Different Fairness Perceptions in Different Fairness problems: Algorithmic Decisions and Strategic Interactions

BY
Faezeh Ataeizadeh

A THESIS SUBMITTED TO
The Faculty of Graduate and Postdoctoral Affairs
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Master of
Human-Computer Interaction

Carleton University
Ottawa, Ontario, Canada

© 2023
Faezeh Ataeizadeh
TITLE: Different Fairness Perceptions in Different Fairness problems: Algorithmic Decisions and Strategic Interactions

AUTHOR: Faezeh Ataeizadeh
B.A., (Psychology)

SUPERVISOR: Dr. Alan Tsang

NUMBER OF PAGES: xi, 103
Abstract

This thesis is divided into two parts, both exploring changes in fairness perception in response to a fairness problem. As algorithms become more integrated into our lives, it is crucial to understand how their decisions are perceived by those affected by them. Therefore, the first component of this thesis compares perceptions of fairness and trust in decisions made by humans versus algorithms.

In the second part, we expand our focus to the societal level, utilizing evolutionary game theory to study how fairness evolves within a society where different perceptions of fairness coexist. Insights drawn from this part could potentially inform the design of more equitable and trustworthy Human-Computer Interaction (HCI) systems.
Acknowledgements

First of all, I would like to thank my supervisor, Dr. Alan Tsang, who has not only been an invaluable source of guidance but also a wonderful person to work with throughout the entire research process. Thank you for giving me the freedom to explore, but also keeping me on track during my research.

I would also like to express my gratitude to my committee members, Dr. Elizabeth Stobert and Dr. Kasia Muldner. I am truly grateful for your expertise, feedback, and support.

I would like to extend a heartfelt thank you to my friend, Hooman Gheshlaghi, who assisted me by being the confederate in the first component of this thesis. I am grateful for your help and support.

I am also grateful to the HCI and Computer Science faculty and all the talented individuals whom I had the honor of learning from.

Finally, my family and my partner have been unconditionally supportive throughout this journey, and I wish to acknowledge their support as well. Thank you for believing in me.

Note: I would like to mention that I used AI-powered tools (Wordtune and ChatGPT) to check possible grammatical, spelling, and punctuation errors, during editing and proofreading this thesis.
Dedication

For mom and dad.
Contents

Abstract iii
Acknowledgements iv
Dedication v
List of Figures ix
List of Tables xi

1 Introduction 1
  1.1 What is fairness? 1
  1.2 Thesis outline and goals 2

2 Human Perception of Unfairness & Trust in Algorithmic vs. Human
  Decisions 5
  2.1 Introduction 5
  2.2 Literature Review 6
    2.2.1 What an algorithm is 6
    2.2.2 Concepts of algorithmic fairness 6
    2.2.3 Distributive fairness 9
    2.2.4 Knowledge types and their impact on fairness perceptions 11
## Fair Play or Foul: Exploring Fairness Perceptions and Strategy Using Nash Bargaining Game

### 3.1 Introduction

39

### 3.2 Literature review

40

#### 3.2.1 Game theory’s role in moral philosophy

40

#### 3.2.2 Evolutionary game theory

43

#### 3.2.3 The evolution of fairness

44

#### 3.2.4 Fair division and different fairness perceptions

45

#### 3.2.5 Nash bargaining game

47

#### 3.2.6 Games and fairness

49

### 3.3 Methodology

51

#### 3.3.1 Model description

51
# List of Figures

3.1 Agents interaction diagram ........................................ 56
3.2 Agents strategy adaptation over time .......................... 59
3.3 Agents strategy adaptation after introducing fair players. All the agents played the fair player’s strategy at least once. .............. 60
3.4 Distribution of agents’ fairness perception after introducing fair players, with standard deviation indicated. ......................... 61
3.5 Moderate players’ proportions with overlapping claims .......... 62
3.6 Low narcissistic agents’ proportions with overlapping claims .. 63
3.7 Moderate players’ proportions with non-overlapping claims .... 64
3.8 Low narcissistic agents’ proportions with non-overlapping claims .. 64
3.9 Moderate players’ proportions with non-overlapping and truncated normal distribution claims .............................. 66
3.10 Low narcissistic agents’ proportions with non-overlapping and truncated normal distribution claims ......................... 67
3.11 Moderate players’ proportions with poisson distribution and non-overlapping claims ............................................. 68
3.12 Low narcissistic agents’ proportions with poisson distribution and non-overlapping claims ............................................. 68
3.13 Moderate players’ proportions with BA network ................. 69
3.14 Low narcissistic agents’ proportions with BA network .......... 70
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.15</td>
<td>Initialization of a BA network with hubs as highly narcissistic agents</td>
</tr>
<tr>
<td>3.16</td>
<td>Mean proportion of each fairness perception after 100 runs with high narcissistic hubs (standard deviations indicated)</td>
</tr>
<tr>
<td>3.17</td>
<td>Initialization of a BA network with hubs as moderate players</td>
</tr>
<tr>
<td>3.18</td>
<td>Mean proportion of each fairness perception after 100 runs with moderate players hubs (standard deviations indicated)</td>
</tr>
<tr>
<td>3.19</td>
<td>Caveman graph configuration</td>
</tr>
<tr>
<td>3.20</td>
<td>Mean proportion of each fairness perception after 100 runs (standard deviations indicated)</td>
</tr>
<tr>
<td>3.21</td>
<td>Changes in fairness perception proportions over time (single run)</td>
</tr>
<tr>
<td>3.22</td>
<td>Caveman graph initialization with high and low narcissistic agents only</td>
</tr>
<tr>
<td>3.23</td>
<td>Caveman graph with low and high narcissistic agents after a single run</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Summary of prior comparative studies involving human subjects. The ‘Term’ refers to the specific terminology employed in the scenario presented to participants. ‘Stakes’ represents the potential outcomes involved in the decision-making process; note that except for studies marked with an asterisk in the ‘Stakes’ column (*), all stakes discussed in these studies were hypothetical. The asterisked studies presented actual tasks with real stakes to the participants. 19

2.2 Summary of key findings and their implications 36

3.1 Payoff matrix for the Prisoner’s Dilemma game. 42
Chapter 1

Introduction

1.1 What is fairness?

Fairness is a fluid concept, and it can be defined differently based on different situations\(^1\). In some contexts, fairness refers to an equal distribution, such as dividing a cake between children. However, it may not be fair to distribute a resource evenly in other situations; for instance, distributing bonuses at work may not feel fair since different individuals may have put in different efforts to earn the bonus. Fairness is therefore a dynamic concept that adjusts according to the context in which we find ourselves\(^2\).

The concept of fairness can become even more complex when we consider who or what is making the decision. One question is whether our perception of fairness changes based on whether a decision is made by a human or an algorithm. The human touch brings a layer of empathy and understanding into the equation, which could change our perception of fairness\(^3,4\). There is a certain level of comfort and familiarity in dealing with another human, someone who can understand the subtleties of our shared human experience.
Yet, when we think about an algorithm making the decision, a different set of factors comes into play. An algorithm, which is not affected by personal biases or emotions, can be used as an impartial judge. Its decisions which are based on pre-defined data and criteria, could be perceived as more fair compared to human decisions, because of this lack of personal bias$^{4,5}$.

Moreover, depending on the stakes or outcomes associated with a particular situation, we might also change our strategy of acting self-interested or not. For example, in financial or economic scenarios, people might naturally want a large profit in the beginning. However, sometimes in order to secure a larger payoff in the long term, we might need to forgo immediate, higher monetary gains in favor of making a fairer decision. For example, an online shopping platform that uses dark patterns$^6$ to deceive its users to buy more products might get more profit in the short term. However, in the long run, customers may become disillusioned with these manipulative practices and leave the platform. In this example, being more fair and transparent about pricing and not using manipulative tactics could actually be more beneficial in the long run.

It is good to think about: how do we define such actions? Are they fair because they lead to fairer outcomes? Or are they unfair because they were initially driven by self-interest? This complexity adds depth to our understanding of fairness, and with these complexities in mind, we aim to examine how humans or agents react to perceived unfairness in different fairness problems.

1.2 Thesis outline and goals

This thesis is an exploration of the concept of fairness, a complex construct that assumes different meanings and interpretations in various contexts. In the first component, we will explore the dichotomy between human and algorithmic judgment. As
we live in a digital era, algorithms pervade almost every aspect of our lives; designing our experiences, shaping our interactions, and influencing our decisions. Due to this ubiquity of algorithmic influence, it is important to understand the differences between human perceptions of fairness/unfairness in the decisions made by algorithms versus humans.

Thus, in the first component, we aim to address the following research questions: When algorithms make decisions, how do we, as individuals, perceive fairness? Do we find it easier to accept an unfair outcome when it is decided by an algorithm, or when a human makes the decision? And in case of previous unfair treatment by humans and algorithms in an identical situation, in which one would we have more trust? The answers to these questions can help us designing more ethically responsible and user-friendly algorithmic systems that respect our collective sense of fairness.

In the second component, we broaden the scope of exploring fairness perception to a societal level, examining it within a game-theoretic framework. This aspect focuses on the dynamics of fairness in interactive scenarios which are involving strategic decision-making. Strategic decision-making is a process of negotiation, where multiple entities or stakeholders with different interests attempt to reach an agreement with or without communicating with each other\(^7,8\). This process is a foundation for many interactions within HCI\(^1\) systems, and it is here that fairness becomes a critical consideration.

For example, the recent rise of remote work and online education have increased reliance on collaborative platforms for resource sharing and task coordination. Thus, the division of resources in collaborative platforms needs to be carried out with fairness to maintain teamwork, cooperation, and productivity. If resources are perceived

\(^1\)HCI stands for Human-Computer Interaction, a field in computer science that focuses on the analysis and design of the interaction between people and computers, examining the ways in which computers are, or are not, developed for effective engagement with human users\(^9\).
as being unfairly distributed, it can lead to decreased motivation, and reduced productivity.

In addition, a part of our study looks into how our perception of fairness can change based on our level of narcissism. Narcissism, as a personality trait, involves a sense of entitlement and superiority, and in certain scenarios, individuals with high narcissistic traits may perceive themselves as more deserving of resources. This individual characteristic can therefore affect one’s fairness perceptions and the subsequent actions they take in a bargaining situation. In the second component, we examine how fairness evolves in a society that individuals have different fairness perceptions (which are based on different levels of narcissism) and how these fairness perceptions might change according to the outcomes of individuals’ interactions.

The sequential approach that we took in this thesis allows us to examine fairness from both individual and societal standpoints. By bridging algorithmic decision-making and strategic interactions, our aim is to clarify the complex nature of fairness perceptions. We are interested in how these perceptions impact, and are influenced by, the settings in which they arise. Our ultimate goal is to enhance our collective understanding of the dynamics of fairness in these contexts. Doing so can help us compare fairness perceptions across various domains and tasks, thereby enriching our understanding of fairness in diverse situations.
Chapter 2

Human Perception of Unfairness & Trust in Algorithmic vs. Human Decisions

2.1 Introduction

As algorithms increasingly intersect with our daily lives, shaping crucial choices in areas ranging from healthcare to justice, their adoption depends on public trust and perceived fairness\textsuperscript{12}. But do humans perceive unfair decisions made by algorithms differently than identical decisions made by humans? If humans and algorithms deliver identical unfair treatment, in which one do people place more trust? Therefore, in the first component of our study we are focusing on addressing these critical questions. Our goal is to explore the differences in how humans perceive decisions made by an algorithm versus a human under identical unfair scenarios, and how this shapes their trust. This is important because in order to develop more fair, trustworthy, and therefore widely accepted algorithmic systems, we must understand human attitudes
toward algorithmic decisions\textsuperscript{13,14}. It is about aligning our digital future with our human instincts of fairness and trust.

In the following sections, we will first provide a comprehensive review of the relevant literature and then delve into the specifics of our research methodology and key findings.

\section{2.2 Literature Review}

\subsection{2.2.1 What an algorithm is}

According to Hill\textsuperscript{15}, an algorithm is a “finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions.” In simple terms, an algorithm is like a well-structured manual that provides clear and effective instructions to solve a particular problem or achieve a specific goal within certain conditions. Algorithms provide the logic and structure for software development, while software is the practical, operational result of implementing those algorithms. In the context of our experiment, an algorithm is an entity capable of making decisions based on the data it has, and despite being programmed by a human, it runs independently.

\subsection{2.2.2 Concepts of algorithmic fairness}

Generally speaking, algorithmic fairness means that algorithm-based decisions should not lead to unjust, discriminatory, or disparate results\textsuperscript{16}. Binns\textsuperscript{17,18} provided insight into this construct by linking it to egalitarianism, a philosophy grounded in the notion of equal treatment and often equal distribution of resources for all, regardless of their individual circumstances.

He explored how egalitarian principles can provide insight into when and why
algorithmic systems may be viewed as unfair, irrespective of whether such unfairness should technically be classified as discrimination. His work emphasized that to truly understand and evaluate algorithmic fairness, it is not enough to say an algorithm is fair just because it produces equal results across different groups. Instead, Binns asserted that we must explore the processes that lead to the outcome, even in the face of seemingly equal outcomes.

An approach to assess algorithmic fairness can be seen in studies in the field of organizational justice. As per Greenberg\(^1\), these studies identify four key dimensions of fairness: distributive, procedural, informational, and interpersonal.

Distributive fairness refers to the equitable allocation of resources, based on principles such as equality, equity, or need\(^2\). It looks at whether benefits and resources are allocated fairly amongst all parties involved.

Procedural fairness, on the other hand, focuses on the decision-making process itself. It considers whether the methods used in making decisions are consistent, transparent, and based on fair criteria\(^1\). Fair criteria may involve ensuring everyone involved in a decision has an equal voice, the process is free from bias or favoritism, and the rules are applied consistently across all individuals and situations.

Informational fairness emphasizes the transparency of algorithmic decision-making systems\(^1\). This entails ensuring that individuals affected by these systems have complete and understandable information about how decisions are made and the logic behind them.

Finally, interpersonal fairness is concerned with how an algorithmic decision-making system interacts with users, particularly in relation to privacy rights and the use of protected data\(^1\). The focus here is on respectful and ethical engagement with users, including attention to confidentiality and respect for individual autonomy.
For algorithmic decisions to be fair, the social context in which they are used should be considered\textsuperscript{16}. This is because these decisions impact human lives, and therefore, studying algorithmic fairness should be underpinned by a human-centric approach. To put it another way, the effectiveness and fairness of algorithmic decision-making should be measured not only by their technical power but also by the experiences, perceptions, and the overall wellbeing of the humans they influence\textsuperscript{21,22}.

This has led to a shift in research focus, with studies examining algorithmic fairness through the lens of human perception\textsuperscript{4,5,23,24}. These studies examine how different people perceive and respond to decisions made by algorithms, how these perceptions influence trust in such systems, and the socio-cultural factors that shape these perceptions. These studies cover a variety of domains and applications, reflecting the pervasive role that algorithms have come to play in our lives. By looking into how humans perceive algorithmic fairness, researchers hope to identify ways in which algorithms can be better designed, regulated, and deployed to respect human values.

While there have been significant advancements in our understanding of algorithmic fairness, the field still struggles with an essential challenge: there is no universally accepted definition of fairness (or unfairness)\textsuperscript{25}. This challenge stems from the complex nature of fairness as a concept. It can vary greatly across different cultures, societies, and individuals. As stated by Hidalgo\textsuperscript{24}, “The world is unfair—not only because people and machines are biased—but because it affords multiple ways of defining a fair outcome”. Therefore, formulating a universally accepted definition is inherently complex and potentially impossible since preferences for different fairness norms are highly context dependent.
2.2.3 Distributive fairness

According to Starke et al.'s comprehensive review of existing literature in the field of human perception of algorithmic decision making, there is a significant emphasis on distributive fairness, which concerns the fairness of decision outcomes. As it was mentioned earlier, this dimension of fairness refers to the unbiased allocation of resources or opportunities, ensuring that no particular group is systematically disadvantaged by an algorithm’s decisions.

According to Deutsch, justice can be broadly perceived as the fair and equitable distribution of conditions and goods that influence individual well-being. This concept of well-being includes psychological, physiological, economic, and social dimensions. Deutsch illuminates that perceived injustice, or unfairness, can arise at different stages of distribution process, which he discusses using a thoughtful example. He discusses a scenario involving a teacher who is responsible for grading students in a classroom. This example vividly demonstrates how the principle of distributive fairness operates and where potential disagreements can arise.

In his illustration of the ‘injustice of values’, a teacher is tasked with assigning grades based on different standards like a student’s final performance, their improvement throughout the year, their effort, innate aptitude, or even their need for good grades. If the teacher decides to allocate grades according to the students’ effort, this principle itself might invite perceptions of injustice—not only from those students who find themselves at a disadvantage but also from those who believe different values should guide the grading process.

The ‘injustice of rules’ occurs when there is a general agreement among the students and the teacher on the guiding principle for grading. Here, the disagreement may stem from how the chosen principle is implemented. If effort is the decided basis for grading, how is it quantified? The parameters could include the length of a term
paper, the number of references used, or even the stress exerted in paper preparation. For instance, a student who prepared a shorter but research-intensive paper may feel slighted if a student who prepared a longer paper with less preparation receives a higher grade.

Thirdly, ‘injustice of implementation’ emerges when the execution of the agreed rules deviates from what was originally intended. To illustrate, if effort is determined to be the measuring stick, then the teacher might decide to assess this based on the time spent by each student on a project. However, if the teacher unknowingly privileges students who had parental assistance, it leads to ‘implementation injustice’. Such a situation might occur when parents have been overly involved in a project, thus the final product may not entirely reflect the student’s individual effort. In this case, despite an agreement on the grading principle and the rules that operationalize it, unfairness can still arise from the way these rules are implemented.

Lastly, the ‘injustice of decision-making procedures’ occurs when stakeholders perceive that the process of determining the values, rules, and their implementation is flawed. For instance, if the teacher alone decided that effort should be the grading principle without consulting the students, it might be deemed unfair. Even if this principle seems to embody fairness, the lack of collective decision-making could still generate perceptions of procedural injustice.

In our research, we are specifically targeting participants’ perceptions of distributive and procedural unfairness, with a particular emphasis on the ‘injustice of values’. We aim to delve into how these feelings of injustice emerge and influence perceptions of both algorithmic and human decision-making processes.
2.2.4 Knowledge types and their impact on fairness perceptions

The work of Sundar and Nass\(^{14}\) offers insights into how people perceive and trust different sources of information, including algorithmic ones. They observed that attitudes towards an information source can shape how individuals assess the credibility and quality of the information provided by that source. Therefore, the underlying trust or skepticism we hold towards an information source can form our perception of the information it distributes.

Literature from cognitive psychology\(^{26,27}\) also delves into the complexities of knowledge and its transferability, offering a valuable framework for understanding the perceived fairness of algorithmic versus human decisions. They differentiate between explicit knowledge - that which can be coded into algorithms - and procedural and tacit knowledge, which cannot be easily coded into algorithms\(^{23}\).

Procedural knowledge refers to the practical knowledge or skills that individuals acquire through experience, such as how to ride a bike or play a musical instrument. Tacit knowledge, on the other hand, refers to the kind of knowledge that is difficult to transfer to another person by means of writing it down or verbalizing it. This could include subjective insights or intuitions, which are deeply personal and hard to formalize\(^{28}\).

It is notable that in algorithms, a function or procedure is a set of coded instructions that the computer can execute. They are explicitly defined and are used to perform a specific task. These tasks are deterministic, meaning they will produce the same output when given the same input\(^{29}\). However, procedural and tacit knowledge, as it was mentioned before, are often difficult to transfer because they are based on context, experience, and learning\(^{28,30}\).
This exploration into the nature of knowledge raises further questions about the perceived fairness and trustworthiness of algorithmic decisions. If algorithms struggle to embody the tacit and procedural knowledge that is often critical in decision-making processes, can they ever be perceived as fair or trustworthy as human decision-makers in situations requiring these forms of knowledge? Or will they always be viewed with a degree of skepticism in such contexts, no matter how advanced their design?

2.2.5 Previous comparative studies

In recent years, a surge of interest has been observed in the Human-Computer Interaction (HCI) research community that examine fairness perceptions from the perspective of those who are affected by algorithmic decisions (including designers who build these systems, end-users who interact with them, decision subjects who are directly impacted by algorithm decisions, and other parties involved). In this section, we will present a comprehensive review of comparative studies investigating the differences between human perceptions of fairness and trust in decisions made by algorithms versus those made by humans. A summary of these works is provided in Table 2.1.

Lee\(^3\) conducted an online between-subject experiment revolving around four different types of managerial decisions. These decisions were categorized based on whether they required ‘human skills’, such as intuition and empathy, or ‘mechanical skills’, which are quantifiable and easier to code into an algorithm.

They found that when it came to decisions that relied on mechanical skills, participants judged the fairness of human-made decisions and algorithm-made decisions as roughly the same. However, when the tasks required human skills, for example the allocation of donations to non-profit organizations, participants perceived human decisions to be fairer than those made by an algorithm. This finding points to a bias in favor of humans when dealing with tasks that involve softer, more subjective skills,
possibly due to the belief that humans possess a level of understanding and empathy that algorithms cannot replicate.

The study also shed light on the factors that shaped participants’ fairness judgments. For tasks assigned by a human manager, participants based their fairness assessments on the perceived authority of the manager or the duties of the employee. On the other hand, when a task was assigned by an algorithm, participants focused more on the characteristics of the algorithm itself, particularly its impartiality and efficiency. This suggests that people appreciate the objectivity that algorithms bring to decision-making processes.

Trust emerged as another key factor influencing participants’ perceptions. Similar to the fairness findings, trust levels were roughly equal for human and algorithmic decisions involving mechanical skills. However, for tasks that required human skills, participants exhibited a higher level of trust in human-made decisions. This finding indicates that despite the growing prevalence of algorithms in decision-making, people still lean towards trusting the judgment of humans in situations that call for a personal touch.

In Hidalgo’s research, an interesting perspective on the human versus machine perception was illuminated. The study showed that humans were judged based on their intentions and considered more responsible for the consequences of their actions, when compared to machines. This greater attribution of intent and accountability to humans meant that they were subject to more critical judgement, whether in positive or negative situations.

For instance, in positive scenarios, when humans corrected biased decisions, they received more credit for their actions compared to machines. Conversely, in negative scenarios, when humans made biased decisions, they received more criticism than machines. Interestingly, despite this harsher judgement, people still exhibited
a preference for human intervention in decision-making processes, even when those decisions had been proven to be flawed. The fact that human-led decision-making is preferred, even in the face of fallibility, is known as algorithm aversion\textsuperscript{31}.

On the other hand, the research conducted by Schoeffer et al.\textsuperscript{5} presented a contrasting perspective. Their study found that automated decisions were typically seen as fairer when compared to those made by humans. There was no significant difference, however, in the perceived trustworthiness of decisions made by humans or automated systems.

Their findings highlighted that people appreciated the objectivity and data-driven nature of automated decision systems (ADS). This appreciation could be tied to the perception that automated systems are devoid of personal biases that humans might have, leading to potentially fairer outcomes.

The study also revealed a link between the level of AI literacy and the perception of fairness and trustworthiness of ADS. Individuals with a higher understanding of AI, or ‘AI literacy’, tended to view ADS as both more equitable and more trustworthy. This implies that with an increased understanding of how AI systems work, people might be more likely to trust and perceive them as fair.

The research by Castelo et al.\textsuperscript{4} sheds light on an interesting divergence in people’s preferences based on the nature of decisions. According to their findings, when it comes to decisions perceived as objective—those that rely heavily on data analysis, numerical comparisons, and algorithmic precision—people tend to favor advice given by automated systems. This likely stems from the recognition of the computational power and unbiased nature of algorithmic systems when dealing with straightforward, data-intensive tasks.

However, when decisions are more subjective—those that require a level of in-
tuition, empathy, or understanding of complex human emotions and dynamics—the preference shifts towards human advice. This reflects the natural human capacity for empathy, intuition, and understanding of social contexts, traits that are currently challenging for AI to replicate fully.

Castelo et al. propose an insightful implication of these findings. They suggest that increasing people’s awareness of the nature of decision-making—whether it is primarily objective or subjective—and considering their perception of AI systems could boost acceptance, particularly in situations where automated decision systems (ADS) have the potential to make decisions that are fairer than those made by humans. This involves transparently disclosing the role of ADS in decision-making and ensuring the systems are designed and implemented with an understanding of user perceptions and expectations.

In summary, as it is shown in Table 2.1, studies examining the perception of fairness in algorithmic and human decision-making can be categorized according to various contexts, but work-related decisions are the most common.

In the context of hiring decisions, Pethig and Kroenung\textsuperscript{32} found that women, particularly those unemployed, prefer algorithmic over male evaluations when they perceive gender bias could impact the outcome, suggesting that algorithmic objectivity plays a key role in women’s preference for algorithmic decisions in stereotype-prone domains.

A variety of studies explored work-related decisions, including those by Lee\textsuperscript{3}, Schlicker et al.\textsuperscript{33}, and Bankins et al.\textsuperscript{34}. These studies found different results; Lee found a distinction between perceptions of fairness in mechanical versus human tasks, with human decisions perceived as fairer in human tasks. Schlicker et al.\textsuperscript{33} also concluded that human decisions were perceived as fairer. Conversely, Bankins et al.\textsuperscript{34} found that when a decision was positive, human decisions were perceived as fairer, but when a
decision was negative, there was no significant difference in perceptions of fairness.

In the context of high impact decisions (particularly health and organizational justice), studies by Harrison et al.\textsuperscript{35}, and Noble et al.\textsuperscript{36} reported that human decisions were perceived as fairer. Whereas, Araujo et al.\textsuperscript{13} found that automated decision-making was perceived as fairer in health related contexts.

Lastly, there were a few studies that provided real stakes and tasks to their participants. For example, Bai et al.\textsuperscript{37} found that that workers perceive algorithmic task assignments as fairer than human-based assignments, resulting in a 15.56%-17.86% increase in efficiency, suggesting the positive potential of integrating algorithmic processes in operations for improved fairness perceptions and productivity.

The contrasting findings across the papers that are mentioned in Table 2.1, highlight the complex nature of fairness perceptions in relation to human versus algorithmic decision-making. It underscores the reality that these perceptions are highly influenced by the context in which decisions are made, making it challenging to make generalizations about the perceived fairness of human decision-making versus algorithmic decision-making.

The context-dependent nature of these perceptions could stem from a variety of factors. These include the type of decision being made, the stakes involved, the cultural or societal norms of the individuals assessing fairness, and their level of familiarity or comfort with algorithmic processes, among others. For instance, in scenarios where decisions are data-heavy and require high precision, people may view algorithmic decision-making as fairer. On the other hand, situations that need empathy, intuition, or understanding of complex social dynamics, human decision-making might be perceived as fairer.

This complexity and variability highlights the need for more comprehensive and
<table>
<thead>
<tr>
<th>Author</th>
<th>Term</th>
<th>Stakes</th>
<th>Perception of Fairness</th>
<th>Perception of Trust</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Plane et al., 2017&lt;sup&gt;38&lt;/sup&gt;</td>
<td>Algorithm</td>
<td>Discrimination in targeted advertising</td>
<td>There was no difference between perception of fairness</td>
<td>No mention</td>
</tr>
<tr>
<td>Lee &amp; Baykal, 2017&lt;sup&gt;23&lt;/sup&gt;</td>
<td>Algorithm</td>
<td>*Social decisions</td>
<td>Human decisions perceived as more fair</td>
<td>No mention</td>
</tr>
<tr>
<td>Wang, 2018&lt;sup&gt;39&lt;/sup&gt;</td>
<td>Algorithm</td>
<td>Criminal justice decision</td>
<td>Human decision perceived as more fair</td>
<td>No mention</td>
</tr>
<tr>
<td>Lee, 2018&lt;sup&gt;8&lt;/sup&gt;</td>
<td>Algorithm</td>
<td>Work related decisions</td>
<td>With mechanical tasks the perception was the same and with human tasks, human decisions were perceived as more fair</td>
<td>With mechanical tasks the perception was the same and with the human tasks human decisions were perceived as more trustworthy</td>
</tr>
<tr>
<td>Acikgoz et al., 2019&lt;sup&gt;40&lt;/sup&gt;</td>
<td>AI</td>
<td>Selection decisions in organizations</td>
<td>AI perceived as more just</td>
<td>No mention</td>
</tr>
<tr>
<td>Araujo, et al., 2019&lt;sup&gt;13&lt;/sup&gt;</td>
<td>Automated decision making</td>
<td>High impact decisions related to justice and health</td>
<td>ADM perceived as more fair</td>
<td>No mention</td>
</tr>
<tr>
<td>Suen et al., 2019&lt;sup&gt;41&lt;/sup&gt;</td>
<td>AI Algorithm</td>
<td>*Assessing interviews</td>
<td>No difference</td>
<td>No mention</td>
</tr>
<tr>
<td>Bai et al., 2020&lt;sup&gt;37&lt;/sup&gt;</td>
<td>Algorithm</td>
<td>*Management decisions</td>
<td>Workers perceived the algorithmic assignment process as more fair</td>
<td>No mention</td>
</tr>
<tr>
<td>Harrison et al., 2020&lt;sup&gt;35&lt;/sup&gt;</td>
<td>Machine learning model</td>
<td>Criminal justice decisions</td>
<td>Human judge perceived as more fair</td>
<td>No mention</td>
</tr>
<tr>
<td>Helberger et al. 2020&lt;sup&gt;42&lt;/sup&gt;</td>
<td>AI / Computer</td>
<td>No context. They just asked participants who do you think is fairer and why.</td>
<td>AI is more fair</td>
<td>No mention</td>
</tr>
<tr>
<td>Marcinkowski et al., 2020&lt;sup&gt;43&lt;/sup&gt;</td>
<td>Algorithmic decision making</td>
<td>Higher education admission</td>
<td>ADM perceived as more fair</td>
<td>No mention</td>
</tr>
<tr>
<td>Bai et al., 2020&lt;sup&gt;37&lt;/sup&gt;</td>
<td>Algorithm</td>
<td>*Management decisions</td>
<td>Workers perceived the algorithmic assignment process as more fair</td>
<td>No mention</td>
</tr>
<tr>
<td>Reference</td>
<td>Type</td>
<td>Domain</td>
<td>Fairness perception</td>
<td>Trustworthiness perception</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------------------</td>
<td>---------------------------------------------</td>
<td>---------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Newman et al., 2020</td>
<td>Algorithm</td>
<td>Human resource decisions</td>
<td>Human decisions</td>
<td>fair</td>
</tr>
<tr>
<td>Wonseok et al., 2021</td>
<td>Robot umpire</td>
<td>Sport related decisions</td>
<td>Human decisions</td>
<td>fair</td>
</tr>
<tr>
<td>Noble et al., 2021</td>
<td>Algorithm</td>
<td>Organizational justice decisions</td>
<td>Human decisions</td>
<td>fair</td>
</tr>
<tr>
<td>Schoeffer et al., 2021</td>
<td>Automated decision systems</td>
<td>Loan approval decisions</td>
<td>Automated decisions</td>
<td>fair</td>
</tr>
<tr>
<td>Schlicker et al., 2021</td>
<td>Automated decision making</td>
<td>Work related decisions</td>
<td>Human decisions</td>
<td>fair</td>
</tr>
<tr>
<td>Miller &amp; Keiser, 2021</td>
<td>Automated decision making</td>
<td>Representative bureaucracy</td>
<td>Red light camera (ADM) perceived as more fair</td>
<td>No mention</td>
</tr>
<tr>
<td>Lee &amp; Rich, 2021</td>
<td>AI health care</td>
<td>Health care decisions</td>
<td>Participants who mistrust human healthcare providers perceived AI healthcare as equally fair as human medical providers</td>
<td>Participants who mistrust human healthcare providers perceived AI healthcare as equally trustworthy as human medical providers</td>
</tr>
<tr>
<td>Hidalgo, 2021</td>
<td>AI</td>
<td>Hiring, admissions, and promotion decisions</td>
<td>Humans were judged more strongly</td>
<td>No mention</td>
</tr>
<tr>
<td>Nagtegaal, 2021</td>
<td>Algorithm</td>
<td>Public management decisions</td>
<td>Algorithm perceived as more fair for tasks that are low in complexity and human decisions were perceived as more fair in tasks that are high in complexity</td>
<td>No mention</td>
</tr>
<tr>
<td>Bankins et al., 2022</td>
<td>AI Algorithm</td>
<td>Human resource management decisions</td>
<td>When a decision is positive, human decisions perceived as more fair and when a decision is negative, there is no significant difference in fairness perceptions</td>
<td>For both positive and negative decisions, trust scores are higher when the decision is made by a human</td>
</tr>
</tbody>
</table>
Participants prefer a decision-making system where humans and algorithms collaborate, specifically in a 60% (human) - 40% (algorithm) partnership. This balance perceived to be the fairest approach.

Women prefer algorithmic decisions over human decisions.

No significant difference in the perception of unfairness. Human perceived as more trustworthy.

Table 2.1: **Summary of prior comparative studies involving human subjects.**

The ‘Term’ refers to the specific terminology employed in the scenario presented to participants. ‘Stakes’ represents the potential outcomes involved in the decision-making process; note that except for studies marked with an asterisk in the ‘Stakes’ column (*), all stakes discussed in these studies were hypothetical. The asterisked studies presented actual tasks with real stakes to the participants.

Diverse research in this field. There is a clear need for systematic comparisons of fairness perceptions across a wide range of domains and tasks. Such research would provide a better understanding of when and why people perceive algorithmic or human decision-making as fair or unfair, thereby enabling the design of decision-making processes that are not only fair but are also perceived as such by those they impact. This is crucial in an era where algorithmic decision-making is becoming more prevalent and influential across various sectors of society.

Additionally, a noticeable gap in the existing literature is the lack of active participant involvement in the tasks within these studies. As it is shown in the Table 2.1, which offers a comprehensive overview of the most recent research from 2017 to 2023, the majority of previous studies, except for three, provided hypothetical scenarios with fictitious stakes, limiting the participant’s engagement with the task.
Active participation often comes with emotional involvement, which can significantly influence fairness perception\textsuperscript{50}. Our research aims to bridge this gap by involving participants more directly in the experimental process.

2.3 Study Design

2.3.1 Participants

Selection criteria

All participants were Carleton University students aged 18 and above. They were also required to be able to communicate in English to ensure clear communication during the experiment.

Recruitment process

The study was advertised on Carleton University’s billboards across different faculties in high-traffic areas to ensure that it reached a diverse audience, as well as through email invitations facilitated by the School of Computer Science’s administrative assistant.

The final sample size

From all participants that expressed interest, 14 of them were present for the experiment, which is considered appropriate for qualitative research based on the concept of data saturation\textsuperscript{51}. Data saturation occurs when no new information or themes are noticed in the data, and any additional data collection becomes redundant. Similarly, in our research, as we approached the final participants, we observed a lack of new information and an increase in repetition of data.
2.3.2 Escape room game development

Game scenario and objectives

We selected an escape room as the task for participants to engage in. An escape room is a team-based game where participants are locked in a themed room and must solve a series of puzzles within a set time limit to ‘escape’ or achieve a specific objective. Similarly, in our escape room, puzzles and challenges were designed to ensure players collaborated to complete the tasks within the given time. This approach aligns with ecological validity principles by providing an authentic and immersive environment, fostering genuine teamwork, problem-solving, and decision-making behaviors that participants might encounter in their everyday lives.

Procedure

Before the experiment, participants were provided informed consent and they were also briefed on the general objectives of the study, without revealing the true purpose.

Participants were then randomly assigned to one of two conditions. In both conditions, they were informed that they would be paired with another participant, and they needed to solve puzzles and find clues through cooperation with their partner. However, they were actually paired with a confederate who was instructed not to engage in the task. Following the game, the decision was announced to the participant and the confederate in each other’s presence. The experimenter told the participant that they would receive 2.93 (which would be rounded up to 2.95), while the confederate would receive 7.07 (to be rounded up to 7.10). For further details, please refer to Appendix A.2.

It was also clarified that this decision was made based on the decision maker’s evalu-

---

1"Confederate" refers to a person who is knowingly and actively involved in a study or experiment but is not presented as a genuine participant. Instead, they play a specific role defined by the researcher to manipulate or control certain variables or aspects of the study.
uation and the information provided by the participants prior to the experiment. In the algorithm condition, this information consisted of both demographic data and the participants’ attitudes towards AI. In the human condition, the information included only demographic data.

After the experiment, participants filled out open-ended questionnaires about their experience, whether they found the decision fair or not and whether they would play the escape room with the same rules and decision maker again in the future (see Appendix B). This helped us to understand whether participants would trust the same decision maker again or not. They were then debriefed about the study’s true purpose, informed about the use of deception, and given the full reward of $10.

**Implementation of Wizard of Oz (WOZ) prototype**

In this study, we implemented the Wizard of Oz (WOZ) prototype by having an unseen experimenter (the author of this thesis) monitor the progress of the game from a separate location, hidden from the participants’ view using a webcam. By doing so, we created an illusion that something or someone is evaluating the participants’ performance throughout the game, whether it is an algorithm or a human. By using the WOZ prototype, we were able to examine participants’ reactions to perceived unfair decisions made by either an algorithm or a human, without the need for actual algorithmic or human evaluations.

### 2.3.3 Ethical Considerations

We recognized that using deception in this study might raise ethical concerns. Therefore, participants were debriefed about the use of deception in the study and its necessity. They were informed of the study’s true purpose and the role of the confederate. They were also given the option to withdraw their data from the study. None of the participants chose to withdraw (refer to Appendix C for the ethics clearance


2.3.4 Study design and experimental conditions

The study is a between-subjects study. The Dependent Variable (DV) is the perceived fairness of the decision. The Independent Variable (IV) is the type of decision-maker, which has two conditions: Algorithm and Human. We told participants in the algorithm condition that their performance would be evaluated by an algorithm based on their performance during the escape room as well as other criteria that would be disclosed only after completing the game. Participants in the human condition were told the same thing, except that their performance in the escape room would be evaluated by a human. We told all participants that, based on these evaluations, $10 would be divided between them.

2.4 Measures

2.4.1 Perception of fairness and trust

Participants were asked to evaluate the fairness of the decision by choosing one of three options: ‘fair’, ‘unfair’, or ‘neutral’. This measurement was selected because it provides a clear and straightforward method for participants to express their perception of the decision’s fairness, reducing the risk of ambiguity and confusion. Additionally, as a measure of trust, participants were asked if they would be willing to play the escape room again under the same rules and with the same decision-maker in the future.
2.4.2 Demographic information

Before starting the experiment, participants provided information about their age, gender, and educational background.

2.4.3 Control variable: Attitude towards AI

Attitude towards AI questionnaire

Before starting the escape room, participants in the algorithm condition were asked to complete a questionnaire regarding their attitudes towards AI. The questionnaire consisted of statements related to AI’s capabilities, trustworthiness, ethical considerations, and potential impacts on society (see Appendix B). By assessing participants’ attitudes towards AI, we tried to control for any potential effects this variable might have had on the study’s results, which allows for a more accurate comparison between conditions.

2.5 Data Analysis

2.5.1 Quantitative results

We did not find any differences in the perception of unfairness between the two groups. In each condition—whether the decision was made by a human or an algorithm—5 participants perceived the decision as unfair and 2 participants were neutral about the decision. This suggests that the source of the decision, human or algorithm, did not significantly influence participants’ judgment of fairness. However, it is important to note that this finding is based on a very small sample size and may not generalize to a broader population. Further studies with larger sample sizes could provide different and more accurate results.

Regarding the perception of trust, three out of the seven participants in the algo-
rithm condition mentioned they would not want to play the escape room again with the same rules and decision maker. In contrast, all participants in the human condition, along with three from the algorithm condition, expressed interest in playing the escape room again. One participant from the algorithm group, who was neutral about the decision’s fairness, left the question unanswered. This suggests a generally higher level of trust toward the human decision-maker.

Furthermore, we found that the attitude towards AI has a minimal effect on the perception of fairness in decisions made by algorithms (coef=0.0477). The coefficient’s low magnitude suggests that changes in attitudes towards AI have only a slight impact on the perception of fairness.

Again, it is noteworthy that future research with larger sample sizes might shed more light on the relationship between attitudes towards AI and perception of unfairness of the decisions made by algorithm.

2.5.2 Qualitative analysis of open-ended responses

We used thematic analysis as our primary analytic tool. Thematic analysis is a widely used qualitative research method that helps in identifying and interpreting patterns of meaning or themes within data.

Data immersion and codebook development

In our research process, we began by immersing ourselves in the data collected from participants’ responses to open-ended questionnaires. This immersion involved a thorough reading and re-reading of the responses to gain a deep understanding of the dataset. During this phase, we identified that among the various coding methods, the ‘Affective method’ was more aligned with our research questions. Within this approach, we particularly used ‘emotion coding’, ‘value coding’, and ‘descriptive
coding’. To elaborate more, consider the following example from our research:

- Researcher: “Would you play an escape room with the same rules and in case of evaluation with the same decision maker in the future? Why?”
- Participant 4 response: “I would play again because it was fun and the reward doesn’t matter to me. But if it did, I would not appreciate the same decision maker as the decision was unfair. Both participants collaborated equally.”

Our coding process involved breaking participant’s response into 5 units and then categorizing each unit into specific codes:

- “I would play again because it was fun.” (having fun - emotion coding)
- “The reward doesn’t matter to me.” (not caring about the reward - value coding)
- “I would not appreciate the same decision maker as the decision was unfair.” (perception of unfairness and unwillingness to have the same decision maker in the future - Descriptive coding)
- “Both participants collaborated equally.” (perception of equal contribution - descriptive coding)

We then proceed to develop a codebook. This codebook served as a reference point for our analysis and included clear definitions for each code. Additionally, it provided guidance on when to apply each code and, when a particular code should not be applied. This ensured consistency and accuracy in our subsequent analysis.

**Thematic extraction and analysis**

The next step was the extraction of broader themes from the coded data by grouping together similar or related codes. These themes were then refined, and organized in a way that represented the data and answered the research questions. The themes are as follows:
Perceived limitations of the algorithm in decision-making and evaluation

This theme emphasizes the perceived limitations of algorithms and their possible influence on fairness perceptions. Participants in the algorithm condition attributed the unfair decisions of the algorithm to its limitations. In their opinion, algorithms can make errors and cannot consider all necessary information. However, no participant in the human condition mentioned anything about the decision maker (human) limitations.

Participant 2: “I think the algorithm was unable to detect enough information”.
Participant 1: “It [the algorithm] is not evaluating everything I did”.

Emotional reactions to algorithmic decisions and perceived unfairness

This theme focuses on the emotional reactions to the decisions made by algorithms. Participants in the algorithm condition expressed emotional responses, and even personalized the algorithm’s behavior, which highlights the complexity of how algorithmic decisions are perceived. For example, Participant 10 commented on the distribution of resources by the algorithm, stating, “The allocation of the money is unfair. Algorithm is confusing”. Meanwhile, Participant 12 anthropomorphized the algorithm, attributing a spiteful intention to it, “AI is mean to me because I said it makes mistakes earlier” (the latter comment refers to the participant’s earlier response in the questionnaire regarding attitudes towards AI).

Overall, the theme emphasizes the emotional impact of perceived unfairness in algorithmic decisions and the tendency to assign human characteristics to algorithm.

Ambivalence and uncertainty regarding the evaluation process

Third theme were found in both the human and algorithm conditions which is participants’ ambivalence and uncertainty about the evaluation process. Some participants
reported a lack of transparency in the decision-making process which led to neutral perceptions of fairness. For instance, Participant 3 in the algorithm condition stated, “I don’t know the reasoning behind it [the result]”, suggesting a desire for more explanation or justification of the decision. Similarly, Participant 14 from the human condition reported, “I don’t really know what the decision is based on (the criteria) so I don’t have a clear judgment”. These quotes highlight the desire for more transparency in the decision-making process, a factor that seems to be crucial to perceptions of fairness in both human and algorithmic decisions.

**Perception of equal contribution**

A recurring theme was the perception of equal contribution. Many participants felt that, despite the lack of active engagement from the confederate in solving the puzzles, the prize should be split equally. For example, Participant 6 in the human condition stated: “We both played the game and should split the prize”. Similarly, Participant 8 highlighted the importance of presence, saying, “I think... the existence of the other person is still cooperation”. This theme demonstrates the significance of social norms (equality) in the perception of fairness, even in situations of unequal task contribution.

**Enjoyment and indifference towards monetary compensation**

We also found a theme related to enjoyment of the experience and indifference towards monetary compensation. In response to a question about whether they would participate in a similar escape room game in the future, most participants expressed interest to participate again due to the fun they had, regardless of the decision maker. Also, monetary compensation was not a key concern for some participants:

Participant 4: “[I’d play again because] it was fun. And the reward doesn’t matter to me”.

Participant 6: “The escape room was lots of fun and I do not need monetary compen-
Participant 13 even expressed surprise about the compensation, saying, "I only come for the escape room and didn’t expect there to be compensation". These comments show the participants’ enjoyment of the task and a lesser concern for the monetary aspect of it, which may have influenced their perceptions of fairness.

Adapting behaviour to influence decision maker

Some participants, from both conditions, intended to adjust their behavior in future interactions hoping for influencing the evaluator’s decision. For example, Participant 12 in the algorithm condition stated, “I’d be nicer to the AI next time to see if there is a different outcome.” Similarly, Participant 14 from the human condition wanted to explore actions that could be more favorable to the decision maker, stating, “I want to find out how to make actions in favor of this particular decision maker”.

Emotional impact of decision revelation

Qualitative observations were also made during the information disclosure to participants. Upon revealing the decision, a notable reaction was observed among participants. There was surprise and shock on the faces of participants, regardless of whether the decision was made by a human or an algorithm. This physical reaction clearly shows how strongly people can be affected emotionally when they perceive something as unfair. It also suggests that participants may have been expecting fair decisions, which shows the importance of fairness in the context of both human and algorithmic decision-making.

Apart from the themes related to ‘algorithmic limitations’ and ‘emotional reactions’, which were particularly seen in the algorithm condition, participants expressed similar perceptions and attitudes across both human and algorithmic decision-making.
2.6 Discussion

In the first component of this thesis, we aimed to explore and analyze the human perception of unfairness in algorithmic decision-making. In this section, we will delve deeper into our findings, to highlight their implications for the field of Human-Computer Interaction (HCI). A summarized overview of our key findings and their implications can be found in Table 2.2 later in this section.

People perceived the unfairness of decisions made by algorithms and humans as largely equivalent.

Within the scope of our experimental setup, our quantitative results showed no significant difference in how participants perceived the fairness of decisions made by algorithms compared to those made by humans. This aligns with the findings of the study conducted by Reeves et al.\textsuperscript{57}, which revealed that people unconsciously treat computers as social actors and give them the same treatment as they would to humans.

Moreover, our study did not found enough evidence to claim that individual attitudes towards AI influences the perception of fairness in algorithmic decision-making.

Our quantitative results highlight the importance of delving deeper to understand the underlying factors that shape perceptions of fairness. We proceeded to examine these factors through qualitative research, which are discussed in the following sections.
People attribute algorithmically unfair decisions to their limitations.

We found that in cases where unfair decisions made by algorithms, users tend to attribute these outcomes to errors or limitations in the algorithm, rather than malicious intent or bias. This could have implications for the design and implementation of algorithmic systems.

It is essential to ensure transparent communication about what these algorithms can and cannot do, the possibility of errors, and the measures in place to mitigate any negative consequences. This could involve developing clear user guides, tooltips, or even interactive features that explain how the algorithm works.

Setting realistic expectations is also important. It is about bridging the gap between the user’s anticipation of the system’s performance and the realities of algorithmic decision-making. This not only increases user trust but also allows users to better comprehend any discrepancies that might occur, which they could attribute to the acknowledged limitations of the system rather than suspecting any form of misconduct.

People can attribute human characteristics to algorithms.

We also found that people can anthropomorphize algorithms, attributing them with human-like qualities such as the capacity to hold grudges.

Participant 12: “AI is mean to me because I said it makes mistakes earlier.”

Therefore, algorithm designers must consider these human-like perceptions when developing systems. If people perceive algorithms as entities capable of holding grudges, it might affect their interaction with the system, perhaps leading to anxiety or fear. As such, it is important for design strategies to mitigate these kinds of misunderstandings by providing clear information about the nature and limitations of the algorithms.
This anthropomorphization also raises ethical questions. On one hand, it can lead to better user engagement and system acceptance as people often feel more comfortable interacting with things they perceive as more human-like\textsuperscript{58}. Therefore, designing algorithms that encourage anthropomorphism might improve user interaction and trust.

However, it could lead to the potential misuse of anthropomorphism in algorithm design. If an algorithm is purposely designed to seem more human-like with the intent to gain trust, users might overestimate its capabilities or misunderstand its limitations, which could lead to misplaced trust or unrealistic expectations.

**People can develop emotional reactions to algorithmic decisions.**

Our research also indicated that people develop emotional reactions to algorithmic decisions. Thus, the psychological impact of these decisions needs to be thoroughly considered and mitigated where necessary. It is not enough for algorithms to be fair; they must also be designed with empathy, taking into account the emotional responses they may evoke.

**Transparency is important in algorithmic decision-making.**

Our findings also emphasize the importance of transparency in algorithmic decision-making. As algorithms grow in complexity, it becomes increasingly difficult for users to understand the principles and processes driving their decisions. However, understanding these underlying mechanics is fundamental to gaining trust in the system. If people can not understand why an algorithm is making a specific decision, they may perceive it as arbitrary or unfair, leading to dissatisfaction or even rejection of the system.

Therefore, it is important to design systems that provide clear, understandable
explanations of how decisions are made. Such transparency can help bridge the gap between the complexities of the algorithm and the user’s understanding, allowing them to better trust and accept the algorithm’s decisions.

**Perceptions of fairness can be influenced by social norms.**

Our study’s observations of self-sacrifice due to social norms can also have implications for the design and implementation of algorithmic systems.

Participant 6: “We both played the game and should split the prize”.

This behavior shows a human tendency to put the group’s interest ahead of individual benefit in social situations.59.

In the context of algorithmic decision-making, this insight could provide a valuable perspective to system designers. As we incorporate more of these systems into our daily lives, it becomes vital that these systems not only mimic human decision-making but also understand and respond appropriately to complex social dynamics such as self-sacrifice for the group’s welfare.

**People might overlook perceived unfairness if the task is enjoyable or the stakes are low.**

We found that an individual might overlook perceived unfairness if the task is enjoyable or the stakes are low. This indicates that in the design of algorithmic decision-making systems, a singular focus on fairness, while critical, might not always be sufficient to ensure user satisfaction. Other aspects, such as the enjoyment derived from the task, the level of engagement it invokes, or the significance of the task outcomes, could potentially alter the perception of fairness.

For example, consider a scenario in a workplace where an AI-powered system is used to distribute bonuses based on an evaluation of each employee’s performance
metrics. While the algorithm aims to be objective and fair, it might be perceived as unfair by some employees. For example, it might overlook the contributions of a team member who has been performing critical but less quantifiable tasks, such as mentoring new employees or encouraging a positive work culture, which might not be accurately captured by performance metrics.

In such a case, the perceived unfairness in the distribution of bonuses could be mitigated if the workplace culture is highly supportive and the employee feels appreciated in other ways. If the employee is deeply engaged with their work and finds it fulfilling, or if the work environment is particularly positive, the perceived unfairness of the bonus distribution might be overlooked. The employee might value these aspects of their job more than a completely fair financial incentive.

Additionally, if the monetary stakes are relatively low, for instance, if the bonus constitutes a small proportion of the employee’s total compensation, they might be less concerned about the fairness of its distribution. The enjoyment derived from the job and the value placed on other, non-monetary rewards could potentially alter the employee’s perception of fairness.

**People believed they can influence the decision-maker’s decisions.**

The belief that people could influence the decision-maker’s decisions was another observation. It implies that users value a sense of control when interacting with algorithmic systems. People have a fundamental need for autonomy and a desire to feel that their actions and decisions matter. This can be true even when the decision-maker is an algorithm, suggesting that algorithms are not perceived merely as passive tools, but rather as interactive entities with which users have a dynamic relationship.

Incorporating this insight into algorithmic system design calls for strategies that
allow user inputs or feedback. This could involve offering options for users to provide feedback on an algorithm’s decisions, and even challenging them when necessary. For example, a recommendation algorithm on a streaming platform could provide users with options to influence its decisions by allowing users to rate or review recommendations, specify their preferences, or provide feedback on the relevance of the recommendations.

The following table summarizes our key findings and their implications:
Finding | Implications
--- | ---
Equivalence in perceived unfairness | There was no significant difference in perceptions of unfairness between human and algorithm decisions.
Attributing unfairness to algorithm limitations | Design implications should include transparent communication about algorithm capabilities and setting realistic user expectations.
Anthropomorphization of algorithms | Potential for both improved user trust and misunderstanding. Ethical considerations arise for designing human-like algorithms.
Emotional reactions to decisions | Algorithms should be designed with empathy to account for potential emotional impacts on users.
Transparency in decision-making | Transparent systems gain more trust. Importance of clear and understandable explanations.
Influence of social Norms | Algorithmic systems should consider complex social dynamics, like self-sacrifice for group welfare.
Overlooking unfairness for enjoyment or low stakes | Enjoyment and task significance can change fairness perceptions.
Belief in influencing decisions | Algorithms should be designed to be interactive and receptive to user feedback.

Table 2.2: Summary of key findings and their implications

2.7 Limitations and Future Research

Our results are preliminary, which invites further investigation, especially with larger sample sizes. Thus, we identified several opportunities for refining our experimental approach. These improvements, derived from both the study’s findings and our ob-
servations throughout the process, aim to address potential limitations and to provide suggestions for future research.

One potential limitation of this study is the generalizability of the findings given the specific context of an escape room game. As this is an enjoyable and engaging task by nature, it could potentially influence the participants’ perceptions and reactions. To address this, future research could consider repeating the experiment but within the context of a more mundane, less stimulating task. Doing so could offer valuable insights into whether the task’s nature and enjoyment level impacts participants’ reactions to unfair decisions made by algorithms or humans. We hypothesize that, a task considered ‘boring’ might increase negative reactions to unfairness.

Moreover, raising the stakes involved could potentially change the dynamics of participant responses. In our study, the stakes involved was $10, but what if we increase it to $100 or even $1,000? A larger sum of money could lead to more significant emotional reactions or different considerations of fairness. This adjustment could show how perceived unfairness and trust may vary with different levels of stakes involved.

Also, changing the setup of the experiment could provide new insights. For instance, informing participants of the unfair decision in a separate room from where the confederate is, might allow for a more honest expression of their opinions.

Lastly, asking about participants’ preferences for future decision-making entities, after they believe they were judged by either an algorithm or a human, would provide deeper insights into their trust levels and comfortability with each decision-maker. For instance, if a participant initially assessed by an algorithm expresses a preference for a human judge in future sessions, it could suggest a lower level of trust or satisfaction with the algorithm’s decision-making process, and vice versa. Moreover, it would be interesting to explore the reasons behind their preferences. Are they based on the
perceived fairness of the decisions, comfort with technology, or some other factors?

Additionally, it would be interesting to examine whether and how these preferences change over time and with repeated interactions. For example, would a person who initially preferred human judgement eventually become more comfortable with an algorithm judge after multiple fair decisions? On the other hand, would continued unfair decisions from a preferred decision-maker (algorithm or human) eventually lead to a switch in preference? Exploring these questions offers a longitudinal perspective on how trust and perceptions of fairness evolve with continued interaction and experience.
Chapter 3

Fair Play or Foul: Exploring Fairness Perceptions and Strategy Adaptation Using Nash Bargaining Game

3.1 Introduction

The purpose of this component is to investigate the impact of different fairness perceptions on strategic behavior and adaptation in Nash bargaining game. We used agent-based modeling as the simulation approach\(^1\). In our model, agents represent individuals with different fairness perceptions and adaptability levels. The simulation environment was created using Python and the SimPy library. We considered various scenarios with different distributions of fairness perception among the agents and claim ranges. These scenarios allow for the observation of how different combinations

\(^1\)Agent-based modeling is a computational approach that simulates the actions and interactions of individual agents within an environment to study complex systems and emergent phenomena\(^61\).
of agent characteristics influence fairness perceptions and strategy adaptation.

We used simulation in this study basically because it allows for the explorations of complex social dynamics in a controlled environment. Agent-based modelling (ABM) helps researchers to analyze the interactions between agents with different criteria and within various scenarios and observe the emergence of behavior based on these specific attributes. It would be difficult and time-consuming to conduct the same process with actual participants.

In the following sections, we will go through a review of relevant literature, followed by a detailed explanation of our methodology.

### 3.2 Literature review

#### 3.2.1 Game theory’s role in moral philosophy

Richard Bevan Braithwaite\textsuperscript{62}, a British philosopher known for his work in the philosophy of science and moral philosophy, was among the early thinkers who appreciated the relevance of game theory for moral philosophy\textsuperscript{63}. During a lecture at the University of Cambridge\textsuperscript{62}, he suggested that quandaries concerning the allocation of resources were intrinsically similar to the structure of John Nash’s bargaining problem. The Nash bargaining problem\textsuperscript{64} is a model in game theory that seeks an optimal solution for two parties trying to negotiate a deal, considering their individual desires and the compromise that might be necessary.

The Braithwaite model of distributive justice\textsuperscript{62} emphasizes the importance of cooperation among individuals or groups with different desires. The model is illustrated using the theoretical example of Luke and Matthew, neighbors engaged in different noisy activities during the same recreational hour. They must determine a fair distribution of time for their activities to maximize satisfaction. Braithwaite suggests
that the best strategy for the neighbors is agree to play alternatively and divide their playing time in a ratio that is based on their individual preferences. However, this method is questioned due to the subjectivity involved in comparing preference scales, a concern which Braithwaite admits to but deems unavoidable.

Overall, the model reflects the balance between individual utility and the imposition of a certain level of utility on the other party. Braithwaite argues that this method is the most fair and effective solution for achieving distributive justice in such contexts.

While the Braithwaite model might not directly address current fairness challenges, by integrating game theory into the concept of fairness, he introduced a new lens through which moral philosophers could examine social equity. Braithwaite’s early recognition of game theory’s significance, has deepened our understanding of fairness by providing a structured framework to explore how individuals negotiate, cooperate, and compete. This approach bridges the gap between individual actions and societal norms, revealing how the interplay between self-interest and morality shapes our collective perception of fairness.

Peter Vanderschraaf, a contemporary philosopher with a focus on game theory, offers a succinct yet comprehensive understanding of game theory: it is, essentially, “the formal logic of interactive decisions”\textsuperscript{63}. This definition illustrates game theory’s mathematical basis as well as its relevance for analyzing complex interactions in a wide range of real-life situations.

The application of game theory to the intersection of morality and rational choice provides important insights. Primarily, game theory considers interactions between different ‘agents’—which could be individuals or entities capable of making decisions and acting on them\textsuperscript{65}. The collective actions and respective outcomes of these interactions constitute a ‘game.’
To simplify, let’s consider the well-known game theory scenario known as the “Prisoner’s Dilemma”\textsuperscript{65}. In this scenario, two people are arrested, but the police do not have enough evidence to convict them on the primary charge, so they plan to sentence both on a lesser charge. Each prisoner is given the opportunity to betray the other by testifying that the other committed the crime. The possible options are as follows:

1. If Prisoner A and Prisoner B both remain silent, they each serve 1 year in prison (cooperate with each other).
2. If Prisoner A betrays Prisoner B but Prisoner B remains silent, Prisoner A will go free and Prisoner B will serve 3 years (and vice versa).
3. If Prisoner A and Prisoner B both betray each other, they each serve 2 years.

The strategies can be summarized in a payoff table\textsuperscript{2}:

<table>
<thead>
<tr>
<th></th>
<th>B cooperates</th>
<th>B defects</th>
</tr>
</thead>
<tbody>
<tr>
<td>A cooperates</td>
<td>A: -1, B: -1</td>
<td>A: -3, B: 0</td>
</tr>
<tr>
<td>A defects</td>
<td>A: 0, B: -3</td>
<td>A: -2, B: -2</td>
</tr>
</tbody>
</table>

Table 3.1: Payoff matrix for the Prisoner’s Dilemma game.

The equilibrium of this game, known as the Nash equilibrium\textsuperscript{66}, is when each player’s action is the optimal response to the actions of others\textsuperscript{3}. In this case, the Nash equilibrium is for both prisoners to betray each other, even though they would collectively be better off if they both remained silent.

However, this Nash equilibrium is not Pareto efficient\textsuperscript{4} because there exists a more desirable outcome (mutual cooperation) where both players would be better off.

\textsuperscript{2}Negative numbers are used because it is a cost for the prisoners to serve time.

\textsuperscript{3}Nash equilibrium is used to identify stable states in strategic interactions where no player has an advantage to unilaterally change their strategy. It provides a framework to analyze situations where multiple agents make decisions and where each agent’s best response depends on the decisions of other agents. By identifying Nash equilibria, we can predict likely outcomes in games and other strategic scenarios.

\textsuperscript{4}Pareto efficiency (or Pareto optimality) refers to a situation where no individual can be made better off without making another individual worse off.
Now, if we could somehow argue that following moral norms (e.g., not betraying each other) leads to a better outcome for involved parties, then game theory could show that acting morally aligns with acting rationally.

### 3.2.2 Evolutionary game theory

Game theory often merges ideas from various disciplines. A notable instance of this is the development of evolutionary game theory, pioneered by biologists John Maynard Smith and G.R. Price. Evolutionary game theory integrates key principles from evolutionary theory into the framework of game theory and uses the natural process of evolution for reaching equilibrium. This process does not rely on conscious choices or deep analysis. Instead, it happens over time due to selection and adaptation across generations.

In evolutionary game theory, each ‘strategy’ corresponds to a different type of behavior, and the ‘game’ represents the survival and reproductive challenges faced by organisms. The ‘strategy’ corresponds to evolutionary fitness, with successful strategies leading to greater survivability and reproductive success. Through repeated rounds of this game—i.e., successive generations—species gradually adjust their behaviors based on the success of different strategies. Over time, this process can lead to the emergence of equilibrium behaviors, a stable state where no individual can improve their fitness by changing their strategy.

For example, in a certain species of bird, there are two strategies for finding food: Scavengers, which look for food left by other birds or animals, and Hunters, which catch insects and small animals. At first, there are equal numbers of Scavenger and Hunter birds. As generations pass, the Hunter birds are more successful because they always find food. They have more offspring, so over time, the number of Hunter birds increases. However, as the number of Hunter birds increases, Scavenger birds start
to do better. Now there is a lot more leftover food from all the successful Hunters, and it is easier for Scavengers to find enough to eat.

After many generations, the bird population reaches a stable state - an equilibrium. There is just the right number of Hunter birds to leave enough food for the Scavenger birds. And there is just the right number of Scavenger birds to eat all the leftover food, so it is not worth it for any more birds to become Scavengers. At this equilibrium, no bird can do better (in terms of survival and reproduction) by changing its strategy.

This perspective has expanded our understanding of how societal norms, behavioral patterns, and even moral codes could potentially evolve. The essence of the process is rooted in nature and is driven by the fundamental biological imperatives of survival and reproduction. The convergence towards equilibrium, in this context, reflects the optimal adaptation of a species to its environment\textsuperscript{67,69}.

### 3.2.3 The evolution of fairness

Evolutionary game theory, suggests that fairness norms can emerge as a result of repeated interactions within a population over time, similar to the way species adapt their behavior through natural selection\textsuperscript{67,70}. A fairness norm can become prevalent if it proves beneficial or ‘fit’ in terms of societal cooperation and cohesion. Over time, a community might gradually converge towards this norm, establishing a societal equilibrium of fairness\textsuperscript{71}.

This approach also provides unique insight into why individuals might adhere to norms of justice, often at a personal cost. From a strictly rational choice perspective, self-interest should discourage such behavior. However, evolutionary game theory argues that these costly actions might represent strategies that, while not immediately beneficial for the individual, contribute to the overall stability and survival of the
group\textsuperscript{71}. Over generations, such cooperative strategies might be more evolutionarily fit and take over a population.

In sum, evolutionary game theory provides a lens to explore complex moral and social issues. It mirrors the evolving nature of society, effectively predicting the growth and stability of ethical norms. As we continue to use and improve this tool, it enhances our understanding of how societies function. Therefore, we intend to incorporate the principles of evolutionary game theory into our research unlocking further potential and insights.

### 3.2.4 Fair division and different fairness perceptions

Aristotle’s endorsement of the principle of fair division\textsuperscript{63,72}, illuminates not just the concept of equality but also the complexities surrounding it. This principle posits that goods should be distributed in accordance with relevant differences, with equals receiving equal shares. This principle summarizes the concept of ‘proportional justice’—the notion that individuals should receive rewards proportional to their contributions\textsuperscript{73}.

“All men think justice to be a sort of equality; ... For they say that what is just is just for someone and that it should be equal for equals. But there still remains a question: equality or inequality of what? Here is a difficulty which calls for philosophical speculation\textsuperscript{72}.”

According to Aristotle, determining the metric of equality or inequality is challenging. In other words, when ensuring fairness or equality, what parameters should we consider? Are we looking at equality of opportunities, equality of outcomes, or some other metric?

The perception of fairness is heavily influenced by individual characteristics, values, and experiences\textsuperscript{74}. Particularly, the role of ‘narcissistic traits’ in shaping per-
ceptions of fairness offers a compelling perspective. Individuals with high levels of narcissism often feel superior to others, possessing an inflated sense of self-importance and entitlement. Consequently, their understanding of fairness tends to be skewed towards their self-interest, believing they are deserving of a disproportionately large share of resources.

Such narcissistic traits can impact the dynamics within a group or society. If unchecked, it can lead to an environment where fairness is distorted to favor the few at the expense of the many. This is especially concerning in contexts involving resource distribution or decision-making processes, where individuals with high narcissistic level may assert their interests over others, leading to unjust outcomes.

Here, the application of evolutionary game theory can be very beneficial. It can provide a structured framework to study these dynamics by modeling them as games with various strategies representing different fairness perceptions. For instance, a strategy could be an individual with high narcissistic traits demanding a larger share of resources, reflecting their inflated sense of entitlement. Another strategy might represent individuals with low and moderate narcissism levels who are more likely to advocate for a more modest allocation of resources.

Over time, through the iterative process of game-playing, we can observe the evolution of these strategies. We can examine how certain behaviors become more or less prevalent based on their success within the game, mirroring societal acceptance or rejection. By modeling this scenario as an evolutionary game, we can study how different strategies compete with each other and which forms of fairness perception predominance over time.
3.2.5 Nash bargaining game

Brian Skyrms, a philosopher and professor at the University of California, Irvine in the book ‘The Evolution of the Social Contract’, uses evolutionary game theory, to examine how cooperative behavior and principles of fairness emerge and stabilize in societies over time. He considers a hypothetical situation in which two individuals are tasked with dividing a cake (a variant of the Nash bargaining game). Each individual makes a claim to a portion of the cake, and if the combined claims do not exceed the available resource, each person receives what they have claimed. However, if their claims exceed the size of the cake, they receive nothing. The trend in such a situation, as supported by empirical evidence, is that most individuals would stake a claim to exactly half of the cake. But why does this trend exist? Why do individuals commonly claim exactly half?

Skyrms attributes this to the principle of ‘informed rational self-interest’. To maximize their payoff, individuals must anticipate the claim that the other participant will make. They must make a claim that cannot be improved by changing it, a condition that must also be true for the other player (Nash equilibrium).

In the context of Nash’s bargaining game, any claim that does not exceed the total resource (100% of the cake) can be considered a strict Nash equilibrium. However, according to Skyrms, only one of these equilibriums—a claim to half of the cake—is seen as just, and perhaps the only rational choice. The reason for this is that if one person claims more, such as 70%, there is no way to ensure that the other person will claim 30%, so both individuals can get their claims.

According to Skyrms, demanding half is the only evolutionarily stable strategy in this game. He illustrates this point by considering the following scenario: consider a group of people who claim either 60% or 30%. Those who claim 60% would not get anything if they encounter each other, but they would get their claimed portion if
they encounter those claiming 30%. However, an agent who claims 50% will get what they asked for when encountering another similar agent.

Skyrms further explains his point by discussing a population in which half are ‘greedy’ (demanding 2/3 of the cake) and half are ‘modest’ (demanding 1/3). The ‘greedy’ individuals risk getting nothing if they meet another ‘greedy’ individual, but they get their claim if they meet a ‘modest’ one. The ‘modest’ individuals always get their claim, regardless of who they meet. If the proportion of ‘greedy’ individuals increases, they tend to meet each other more often, leading to lower average payoff.

Skyrms takes the analysis further by considering a population including individuals who are ‘super greedy’ (claiming more than the ‘greedy’ group), ‘super modest’ (claiming less than the ‘modest’ group), ‘moderate’ (claiming more than the ‘modest’ but less than the ‘greedy’), and ‘fair-minded’ (claiming exactly half). Interestingly, he suggests that in such a population all groups including the ‘fair-minded’ one would eventually go extinct. This is because whenever they meet ‘greedy’ individuals, they would receive nothing, and when they meet the ‘modest’ ones, their payoff would be less than what the ‘greedy’ individuals would get.

Skyrms also tested ten possible ways of dividing the cake to see which approach society would most likely adopt over time. Interestingly, in 62% of cases, the society converged towards claiming half. When this did not occur, the outcomes were described as ‘polymorphic traps’ which are situations where the society gets caught in a state of equilibrium with multiple stable strategies, rather than converging on one single strategy. For example, Skyrms argues that as the number of possible strategies for dividing the cake increases, society tends to converge more often on strategies close to claiming half.

Skyrms’ analysis of the dynamics of fair distribution and claim strategies provides valuable insights into understanding how notions of fairness and self-interest interact.
Yet, it also opens up new questions and challenges that invite further exploration. For instance, the degree of randomness in the agents’ claims near equal division and its effect on the overall distribution pattern is an area that needs more clarity. How much deviation from equal division can occur before it impacts the societal trend towards claiming half? Understanding this can help in creating models that are more accurate and representative of real-world scenarios. Also, randomness could lead to the exploration of new strategies that could potentially be more beneficial. If agents strictly follow a defined strategy, the opportunity to discover these potentially better strategies could be lost.

It would also be worthwhile to determine the minimum number of moderate players needed, with their claims’ average equating to half, for society to converge towards these agents’ strategy. Understanding these aspects would not only provide a more comprehensive view of the factors influencing the societal norms of fairness but would also enhance our ability to predict and guide the evolution of these norms. Therefore, our research aims to delve deeper into these questions, using evolutionary game theory as a tool to explore the complex interplay between individual behaviors and societal norms.

3.2.6 Games and fairness

In another part of ‘The Evolution of the Social Contract’ book, Skyrms also mentions that being just is rational but not self-interested. This perspective is somewhat contentious, as it can be argued that even someone is claiming more than half the cake, for example—might eventually realize that their payoff would be better by claiming half and adjust their strategy accordingly. This modification could be seen as a self-interested act, illustrating the blurred lines between self-interest and fairness.

Justin D’Arms critically examines Skyrms’ approach to the symmetric bargaining
game. He argues that terms like ‘fair division’ can be misleading. Instead, he prefers the more accurate phrase ‘demand 1/2’, emphasizing that an individual player’s strategy is not necessarily about equalizing payoffs but maximizing their own.

He questions the association of fairness with the symmetric bargaining game by challenging the moral judgment of deviating from the ‘demand 1/2’ strategy. He provides an example of a Nash bargaining game where player 1 bids only 40%. If player 2 responds by bidding 60% and subsequently attains it, is this a demonstration of injustice or merely strategic opportunism? According to D’Arms, while such behavior might be deemed irrational and opportunistic, labeling it as morally unjust seems out of place. He questions whether exploiting another player’s underbidding could be seen as unfair and whether there is a moral obligation to limit one’s claim to 50%, or even lower it to 40% to ensure equal payoffs.

D’Arms argues that the rules of a game might be fair, but once they are established, players should be morally free to pursue any strategy they believe will be most profitable. He challenges us to test the idea of equal division as a justice principle in real-life scenarios that reflects Nash bargaining game. However, he admits that it is difficult to identify such cases due to the complexities of real-world situations.

D’Arms also challenges Skyrms’ assumption that humans would often find themselves in situations where an equal division of resources would be considered fair. He argues that these scenarios are rare and complicated, making it difficult to accept that our tendency to choose half as a bargaining strategy has an evolutionary origin.

Furthermore, D’Arms draws attention to the friction between the neatness and abstraction of game-theoretic models and the complex, unpredictable nature of human behavior. He highlights that reality rarely aligns neatly with theoretical models, and human decisions are influenced by many factors other than rational self-interest. As D’Arms argued, using evolutionary game theory to analyze moral and social norms
has a number of limitations and challenges. While it can offer valuable insights, it is essential to remember that human behavior is complex, and it cannot always be simplified into clear-cut strategies and outcomes.

Nevertheless, it is the adaptive nature of evolutionary game theory that makes it particularly suitable for the study of social evolution. It mirrors the dynamic nature of societies, reflecting how norms and perceptions can change over time in response to various pressures and influences. This theoretical framework, therefore, helps to bridge the gap between the individual and the societal level, providing insights into the emergence and stability of societal norms and behaviors.

3.3 Methodology

3.3.1 Model description

In this simulation model, agents are individuals with different fairness perceptions that adapt to their environment over time. The game that agents playing is an alternated version of the Nash bargaining game, which models the interactions between agents as they make claims on a shared resource. The model assumes that agents have bounded rationality and are influenced by their fairness perception, level of the adaptability, and the number of consecutive losses. If agents experience losses for five consecutive rounds, they change their strategy by adopting the strategy of an agent with higher reputation, which is determined by the average of all the money that particular agent has earned.

3.3.2 Model components and variables

1. Agents: Agents \(a\) in the model represent individuals with unique ID \(i \in \mathbb{N}\), fairness perceptions which are either high narcissistic, moderate player or low
narcissistic \( f_i \in \{H, M, L\} \), and adaptability levels which are either high or low \( AD_i \in \{H, L\} \). Each agent keeps track of their money earned \( m_i \in \mathbb{R}_{\geq 0} \), claims made \( C_i \in \{0, 1, 2, ..., 100\} \), number of games played \( g_i \in \mathbb{N} \)^5, number of consecutive losses \( l_i \in \mathbb{N} \), and strategy changes \( SC_i \in \mathbb{N} \).

2. **Interactions:** During each time step, agents are randomly paired with each other and interact within a game where they make claims on a shared resource based on their fairness perception. The total amount of the shared resource is fixed at 100 dollars. If the sum of the claims made by both agents is less than or equal to $100, each agent receives their claim (winning). But if it exceeds $100, neither agent receives any resource (losing). For example, \( a_i \) and \( a_j \) are two agents in this society:

If \( c_i + c_j \leq 100 \), \( m_i \) and \( m_j \) will be increased, otherwise (if \( c_i + c_j > 100 \)) \( m_i \) and \( m_j \) remain the same.

Also, if an agent loses a claim \( l_i \) will be updated, otherwise it will be set to zero:

If \( c_i + c_j > 100 \) then \( l_i' = l_i + 1 \)

Otherwise if \( c_i + c_j \leq 100 \) then \( l_i = 0 \).

If \( l \geq 5 \), then the agent evaluates a possible strategy change.

3. **Fairness Perception:** We will provide more details on the claims themselves in a later section in this chapter, but this variable represents an agent’s tendency to make claims that are either moderate, high, or low. Differences in fairness perception affects individual decision-making behaviors and it is an essential variable in understanding the dynamics of a society.

4. **Adaptability (AD):** This parameter indicates an agent’s ability to change their strategy based on the number of their consecutive losses. High adaptability shows a greater likelihood of strategy changes, while low adaptability means

\(^5\)Agents each track the number of games they have individually played.
a lower likelihood of changing strategy. The adaptability parameter helps to investigate the impact of learning and adaptation on the evolution and changing fairness perception in the society.

5. **Strategy Change Probability (SCP):** This variable represents the likelihood of an agent changing their strategy based on their adaptability level. Agents with high adaptability have a higher strategy change probability (0.8) than those with low adaptability (0.2):
   
   If \( AD_i = H \) then \( SCP = 0.8 \) otherwise if \( AD_i = L \) then \( SCP = 0.2 \).

   As defined earlier, if the agent’s consecutive losses \( (l_i) \) is less than 5, the strategy change probability is 0 (the agent does not change its strategy).

6. **Reputation (REP):** An agent’s reputation is calculated as the average of all the money they have made during the simulation. It influences what strategy agents adopt during the simulation. The reputation of an agent can be calculated as:

   \[
   REP_i = \frac{1}{g_i} \sum_{n=1}^{g} m_n
   \]

   \( m_n \) is the money earned by the agent in the \( nth \) interaction, and \( g_i \) is the total number of games the agent has played.

### 3.3.3 Initial conditions

The initial state of the simulation consists of 100 agents, each with an ID, fairness perception, and adaptability level. The agents are categorized into three groups based on their fairness perception: high narcissistic, low narcissistic, and moderate player. The adaptability of each agent was randomly assigned as either high or low \( (AD_i \in \{H, L\}) \). All agents start with zero money \( (m_i = 0) \),
and their initial reputation is calculated based on their earnings in the early stages of the simulation.

### 3.3.4 Running interaction example

Consider Agent A, with high adaptability, playing a series of games:

Game 1: Agent A and another agent play a game, and both agents, including Agent A, lose (the sum of money exceeds $100). This marks the first consecutive loss for Agent A.

Game 2: Agent A loses again.

Game 3: Agent A loses once more.

Game 4: Agent A loses yet again.

Game 5: Agent A loses one more time, marking the fifth consecutive loss. At this point, Agent A, with high adaptability, triggers the adaptation rule due to the series of losses. There is an 80% chance that Agent A, with high adaptability, decides to change its strategy due to the loss. If it does so, it modifies its fairness perception to match that of an agent with a higher reputation, chosen randomly in the fully connected network and from within the agent’s immediate social network (neighbors) in the BA and caveman graphs. If there is not an agent with a higher reputation available, the agent will not change their strategy.

### 3.3.5 Simulation setup and implementation

The entire conceptual framework behind the design of this simulation model is visually represented in Figure 3.1. Next, I will present the pseudo-code for the most important components of the simulation.\(^6\)

---

\(^6\)Code available at [https://github.com/Fae97/Simulation_Code.git](https://github.com/Fae97/Simulation_Code.git)
Agent

class Agent:
    Initialize env, name, fairness_perception, adaptability, and other attributes

make claim method

Generate a random claim based on fairness_perception
    Append claim to claims list
    Return claim

update reputation method

    Calculate reputation as the average of claims
    Update reputation attribute

copy strategy method

    If consecutive_losses >= 5 and random probability allows:
        Calculate average money for all agents
        Select more successful agents based on average money
        Copy strategy from a randomly selected successful agent
        Reset consecutive_losses to 0

play method

    While True:
        Select a random opponent
        Make claims and update game statistics
        Update money and reputation
        Copy strategy if conditions are met
Figure 3.1: Agents interaction diagram
3.3.6 Manipulation of model parameters and variables

In this study, we manipulated several key model parameters and variables to explore the impact of different fairness perceptions and network structures on the system’s behavior. The manipulations are as follows:

**Distribution of fairness perception**

We examined various distributions of fairness perception among agents and then created a heatmap to show the results. Each cell in the heatmap represents the proportion of a specific fairness perception in the society (we captured moderate player and low narcissistic perceptions only, assuming that the rest of the proportions are agents with high narcissistic traits).

We varied the number of moderate players along the left side of the heatmap (see Figure 3.4 as an example), with values ranging from 2 to 16. The remaining agents were assumed to be distributed equally between high narcissistic and low narcissistic agents. That is, for example when there are 16 ‘moderate players’, there is 42 agents being high narcissistic and another 42 being low narcissistic.

**Claim ranges**

We manipulated the claim ranges along the bottom side of the heatmap. This allowed us to analyze how varying claim ranges affect the dynamics of the simulation and the agents’ fairness perceptions.

**Network effects and topologies**

We also explored the impact of network effects on the dynamics of the game by applying the Barabási-Albert (BA) network and caveman graph into the model.

By manipulating these parameters, we were able to analyze the effects of differ-
ent fairness perceptions, claim ranges, and network structures on the system’s behavior. The detailed descriptions of these manipulations and the results obtained from each scenario will be presented in the “Model scenarios and analysis” section which we will delve deeper into the specific scenarios and their outcomes, shedding light on how the model’s parameters and variables interact and shape the overall dynamics of the system.

3.3.7 Data collection

Data was collected during the simulation runs, capturing each agent’s claims, money, reputation, fairness perception, adaptability, and other relevant variables. For most of the experiments, we conducted a total of 50 simulation runs for each data point so we can have a rich dataset to analyze the dynamics and trends in agents’ behavior. We calculated the proportion of agents with different fairness perceptions and the number of strategy changes made by agents over time.

After running the simulation and collecting the data, we used matplotlib library in python to visualize the data. By transforming the proportions of each fairness perception into a heatmap, we were able to see the dynamics and trends in agents’ behavior (see Figure 3.4 as an example).

In the following section, we will present our analysis of the simulation results and the comparison of different scenarios and parameter settings.

3.3.8 Model scenarios and analysis

Experiment 1: high and low narcissistic agents

In the first experiment, we considered a population of 100 agents, with 50 high narcissistic and 50 low narcissistic agents. The high narcissistic agents claimed
a range of $50 to $100, while the low narcissistic agents claimed a range of $0 to $50. We observed that the society gradually converged towards the low narcissistic group, indicating a shift in strategy. This result was predictable; when individuals with high narcissistic traits interact with those with low narcissistic traits, they get more payoff if they win. However, when high narcissists interact with each other, they usually receive nothing unless they claim half, which is rarely the case.

On the other hand, individuals with low narcissistic traits always receive what they claimed for when they interact amongst themselves or with high narcissistic agents. Therefore, over time, society will gravitate towards low narcissistic agents’ strategy (see Figure 3.2).

![Figure 3.2: Agents strategy adaptation over time](image)
**Experiment 2: introduction of fair players**

We then introduced an additional group of agents, the fair players, who only claimed $50. We distributed the agents into 20 fair players, 40 high narcissistic, and 40 low narcissistic agents. We observed that all agents rapidly evolved to adopt either low narcissistic or fair players’ strategy (see Figures 3.3 and 3.4). This result can be interpreted as the high narcissistic agents receiving nothing when they are playing with fair players due to their excessive claims. On the other hand, low narcissistic agents receive what they claimed for every time they encounter fair players, but it is still less than half. Thus, in this society, any strategy other than claiming half performs worse, which is in line with Skymrs’ findings\textsuperscript{71}.

![Claims made by each agent over time](image-url)

**Figure 3.3**: Agents strategy adaptation after introducing fair players. All the agents played the fair player’s strategy at least once.
Figure 3.4: Distribution of agents’ fairness perception after introducing fair players, with standard deviation indicated.

**Experiment 3: introduction of moderate players**

In this experiment, we introduced an element of randomness to the fair players’ claims and introduced another group of agents called moderate players. The mean for moderate players claims ranges is always 50. We then conducted 50 simulation runs with varying numbers of moderate players and range of claims. This allowed us to capture the uncertainty present in real-world scenarios. Therefore, by gradually increasing the number of moderate players in the simulation, we aimed to examine how the presence of these agents with moderate claims affects the dynamics of the system and the behavior of the agents. Also, maintaining an equal distribution of high narcissistic and low narcissistic agents, helped us identify the specific contributions of moderate players in shaping the outcomes. This gave us an insight into how agents making claims closer to claiming half ($50) affects agent interactions.

In the following sections, we discuss and compare the results in different sce-
Experiment 4: overlapping claims scenario with fully connected network

In the first set of scenarios, we considered overlapping claims for the three types of agents:

High narcissistic agents claimed between 50 and 100 dollars.
Low narcissistic agents claimed between 0 and 50 dollars.
Moderate players’ claims overlapped with the claims of the other two groups.

To explore the impact of moderate player agents deviating from claiming half (50 dollars), we ran multiple simulations with different claim ranges for moderate players. We started with a narrow range of 49 to 51 dollars and gradually widened the range in subsequent runs. Each time, we increased the claim range by one dollar on both sides of the range, moving from 48 to 52, 47 to 53, and so on, until we reached the final range of 25 to 75 dollars.

Figure 3.5: Moderate players’ proportions with overlapping claims
We found that when claims are overlapped, as the level of deviation from the fair claim ($50) increases, the proportion of moderate players in society slightly decreases. And if there are at least six moderate players in the society, it converges towards moderate players’ strategy almost all the time. As you can see in the Figures 3.5 and 3.6, if the society did not converge to moderate players strategy, it mostly converged to high narcissistic agents’ strategy. The proportion of low narcissistic agents is very small (less than 0.1), which is predictable, as in this situation, agents with low narcissistic traits struggle to compete against both types of opponents. This is particularly true when moderate players’ claims get narrower. In such cases, even when these low narcissistic agents successfully secure their claims, they receive less payoffs compared to high narcissistic opponents. They also often receive less compared to moderate players.
Figure 3.7: Moderate players’ proportions with non-overlapping claims

Figure 3.8: Low narcissistic agents’ proportions with non-overlapping claims
Non-overlapping claims scenario with fully connected network

In this simulation scenario, we aimed to investigate the behavior and interactions between agents when their claims do not overlap. Unlike the overlapping claims scenario, the high and low narcissistic agents’ claims were designed not to overlap with the moderate players’ claims. For instance, if moderate players claimed amounts in the range of $25 to $75, high narcissistic agents claimed amounts from $76 to $100, and low narcissistic agents claimed amounts from $0 to $24. By comparing the results of this scenario with the overlapping claims scenario, we aimed to understand the influence of claim distribution on the interactions between agents and the overall outcomes of the simulation.

In contrast to the overlapping scenario, this non-overlapping scenario exhibits a decrease in the proportion of moderate players as they move closer to claiming half and we often see that low narcissistic agents take over the society. This pattern can be seen on the right side of the heatmap (see Figure 3.8). This is in contrast to another heatmap which displays the proportions of moderate players in the society (Figure 3.7).

The observed trend can be explained by considering the effects of non-overlapping claims on players. For moderate players, as the bounds of their claims decrease, the range for low narcissistic claims expands. This implies that when individuals with low narcissistic traits interact, they are more likely to receive higher payoffs as their claims does not overlap with moderate players. A similar pattern can be identified for moderate players. Therefore, we can say when agents’ claims do not overlap, the wider the range of a particular fairness perception claim gets, the more prevalent it becomes in society. However, we cannot see this pattern for high narcissistic agents. Due to their excessive claims, when they interact with each other, they often end up with no payoff.
Following the experiments with both overlapping and non-overlapping claims, we observed that the non-overlapping version generated noisy results. To address this issue, we used truncated normal distribution and Poisson distribution for claims, as these distributions allow us to represent claims with a defined range to avoid extreme values that could happen in the non-overlapping scenario.

**Truncated normal distribution**

![Figure 3.9: Moderate players’ proportions with non-overlapping and truncated normal distribution claims](image)

We modeled the claims made by moderate players using a truncated normal distribution. The truncated normal distribution is a variation of the normal distribution, where values are limited to a specific range. In our case, this range was chosen to represent the reasonable claims that a moderate player would make. By using the truncated normal distribution, we ensured that the claims made by moderate players are centered around a specific value (i.e., the average of the claim bound which is 50) while still allowing for some variability in the
Figure 3.10: Low narcissistic agents’ proportions with non-overlapping and truncated normal distribution claims

As it is shown in the Figures 3.9 and 3.10, the proportions stayed roughly the same.

**Poisson distribution**

We also modeled the claims made by moderate players using a poisson distribution. Again, the poisson distribution was employed to represent the claims made by moderate players, with the mean value (lambda) set to a specific value (50), which reflects the average claim that a moderate player would make. By using the poisson distribution, we ensured that the claims made by moderate players are centered around this mean value while still allowing for some variability in the claims. The choice of a poisson distribution allowed us to capture the natural variation in agents’ claims while keeping them within a reasonable range. However, there were not much of a change when we change the distribution to Poisson.
Figure 3.11: Moderate players’ proportions with poisson distribution and non-overlapping claims

Figure 3.12: Low narcissistic agents’ proportions with poisson distribution and non-overlapping claims
Experiment 5: Barabási-Albert (BA) network

Figure 3.13: Moderate players’ proportions with BA network

To investigate the impact of network structure on the dynamics of fairness perception, we transitioned from a fully connected network to a Barabási-Albert (BA) network, where agents interact only with their social network. We observed that the BA network introduced more noise into the system (see Figures 3.13 and 3.14). BA network structure can introduce more degrees of freedom, as nodes have variable number of connections, compared to a fully connected network where each node has the same number of connections. The process of generating a BA network also involves randomness. New nodes preferentially attach to existing nodes with high degrees, but there is still a probabilistic element to which nodes are chosen. This variability can introduce more randomness and unpredictability into the system, leading to noisier results.

We then enforced the hubs (top 20% of the high-degree nodes) in the BA network to be high narcissistic, and we saw that the society converges to either
Low narcissistic or moderate players’ strategy with the proportion of low narcissistic agents being more (Figures 3.15 and 3.16). Thus, we observed that high narcissistic agents’ advantage as a hub does not provide any long-term benefit in the BA network structure.
Figure 3.15: Initialization of a BA network with hubs as highly narcissistic agents

Figure 3.16: Mean proportion of each fairness perception after 100 runs with high narcissistic hubs (standard deviations indicated)

We also forced hubs to be moderate players and we observed the same results
(see Figures 3.17 and 3.18). When the hubs are moderate players, low narcissistic agents’ strategy will gradually take over the society. Although these low narcissistic individuals may seem to have less payoff compared to other groups, their modest and less risky claims help them to gain a higher reputation over time, which contributes to their sustainability (see Figure 3.18).

Figure 3.17: Initialization of a BA network with hubs as moderate players
Figure 3.18: Mean proportion of each fairness perception after 100 runs with moderate players hubs (standard deviations indicated)

**Experiment 6: caveman network**

In our research, we also examined the implications of the configuration of a caveman graph, a network structure that represents a society of tightly-knit groups, or cliques, where each group member is connected to every other member in the same group, and each group is interconnected with others by only one or a few links. We focused on scenarios where there exists a proportion of high narcissistic agents within some communities, while the rest of the communities are populated by either moderate players or low narcissistic agents (see Figure 3.19).

Despite changing the network structure, we observed that high narcissistic agents will change their strategy after some time, while low narcissistic agents and moderate players’ strategy dominated over the society (see Figures 3.20 and 3.21). These agents are more likely to win their interactions due to the modest
claims they make, allowing them to get more reputation over time and grow.

In contrast, high narcissistic agents make excessive claims during their interactions. These claims are more likely to exceed the shared resource ($100), leading to a lower probability of winning these interactions. Consequently, high narcissistic agents struggled to maintain or grow within their communities. Despite being connected to a set of moderate players or low narcissistic agents at the start of the simulation, high narcissistic agents ultimately failed to keep their strategy due to their excessive claims.

Figure 3.19: Caveman graph configuration
Figure 3.20: Mean proportion of each fairness perception after 100 runs (standard deviations indicated)

Figure 3.21: Changes in fairness perception proportions over time (single run)

We observed that high narcissistic agents might remain in a society that does
not have moderate players and is only populated by low and high narcissistic agents (see Figures 3.22 and 3.23). However, their persistence is limited; after 100 simulation runs, we found that the proportion of high narcissistic agents remaining in the society was significantly small (0.03 %). This indicates that while high narcissistic agents' strategy might sustain in the absence of moderate players, their dominance or survival is not assured in the long term. This observation reveals the importance of cooperative behavior for survival and growth within a networked society. Even though network structure can influence the dynamics of a society, the behavior and strategy of the agents within the network can play a crucial role in determining their long-term survival.

Figure 3.22: Caveman graph initialization with high and low narcissistic agents only
3.4 Discussion

In the second component of our research, we addressed a key area that Skyrms’ model\textsuperscript{71} did not fully explore: the role of randomness in agents’ claims. In real life, people do not always act in a predictable way. Their actions can be influenced by their personal beliefs, values, or even by how they are feeling in the moment. These behaviors can be hard to predict, and they do not always follow the clear-cut strategies that Skyrms’ model suggests.

To better mirror this unpredictability in our study, we added an element of ‘randomness’ to the fair players’ strategy. This change lets us consider a wider variety of strategies, allowing for a more flexible and realistic representation of how people make decisions in the real world. By introducing this element of unpredictability, we could examine a broader range of potential outcomes and their impacts.
In the following sections, we will discuss the main findings of our study.

**Rational strategies take over the society over time.**

Our observations highlighted that over time, certain strategies led to better outcomes, and these strategies gradually took over the society. We found that strategies with claims near half or lower were more successful in the long term, while strategies with excessive claims tended not to endure in society. This may be due to the simple wisdom that actions driven by short-term self-interest (like claiming more than half) often lack rationality and therefore do not yield favorable results in the long run.

Therefore, to answer the question of how fairness evolves in a society with different fairness perceptions or strategies, we can say that within the scope of our simulation setup, certain strategies led to better results and remained prevalent in the society over time. These particular strategies were the more rational choices, and over time, these rational actions came to be known as fairness.

**Unfairness or Opportunism?**

One might ask, “should self-interested decisions be labeled as unfair”? In a situation where two symmetrically positioned players make different claims - one claims 40 and the other claiming for a payoff exceeding 50 - is the latter player acting unfairly or simply exhibiting opportunism? As it was mentioned in the literature review, D'Arms invites us to consider the potential benefits of allowing individuals to pursue their interests, rather than strictly adhering to a rule of equal division. However, we observed that the pursuit of personal gain at the expense of others is not a sustainable strategy.
Regardless of the network configuration, cooperative behavior is crucial for long-term survival.

Our study also highlighted the importance of cooperative behavior for long-term survival and growth within a networked society. While the structure of the network undeniably influences the dynamics of society, it is the behavior and strategy adopted by the individual agents within this network that significantly determine their survival.

We found that agents who consistently made excessive claims were eventually compelled to change their strategy in the long run. This was observed in both the Barabási-Albert (BA) and caveman network models we studied. These models, known for their distinctive characteristics - the BA model’s scale-free network leading to a small number of highly connected nodes, and the caveman model’s community structure - were observed to be challenging environments for agents who consistently made excessive claims. Thus, agents with high levels of narcissism were essentially forced to adjust their strategies and make lesser claims to ensure their survival.

**Implications for Human-Computer Interaction (HCI)**

From the perspective of HCI, our study emphasizes the importance of designing HCI systems, particularly collaborative platforms, in a way that fosters perceived fairness in resource allocation. This could be achieved by incorporating transparent decision-making processes, ensuring equal access to opportunities, or integrating mechanisms that encourage balanced contributions from all team members.

Additionally, fairness in the system, as our research shows, not only promotes cooperation but also contributes to the survival and growth of the networked
society. As such, collaborative platforms with fairness embedded in their design are likely to see increased user satisfaction, long-term engagement, and sustained productivity.

In conclusion, a deeper understanding of fairness, its implications, and how it has evolved in society should guide the design of HCI systems, particularly collaborative platforms. Embedding fairness into the design process could have profound positive effects on teamwork, cooperation, and productivity in remote work and online education environments.

### 3.5 Limitations and future work

Following are the key limitations of this study and potential future work:

**Simplified representation of agents**

The agents in this model may not capture the full complexity of human decision-making processes, as they are based on a limited set of predefined variables.

**Limited interactions**

The model considers only pairwise interactions between agents through a series of games. However, in real-world situations, individuals might engage in more complex interactions which involves multiple parties, communication, and indirect effects.

**Absence of external factors**

The model does not account for external factors that may influence agents' fairness perceptions and behaviors, such as cultural, social, economic, or political differences.
Future research directions

Adaptability and factors related to it (number of consecutive losses and strategy change) were added to the model as they are realistic aspects of human behavior and decision-making. However, the model was intentionally designed to emphasize the impact of fairness perception on behavior. Future research can explore scenarios where adaptability plays a more significant role or investigate the interactions between adaptability and fairness perception in greater detail.
Chapter 4

Conclusion

This thesis was a comprehensive exploration of the concept of fairness, in the context of human-computer interaction (HCI). Our study was divided into two major components. In the first, we delved into comparing human perception of unfairness in decisions made by humans and algorithms, with an intent to identify any possible differences between the two.

The second part of our study extended our investigation to a societal level, looking into how different fairness perceptions could shape societal evolution. Insights from this comprehensive investigation enhanced our knowledge about the complex nature of fairness and its implications for the design and implementation of HCI systems.

Future research and practice in the field of HCI should actively try to design algorithm systems that are not only fair in their operation but are also perceived as fair by those who use them. This means looking beyond the underlying algorithm and considering the complete user experience, which includes aspects such as interface design, transparency of operations, and user education.
In terms of interface design, the use of clear, easy-to-understand visual cues can impact a user’s perception of fairness. For example, representing algorithmic decisions in a visual manner can make the system’s workings more transparent to the user, leading to an increased sense of fairness and trust. Conversely, failing to communicate key information can lead to distrust and a perception of unfairness, even if the underlying algorithm is unbiased.

Transparency has also a crucial role in perceptions of fairness. People are more likely to trust and perceive an algorithm system as fair if they can understand its decision-making process. Explaining how an algorithm makes decisions in a simple and understandable way can be a challenging task, given the complexity of many algorithms; however, the more insight users have into how decisions are made, the more likely they are to trust the system and perceive its outcomes as fair.

Moreover, user education is another aspect of the user experience that can impact perceptions of fairness. Informing users about how algorithm systems work, their limitations, and potential biases can equip users with the knowledge to make informed judgments about the fairness of these systems. This can help to set realistic expectations and bridge the gap between the capabilities of algorithm systems and user perceptions.

While it is important to study the individual’s perception of fairness, as our societies become more interconnected and algorithm systems more prevalent, it is crucial to broaden our scope. Examining fairness at a societal level can provide insights into how collective perceptions of fairness are formed and how they evolve over time. This can be particularly important in the design of public facing algorithm systems (e.g. Google’s search algorithm), where decisions made by the algorithm can have widespread impacts.
This expansion of focus would also allow us to better understand how different social, cultural, and demographic groups perceive fairness, which in turn can help to identify and mitigate potential biases in algorithm systems. This societal level perspective could provide insights into group dynamics and norms that influence perceptions of fairness, potentially informing the design of algorithm systems that are more equitable and just.

Overall, the concept of fairness in HCI would benefit from both individual-centric approach and societal perspectives. Such a holistic approach can help us create the foundation for the design and implementation of algorithm systems that are not just technically fair, but are also perceived as fair and trustworthy by the individuals and societies they serve.
Appendix A

Task Explanation

A.1 Escape Room

A.1.1 Scenario

My supervisor (Dr. Tsang) is missing; You need to find his location in university through solving puzzles and teamwork.

A.1.2 Puzzles and set up

The room will consist of pictures, books, a flowerpot, a white board, some colorful gummies, a mirror and a speaker.

Puzzle 1: There will be a note on the white board saying:

SOLVE THIS PUZZLE FIRST IN ORDER TO GET A HINT AND UNCOVER A MIND BLOWING AND CREEPY SECRET ABOUT THE GAME DESIGN

Puzzle 2: Bold letters next to each other will form the following sentence: Read binary code

Puzzle 3: There will be a key under the flowerpot that players need to find it
in order to open a locked box that has the binary code inside it.

*Puzzle 4:* When they put the binary code together this word be made: Listen (010011000110100101110110011010100011001011011110)

*Puzzle 5:* From the previous puzzle participants will understand that they need to listen to the speaker. The speaker will be placed next to the mirror and will play a music (Breathe by Anna Nalick) which contains “breathe” in its lyrics.

[lyrics: And breathe, just breathe Oh breathe, just breathe]

*Puzzle 6:* Participants will need to breathe into the mirror in order to read a code which is written by a cotton swab and denatured alcohol; the code is: Red, Green, Blue.

*Puzzle 7:* There will also be a bunch of colorful gummies in the locked box on the table, participants need to count red, green and blue gummies in order to reach the code: 247.

*Puzzle 8:* There will be a picture of Dr. Tsang on the wall reading a book which will be placed in the room, players need to read page 247 of that book in which some words are highlighted.

*Puzzle 9:* Participants need to put first letters next to each other in order to reach the answer of the escape room: Tory.

### A.2 Oral Script

**Before the experiment:**

[Lead Researcher]: Hello and thank you for your participation. It is a 2-player escape room, so you need to work with your partner to solve the puzzles and find clues. During the experiment, you will be monitored by a Wi-Fi camera. If you wish to leave the experiment at any time, simply look at the camera and say something.
Human condition: There will be a human evaluating your performance in this escape room. You may earn up to $10 based on the evaluation and some other criteria.

Algorithm condition: There will be an algorithm evaluating your performance in this escape room. You may earn up to $10 based on the evaluation and some other criteria.

After the experiment:

Human condition: [Lead Researcher]: The decision is made by a human, you [the participant] get 2.93; we rounded it up to the nearest five cents, which is $2.95. You [the confederate] will receive 7.07, which will be rounded up to $7.10. The decision was made based on your performance and the information you provided in the survey before the experiment.

Algorithm condition: [Lead Researcher]: The decision is made by an algorithm, you [the participant] get 2.93; we rounded it up to the nearest five cents, which is $2.95. You [the confederate] will receive 7.07, which will be rounded up to $7.10. The decision was made based on your performance and the information you provided in the survey before the experiment.
Appendix B

Questionnaires

B.1 Attitude towards AI questionnaire

Please read every statement carefully and select the answer that applies to you best.

(a) There are many beneficial applications of Artificial Intelligence.
   • Agree
   • Disagree
   • Neutral

(b) Artificially intelligent systems can perform better than humans.
   • Agree
   • Disagree
   • Neutral

(c) The rise of Artificial Intelligence poses a threat to people’s job security.
• Agree
• Disagree
• Neutral

(d) I am interested in using artificially intelligent systems in my daily life.

• Agree
• Disagree
• Neutral

(e) Artificial intelligence is used to spy on people.

• Agree
• Disagree
• Neutral

(f) I think Artificial Intelligence is dangerous.

• Agree
• Disagree
• Neutral

(g) An artificially intelligent agent would be better than an employee in many routine jobs.

• Agree
• Disagree
• Neutral
(h) Some complex decisions are best left to artificially intelligent systems.

- Agree
- Disagree
- Neutral

(i) Artificial intelligence is limited on its abilities.

- Agree
- Disagree
- Neutral

(j) I think artificially intelligent systems make many errors.

- Agree
- Disagree
- Neutral

B.2 Questionnaire that was provided after the task

Please answer the following questions to the best of your knowledge. Do not include your name anywhere in the questionnaire.

(a) I believe the decision maker made a fair decision.

- Agree
- Disagree
- Neutral
(b) Please explain the reason of your answer to the previous question.

(c) Would you play an escape room with the same rules and in case of evaluation with the same decision maker in the future?
   - Yes
   - No

(d) Please explain the reason of your answer to the previous question.
Appendix C

Ethics Clearance Certificate

Office of Research Ethics
400 ARRE Building 1125 Colonel By Drive
Ottawa, Ontario K1S 5B6
613-520-2600 Ext. 4065
ethics@carleton.ca

CERTIFICATION OF INSTITUTIONAL ETHICS CLEARANCE

The following research has been granted clearance by the Carleton University Research Ethics Board (CUREB). CUREB is constituted and operates in compliance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS2).

Ethics Clearance ID: Project # 118331

Project Team Members: Faezeh Ataeizadeh (Primary Investigator)
Arun Tsang (Academic Supervisor)

Study Title: Human perception of unfairness and trust in decisions made by algorithms versus humans

Funding Source: (If applicable):

Effective: November 03, 2022   Expires: November 30, 2023

This certification is subject to the following conditions:

1. Clearance is granted only for the research and purposes described in the application.
2. Any modification to the approved research must be submitted to CUREB via a Change to Protocol Form. All changes must be clarified prior to the continuance of the research. An Annual Status Report for the renewal or closure of ethics clearance must be submitted and cleared by the renewal date listed above. Failure to submit the Annual Status Report will result in the closure of the file. If funding is associated, funds will be frozen.
3. During the course of the study, if you encounter an adverse event, material incidental finding, protocol deviation or other unanticipated problem, you must complete and submit a Report of Adverse Events and Unanticipated Problems Form.
4. It is the responsibility of the student to notify their supervisor of any adverse events, changes to their application, or requests to renew/close the protocol.
5. Failure to conduct the research in accordance with the principles of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans 3rd edition and the Carleton University Policies and Procedures for the Ethical Conduct of Research may result in the suspension or termination of the research project.

Upon reasonable request, it is the policy of CUREB, for cleared protocols, to release the name of the PI, the title of the project, and the date of clearance and any renewal(s).

Please email the Research Compliance Coordinators at ethics@carleton.ca if you have any questions.

CLEARED BY:   Date: November 03, 2022
Natasha Artemeva, PhD, Chair, CUREB-8

Bernadette Campbell, PhD, Chair, CUREB-8
Bibliography


[41] Hung-Yue Suen, Mavis Yi-Ching Chen, and Shih-Hao Lu. Does the use


