasks they were instructed to do. Literature, however, warns against relying on algorithms to do things they weren't specifically programmed to do—and, likewise, not to do things they weren't instructed not to do (Kearns *et.al.*).

3.2 Algorithmic decision-making

Algorithms, especially the machine learning ones, are often involved in systems that are used to support decision-making (STOA 2019). Such systems are heavily reliant on the analysis of large volumes of data. They are increasingly used in such areas as e-commerce, recommendation systems, employment, health, justice, policing, banking and insurance as they provide great benefits for individuals and organizations—both in the public and the private sectors.

Depending on the system, the degree of human intervention involved in the overall process may vary. Semi-automatic systems assist humans in making decisions (e.g. assisting doctors in diagnosing a disease). Some systems may be legally obligated to include human oversight in the process: for instance, Article 22 of the General Data Protection Regulation states that "automated individual decision making, including profiling, prohibits any "decision based solely on automated processing…" (GDPR 2016). Others may function entirely without any human intervention (e.g. recommender systems in entertainment) (*ibid*).

The same goes for impact: in certain instances, the impact on the end user may be barely noticeable (e.g. a recommender system deciding which song is going to be played next in the web music player)—and in others, it could potentially cost someone their life such as one of the scenarios that is going to be used in this

study (an algorithm needs to decide which of the critical patients get an allocation at the intensive care unit when spots are limited).

A distinction may also be drawn between predictive and prescriptive algorithmic decision-making although the line between the two may be quite blurry. (STOA 2019). An example of a predictive system can be one used in the judicial sphere to determine whether a criminal is likely to reoffend. A prescriptive system, in turn, can be exemplified by automated recruitment practices or university admissions.

Due to the ability of machine learning algorithms to infer correlations from large amounts of data, they can lead to better informed decisions (e.g. more efficient resource use or proposing a better patient treatment plan). At the same time, the use of algorithmic systems also gives rise to many potential risks. Among them are such downsides as discrimination, unfairness, manipulation or privacy breaches.

The judicial system, credit scoring, targeted advertising or automated employment practices, for instance, have clearly demonstrated the potential for discrimination borne by algorithmic decision-making systems. It can arise from different types of biases present in the training data, as well as due to certain technical constraints or already existing societal and individual biases.

Algorithmic decision-making is also often said to pose a threat to privacy and data protection. First, there is clearly an issue when collection of massive amounts of personal data is taking place on a regular basis such as when individuals use search engines or social media. Even in the absence of a clearly targeted personal attack on someone, just the mere fact of collecting this data is viewed by many as a pandora's box which can be opened at any moment, potentially not with the data owner's interests in mind.

Lack of transparency is also commonly discussed in relation to the risks created by algorithmic decision-making. Some authors claim that it opens doors to many different kinds of potential manipulation of data making it extremely difficult to

contest a decision carried out by an automated system. Scoring is another area that raises numerous concerns since it fosters a tendency of regarding humans merely as numbers which can be seen as an attack on human dignity (STOA 2019).

All of these may arise intentionally, accidentally or as a side-effect. Errors and inaccuracies are also a common occurrence. These negative effects of algorithmic decision-making raise a plethora of ethical, political, legal, or technical issues, where great care must be exercised in addressing them. Failure to tackle these issues properly may exacerbate them even further (STOA 2019).

As the use of machine learning algorithms in our daily lives is becoming increasingly widespread, legislators and consumer rights organizations have already made multiple attempts in addressing questions about the numerous ethical risks posed by such technology (Schmid 2022). Considering the potential for such algorithms to have a major impact on our society, many authors claim that they must also be subject to public debate (STOA 2019). And with the highlighted risks and challenges in mind, researchers and lawmakers alike concur on the need for multiple requirements to be set for the ethical implementation of algorithmic decision-making systems. Fairness, or absence of undesirable bias, is among one of the requirements for such systems which many authors at this point consider to be intrinsic.

One must keep in mind that the process of algorithmic decision-making includes multiple potential sources of unfairness. Among them is the actual content of the training data, as well as the way that data may be labeled, or which selection ends up being featured. In addition, there are simply many different definitions of fairness some of which may be incompatible with one another, as it will be shown below.

3.3 Algorithmic fairness

To better understand the issues at hand when discussing algorithmic fairness, one must have a conception of fairness in a more general sense. Below, fairness is first going to be discussed from a social and psychological perspective (and maybe a bit philosophical as well). Then, we will briefly touch upon the legal understanding of fairness; and finally, fairness will be looked into as understood in data science.

3.3.1 Fairness as a social construct

Fairness is a fundamental ethical conception, and the arguments about it have a long tradition in Western civilization (Rescher 2002; Velasquez *et.al.* 1990). A sense of fairness is a basic human reality, manifested even by the youngest. (Rescher 2002). The concept of fairness is very closely related to the concept of justice: while often used interchangeably (and they will be on a few occasions below), some authors note that there is nonetheless a difference between them. If justice more commonly refers to the standard of rightness, fairness points to the ability to judge someone or something without regard for one's own feelings or interests. Irrespectively, both highlight the idea of how one deserves to be treated and so they overlap substantially (Velasquez *et.al.* 1990).

Fairness is a concept that deals more with what is right rather than what is advantageous to the involved parties in a given context, with its very reason for being residing in the area of impartial justice—as opposed to that of human satisfactions (Rescher 2002). From ancient times to modern day, many have posited that justice and fairness form the foundation of ethics and morality, and that they underpin the very social stability and human dignity:

In antiquity, Aristotle formulated his initial conceptions of justice in the following way:

"Equals should be treated equally and unequals unequally"

Later, Kant expressed the following:

> *"Human beings are all equal in this respect: they all have the same dignity, and in virtue of this dignity they deserve to be treated as equals. Whenever individuals are treated unequally on the basis of characteristics that are arbitrary and irrelevant, their fundamental human dignity is violated."*

As another example, an American moral and political philosopher John Rawls writes:

> *"The stability of a society—or any group, for that matter—depends upon the extent to which the members of that society feel that they are being treated justly."*

(Rawls 2001)

"Justice, then, is a central part of ethics"

(Rescher 2002)

Questions of justice and fairness inevitably arise when individuals are of a different opinion regarding how various benefits and burdens are to be distributed when the appropriate decision needs to be made. In fact, most ethicists would argue that there would be no point of talking about justice or fairness if it were not for the conflicts of interest that arise when goods and services are scarce, and people differ in their views of who should get what. (Velasquez *et.al.* 1990). No stable society can function without some sort of principles of fairness that the vast majority accepts as reasonable.

According to the literature, there are three major types of justice:

- Distributive justice refers to the extent to which society's institutions ensure that benefits and burdens are distributed among society's members in ways that are fair and just; when the institutions of a society distribute benefits or burdens in unjust ways, there is a strong presumption that those institutions should be changed

- Retributive justice refers to the extent to which punishments are fair and just
- Compensatory justice refers to the extent to which people are fairly compensated for their injuries by those who have injured them; just compensation is proportional to the loss inflicted on a person.

(Rescher 2002)

Distributive justice can be said to be at the forefront of the conversation in the vast majority of cases involving fairness. It is common to hear discussions about who deserves to get a raise when the department can only allocate one; whether inheritance should be split equally between the two children of vastly differing financial status; or, if users should be getting their share from the corporations that make billions of dollars off their personal data that is being shared by individuals on social media.

Deutsch identifies the following general values involved in justice:

All individuals should be treated:

1. so that all receive outcomes proportional to their inputs
2. so that all are regarded as equals
3. according to their needs
4. according to their ability
5. according to their efforts
6. according to their accomplishments
7. so that they have equal opportunity to compete without external favoritism or discrimination
8. according to the supply and demand of the marketplace
9. according to the requirements of the of the common good
10. according to the principle of reciprocity
11. so that none falls below a specified minimum

(Deutsch 1975)

Here, it becomes quite evident that the above may conflict:

> *"The most needy may not be the most able, those who work the hardest may not accomplish the most, equal opportunity may not lead to equal reward, treating everyone as equals may not maximize the common good."*
>
> (Deutsch 1975)

With this daunting observation in mind, the author continues:

> *"Such questions as these have preoccupied scholars throughout history, and so far no completely satisfying answer has been given. I do not have one to offer either."*
>
> (Deutsch 1975)

One major conflict here is that of the individual and the group: depending on the context, there may not be sufficient resources in the pool to grant every individual in the group with what one deserves based on their inherent or earned merits. At the same time, dividing the available resources in equal parts might not seem fair to those who may have invested more effort than others. That, in turn, could lead to social tensions within the group or the tragedy of the commons (Foster Lloyd 1883; Ostrom 1990).

In response to this dilemma, Deutsch notes:

> *"...there is usually a positive, circular relation between the well-being of the individuals in a group (or society) and the well-functioning of that group: The more satisfied individuals are, the better their group functions, and vice versa"*
>
> (Deutsch 1975)

The author continues by stating that such a proposition is supposed to suggest that "justice is intrinsically concerned with both individual well-being and societal

functioning", and that "the natural values of justice are thus the values which foster effective social cooperation to promote individual well-being" (*ibid*.).

The specific ways in which justice should be served in each particular instance are highly contingent upon the circumstances: in some cases, distribution according to need may be the most fair; in some—according to effort, and so on. (Deutsch 1975). Likewise, use of quotas may be considered discriminatory where practiced to exclude certain individuals from the group, and highly commendable where practiced to ensure inclusion of disadvantaged groups (Deutsch 1975). While each case may be unique, Deutsch proposed the following general classification of which principles he believed apply in different types of social relations:

1. In cooperative relations where economic productivity is the primary goal, equity (rather than equality or need) are likely to be the dominant principle of distributive justice.
2. In cooperative relations where the fostering or maintenance of enjoyable social relations is the common goal, equality is likely to be the dominant principle distributive justice.
3. In cooperative relations where the fostering of personal development and personal welfare is the common goal, need will most likely be the dominant principle distributive justice.

(*ibid*.)

Another fundamental prerequisite for any discussion regarding fairness is having a valid claim to what it is that is being divided (Rescher 2002). For instance, one must be a lawful heir to claim inheritance, or a participant of a bet or a game to claim their share of the winnings. When we are the ones presenting such a claim, the situation has to do with rights; when the claims belong to others, we are dealing with obligations. (Rescher 2002). Both are quite evidently important to social stability.

Claims need to be set with objective—or at least neutrally impartial—standards, and can be said to have three major corresponding factors:

1. Equity: having people's shares be proportionate to their claims (and, accordingly, allocating equal shares to them in the case of equal claims)
2. Impartiality: avoiding favoritism and treating claimants with even - handedness, i.e. "without fear or favor"
3. Uniformity: proceeding via the uniform application of appropriate principles where everyone deserves the privilege of "due process"

(Rescher 2002)

The above factors reflect the general principles which will undoubtedly sound fair to most, although in practice, the distribution scheme might not always be very straightforward, with many different distribution scenarios being possible depending on the context.

One scenario could be to divide the goods (or bads) in equal parts, which would apply where the goods are identical and their number is divisible by the number of claimants (e.g. dividing a pizza of 8 pieces between 4 people so that each gets 2). Another could be distributing goods on a first-come first-served basis, for instance if a movie theater is giving away a limited number of free tickets as a promotion. Goods could be divided randomly which is commonly applied in situations where such goods are different and not fissionable (e.g. a host giving away different kinds of party favor gifts to the guests).

Where goods are different, with some clearly superseding the other one in quality or value, each of them could be split between the claimants evenly. For instance, where two siblings may inherit two land plots from their parents one of which has fertile soil and the other one does not, it could be seen as a fair solution for each of the siblings to take over half of each land plot. There are many more potential scenarios, and these few examples are simply provided here to demonstrate the contextuality of distributive justice.

With the above in mind, one could say that fairness hinges critically on the comparative magnitude of the claims that are at issue in relation to the magnitude of the good being divided (Rescher 2002). But on the other hand, the de-facto distribution practices are a whole other issue that may pose limitations or a plethora of additional considerations. While it is an absolute prerequisite for an individual to have a valid claim to even be considered for the potential distribution, it is not at all a guarantee one will get what one has a theoretical claim over.

Here, one could say that even where claims are equal, i.e. one individual does not have any more right to the claimed good (or bad) than someone else, the actual distribution can still be carried out differently in different instances. Consider the example of organ transplant allocation in the medical context: one could make a case that all life is equal, and therefore each patient has an equal claim to an organ when one becomes available. But because there are not enough for all, a hospital will have to make a decision regarding the allocation order. As discussed by Li *et.al.,* some hospitals have employed the first-come first-served principle while others exercise the maximum years saved approach (i.e. striving to prioritize younger patients) (Li *et.al.* 2019).

Considering that the involved individuals have been established to possess the appropriate claims, injustice may occur on the following levels:

> a. the values underlying the rules governing the distribution (injustice of values);
>
> b. the rules which are employed to represent the values (injustice of rules);
>
> c. the ways that the rules are implemented (injustice of implementation);
>
> d. the way decisions are made about any of the foregoing (injustice of decision-making procedures)"
>
> (Deutsch 1975)

Very often, injustice lies in the methods by which the decisions are made rather than in the substance of the actual decisions, and there is much social psychological research suggesting that this type of injustice is in fact the most fundamental (Deutsch 1975).

3.3.2 Legal fairness

The general conception of fairness lies at the foundation of any legal system. Here, fairness is primarily reflected in the due process the standards for which must be fulfilled both through procedural due process (i.e. whether the right procedures have been followed) and substantive due process (i.e. whether government's action is justified by due purpose) (Hiller 2021). While there is acceptance of the fact that there may be instances where the government may need to exercise its power over individuals, the important aspect here is for this to be carried out in a fair manner. It is often said that AI, similarly, has such a power over individuals' lives, and such power only keeps increasing.

One major difference between the legal notions of fairness and other types of fairness is that in the case of the former, there may be additional considerations alongside attempted non-discrimination or ethical outcomes such as different legal considerations and policy measures that could steer preference for a certain algorithmic model depending on the domain of application (Abu-Elyounes 2020). In certain instances, it could be a sensible approach to focus on short-term and long-term policy goals as opposed to individual cases.

Also, there may be additional obstacles involved: for instance, because it is one of the functions of the legal system to provide public order and safety, in some cases there may be a need for a tradeoff in favor of the latter (once again, as opposed to individual situations). So, while both algorithmic fairness from a computer science perspective and legal fairness involve tradeoffs, the issue may be with the overall goals of each which may be simply incompatible (*ibid*.).

Enormous work is currently being carried out by a variety of stakeholders on policy design for AI, with the initiatives predominantly focusing on establishing such requirements that would ensure that algorithms are fair. This means that algorithmic fairness must be translated into legal and policy terms (*ibid*.). To better understand the associated challenges, fairness as understood in data science is going to be briefly addressed below.

3.3.3 Fairness in data science

With over 20 different definitions of fairness in computer science literature (Abu-Elyounes 2020) and over 25 in the field of artificial intelligence (Hiller 2021), data science has not progressed in putting its finger on the essence of the term much further than philosophers or social psychologists. The focus of any methods being used in this field is typically on mitigating one impact at a time (Simons 2020). There is currently a disagreement in the data science community regarding the terms and goals, and several different approaches exist (Hiller 2021), as presented below.

3.3.3.1 Mathematical fairness

Some authors interpret fairness as a "concrete mathematical embodiment of some rule provided by an external party such as a government and which must be imposed on a learning algorithm" (Wick *et.al.* 2019). The two significant components here are (1) the mathematical expression and (2) the externally imposed rule-based nature of fairness. Because algorithms are good at optimizing what they are asked to do (and not good at optimizing what they were not asked to do), as established earlier in the discussion (Kearns *et.al.* 2019), the fact of external specification of rules is highly critical and is essential for the establishment of legal standards and principles (Hiller 2021).

To return to the conflict that has already been pointed out above, the mathematical conception of fairness also distinguishes between group and individual fairness. Group fairness is often referred to as statistical fairness, and it

seeks to foster equality across groups based on a particular mathematical attribute. A common example of such attributes could be demographics (and here, it is important that they are statistically similar). Achieving equality in such metrics is regarded as a way of avoiding discrimination. One must keep in mind, however, that the specific measures of group fairness will inevitably conflict with one another, and fairness is thought to be achieved when the needs of an average group member are met, as opposed to those of each particular individual within the group (*ibid*.).

The measure used in individual fairness, in contrast, is ensuring equal treatment among individuals with similar characteristics (as opposed to establishing a group average). One difficulty with individual fairness that some authors are pointing to is that there is a presumption of an external authority which determines the similarity (or differences) of individuals' attributes upon which the distinction among them is going to be made (Chouldechova & Roth 2018).

When comparing the two, Dwork and co-authors note that meeting the group conception of fairness is blatantly unfair on an individual level (Dwork *et.al.* 2011). The scientific community is presently in disagreement on the subject of which type of fairness is the most optimal, but in either case, one may say that mathematical fairness and predictive accuracy are considered to be opposites (i.e. by increasing one, the other decreases, and vice versa) (Hiller 2021).

3.3.3.2 Causality fairness

Another approach to fairness that has recently entered the data science field is through the lens of causality. Instead of addressing the question of whether the algorithm is producing a fair result, the causality approach is addressing the question of whether specifying a protected group causes a discriminatory result (Khademi *et.al.* 2019). This approach is fundamentally different from the group or individual fairness metrics, and not a common one to be used for mathematical fairness.

3.3.3.3. Quasi-mathematical fairness: socio-technical design

Despite thorough effort of the data science community to express fairness in mathematical terms, many authors still argue that mathematical methods are not an appropriate measure of fairness (or, to the contrary, of discrimination) in algorithms and thus, the technical approach alone is not sufficient. Potential unintended consequences of applying algorithms designed to satisfy such mathematical measures can harm the minority and the majority communities alike (Corbett-Davies & Goel 2018).

In a socio-technical approach to fairness, the mathematical measures are only one of the aspects, with social systems being no less than another equal component (Hiller 2021). There is a recognition of the fact that the socio-technical approach may not necessarily provide definitive answers but it is nonetheless a solid attempt in the right direction through strong frameworks that can enable process and order (Selbst *et.al.* 2019). There is also not one common way of administering the socio-technical approach; instead, different experts have proposed different ways of going about it in varying contexts.

Some researchers, for instance, have proposed a system design that would include collecting public feedback on the fairness of a model (Saxena 2019). Another proposal is to create a system from which data is submitted to a regulator for further correction of bias, and then returns to the appropriate decision maker to choose from a sliding scale of mathematical fairness (Zehlike *et.al.* 2020). In such solutions, the mathematical attributes are still at the core of the decision-making algorithm and design but the social lens is what allows to achieve a meaningful degree of fairness.

4 Method

This section is going to provide an insight into the various considerations throughout the process of method selection and study design deemed most appropriate to answer the research questions of this thesis. The section is divided into six parts, and will include (1) a brief overview of the general method approach, (2) a nutshell description of the research setting, (3) an in-depth explanation of the study design, as well as a few words on (4) data collection, (5) data analysis and (6) ethical considerations.

4.1 Method approach

Because this thesis intends to gain insight into the public opinion on the subject of automated decision-making and fairness of algorithmic models, conducting a survey appeared to be an appropriate form of research. Even though the questions of the survey constitute the main part of the study, they were nonetheless complemented with a focus group in order to both obtain more in-depth insights of the audience's opinions and provide the interactive element which is absent in the survey. For that reason, the methodological approach of this study can be described as a mixed methods approach since it combines both the quantitative aspect through the multiple-choice questions of the survey and the qualitative aspect through the open-ended questions of the survey and the feedback received during the focus group discussion.

It can be noted here that some of the literature on research methods regards the mixed methods approach as more validating than either quantitative or qualitative methods administered individually as the two complement each other (Bryman 2012). First, one method allows to confirm (or disprove) the results of the other. And second, while the quantitative aspect points to the general trend, the qualitative aspect allows to better understand the reasons behind such a

trend. The latter could also reveal weaknesses in the design of the former, and provide clues on improving it.

4.2 Research setting

Originally, this study emerged as a result of the author's collaboration with a research center in Stockholm which works on addressing ethical issues arising from the use of AI in various societal domains (for the sake of anonymity, this organization will be referred to as the Research Center hereinafter). The collaboration began in July 2021 but in March 2022, it was announced that the Research Center had undergone a reorganization as a result of change of investors and could no longer continue with this project. The author has chosen to continue with the same topic given a short time frame that was remaining but some adjustments had to be made to accommodate for the limited resources. Even though the Research Center did not go through with conducting the study as it had been planned, its background will be provided since it served as the foundation for picking this particular direction.

The Research Center is an organization located in Stockholm which prior to the aforementioned modifications consisted of a group of researchers and developers tasked with creating and testing AI systems to ensure their ethical design and sustainability of future applications. Its work included both commercial and non-commercial projects. The latter could be exemplified by some of the Swedish public institutions and agencies such as the Swedish Tax Agency (Skatteverket).

As for the project that this thesis is dedicated to, this was a task that the Research Center needed to complete in order to make a case before a law firm and demonstrate the reasons behind a choice of a specific algorithmic model for its client in healthcare. The latter intended to implement an AI system, and the Research Center was tasked with designing it (the exact details of the system to be designed had not been disclosed to the author). While the technical aspect of the task constitutes little difficulty (relatively speaking), the ethical aspect required

further research. The study to be conducted for this project was intended to serve as a pilot study with the aim of developing a framework for practitioners which could later be applied across different fields.

The author joined this project as a remote intern in July of 2021. The work primarily consisted of Zoom calls, emails and exchanges of literature. The author was introduced to the overall subject and was first tasked with investigating the latest state of research on the topic. Later, the author was introduced to the algorithmic models themselves, and needed to do extensive reading in order to properly understand the upsides and downsides of each, and by that, the conflict itself and the necessary tradeoffs.

After about two months of reading and researching the topic, the author was expected to begin preparing the survey where participants would be presented with several options corresponding to different algorithmic models. Because there are so many nuances to this subject, this task resulted in several failed attempts at formulating questions and creating visualizations. Several rounds of proposed updates and feedback took about two more months.

The final format of the survey was ready around November 2021. In December 2021, the author had an in-person meeting with the representatives of the Research Center in Stockholm where the resulting work was presented and approved for further use. The questions of the survey included a condition matrix visualization (to be explained below), a detailed description and a brief description. The thought behind presenting the same information in different ways was to attempt to appeal to individuals' different types of information processing and ensure that the questions are as understandable as possible to different audiences. The draft of the questions was tested with several representatives of the Research Center: it did take a while for them to read through them, but all concurred they were reasonably understandable.

It was agreed to allow for some additional time for the author to finalize the details of the survey (the condition matrix visualizations included lots of numbers which had to be manually adjusted to demonstrate the point in question). During the meeting, it was decided to conduct a focus group with other representatives of the Research Center once the contents of the survey were finalized. The main purpose of such a focus group would be receiving feedback on the survey design and making any appropriate adjustments before actually publishing the survey.

As it was mentioned above, it was communicated to the author in March 2022 that the Research Center had undergone a reorganization due to change of investors and leadership leading to the organization's changed status. The main point of contact (POC) that the author had been previously in touch with was moved to the commercial division of the organization and no longer had the appropriate status to continue with this project.

A list of other researchers affiliated with the Research Center had been provided to the author to check for potential interest for further collaboration as a replacement for the previous POC. After reaching out to several of them throughout March-April 2022, some responded that their current engagements did now allow for any additional projects and there were a few others who a response never followed from. In view of this, the author made the decision to make adjustments to the existing project and finalize it at own discretion.

4.3 Study design

Below, the design considerations for both the survey and the focus group will be discussed. Please note that data collection and analysis are going to be discussed as separate items after this subsection since they apply to both the survey and focus group design.

4.3.1 Survey

The questions of the survey were logically broken down into several sections for ease of processing. The survey had two main questions, and the rest of the questions were centered around them to help to understand the answers to the above. The description of the design of the survey is going to begin with the discussion of the main part, and will then move to the explanation for the choice of the remaining supporting questions.

Main part of the survey

The main part of the survey is presenting participants with two different scenarios in the healthcare context, and asking them to choose between four different algorithmic models that they feel achieve the best fairness in the given situation. The four algorithmic models are Demographic Parity, Equal Accuracy, Equal Odds and Positive Predictive Parity (to be explained in more detail below), and the two scenarios are formulated as follows:

> *Scenario 1. Imagine that a hospital is using an AI-based monitoring system to warn the rapid response team about patients at a high risk for deterioration, requiring their transfer to an intensive care unit (ICU) within 6 hours.*
>
> *Scenario 2. Imagine that a hospital is using an AI tool that allows to identify patients who are likely to develop vitamin deficiencies from the existing patient health data, and may subsequently recommend such patients to get a blood test*

The above scenarios were inspired by a case study presented in Rajkomar *et.al.* (2018): the first high-stake condition was borrowed from the paper unchanged, and the second one was adapted to convert it into a low-stake condition.

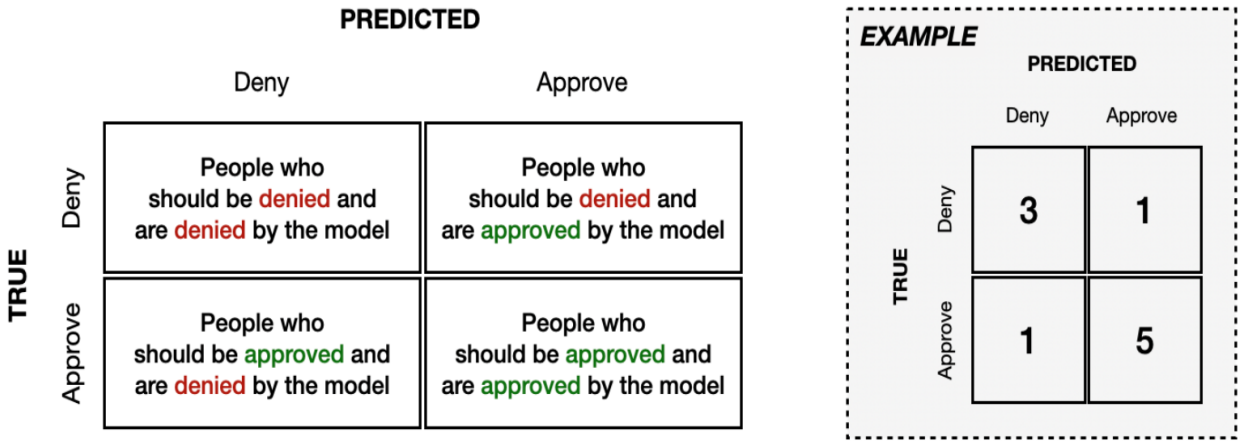
The four algorithmic model option presented under each scenario are exactly the same, and the only thing that is different is that the first scenario is a high-stake condition where human lives are at stake, and the second one is a low-stake

condition where no lives are at stake as a result of the algorithmic decision (i.e. it deals with a "nice-to-have" option rather than a "must-have"). The aim of contrasting these two conditions is to see (1) if there is any difference at all between participant's preferences depending on the context, and (2) if there is, then what does that difference look like exactly and which factors could it possibly be attributed to.

Also, the study contrasts two different groups of patients: a high-income patient group and a low-income one (and for the sake of simplicity, the study presumes that the groups are equally-sized). The low-income group is identified as the protected group (PG) as it could potentially be disproportionately negatively impacted by the algorithmic decisions due to quality of data. For instance, historically, individuals from economically disadvantaged backgrounds could be getting fewer allocations (whatever the resource might be) due to being unable to afford the needed care—not due to not needing one. But from a computer science perspective, the algorithm is much more likely to make a prediction that will conform to past patterns.

➢ Condition matrices

Even though not directly involving them in the survey (not their graphical representation at least), its questions were nonetheless based on confusion matrices that are typically used in algorithmic models:



The above image contains two rows and two columns resulting in four cells. “TRUE” represents cases where the allocation is indeed needed (e.g. a patient who is deteriorating rapidly and needs to be transferred to the ICU). Such cases can also be referred to as “qualified” (the term that will be used in the model descriptions below and in the actual survey questions), since these are the patients who have a legitimate claim to the resources in view of their health condition. “PREDICTED” indicates the results generated by the algorithm, which may or may not correspond with reality.

In sum, each algorithmic model produces four types of output:

(1) True Negative: individuals who are not in need of allocation, and they are correctly identified by the model as not needing one;
(2) False Positive: individuals who are not in need of allocation, but they are incorrectly identified by the model as needing one;
(3) False Negative: individuals who are in need of allocation, but they are incorrectly identified by the model as not needing one;
(4) True Positive: individuals who are in need of allocation, and they are correctly identified by the model as needing one.

➢ Algorithmic models

While there are quite a few different models out there, this study uses 4 which are often regarded as being among the most common ones: they are Demographic Parity, Equal Accuracy, Equal Odds and Positive Predictive Value. Each of these models, just like any other algorithmic model, will have at least one condition which it alleviates (e.g. equity), and at least one condition which it deteriorates (e.g. accuracy). This is perfectly in line with the point that has been reiterated throughout this thesis regarding the fact that every algorithmic model involves several conditions which are impossible to satisfy at the same time purely from a mathematical standpoint.

These four models will be briefly explained below with a visual example. These examples were originally supposed to be used in the survey when the work with

the Research Center was still ongoing, but they were no longer kept in the final survey that was conducted. They will be used below as a part of the explanation of the method of this study, and they were also used throughout the focus group (which will be discussed in more detail below under the appropriate item).

It should also be noted here that the numbers used in these sample illustrations are completely arbitrary. They are intentionally simplified to ease the cognitive burden of the study participants, as well as intentionally steered to favor one group over the other in respect to whichever condition that particular algorithm deteriorates. Needless to say, the de facto allocations in real-life cases could include very different kinds of representations; the purpose of these examples, however, was to exaggerate one possible type of a negative outcome. The idea here is not to suggest that this is the type of outcome that will happen every time these models are used (nor in the majority of cases), but merely that this is one realistic possibility that could take place. As it was discussed in the previous sections, the de-facto results could depend not only on the actual algorithmic model being employed but also on the quality of the data.

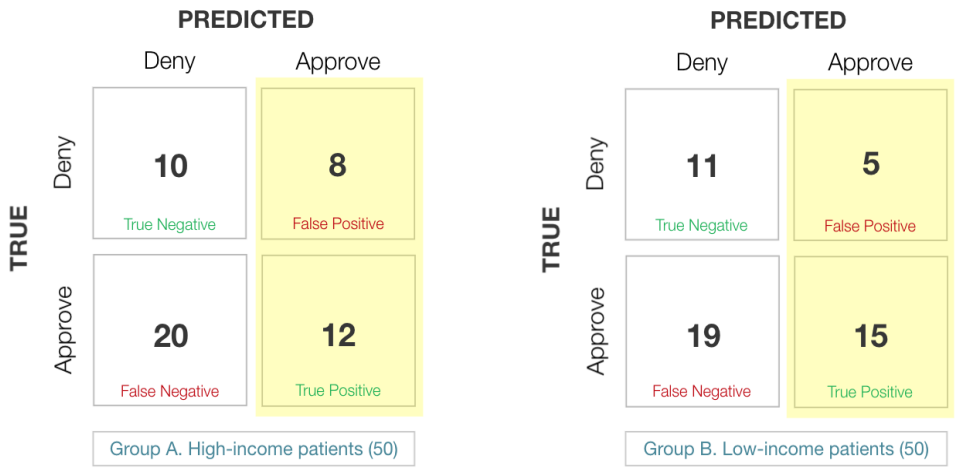
Specifically in these examples below, the allocation is being steered in favor of the high-income group, which is an outcome that is possible where historical bias is present in the data being utilized, and no protected group has been specified. This would represent one scenario of unfairness: here, one could be asking the question, *Why should the high-income patients be getting more resources while the low-income patients are already struggling with so many other areas of their life, and might not have alternatives to this type of healthcare as opposed to the high-income patients?*

The opposite scenario of unfairness is possible where a protected group is in fact specified, and the allocation is being steered in its favor. Some might ask, *Why is it fair for the low-income patients in need of urgent care to receive it in a much higher proportion compared to the high-income patients? A dying patient is a dying patient, regardless the financial status; all in need should be treated equally.*

To briefly sum up, the point with these specific numerical representations is not to place focus on either of the above two possible scenarios of unfairness, but just to use one particular example to underline a potential for enhancing inequality. This study does not attempt to test for participants' attitudes to algorithmic outcomes for the low-income patients as opposed to those for the high-income patients. In reality, there could be a different set of groups whose interests are at stake (older vs. younger applicants in the recruiting process; white vs. black offenders in the criminal justice example etc.), and the allocation could be steered in either direction depending on the algorithm design. The below examples are merely samples.

1) Demographic parity

Google's Machine Learning Glossary defines demographic parity as a fairness metric that is satisfied if the results of a model's classification are not dependent on a given sensitive attribute. What this would mean in our case as applied to either of the two presented scenarios, is that an equal percentage of patients receives the allocation irrespective of qualification. Here, the division is made based on group attributes (in our case, whether an individual belongs to a high-income or a low-income group)—rather than individual needs of receiving an



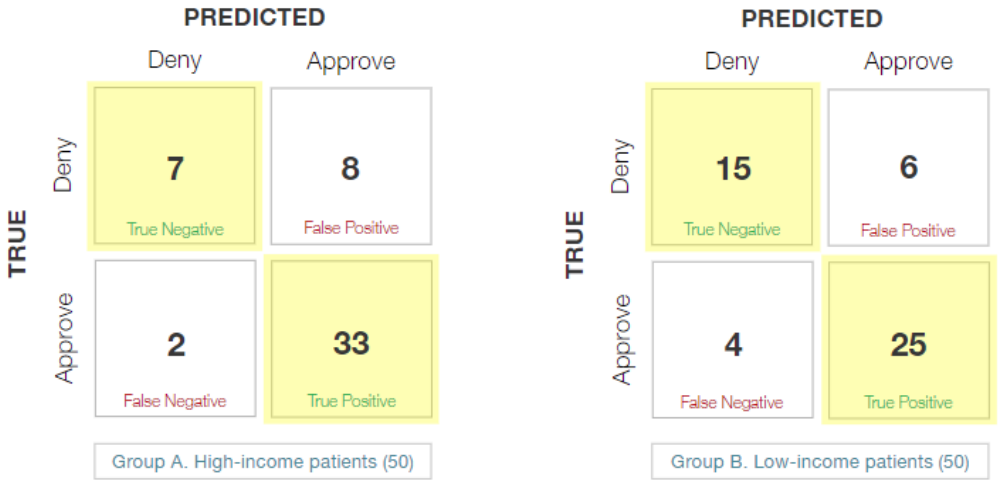
allocation of appropriate healthcare resources.

Below is a sample representation of a confusion matrix filled out for our case according to the demographic parity model using the first high-stake scenario:

Here, the groups are equally-sized (50 high-income and 50 low-income patients); and even though the de-facto numerical distribution looks slightly different across the two groups, what satisfies the demographic parity measure is that out of a total of 100 patients (50 patients with high income, i.e. Group A, and 50 patients with low income, i.e. Group B), the same percentage of patients from each group (in this case, 40%, or 20 out of 50 patients per group, highlighted in yellow), get transferred to the ICU. The issue here, however, is that the model accuracy for the high-income patients (Group A) is 60%, whereas it is 44% for the low-income patient group (Group B).

2) Equal Accuracy

While there may be different ways of defining accuracy depending on the context, in our case accuracy can be described as the proportion of the sum of True Negatives and True Positives from the total number of patients.

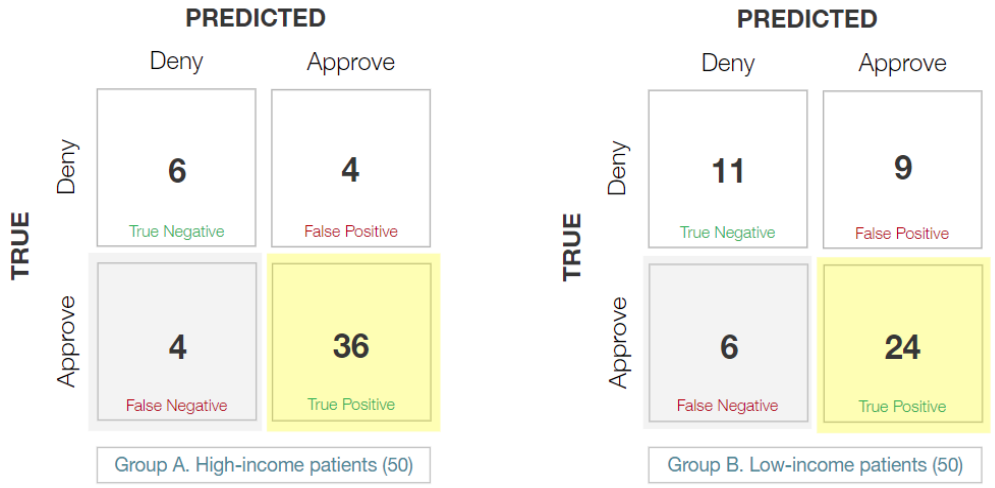


Here, the model is 80% accurate for both groups (a total of 40 out of 50 correct results in each of the groups, highlighted in yellow)—which is both even across groups (and 80% can be said to be quite a high level of accuracy overall). The issue here, however, is that a higher percentage of high-income patients are

receiving spots at the ICU than low-income patients: 82% (i.e. 8+33=41 out of 50 patients) vs. 62%, (i.e. 6+25=31 out of 50 patients), respectively.

3) Equal Odds

Put simply, the fairness condition in this model is satisfied if the same percentage of qualified patients from both groups receive spots at the ICU. "Qualified" refers to the patients who are actually in need of a transfer, i.e. the sum of False Negatives and True Positives (or, as represented by the TRUE values of the condition matrix occupying the bottom row).

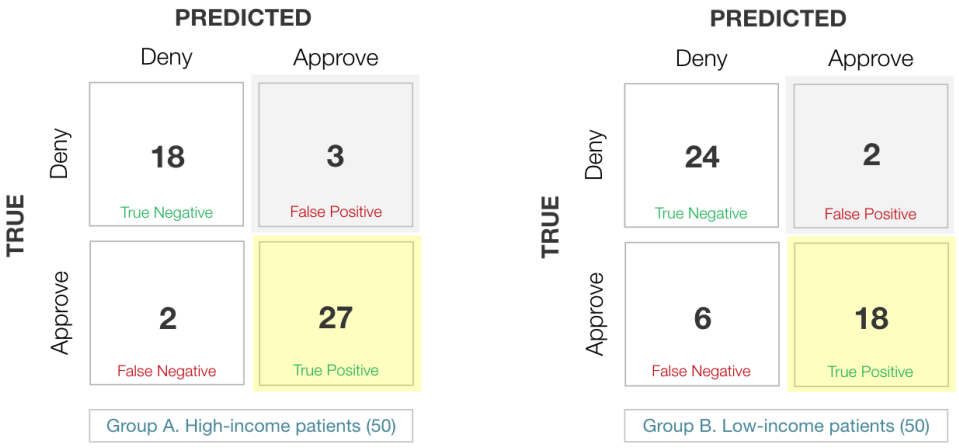


In this representation, we can see that out of all the patients who actually require a transfer to the ICU (4+36=40 from Group A, and 6+24=30 from Group B, as highlighted), 80% are identified correctly in both groups (36 out 40 in Group A and 24 out of 30 in Group B). At the same time, we see that a higher portion of all qualified high-income patients end up getting a spot (36 out of 50, or 72%)—as opposed to the low-income group (24 out of 50, or 48%), as highlighted in yellow; and, a higher percentage of high-income patients overall receive a spot at the ICU (34+4=40 out of 50, or 80%, as opposed to 24+9=33 out of 50, or 66%).

4) Positive Predictive Value

This algorithmic model looks at the accuracy of the prediction made by the algorithmic model. In other words, fairness is satisfied when the same percentage of qualified patients from each group is identified correctly out of all of the patients

who received a spot at the ICU. So, this model looks at the right column of the condition matrix, and is calculated as the percentage of True Positives from the sum of False Positives and True Positives.



In this sample representation, out of all the patients who get approved for a transfer to the ICU (3+27=30 from Group A and 2+18=20 from Group B, as highlighted), the probability that the algorithm targeted the right patients is 90% for both groups (i.e. 27 out of 30 in Group A and 18 out of 20 in Group B) but also, that means that a higher percentage of high-income patients gets transferred to the ICU compared to the low-income group (30 out of 50, or 60%, vs. 20 out of 50, or 40%, respectively). Additionally, there is a substantial difference in accuracy across the two groups (45 out of 50, or 90% for the high-income group vs. 38 out of 50, or 76% for the low-income group).

In the resulting survey, it was attempted to convert the main gist of the above four models into a verbal expression. The results were as follows:

Demographic Parity	A. The same percentage of patients from both groups, regardless of whether they actually need a transfer or not (i.e. if they are qualified), get transferred to the ICU—but the algorithm may not be equally accurate across the two groups (i.e. it may happen that the accuracy is much lower for either of the groups)

Equal Accuracy	B. The model is equally accurate for both groups (i.e. same percentage of True Negatives + True Positives)—but it may be so that a smaller portion of patients from either of the groups overall ends up getting placed at the ICU (in this case, irrespective of being qualified)
Equal Odds	C. The same percentage of patients actually requiring a transfer (i.e. those who qualify) end up getting one across both groups—but it could happen that a much smaller overall portion of patients from either group get the allocation
Positive Predictive Parity	D. Out of all the patients who end up getting a spot at the ICU (i.e. True Positive + False Positive), the same portion is identified correctly by the algorithm across both groups—but that could include a much lower portion of patients from one of the groups overall, as well as it may imply lower accuracy for one of the groups

These resulted as the multiple choice options for the two scenarios presented at the beginning of this section. In other words, the options for both questions are identical, with the only difference being that the first question involves a high-stake condition where people could potentially lose their lives, and the second question involves a low-stake condition where the advantageous outcome could be described more like a “nice-to-have”. The initial idea was to see if a difference in preference for a particular option can be detected depending on the context since, as it has been pointed out above, fairness is contextual). As it was becoming increasingly obvious throughout the preparation of this study, however, a much bigger concern is whether participants can understand the options they are being presented with in the first place.

For this reason, the author has decided to include option E for the survey participants to be able to state that they do not fully understand these four options above. Even though this option was not originally discussed with the Research Center in the process of designing the survey, the author has deemed it necessary to maximize response honesty. Forcing participants to choose one of four options that they do not truly understand does not seem to contribute to fairness or

transparency in this domain. The entire study rests precisely on people choosing what they think is fair in a given context so they need to understand what they are choosing from.

Survey description

The survey begins with a description that gives some background into the subject. The main aim of this description is to help participants understand the above two questions that constitute the main part of the survey. The description reads as follows:

> ***
>
> *Algorithmic decision-making is increasingly used in different socio-economic contexts such as when granting bank loans, hiring employees or grading student papers. It is often said that algorithmic decision-making requires a tradeoff: what it means is that typically, there are several conditions involved but they cannot be satisfied all at the same time purely from a mathematical standpoint. For instance, the model can be trained to achieve maximum accuracy—but then, it can steer the allocation towards one particular group; or, it can be trained to divide resources equally across groups, but the results may not be very accurate. Additionally, the output may be impacted by the quality of available data which may include historical bias or inaccuracies depending on how it was collected.*
>
> *All of the above may lead to automated decision-making which many individuals would consider unfair. With this in mind, there may be contexts in which it is appropriate to identify a protected group (i.e. individuals who could be disproportionately affected in a negative way), and adjust the algorithm to enhance the resource allocation in its favor. This could be, for instance, a low-income or a minority group. At the same time, because fairness is highly contextual, there may also be situations where not everyone would agree that this is a fair course of action.*

In this survey, you will be presented with two different scenarios taking place in the healthcare context. An AI system needs to make a decision resulting in the allocation of resources across 2 equally-sized groups: a high-income patient group, and a low-income one. In our case, the choice needs to be made between 4 different algorithmic models which prioritize different conditions. The low-income patient group is regarded as the protected group.

Keep in mind that every model will create four types of output: (1) patients who do not actually require an allocation and they correctly do not get one (True Negative); (2) patients who do require an allocation but do not get one in error (False Negative); (3) patients who do not require an allocation but do get one in error (False Positive); and, (4) patients who do require an allocation and do correctly get one (True Positive).

From the above, False Negative (2) and True Positive (4) are the qualified patients since they are actually in need of the allocation, regardless of the final decision made by the AI system. It is their well-being that is at stake.

***

Of course, the above constitutes a rather heavy cognitive load for an average participant even in its simplified form. This description can be considered simplified in the sense that it attempts to use simple everyday language, although, to clarify some bits that are perhaps a bit more technical such as the types of output, extra clarifications needed to be provided. This did make the language easier to understand but it also expanded the overall amount of text. The resulting description was a balancing act between including the minimum of information that is critical to understanding the main questions of the survey, keeping the language simple, including clarifications to address different audiences yet at the same time keeping it as brief as possible.

Additionally, an audio was recorded by the author containing the same description as presented above. This was done both to accommodate people's different modes of information perception, and also the practical aspect of it being more convenient for some of the participants to be able to listen to the description on the go instead of having to sit and read the text. The resulting audio was a little over 3 minutes long.

Supporting questions

Before presenting the participants with the main two questions discussed above, the following two introductory questions were asked:

> *Do you have any familiarity with the subject of algorithmic fairness / AI bias?*
>
> and,
>
> *How do you feel about automated decisions used in the social sphere, where a direct impact on human lives is possible?*

The options available under each of these questions will be discussed in more detail in the following sections of this thesis, but it shall be mentioned here that the purpose of including these questions was to attempt to place the answers to the main two questions of the survey into some context. For instance, one can try to correlate option E being selected as a response to the main questions of the survey with the participant's general level of awareness on the subject.

After the main part of the survey, participants are presented with a few basic demographic questions such as age, gender, occupation and place of residence. These questions could provide some insight into the overall trend of preference of a specific algorithmic option based on some individual characteristics. At the very end, a comment field was provided where participants could freely include any thoughts or considerations they had.

To sum up, it can be said here that the main part of the survey seeks to answer RQ1 while the supporting questions aim to answer RQ2. Although, partially, the main part will contribute to answering RQ 2 as well.

Distribution

The survey was primarily distributed through a multitude of online channels:

- the survey was shared in several Facebook groups, both local community groups and global interest groups dedicated to the topic of AI (each has several thousand members);
- the survey was shared on the author's LinkedIn profile with over 500 connections, and reshared by one of the professors from the Department of Applied IT at the University of Gothenburg also with over 500 connections;
- the survey was shared by the author's thesis supervisor on his Twitter page with over 2,000 followers;
- the survey was sent out to student mailing list of the University of Gothenburg;
- the survey was shared on author's three different Instagram accounts with over 1,000 followers combined

The author also shared the survey personally with several contacts working in the field of AI ethics.

Limitations

There were no limitations imposed throughout the response collection.

4.3.2 Focus group

Even though the original intention was to conduct a focus group with the representatives of the Research Center prior to publishing the survey, the unexpected turn of events mentioned earlier did not allow for this to happen. The best that the author could arrange given the circumstances was a focus group at their place of employment which was conducted shortly after the survey was published.

To mention a few words about this organization, this is a language and technology company providing services to world's leading streaming platforms. AI is used throughout many stages of the process, and issues of fairness of the platform do come up in the daily work as well. While the participants were not knowledgeable about the specific algorithmic models prior to the study, most of them had some degree of familiarity with the overall subject.

There were a total of 12 participants, and the study took place in a hybrid form: about half the participants gathered at the company's office space, and the other half joined via Teams. The session lasted for 30 minutes and was recorded. Essentially, the participants were asked the same questions (and were expected to fill out the same survey at the end), with the only difference being that the case description was given to them by the author verbally and in much more detail including the condition matrix visualization. The overall aim of this focus group was to allow for interactivity to ensure that participants understood the questions, but also to see what exactly was unclear to them.

4.4 Data collection

The main part of data collection was done quantitatively by conducting a survey. The survey has been published online as a Google Form, and the responses were collected throughout most of May 2022.

The focus group took place on May 13th, 2002 in Gothenburg, Sweden, and this data collection was qualitative.

4.5 Data analysis

For RQ1, the responses to the two main questions of the survey have been analyzed qualitatively using a comparative approach. The independent and dependent variables are identified, and the differences in the results are discussed.

For RQ2, a thematic approach has been applied to the studied literature to derive at a list of points which then serve as the unit of analysis for the data obtained from the study.

4.6 Ethical considerations

The study aimed to preserve participant anonymity which was announced at the very beginning of the survey description. There was still an option for participants to disclose their identity in comments if they so wished (some did because they wanted to keep contact and see the study results, for instance). But there was no prerequisite for participants to provide any of their contact information. While this is generally an ethical approach, an additional reason for pursuing this path was also a psychological one: because one of the response options to the main questions implied admitting to not understanding the presenting options, it was important to provide safe space for participants to answer honestly.

The study did not impose any demographic limitations for the participating audience as all opinions were highly sought after.

5 Results

In this section, the results of the study are going to be laid out. Below, first the direct responses to the survey questions will be presented, followed by their analysis. Then, the comments received throughout the focus group will be discussed.

5.1 Survey

First, the response count for the survey and essential demographic highlights will be provided. Then, similarly to the Methods section, the presentation of survey responses will begin with the two questions that ask participants to choose the algorithmic model that they think is the most fair in the described scenario since they comprise the main part of the study. This part will be followed by the discussion of the responses to the remaining questions and attempting to tie them to the responses to the two main questions.

A total of 132 responses have been collected for the survey within the allocated time frame. There have been a little over half of male respondents, a little under half of female respondents and a few percent who identified as non-binary or prefered not to disclose their gender. All ages were represented, with the average age being 32. Over 60% of respondents work full-time, and almost 40% are students. 76% of respondents reside in Sweden; 19% reside in the EU outside of Sweden, and a few percent reside in several other countries around the world.

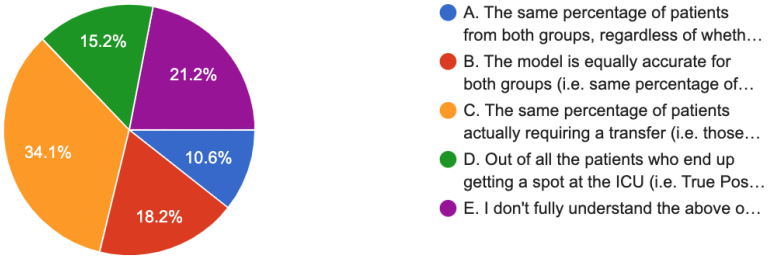
Main part of the survey

As far as the main part of the survey is concerned, the responses to the two “big” questions were as follows:

Scenario 1. Imagine that a hospital is using an AI-based monitoring system to warn the rapid response team about patients at a high risk for deterioration, requiring their transfer to an intensive care unit (ICU) within 6 hours.

Which of the following patient distribution models do you think achieves the best fairness?

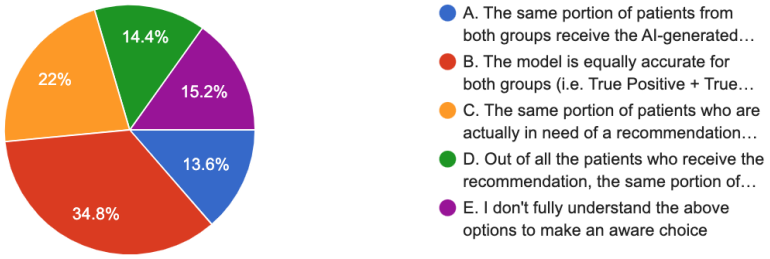
132 responses



Scenario 2. Imagine that a hospital is using an AI tool that allows to identify patients who are likely to develop vitamin deficiencies from the existing patient health data, and may subsequently recommend such patients to get a blood test

Which of the following patient distribution models do you think achieves the best fairness?

132 responses



While every option is represented, we can see that there is a clear preference for a particular algorithmic model in each case: that is, Option C in Scenario 1, and Option B in Scenario 2. In other words, the majority of respondents preferred the outcomes corresponding to the Equal Odds model in a high-stake scenario, and the outcomes corresponding to the Equal Accuracy model in a low-stake scenario.

As it was pointed out earlier under 3.3.3.1, mathematical fairness and predictive accuracy are considered to be opposites when discussing the fundamental differences between algorithmic models. To translate this into the language of fairness as a social construct, algorithmic models could be broadly divided into those that prioritize equality and those that prioritize equity. In this particular example, equality means ensuring inclusion of members of both groups, while equity can be expressed as accuracy. Each of these broad categories can also be broken down further into models that achieve equity and equality in different ways.

To analyze the four options of the models selected for this study, only one of them can be classified as the "equality algorithm" (and that is Option A, the Demographic Parity model). The other three models (Equal Accuracy, Equal Odds and Positive Predictive Parity) are three different variations of the "equity" algorithm: they all prioritize accuracy but do so using a different logic.

Firstly, the result of this study has shown that participants do prefer accuracy over equality regardless of whether we are looking at a high-stake or a low-stake scenario. The author's interpretation of this finding is that irrespective of a scenario, it is still people's health that is at stake (as opposed to monetary resources), and that in itself is enough to opt for accuracy. This can be said to be in line with the findings of the study by Li *et.al.* (2019) which concluded that individuals prefer accuracy when people's lives are at stake, and equality when monetary resources are at stake. It is also in line with the findings of the study by Srivastava *et.al.* (2019) which showed a tendency to prefer the Demographic Parity model in most contexts except for healthcare where accuracy is more preferred.

Secondly, even though it was accuracy (and not equality) that the participants cast their preference for in both cases, there was still a difference in preference for a specific model. While both Equal Accuracy and Equal Odds prioritize accuracy, the latter can be said to be more focused on the final outcome for those

requiring the allocation (i.e. those who are “qualified”). The Equal Odds model seeks to ensure that those who truly need help are going to get it.

Equal Accuracy, on the other hand, focuses more on the accuracy of the outcome for everyone—in other words, this model seeks to ensure that those who do need help, do get such help; and, at the same time, that those who do not need help, do not get it. In that sense, even though this model prioritizes accuracy, there is an element of equality present here as it is trying to do right by everyone involved rather than only focusing on delivering help to those who need it the most.

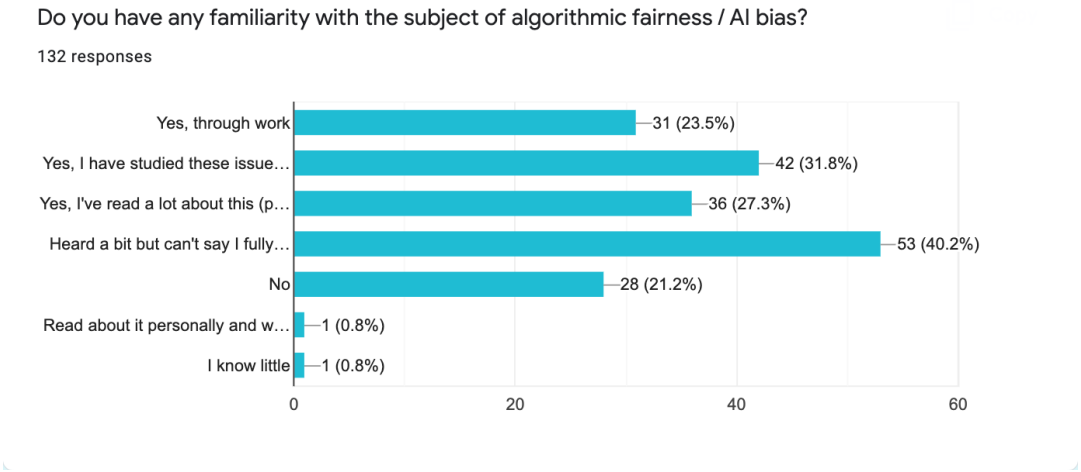
If one were to apply this lens, this could offer a potential explanation of the difference in the resulting preferences across the two scenarios. In the high-stake scenario, where human lives are at stake, it is of utmost importance to ensure that as many of those who need help end up getting it. In the low-stake scenario, it could be more reasonable to consider the final outcome for everyone involved, even if still trying to achieve maximum accuracy.

Supporting questions

As it was mentioned in the Methods section, before presenting the participants with the two main questions of the survey, they were asked two general questions about their familiarity with the subject and their general opinion on the use of algorithmic decision-making in the social sphere. Below are the results preceded by a full formulation of the questions and the options for each since they are not fully visible in the graphics:

(1) *Do you have any familiarity with the subject of algorithmic fairness / AI bias?*

- *Yes, through work*
- *Yes, I have studied these issues at a university*
- *Yes, I've read a lot about this (personal interest)*
- *Heard a bit but can't say I fully understand the issues*
- *No*

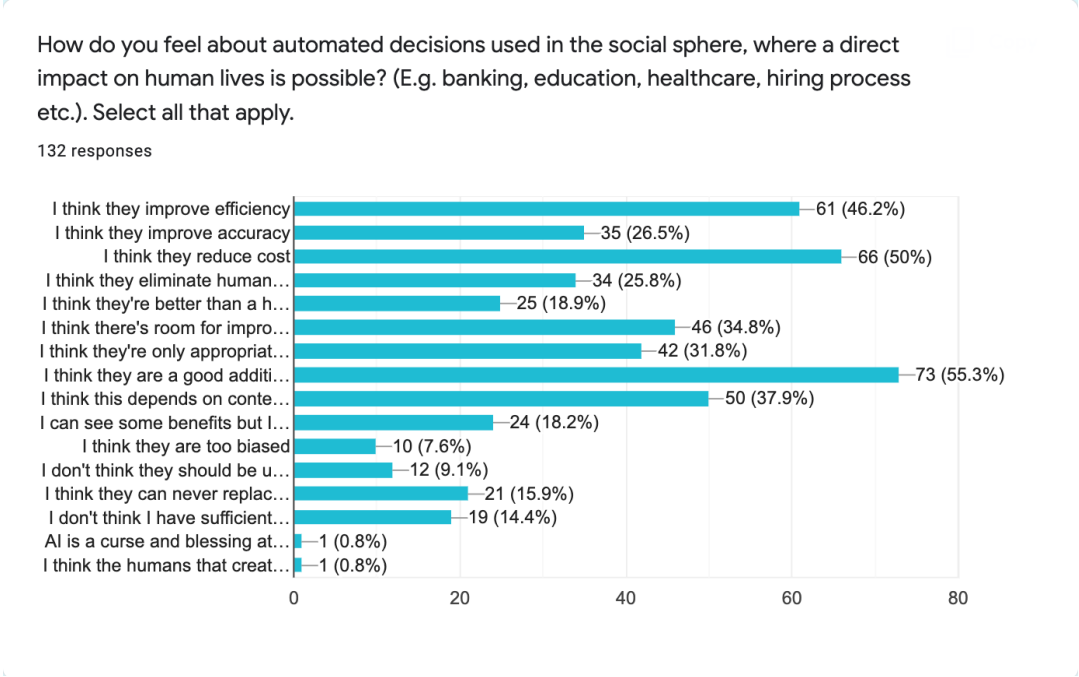


This question stemmed from the author's own suspicion that the level of awareness of the general public on the subject is rather low, and the results have confirmed it. It should be mentioned here that the percentages do not add up since not all of the options necessarily contradict each other (one could be familiar with the subject through work or school but still not be very knowledgeable on it). What is important here, however, is that 40% have responded that they do not fully understand the issues despite having some degree of familiarity with them, and 21% are not familiar with this subject at all. Combined, it can be said that over 60% of respondents have little to no understanding of the issues at hand.

(2) *How do you feel about automated decisions used in the social sphere, where a direct impact on human lives is possible? (E.g. banking, education, healthcare, hiring process etc.). Select all that apply.*

- *I think they improve efficiency*
- *I think they improve accuracy*
- *I think they reduce cost*
- *I think they eliminate human bias*
- *I think they're better than a human counterpart because they're objective*
- *I think there's room for improvement but they nonetheless offer many benefits and are therefore needed*
- *I think they're only appropriate in a very limited range of contexts*
- *I think they are a good addition to a human counterpart—not a replacement though*

- *I think this depends on context: they could be a good thing, or could be a bad thing. This should be evaluated on a case-by-case basis*
- *I can see some benefits but I don't think they outweigh the risks*
- *I don't think they should be used in the social sphere*
- *I think they can never replace a human counterpart*
- *I don't think I have sufficient knowledge to form an opinion on the subject yet*



As it was discussed in the Methods section, the aim of this question was to place the responses to the questions from the main part into perspective and try to evaluate people's general opinion on the overall subject. The options included the extreme positive and negative evaluations, and a multitude of milder options in-between. Participants were also free to write their own answers under "Other".

With 55% of responses, the most popular answer was, *"I think they [automated decisions] are a good addition to a human counterpart—not a replacement though"*. This can be said to show fundamental distrust for algorithms as making decisions that impact people's everyday lives. While the other two prominent options ("*I think they reduce cost"* and *"I think they improve efficiency*", with 50 and 46% respectively) are on the positive side of the opinion spectrum, they nonetheless do not pertain so much to the issue of fairness but rather the

economic benefits that automated decisions may offer due to their technical capability.

Other notable responses include *"I think this depends on context: they could be a good thing, or could be a bad thing. This should be evaluated on a case-by-case basis*", and *"I think there's room for improvement but they nonetheless offer many benefits and are therefore needed*" with 38 and 34%, respectively. Both of these show acknowledgement of issues that the use of algorithms raises but also their acceptance to a certain degree.

Strongly charged negative opinions regarding automated decision-making in the social sphere were in the minority: 16% of respondents believe that algorithms can never replace a human counterpart, 8% believe they are too biased, and only 9% think that they should not be used in the social sphere at all. So this further points to very little rejection of algorithmic decisions by the people which appears understandable considering the magnitude of their embeddedness into our daily lives.

Notably, 14% have responded that they don't think they have sufficient knowledge to even form an opinion on the subject. This can be further tied to the discussion of the previous question regarding the low level of awareness on the overall subject.

5.2 Focus group

The primary aim of the focus group was to gain insight on participants' ability to understand the questions of the survey and by that, to contribute to the answer of RQ2. The meeting took place in a hybrid form but it needs to be mentioned that irrespective of whether a participant joined online or in-person, each had the survey open on either a computer or a projector in front of them.

As it was mentioned in the Methods section, the focus group participants were presented with the same survey questions, but instead of having the participants

read the description at the beginning of the survey, the author walked them through the subject in much more detail verbally. “Detail” here refers not only to the information itself coupled with more abundant examples but also to visualizations. Even though the condition matrices were originally going to be used in the actual survey while the collaboration with the Research Center was still ongoing, the author made the decision to not include them in the final survey out of fear of potentially confusing the participants even further. This was one of the things that a focus group provided some further insight on.

After introducing the participants into the subject using general information and plenty of verbal illustrations of algorithmic decision-making used in our everyday lives, there seemed to be a very comfortable level of understanding of the subject in the room. When speaking purely in hypothetical terms, there did not seem to be an issue grasping the issues at hand. More difficulty arose when addressing the main two questions of the survey and looking at the specific scenarios.

When formulating the text of the survey, the author’s concern was that keeping the descriptions short could risk omitting information important to the understanding of the questions, while including overly-abundant information could overwhelm some of the respondents. With the focus group, however, the expectation was that this issue would be solved since such descriptive information delivered verbally is perceived much more easily through conversation than in writing. And this was true for the general background information although not so much for the actual questions from the main part of the survey. When it came to the main two questions, it was becoming difficult for the participants to follow the author’s clarifications without having to constantly look into the screen and remind themselves of what, for instance, “True positives” or “True negatives” were.

Turning to confusion matrices, however, produced mixed reactions. There were some participants in the room who said the tables were quite useful to their understanding of the two main questions of the survey, while others said that they

became even more confused than they already were after looking at them. To provide an example here, there were two participants who were most vocal in expressing their reaction to the matrices: the person who found them helpful works at the finance department, while the person who found them confusing (and who also noted that the survey description was perfectly understandable to him) works with communications. Clearly, people take in information differently just in view of their natural proclivities. Also, there were some developers in the room who did not find any difficulty in understanding condition matrices since they are very much familiar with the field of ML.

This indicates an important finding regarding conducting this type of a study. This kind of a survey includes two different types of information: the general background information and a specific case information; each requires different modes of presentation due to people's inherent differences in perception of information of different types. This study has indicated that that the general background information could be better perceived when delivered verbally with the use of ample examples. The reason for that is that this would provide sufficient context and understanding of the different aspects of the subject without resorting to a cognitive overload if participants were to read the same information as text since it would be far too long. The specific questions, in turn, could be better perceived in a visual form (whether textual or graphical) since there are a lot of specific details included which can get easily mixed up when delivered verbally. Additionally, one must consider individual differences in informational perception regardless of information type which was clearly demonstrated through the use of condition matrices.

The author would like to stress an additional aspect regarding the information presentation that became clear upon conducting this focus group, namely, interactivity. There may be an overwhelming number of questions arising among participants all of which are impossible to predict (and even if it was possible, including all of this detail into the survey description would lead to an information overload). This could be both because the subject of algorithmic-decision-making

is relatively new for the general population, and due to the differences in individual perception. Also, there are numerous nuances which may not be directly relevant to be able to answer the questions of the survey but some people might want to understand, for instance, why these are even the options they are being presented with in the first place.

6 Discussion

This section tries to place the results of the study discussed above into a broader context. First, answers to both research questions are going to be provided, followed by a discussion of the study's implications for practitioners and proposed avenues for future research.

6.1 Research questions

The discussion below attempts to provide a direct answer to the research questions of this thesis, and to connect them to some of the conceptual framework presented earlier.

RQ1. Does the public opinion on algorithmic fairness depend on contextual factors, and if so, what are some of these factors?

Similarly to so many other studies exploring fairness, the study conducted for this thesis also finds that people's opinion specifically on algorithmic fairness (as opposed to fairness in other domains) does depend on contextual factors. Before discussing such factors, it should be mentioned here that when speaking of fairness, "context" can be viewed on many different levels. Below is the author's classification synthesized from the literature processed while writing this thesis.

(1) Context could be categorized based on the ***source of fairness principles***. This could be: (1) humans' inherent sense of fairness resulting in its social conceptions aimed at resolving any potential conflict of interests through maximizing the satisfaction of both individual and group interests as much as possible (see e.g. Deutsch 1975); (2) legal fairness where the government and the legislative system arrive at carefully negotiated fairness principles aimed at maintaining the overall social balance, which does not always correspond to humans' inherent sense of fairness (see e.g. Hiller 2021); (3) and, finally, data science fairness where fairness principles are established predominantly in a

mathematical way by those creating the autonomous systems in question (*ibid*.).

(2) Context could be categorized based on the ***sphere of application***. This could be, for instance, the criminal justice system, banking, education, healthcare, insurance, and so on. Each of these spheres could also be broken down further based on different criteria. For instance, the educational context could be broken down into primary, secondary and higher education; it could be broken down by the area of study (e.g. humanitarian vs. technical); one could look at private vs. public education etc.

In the same key, healthcare could be broken down into public and private care, pediatric and adult care, or urgent and non-urgent care. Healthcare could be analyzed by country since each will have different legislation that applies. Healthcare could be looked into in terms of the allocated resources, for instance, monetary resources vs. organ transplants, as discussed in Li *et.al.* (2019). Healthcare could be viewed from the standpoint of the government passing the relevant legislation on resource allocation as discussed by Sheldon and Smith (Sheldon *et.al.* 2000), or from the perspective of hospital management deciding how they are going to distribute the limited resources among their patients (e.g. Li *et.al.* 2019).

(3) Context could be categorized based on the "***end recipient***", i.e. those whose interests are impacted by the involved fairness decisions. The most obvious classification here is individual vs. group fairness, although the latter could be broken down further into all sorts of subgroups. Additionally, groups can be divided based on very different principles which will not always have clear boundaries and may overlap substantially. This point has been addressed by some authors when discussing identity politics (Peterson 2018): for instance, one can be a female, belong to a minority group, coming from a financially

disadvantaged background, and represent the LGBTQ community all at the same time. In cases such as these, it may be very difficult to make the placement of an individual within the appropriate group and grant the relevant allocations.

(4) Context could be categorized based on what is at ***stake***. This could be described very broadly (e.g. economic vs. non-economic resources; high-stake vs. low-stake conditions), or defined in much more concrete terms within each larger category (e.g. monetary resources vs. real-estate; a patient's overall health and long-term well-being vs. their immediate survival (such as after an accident or as a result of grave illness); a student's grade vs. college admission etc.). This could also be analyzed by sphere such as health, privacy (such as in the case of social media and search engine use), dignity (such as employers tracking their employee performance, or parents sharing photos of their children on social media without their consent), financial solvency, freedom (such as in case of the decisions made in the criminal justice system) etc.

(5) Context could be categorized based on the ***resulting outcome***. In ML, one aspect that is distinguished across different algorithmic models is whether the outcome is advantageous or disadvantageous (e.g. Verma *et.al.* 2018). For instance, in the example of the COMPAS system, the outcome is disadvantageous since an individual who is predicted to recidivate is likely to remain in jail longer. An algorithm making a decision leading to a student's college admission, in contrast, is an advantageous one.

There are likely more ways of looking at contextual factors, however, the author has considered these five to be the most fundamental. Also, it should be mentioned here that the above selection was tailored to this study. For instance, in his book on the general conceptions of fairness in its common social understanding, Rescher is very detailed in his discussion of the numerous contextual factors that can arise in fairness conflicts (2002). Many of them,

however, do not apply either to algorithmic fairness more broadly or specifically to algorithmic fairness in healthcare and so they have not been included above.

Below, the results of the study will be analyzed through this lens of context.

As it has been demonstrated above, understanding the various dimensions of context can be extremely complex. There are numerous ways of categorizing it, with an almost infinite number of possibilities to break it down further into smaller subcategories. The study conducted for this thesis looks at only one particular slice of contextual factors which will be named below in accordance with the five-point context model provided above:

(1) *Source of fairness principles*: in our case, participants were asked to apply their own inherent sense of fairness to the data science type of fairness.

(2) *Sphere of application:* firstly, and obviously, the sphere that our study has looked at is healthcare. The two different scenarios have contrasted urgent and non-urgent care. The study looked at the decisions intended to be made by the hospital management (as opposed to the government passing the appropriate legislation), and it was the allocation of hospital resources that was involved (which is technically a material resource although it has a direct impact on patients' lives or health more broadly). Other dimensions have not been clearly stated here (e.g. whether we are looking at private or public healthcare).

(3) *End recipient:* the end recipients are hospital patients so there are first and foremost individual interests at stake here. Although, depending on which condition a specific algorithmic model prioritizes, one could argue that group interests are also at stake here when vulnerable populations may be involved (which is the whole reason for identifying a PG).

(4) *What is at stake:* in the first scenario, human lives are at stake, and in the second one, it is patients' overall health and well-being.

(5) ***Resulting outcome***: receiving either a spot at the ICU or the AI-generated recommendation is the desired (and by that, an advantageous) outcome. It can be mentioned here that the two scenarios were designed with this condition in mind: they were intentionally brought to the "common denominator" for ease of processing, better understanding by the participants and clarity of analyzing the results.

From the above, we can see that there are several contextual factors that applied to both scenarios (i.e. they served as independent variables) with one that differed (i.e. our dependent variable). The one factor that differed across the scenarios was the high-stake outcomes as opposed to the low-stake outcomes. And as it was discussed in the Results section, the result did differ which points to the influence of the contextual factor in question. At the same time, as it was also pointed out there, the difference was not enormous which could be explained by the similarity of the other factors.

RQ2. What are some of the challenges of collecting public opinion on the subject of algorithmic fairness?

In the author's assessment, the challenges of collecting public opinion on the topic of algorithmic fairness are threefold:

(1) Overall knowledge on the subject

This study has demonstrated a low level of awareness on the subject, with over 60% of survey respondents reporting little to no knowledge on the topic. Also, 6 people wrote in comments at the end of the survey that the questions were too difficult to understand for an ordinary person. This is perhaps not surprising as there is a lot of background information which needs to be understood by study participants prior to them being able to express their true opinion regarding what they believe is fair in a given context and what is not. At the same time, there are practical limitations

which do not allow to include exhaustive relevant information into the study description.

(2) Complexity of the subject

Filling the above-mentioned knowledge gap is no easy task: the issues at hand are rather complex and involve some specialized knowledge which may not be easily understandable for an average audience. One needs to have a mathematical understanding of each individual algorithmic model and subsequently, the difference between them—which is not an impossible task but could prove more challenging for some individuals than others depending on their natural proclivities or sphere of occupation. There needs to be some understanding of why it is even being proposed to apply algorithmic decision-making in place of a human counterpart in the first place (something pointed out in comments by one of the survey respondents); what are the pros and cons of each, and why one outweighs the other in a particular instance. A common question is, why the variables the discrimination by which it is attempted to avoid are used at all in an algorithmic model (and the answer to that is, because it doesn't necessarily prevent the problem to remove certain variables that could cause discrimination such as race because other variables such as zip code could still produce a discriminatory result indirectly). One participant mentions in comments at the end of the survey, "even the notion of accuracy is not straightforward"; another person asks if one was supposed to apply one's own understanding of fairness in the presented scenarios or any particular definition as there are many.

(3) Individual peculiarities of informational perception

While there are additional forms of information presentation which could aid study participants in better understanding the issues at hand, they may not be equally effective for different people. Some people may

perceive information better in a textual form; some may perceive it better in an audio form such as an audio recording; others may prefer an interactive conversation and others may respond better to graphical visualizations. Additionally, people might have different reactions to the same format: for instance, one person left a comment at the end of the survey stating that they thought the formulation of the questions was too wordy; at the same time, another participant wrote that they would appreciate an even more detailed clarification in parenthesis after each option, and one more person wrote that *"the issues need to be teased out"*.

So, not only is the subject of algorithmic decision-making highly complex even for someone who is familiar with it (e.g. one survey respondent mentions in comments that his mother is a doctor who endorses algorithm use in many healthcare contexts—and yet he himself recognizes that the subject is still very *"tricky"*), but there is also a very low level of knowledge on this difficult topic among the general population. Additionally, this type of a study could be said to be based on quite a few presumptions (e.g. that an algorithm is better than a human counterpart in this particular case), and one would require a lot of background information in order to understand why these presumptions are in place before one can express their opinion on fairness in the presented scenarios with a high degree of awareness of what one is choosing.

6.2 Implications for practitioners

Below, a brief comment will be offered on how some of the insights derived from this study could be of use to practitioners. Here, "practitioners" can refer to both commercial or public entities administering autonomous decision-making systems that directly impact people's lives.

Before proceeding, the author would like to briefly return to the point that was raised in the Introduction. One should keep in mind here that there is a difference

between the type of issue discussed in this thesis, and the broader issue of AI transparency. In many cases, it suffices to administer the decision-making algorithm accompanied by an appropriate explanation to achieve transparency (e.g. grading process, see Hosanagar *et.al.* 2018). In the case of this study, however, the issue is not about making an automated decision and being able to explain it later but rather including the end user into the design process before any decisions are made because in some instances, the cost of the wrong decision can be far too high. Healthcare or criminal justice systems are among the most obvious examples of such areas. With this in mind, this study provides several practical implications for collecting public opinion on algorithmic fairness.

First, it is important to specify the context as precisely as possible. The narrower the studied context, the more accurate the results will be. This study has shown that the algorithmic models chosen for the two different scenarios were not very far from each other in terms of their fairness logic which can be explained by a rather narrow contextual niche being examined. And at the same time, the difference was still detected which can be directly attributable to the one differing variable (i.e. the high-stake vs. low-stake condition). This is also supported by the findings of Li *et.al.* that people's preferences for resource allocations in the healthcare context oscillate between equity and efficiency depending on the specific issue at hand (Li *et.al.* 2019).

Second, considering the vast knowledge gap among the general population and the complexity of the subject, it is critical to conduct any public consultations in an interactive form. Purely from a practical perspective, it is impossible to include every single relevant detail into the survey description as that would make such a description far too long for an average person to process as it would become too cognitively demanding. Also, it is impossible to anticipate every single detail that may appear as relevant to any given participant. It appears sensible to provide the bare minimum of relevant information in the survey description, and then to conduct the study in an interactive form where more detail is provided verbally, and participants have the possibility to ask any follow-up questions. Further

analyzing such follow-up questions can also provide additional insight into how individuals perceive the presented subject, and how further studies could be improved. And generally speaking, the work on addressing this knowledge gap should be tackled from both directions: on the one hand, general education on the subject needs to be increased in schools and at universities (which will take time but is not any less necessary in view of that), and on the other hand, extra effort needs to be put into carefully crafting the study information to compensate for the lacking background knowledge.

And third, different modes of information presentation should be given serious consideration when conducting this sort of studies. This could be approached in a modular way: on the one hand, there are some forms of information presentation that most people are comfortable with; on the other hand, there may also be individual differences in preference for a specific type of information presentation. In addition to specifying the context, it could also be beneficial to specify your audience.

6.3 Future research

As it was pointed out above, there is a near-infinite number of constellations of contexts that are possible in a given scenario, and this study has analyzed only one of them. Therefore, there are many more to go!

The five-point context model provided earlier could be a starting point for mapping all of the major types of context that apply in a given situation, and a study could be conducted for each of the resulting niches. If we continue with our case, below are some examples of studies that could be undertaken:

- this study could be replicated but testing for different combinations of algorithmic models to choose from;
- this study could be replicated but using different scenarios. This could still include a high-stake / low-stake pair, but perhaps there could still be differences in the resulting preference depending on the specific case in

question. Also, one could try to detect a correlation between someone's responses and personal experiences in the healthcare context.

- another study could try to look at a more specific sphere of application within healthcare, i.e. public vs. private care, or specific departments within a healthcare system;
- a meta-analysis of all of the above could be conducted to have a map of individual preferences for different algorithmic fairness models within healthcare depending on a variety of contextual factors.

All of the above could be replicated for a multitude of other domains where algorithmic decision-making is impacting human lives. And prior to doing so, it is important to map out all of the applicable types of contextual factors, and see how they can be combined in one study.

7 Conclusion

This thesis has addressed some of the challenges that arise when implementing autonomous decision-making systems across a multitude of social domains where people's lives are impacted. Depending on the sphere of application, simply providing an explanation of an automated solution may not always be sufficient to adhere to ethical practices. Where the stakes are too high, a public consultation is needed to offer some reconciliation of the competing notions of algorithmic fairness—a mathematical conflict which does not have a technical solution.

The focus of this thesis has been twofold: RQ1 had a narrower scope and sought to detect people's preference for a specific algorithmic model in the two presented scenarios in the attempt to evaluate the role of context, while RQ2 looked at broader challenges of conducting this type of studies such as named in RQ1.

It should be mentioned that both the direct result of the study and its broader implication conform to the existing literature. Firstly, the results pointed to the fact that contextual factors do matter in defining people's preferences for a specific algorithmic model. This is precisely what traditional literature on fairness in its social understanding has been attempting to pronounce for decades—and one may see that this perfectly translates into the domain of algorithmic fairness as well. And secondly, the result of this study is in line with Li *et.al.* (2019) and Rajkomar et.al. (2018) who have earlier detected preference for accuracy (over parity) in the general healthcare context and the healthcare context involving algorithmic decision-making, respectively.

Throughout the analysis of the findings, special attention was paid to contextual factors. A five-point context model has been proposed as a result, and it includes (1) source of fairness principles, (2) sphere of application, (3) end recipient, (4) stakes, and (5) the resulting outcome. These are the most general criteria that can

help to map our different dimensions of context which can be highly complex in some instances. Going by such a classification in further work would help to map out everything that has been done and what remains to be done, while keeping the involved variables at a common denominator to avoid overlapping values as much as possible. Naturally, each of the above 5 points can be broken down into different subgroups, possibly using very different principles.

Another issue that this thesis has addressed is the practical aspect of information presentation throughout studies of this kind. A vast knowledge gap on the subject among the general population is a very pressing issue, and it impedes prospects of successful public consultations in the future. While so many experts make bold claims regarding the urgent need of increasing transparency and explainability, few address the practical barriers in the way of doing so.

The author proposes that shrinking this knowledge gap should be achieved through a combination of a long-term and short-term approaches. On the one hand, the pervasiveness of AI applications across a multitude of social domains calls for improved general education to ensure a due level of understanding of the related issues by the general education. Such a strategy, however, will require quite a long time before yielding results while people's lives are increasingly impacted by AI in general and algorithmic decision-making more specifically every day. This is why this long-term vision needs to be coupled with a short-term solution which could partially mitigate the concurrent issues.

So, on the other hand, studies such as the one presented in this thesis need to continue. While they would not fully solve the problem since they could only provide limited knowledge on the subject to relatively small groups of people (as opposed to the general population at large), they are nonetheless an important step in the right direction. Not only such studies would be contributing to the education of the public on the topics of AI and algorithmic decision-making (even if only covering a small part of the population), but they could also generate valuable feedback which could be used both for designing even better studies of

this kind moving forward, and for designing better educational curriculums for schools and universities on the topic of AI. Pinpointing which aspects are not clear or are especially difficult to understand, as well as which methods of information delivery different groups of people better respond to, is critical in the endeavor of improving the general knowledge on the subject and, therefore, in being able to implement AI systems into the social domain in an ethical fashion by collecting public opinion first.

8 References

Abu-Elyounes, D. (2020). Contextual fairness: legal and policy analysis of algorithmic fairness. University of Illinois Journal of Law, Technology & Policy, 2020(1), 1-54.

AIES (2019). Towards Algorithmic Definitions of Fairness. Proceedings of the 2019 AAAI/ACM, Conference on AI, Ethics, and Society 99 (2019), https://dl.acm.org/doi/10.1145/3306618.3314248.

Alzubi, J., Nayyar, A., Kumar, A. (2018). Machine Learning from Theory to Algorithms: An OverviewJ. Phys.: Conf. Ser. 1142 012012

Aristotle (1942) Nicomachaen Ethics, tr. W.D. Ross (Oxford: Oxford University Press)

Boddington, P. (2017). Towards a Code of Ethics for Artificial Intelligence. Cham:Springer

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. American Economic Review, 90(1), 166–193.

Bryman, A. (2012). Social Research Methods. 4th edition. Oxford: Oxford University press

Chouldechova, A., Roth, A. (2019). The Frontiers of Fairness in Machine Learning, ARXIV (Oct. 28, 2018) https://arxiv.org/abs/1810.08810

Coeckelbergh, M. (2020). AI Ethics. MIT Press

Colby, H., DeWitt, J., & Chapman, G. B. (2015). Grouping promotes equality: The effect of recipient grouping on allocation of limited medical resources. Psychological Science, 26, 1084–1089. doi:10.1177/0956797615583978

Corbett-Davies, S. (2018). Identifying bias in human and machine decisions (Order No. 28115054).

Corbett-Davies, S. et al., (2016). A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually not that Clear., WASH. POST (Oct. 17, 2016, 4:00 AM), https://www.washingtonpost.com/news/monkey-

cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysisis-more-cautious-than-propublicas/?utm-term=.ef3l9d030999.

ARXIV (2018). Review of Fair Machine Learning https://arxiv.org/pdf/1808.00023.pdf

T. H. Corman, C. E. Leiserson, R. L. Rivest, and C. Stein (2002). Introduction to Algorithms, Second Edition. McGraw-Hill.

Deutsch, M. (1975). Equity, Equality and Need: What Determines Which Value Will Be Used as the Basis of Distributive Justice? In Journal of Social Issues 31(3), 137-149.

Downey, L. (2022). What is an Algorithm? Investopedia

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. (2011). Fairness Through Awareness, ARXIV (Nov. 29, 2011), https://arxiv.org/pdf/1 104.3913.pdf

El Naqa, I. (2015). Machine Learning in Radiation Oncology: Theory and Applications

Encyclopedia Britannica: https://www.britannica.com/technology/artificial-intelligence

Fehr, E., Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. Quarterly Journal of Economics, 114(3), 817–868.

Frankenfield, J. (2021). Artificial Intelligence, Investopedia https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp#:~:text=Artificial%20intelligence%20(AI)%20refers%20to,as%20learning%20and%20problem%2Dsolving.

Frankish, K., Ramsey, W.M. (2014). Cambridge Handbook of Artificial Intelligence. Cambridge, UK : Cambridge University Press

Gordon-Hecker, T., Choshen-Hillel, S., Shalvi, S., Bereby-Meyer, Y. (2017). Resource allocation decisions: When do we sacrifice efficiency in the name of equity? In M. Li & D. P. Tracer (Eds.), Interdisciplinary perspectives on fairness, equity, and justice (pp. 93–105). Cham, Switzerland: Springer International.

Halevy, N., & Chou, E. Y. (2014). How decisions happen: Focal points and blind spots in interdependent decision making. Journal of Personality and Social Psychology, 106(3), 398–417.

Hao, K. (2019). AI is Sending People to Jail-and Getting it Wrong, MIT TECH. REV., https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/.

Hiller, J. S. (2021). Fairness in the eyes of the beholder: Ai; fairness; and alternative credit scoring. West Virginia Law Review, 123(3), 907-936.

Holm, S. (2022). The Fairness in Algorithmic Fairness. Res Publica, pp. 1356-4765

Hosanagar, K., Jair, V. (2018) We Need Transparency in Algorithms, But Too Much Can Backfire, HBR, https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire

IBM (2021): https://www.ibm.com/cloud/learn/machine-learning

Jarrahi M.H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making, in Business Horizons #61, 577-586

Kearns, M., Roth, A. (2019). The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, Inc., USA.

Khademi, A., Lee, S., Foley, D., Honavar, V. (2019). Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality, in PROCEEDINGS OF THE 2019 WORLD WIDE WEB CONFERENCE 2907.

Kizilcec, R.F. (2016) How Much Information? Effects of Transparency on Trust in an Algorithmic Interface, https://rene.kizilcec.com/wp-content/uploads/2016/01/kizilcec2016information.pdf

Köchling, A., Riazy, S., Wehner, M.C. et al. (2021). Highly Accurate, But Still Discriminatory. Bus Inf Syst Eng 63, 39–54. https://doi-org.ezproxy.ub.gu.se/10.1007/s12599-020-00673-w

Marsland, S. (2014). Machine Learning: An Algorithmic Perspective, Second Edition, CRC Press LLC

Li, M., Colby, H.A., Fernbach, P. (2019). Efficiency for Lives, Equality for Everything Else: How Allocation Preference Shifts Across Domains. In Social Psychological and Personality Science, 10(5), 697-707

Li,M., & DeWitt, J. (2017). Equality by principle, efficiency by practice: How policy description affects allocation preference. InM. Li & D. P. Tracer (Eds.), Interdisciplinary perspectives on fairness, equity, and justice (pp. 67–91). Cham, Switzerland: Springer International.

Li, M., Vietri, J., Galvani, A. P., & Chapman, G. B. (2010). How do people value life? Psychological Science, 21, 163–167. doi:10.1177/0956797609357707

Li, M., Tracer, D.P. (2017). Interdisciplinary Perspectives on Fairness, Equity, and Justice, Springer

Loewenstein, G. F., Thompson, L., & Bazerman, M. H. (1989). Social utility and decision making in interpersonal contexts. Journal of Personality and Social Psychology, 57(3), 426–441.

Mohri, M., Rostamizadeh, A., Talwalkar, A. (2018). Foundations of Machine Learning, Second Edition. The MIT Press.

Nemitz, P.F. (2018). Constitutional Democracy and Technology in the Age of Artificial Intelligence. In Philosophical Transactions of the Royal Society. Vol.376, Issue 2133 https://royalsocietypublishing.org/doi/10.1098/rsta.2018.0089

Okun, A.M. (1975). Equality and efficiency: The big tradeoff. Washington, DC: Brookings Institution Press

Ostrom, E. (1990) Governing the commons. Cambridge University Press

Peterson, J. (2018) Munk Debates https://www.youtube.com/watch?v=chP_xFklSjQ

Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med. 2018, 169(12), 866-872

Rawls, J. (2001). Justice as Fairness. Cambridge, Mass.: Belknap

Salovaara, A., Lyytinen, K. & Penttinen, E. (2019). High Reliability in Digital Organizing: Mindlessness, the Frame Problem, and Digital Operations. MIS Quarterly, [s. l.], v. 43, n. 2, p. 555–578, 2019. DOI 10.25300/MISQ/2019/14577.

Samuel, A.L. (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of research and development. 1959 Jul;3(3):210-29.

Samuel, S. (2022) Why it's so damn hard to make AI fair and unbiased, Vox https://www.vox.com/future-perfect/22916602/ai-bias-fairness-tradeoffs-artificial-intelligence

Saxena, N.A. (2019). How Do Fairness Definitions Fare? Examining Public Attitudes

Schlicker, N., Langer, M., Ötting, S.K., Baum, K., König, C.J., Wallach, D. (2021). What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. Computers in Human Behavior, Volume 122. ISSN 0747-5632, https://doi.org/10.1016/j.chb.2021.106837.

Schmid, F. (2022). Understanding the Importance of Algorithmic Fairness https://www.genre.com/knowledge/blog/understanding-the-importance-of-algorithmic-fairness-en.html

Schulz, J. F., Fischbacher, U., Thöni, C., & Utikal, V. (2014). Affect and fairness: Dictator games under cognitive load. Journal of Economic Psychology, 41, 77–87.

Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. (2019). Fairness and Abstraction in Sociotechnical Systems, in FAT* '19: PROCEEDINGS ON THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 59, 63(2019), https://dl.acm.org/doi/10.1145/3287560.3287598

Simons, T. (2020). Addressing Issues of Fairness and Bias in AI. THOMSON REUTERS (Nov. 30, 2020) https://blogs.thomsonreuters.com/answerson/ai-fairness-bias/.

Srivastava, M., Heidari, H., Krause, A. (2019). Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning, https://arxiv.org/abs/1902.04783

STOA (2019). Understanding Algorithmic Decision-Making: Opportunities and Challenges https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf

Tona, O., Leidner, D.E., Wixom, B., Someh, I. (2021). Make dignity core to employee data use. MIT Center for Information System Research. Access https://cisr.mit.edu/publication/2021_0601_DignityDataUse_TonaLeidnerWixomSomeh

Ubel, P., Baron, J., & Asch, D. A. (2001). Preference for equity as a framing effect. Medical Decision Making, 21, 180–189. doi:10.1177/0272989x0102100303

Ubel, P. A., DeKay, M. L., Baron, J., & Asch, D. A. (1996). Cost-effectiveness analysis in a setting of budget constraints—Is it equitable? New England Journal of Medicine, 334, 1174–1177. doi:10.1056/NEJM199605023341807

Upadhyay, S. (2022). What Is An Algorithm? Characteristics, Types and How to Write It.

Van Lange, P. A. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. Journal of Personality and Social Psychology, 77(2), 337–349.

Van Lange, P. A., De Bruin, E., Otten, W., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. Journal of Personality and Social Psychology, 73(4), 733–746.

Velasquez, M., Andre, C., Shanks, T., Meyer, M.J. (1990). Justice and Fairness. Issues in Ethics, 3(2)

Verma, S., Rubin, J. (2018) Fairness Definitions Explained. In FairWare'18: IEEE/ACM International Workshop on Software Fairness, May 29, 2018,

Gothenburg, Sweden. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3194770.3194776

Walster, E., Berscheid, E., & Walster, G. W. (1973). New directions in equity research. Journal of Personality and Social Psychology, 25(2), 151–176.

Wick, M., Panda, S., Tirstan, J.B. (2019). Unlocking Fairness: A Trade-Off Revisited, 33RD CONF. NEURAL INFO. PROC. Sys., https://papers.nips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf.

Wischmeyer, T., Rademacher, T. (2020). Regulating Artificial Intelligence

Yampolskiy, R.V. (2019). Artificial Intelligence Safety and Security. Chapman and Hall/CRC, 2019. APA 7th Edition (American Psychological Assoc.)

Yanofsky, N.S. (2011). Towards a Definition of an Algorithm

Zehlike, M., Hacker, P., Weidemann, E. (2020). Matching Code and Law: Achieving Algorithmic Fairness with Optimal Transport, 34 DATA MINING & KNOWLEDGE DISCOVERY 163 (2020).

Zuboff, S. (2015). Big other: Surveillance Capitalism and the Prospects of an Information Civilization. Journal of Information Technology. 30(1):75-89. doi:10.1057/jit.2015.5