Integrating fairness in opponent modeling

Ilinca Trestioreanu
Supervisor(s): Luciano Cavalcante Siebert, Sietze Kai Kuilman
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Is there a way to incorporate fairness in the opponent modeling component of an automated agent? Since opponent modeling plays an important role in a negotiation strategy, it is reasonable to research how fairness can be integrated into this component, as it influences the outcome of the negotiation. A first step towards finding an answer to this question is to define fairness and how this definition can be translated to algorithmic fairness. The next step is to investigate already available opponent models and assess whether their strategy can be considered fair or not. This paper analyses the process of Bayesian learning in the context of opponent modeling, and tries to reveal possible flaws or strengths that the model has embedded in it, with the hope to add relevant information in the area of automated negotiation.

1 Introduction

Negotiation has its place in contexts such as economics or psychology, but also in the area of artificial intelligence, in which the aim is to automate the aforementioned processes. The common goal in automated negotiation is to achieve the best outcome possible for the parties involved while maximizing their utility, but what does the best mean in this context? There are different plausible answers to this question, but an intuitive response is that the best outcomes are the fairest outcomes. Since there are many contexts in which negotiation can take place, the notion of fairness varies accordingly. Despite this, by achieving adaptability and by integrating fairness, automated negotiation can achieve user trust, and consequently, it can reduce the time spent on such matters, costs and cognitive effort (Barslaag, Kaisers, Jonker, Gerding, & Gratch, 2020).

In a negotiation between two or more agents, the goal is to reach a maximum gain for the represented party, while keeping their preference profile private to avoid exploitation. This represents "a major challenge" in the area of automated negotiation (Baarslag, Hendrikx, Hindriks, & Jonker, 2013), since the lack of information about the preferences of the opponent may lead to a sub-optimal agreement for the parties involved. To mitigate this, the agents can build an opponent model to estimate the preferences of the opponents, which leads to optimizing the overall negotiation. Moreover, the availability and quality of the information gained about the opponent are vital to obtaining an efficient negotiation (Hindriks & Tykhonov, 2008). This shows that the opponent modeling component plays an important role in an automated negotiation, and leads to the question of how fairness can be integrated by way of opponent modeling within an automated negotiation.

1.1 Background and related work

The topic of fairness and its application in this area has already started to be researched. There have been attempts to study multiple definitions of fairness, such as (Verma & Rubin, 2018), which resulted in obtaining a more intuitive understanding of fairness, and in proposing alternatives for future work. Moreover, research has been done regarding how could fairness be measured and what is its form at an individual or group level. Studying ways to measure fairness can help researchers "anticipate and mitigate fairness-related harms arising from computational systems" (Jacobs & Wallach, 2021). There is vast research on fairness and automated negotiation, and they serve as a good foundation for

understanding and exploring these topics. Nevertheless, there is a lack of research on how can fairness be incorporated into the strategy of the opponent model of an automated agent.

Bias in artificial intelligence is a subject that is discussed more and more nowadays, and this is due to the role that such intelligent systems must fulfill in the area of decision-making problems. In the context of automated negotiation, bias can occur in opponent modeling, as the learning phase from this component may lead to biased behavior. The notion of bias is a phenomenon that occurs when machine learning algorithms often produce erroneous results, that show a discriminating behavior or reflect unfairness. At the same time, bias can cause the AI system to act sub-optimally, and lead to undesired outcomes, that does not benefit the user (Roselli, Matthews, & Talagala, 2019). In close relation to bias, unfairness in an algorithm describes similar issues that arise when it does not guarantee the "absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021). This means that when the system is developed in such a way that its behavior is prejudiced towards a group of individuals that hold some specific set of characteristics, it leads to unfair or unjust results. This is an important problem because it affects the advancements of artificial intelligent programs in real-world applications, and more specifically, it affects the trust relationship between humans and computers. Bias can occur for many reasons, such as when in classifiers the used training data set is imbalanced, that is, the number of observations for all classes is not uniformly distributed. This can happen for two reasons, the majority class is under-sampled, or the minority class is over-sampled (Chakraborty, Majumder, & Menzies, 2021). Another example would be the presence of protected attributes, that represent individual characteristics such as gender, race, or social status, and these kinds of attributes can lead to skewed results.

Regarding the structure of this paper, in the following section, we will discuss the importance of the opponent model component and what value it adds to automated negotiation. In section 3, I will present different perspectives on fairness which will help in creating a more specific context for the main research question and will provide a good reference point for fairness in the context of automated negotiating systems. In section 4, I will present a strategy used in the construction of the opponent model, and this will complete the context in which I will be researching and analyzing if there is a way to integrate fairness in this component. Followed by that, section 4.1 will present an analysis of the chosen model strategy, and in section 5 I will present arguments as to what are some aspects of the strategy that indicate unfairness. Lastly, in sections 6, 7 and 8 I will discuss more extensively the arguments brought in the previous sections, I will present why this work follows the guidelines of responsible research and then I will give ideas for future research on this topic and will conclude this paper.

2 Importance of the Opponent Model

The BOA framework, presented in (Baarslag, 2016, ch. 3), is an architecture that can be incorporated in the construction of an automated negotiating agent and it is based on three main strategies that are commonly used in a negotiation context: **B**idding strategy, **O**pponent model and **A**cceptance strategy. Although the opponent model component may not be present in the strategy of some agents, it has been shown that "reasonable" knowledge of the opponent's preference profile is required to approximate the Pareto frontier and,

thus, achieve Pareto efficiency (Hindriks, Jonker, & Tykhonov, 2009, p. 2). As mentioned earlier, typically in an automated negotiation, agents keep their preference profile private to avoid taking advantage for selfish purposes and in order to reach an efficient negotiation that maximizes the outcome of the parties involved, there is the need to estimate the opponent's profile. The solution to this problem is to use some kind of algorithm that does exactly that.

Many different such algorithms can be used in the opponent model strategy, and until 2012, there has not been done much research regarding which strategy is better than the other. In (Baarslag, Hendrikx, Hindriks, & Jonker, 2012), valuable extensive research is performed regarding what types of models can be used to extract information about the adversary party, and measurements are performed to test different possible theories about the benefits of using the opponent model component. In (Baarslag et al., 2012), it is demonstrated the need and benefits of using opponent models, by using various types of agents and by creating different negotiations between them, while observing their behavior and analyzing the outcome of the negotiations. The following three hypotheses were validated in (Baarslag et al., 2012, p.9-11) and show that competitive opposing agents and high sizes of bid spaces or bid distributions are factors that lead to the need of implementing an opponent model, to reach good gains in a negotiation.

- *H3. The more competitive an agent, the more it benefits from using an opponent model.*
- *H4. An agent benefits more from an opponent model against competitive agents.*
- *H6. The higher the amount of bids, bid distribution, or opposition of a scenario, the more an agent benefits from using an opponent model.*

3 Perspectives on fairness

It is useful to first discuss some important philosophical approaches to fairness, as they not only intertwine with the foundations of computational social choice but also with the duty that we, computer scientists, have a sense of what is moral and a sense of how the algorithms that we develop affect humans and the reality in general. Questioning what is fairness is a result of "centuries of oppression of people based on race, gender, and social class" (Cerbone, 2021, p. 1), and looking for answers in the philosophy area can lead to useful answers, that may serve as good reference points when it comes to deciding what is fair and what is not in computational problems. I will then present two types of fairness, distributive and procedural fairness, which are more abstract and which will convey more specifically the methodology of this research.

3.1 Rawls vs Nozick

John Rawls and Robert Nozick were two influential American philosophers, in the 20th century. They both contributed with theories about the role of the government and the responsibilities of individuals (Wilkerson, 2012). Rawls presented his view about justice in (Rawls, 1971), and he recommended equality of opportunity among individuals and maximizing the benefits of the most disadvantaged members of society. The latter is also known

as the Difference Principle, and it essentially transmits that any possible inequalities of social or economic type should be in favor of groups of individuals that are least advantaged, and, nowadays, this Rawlsian perspective is one of the most chosen approaches in algorithmic fairness (Franke, 2021). He proposed a thought experiment, which refers to the "Veil of ignorance" reasoning, in which decision makers are deprived of information such as their social status, race, gender, or health. As a consequence, this will lead people into making impartial decisions and think more rationally. Another perspective that he previously presented in (Rawls, 1957) was that in a society in which the members of it create rules under which they are to be judged, the other remaining members shall be judged using the same rules and this should be considered as fair. Translating this notion to the computer science area, when an algorithm is chosen as a metric by some group of people and the algorithm attempts to assign roles to each individual from that group, this must be a fair result that demonstrates justice. (Franke, 2021).

In response to this theory, Robert Nozick presented a different perspective on the same matter. He argued that if individuals agree upon a certain distribution of goods and if they freely participate, then this process can be viewed as fair or just, despite other inequalities that might arise. His view supports the idea that any individual is entitled to what they currently possess, which is in contrast to Rawls's view on the problem of the distribution of goods. Although the implications of this perspective are in some contexts harmful, Nozick's view does not deny the presence of unjust inequalities, such as race or social status, and focuses more on the idea that if present injustices are dealt with, then "the problems of the past will eventually disappear" (Wilkerson, 2012).

The analysis of the Bayesian learning strategy, from section 5, uses a Rawlsian approach, due to its benefits of providing a fair way to distribute goods among people who have distinct morals. Additionally, it provides an objective process of decision making and it is not biased towards a particular social norm or moral perspective, as presented in (Doorn, 2009). Similarly to humans, agents can have different strategies and values, and for this reason, the Rawlsian perspective serves as a solid reference point to the judgment of what is fair and what is not, pertaining to the process of opponent modeling.

3.2 Distributive vs procedural fairness

In the context of automated negotiation, the user can question how fair is the outcome, but also how fairly did the opponent play. Distributive and procedural fairness are two aspects of social fairness, which rely on social norms (Ferguson, Ellen, & Bearden, 2013), and they indicate which parts of a negotiation process are subject to the investigation of fairness.

Distributive fairness represents the comparison between the outcomes of the parties involved in a negotiation. Although equity theory suggests that all members should have a fair and just outcome (Ferguson et al., 2013), distribution can lead to unequal payoffs for the involved parties (Diekmann, Soderberg, & Tenbrunsel, 2013). Additionally, it has been shown that in many negotiation contexts, in which inequality is unavoidable, individuals tend to opt for a more favorable outcome compared to the opposing party's outcome and this is viewed as "egocentric fairness bias" (Diekmann et al., 2013). To quantify distributive fairness in an automated negotiation, the Nash bargaining solution represents a well-known metric to estimate the optimal solution, in which social welfare is maximized (Fujita, Ito, &

Klein, 2010).

Procedural fairness represents the degree to which the processes that are used to arrive at an agreement are consistent and are inclusive of the interests of all members who take part in the negotiation (Ferguson et al., 2013). In a physical negotiation, procedural fairness is often related to the voice, neutrality, respect, and trust that is used throughout the entire process (Diekmann et al., 2013). In an automated setting, procedural fairness concerns the negotiation strategy, more specifically, the bidding, acceptance, and opponent modeling strategy. This type of fairness is harder to quantify, and this paper will focus on the opponent model process, to establish good practices concerning procedural fairness.

4 Opponent Model

When building an opponent model, the construction of it revolves around three concerns: what does the opponent want, what will the opponent do and what type of player is the opponent (Baarslag, Hendrikx, Hindriks, & Jonker, 2016). The first two questions imply that an agent aims to find out what are the true preferences and what are the next moves of the adversary party. The last question is related to classifying the opponent, that is, the agent tries to estimate an inclusion relation between the opposition party and a certain class. The difference between the last question and the other ones is that the accuracy of the classification impacts the efficiency of the negotiation. Namely, if an agent classifies the opponent wrong, then the agreement of the parties involved may not be as beneficial as possible, due to subjective decisions of the agent, which can be tailored to a specific type of opponent. The Bayesian learning technique, a supervised machine learning algorithm, tries to establish a likelihood function between the preference profile of the opponent and certain classes which hold multiple preference relations (Buffett & Spencer, 2007). There are other classification methods used in this context, such as frequency models and value models, which aim to estimate the issue and value weights of the opponent's profile, based on how often bids with certain issues and values are placed (Baarslag et al., 2013). Since the Bayesian technique seems to be widely used (Baarslag et al., 2012, p. 9), this paper aims to answer the question of how fair is the process of Bayesian learning in an opponent modeling strategy.

4.1 Strategy Analysis

The Bayesian learning approach can be used for different purposes in the opponent model. For example, in (Lin, Kraus, Wilkenfeld, & Barry, 2006), the model assumes a fixed possible set of profiles in a specific domain that the opponent can have and the Bayesian technique is used to estimate the likelihood of the opponent's membership in one of those classes. In (Zeng & Sycara, 1998), the Bayesian strategy is used to approximate the opposing party's reservation value, which is the minimal utility value below which other offers will not be accepted. In this paper, we will examine the general Bayesian model proposed in (Hindriks & Tykhonov, 2008), which aims to learn the issue preferences and issue priorities, without using a fixed set of preference profile classes. The agent can have some knowledge about the opponent in certain negotiations, but this model works as efficiently with incomplete information.

Table 1: The hypothesis space composed of three function types (Hindriks & Tykhonov, 2008)

Function type	Definition
Downhill shape	Minimal issue values are preferred over other issue values, and the evaluation of issue values decreases linearly when the value of the issue increases
Uphill shape	Maximal issue values are preferred over other issue values, and the evaluation of issue values increases linearly when the value of the issue increases
Triangular shape	A specific issue value somewhere in the issue range is valued most, and evaluations associated with smaller and bigger issue values linearly decrease

4.1.1 Assumptions

In this strategy, there are two kinds of assumptions that are to be made about the opponent party, and this is done to maximize the chance of reaching an estimation that reflects the ground truth. The first type of assumption is concerning the structure of the opponent's preference profile, while the second type is related to the opponent's negotiation strategy, also called rationality assumptions.

The opponent model assumes that the profile of the opposing agent has a linearly additive profile, meaning that the utility of a bid can be calculated by a weighted sum of evaluation functions and priorities associated with each issue, using the following formula:

$$u(b_t) = \sum_{i=1}^{n} w_i * e_i(x_i \in b_t) \tag{1}$$

in which $u(b_t)$ is the utility of a bid at round t, w_i is the weight associated with an issue and $e_i(x_i)$ is the evaluation function of the value of issue i.

The next assumption that the model makes, is regarding the hypothesis space that is used in the Bayesian technique, more exactly, it is assumed a fixed set of possible evaluation function types, as defined in table 1. What is interesting about these three types of functions is that they serve as a mathematical way to estimate other function shapes, by assigning different probability distributions and by using a "composition of several simple evaluation functions from the hypothesis space" (Hindriks & Tykhonov, 2008, p. 4). Thus, the agent can estimate more complex behavior and is not restricted to a prefixed set of preference profiles, as the Bayesian Strategy presented in (Lin et al., 2006).

So far, I have described the structural assumptions that are used in this strategy. There are two rational assumptions made, about the general negotiation strategy of the opponent. The premises state that the opponent follows a concession-based strategy and uses a time-dependent tactic. This means that the opponent starts with the highest utility offer and towards the deadline of the negotiation, it approaches its reservation point. These types of assumptions are also highlighted in (Baarslag et al., 2012, p. 4), as a result of an analysis of different opponent models and their assumptions. The reasoning behind these presumptions

is that the agent needs to make an educated guess about the opponent's behavior, and it represents a reference for the observed data. Many opponent model strategies use these assumptions, and it can happen that in some cases they do not lead to the correct result, which in turn, does not reflect the reality. Although the Bayesian strategy discussed uses these assumptions, it is argued that the model does not exactly presume an exact behavior, and permits for a different range of tactics, by using probability distributions assigned to a range of tactics. As a consequence, these assumptions are used to calculate a possible utility of the bid made by the opponent, and the utility calculated gets assigned a certain probability.

These assumptions will be revisited in section 5 and will help in the investigation concerning the fairness of this technique.

4.1.2 The Bayesian Learning Algorithm

Bayesian learning starts with assigning a uniform distribution to the elements in the hypothesis space, and if there is any available information about the opponent, it then uses a probability distribution that fits the known data. The heart of the algorithm is based on each received bid from the opponent, and these bids are used to find the hypothesis which most likely describes the behavior of the opposing party. Bayes rule represents the core mechanism, and it is used to integrate the information regarding the bids, by updating the probability of a certain hypothesis, given an offer. Equation 2 illustrates the Bayes theorem, which uses prior available information regarding the hypothesis space, and which affects the final result of the process. The algorithm will increase the likelihood of a hypothesis that best reflects the offers received, and thus, the agent can better estimate the utility of the opponent's bid. The increase in the probability of a certain hypothesis can also be used as a measure of how well the agent performs, as the hypotheses that do not align with the observed behavior will decrease in time.

$$P(h_j|b_t) = \frac{P(h_j) * P(b_t|h_j)}{\sum_{i=1}^{m} P(h_k) * P(b_t|h_k)} \tag{2}$$

5 Unfairness

This section is centered around the idea of establishing whether the Bayesian learning opponent model strategy is a fair process with respect to the opposing party and the social norms involved. This will be done by assessing the assumptions presented in the previous section, and by discussing the possible presence of bias in this algorithm. The algorithmic fairness will rely on the Rawlsian approach.

5.1 Settings of the model

The negotiating strategy can be influenced to a great degree by the presence of information about the opponent's preference profile or strategy, as exploitation is the main concern, and this applies to all types of opponent models. Sharing preference profiles may only lead to acts of selfishness and will minimize the social welfare of the negotiation and, in turn, the utility of the opponent's outcome. From this perspective, it seems as though the social welfare value affects the level of procedural fairness, in the sense that a lower social welfare value leads to an unfair negotiation, while the vice versa is also true. As the social welfare value

takes into account the utilities of the accepted offers of the parties involved when agents act selfishly, the welfare value will decrease, due to one party having maximized utility, and the other party having a minimized utility. Such an outcome can reflect the procedural unfairness of the negotiation Additionally, this value can be influenced in parallel by other factors, such as the bidding and acceptance strategies that are used, which indicates that when assessing the results at the end of the negotiation, all causes need to be taken into account. The aforementioned arguments indicate that, when building an opponent model, the way to reach a fair negotiation process is to not share the parties' preference profiles or other information regarding their strategies, unless some external factor or agreement between parties imposes such behavior.

The structural and rational assumptions presented in the previous section represent the core values of the agent. All the discussed structural assumptions are needed in the learning process, as they give meaning to the observed data during the negotiation, and without them, the process may become ineffective and lead to unwanted results. The assumption about the hypothesis space reflects that the agent is not relying on some exact information about the opponent. This idea promotes fairness and unbiasedness, in the sense that the agent does not assume an exact opponent's behavior which leads to more objective and more accurate estimates. The rational assumptions work in the same direction. As described in section 4.1, the agent does not assume that the opponent follows a specific time-dependent or concession-based tactics. The implementation leaves room for other kinds of tactics, by using probability distribution which in turn will show the likelihood of all tactics. By making an analogy between the act of making certain presumptions about an individual's race, gender, or social class, and the act of assuming certain tactics about the opposing party, and by making a connection with the Rawlsian perspective on fairness, the assumptions that are used in this learning phase can be seen as fair and just.

On the other hand, the Veil of Ignorance and the Difference principle discussed in section 3.1, imply that in the problem of distribution of goods, the humans involved would need to be aware of who is the most disadvantaged to arrive at a fair outcome, while discarding sensitive inequalities. Analogously, agents would need to know, in a negotiation setting, who is the "worst-off" player, so that optimal bids are placed, with the goal to reach a beneficial agreement and an efficient negotiation. While this may entail sharing information among agents, such as preference profiles, the use of opponent modeling tackles the problem of knowing who is the least advantaged player, by trying to estimate what is the opponent's preference profile or what strategy is the opponent using. The argument brought to the table is that instead of sharing prior knowledge about the parties, the agent can make use of the opponent model and become aware of what are the disadvantages of the parties involved. By exploiting the benefits of the opponent model, procedural fairness can be increased by taking into account the adversary party's disadvantages, while avoiding the possible negative effects of sharing knowledge between parties. Nevertheless, this would only work if the estimations of the opponent model are as close as possible to reality, that is the agent can approximate accurately what are the values or strategies used by the opponent.

5.1.1 Bias

The Bayesian learning strategy in opponent modeling uses data from the received bids for training purposes during the negotiation rounds, and, additionally, it can make use of data obtained from previous encounters with other agents to train the agent before the start of the negotiation. In the latter case, the training data alters the behavior of the agent, in a way that makes the negotiation effective only when the agent has encountered before the strategy of the current opponent, and this leads to hidden biases which may not favor the involved parties. In the case of the strategy discussed in section 4.1, the settings of the model can lead to biased results, because they have a direct impact on which type of parties are privileged, and, in this circumstance, the agent will only respond with fair counter-offers if the behavior of the opponent fits in a specific pattern that the agent can recognize and respond to. The rational assumptions impose hard restrictions on the effectiveness of the model, meaning that parties that do not follow a concession-based and time-dependent strategy will not benefit from a negotiation with an agent that assumes these kinds of tactics. In (Hindriks & Tykhonov, 2008), it is stated explicitly that these assumptions are not realistic in certain scenarios, and it is explained how the model does assume a range of tactics under the umbrella of time-dependent strategies, but this does not mitigate the general rational prejudice that is used.

In line with the perspective of algorithmic unfairness presented at the beginning of this section, an analogy can be made between this view and the posed rational assumptions. In essence, the Bayesian model does not guarantee effectiveness and successful estimations regardless of the characteristics of the opponent, and this can lead to unfair results and an overall unfair negotiation process. This happens because the agent's behavior is tailored towards rational players, which causes the agent to inaccurately estimate the model of the opponent.

6 Discussion

The arguments presented so far indicate that the Bayesian learning strategy poses some unfair constraints on the model, and this leads to an unfair process of negotiation. The key implication of this conclusion is that many other opponent models use the same type of assumptions, as discussed in (Baarslag et al., 2012, p. 4), which means that they also hide the same type of unfairness. Some models use even more fixed assumptions about the opposing party, as seen earlier in (Lin et al., 2006), in which the strategy leads to desired outcomes only if the requirements of the model are met. Using fixed constraints on the model may lead to effective and useful agreements, but in other cases, it does not, and this may be a relevant argument as to why fully autonomous negotiation agents are not predominantly used in the area of real-world decision-making and negotiating contexts.

On the other hand, there are clear explanations given regarding the need for such assumptions, and most of them have mathematical implications. What this entails is that the algorithm needs a way to interpret the data obtained from the offers made by the opposing party, but this does not necessarily reflect a fair process of arriving at a beneficial agreement. Moreover, one could argue that, in general, unfounded assumptions reveal biases and unjust mindsets, and this is a similar perspective to the one presented in section 5.

The rational assumptions presented in section 4.1.1, impose hard constraints on opponent models, and as argued in section 5, they may decrease the overall procedural fairness of the agent. In (Baarslag et al., 2012, p. 9-10), it is shown how these rational assumptions in Bayesian models have a negative effect in negotiations with agents who do not follow a concession-based strategy or a time-dependent tactic. In contrast, it is also demonstrated that Frequency models, which do not use such assumptions, can perform efficient negotiations with a larger range of types of opponents. The downside of the frequency models is that they do assume that the opponent will offer more frequently bids that are higher valued, thus, they also hide a certain level of procedural unfairness. From my perspective, I think that the assumptions of the Bayesian model can be adjusted so that they do not pose hard constraints on the negotiation strategy, and, although the following proposition is not tested, nor based on actual evidence, it may lead to possible relevant future work. If an agent uses the aforementioned rational assumptions, then it may be beneficial to not rely confidently on the fact that the opponent follows these negotiation strategies. What this means is that the agent may keep the rational assumptions, but when confronted with a different type of player, it should become aware that the adversary agent is not a rational player. Its next step should be to adapt and follow a different negotiation strategy that does not assume anything about the player, similar to the frequency models. My proposition suggests a middle ground between opponent models like Bayesian and Frequency models. The reason for keeping the rational assumptions in the agent's model is that when the agent's opponent follows these rational assumptions, then the negotiation is more efficient and more beneficial for the final agreement.

7 Responsible Research

This research is concerned with establishing whether an algorithm leads to fair automated negotiations, with the hope to bring relevant and useful information that may help in the process of creating autonomous agents who value morality and promote fairness. Additionally, this research did not involve development, only assessing and arguing if a strategy of opponent modeling is using settings and values that may or may not lead to a fair process while relying on relevant work performed by other scientists.

8 Conclusions and Future Work

Opponent modeling plays an important role in automated negotiation, mainly due to its benefits of optimizing the negotiation and increasing the chance of arriving at a final agreement that is beneficial for the parties involved. Since it is our duty, as computer scientists, to implement algorithms that take into account what is moral and what is right with respect to a set of social norms, it is then relevant to research how we can integrate fairness in opponent modeling. The algorithmic fairness in this paper relies on the Rawlsian approach because it tackles the problem of the distribution of goods among a group of people who may have different values and characteristics, which can lead to inequalities. Since there are many types of opponent models, it is useful to investigate how fair is the process on these strategies individually, as it may reveal general problems or strengths of such models. This paper analyzed the Bayesian learning strategy, by taking a closer look at the settings of the model and at the overall approach of this algorithm.

Key aspects of the Bayesian learning algorithm were highlighted, such as the rational and structural assumptions, which enforce the agent to negotiate efficiently with players that satisfy these assumptions. It was argued that these settings affect negatively negotiations with agents who violate these requirements, and this leads the system to be unable to negotiate optimally with these types of parties. It was also rationalized that these assumptions reveal biases that the Bayesian model has, even though the hypothesis space is designed to support a range of tactics under the umbrella of rational assumptions. These arguments indicate that the Bayesian learning technique holds unfair aspects, that lead to an unfair negotiation process, and ultimately, they decrease the overall procedural fairness of this strategy.

This research did not attempt to measure the fairness in the Bayesian model, and this represents an important task for future work, which will motivate the creation of other opponent models that integrate fairness. There is still an open discussion if other available opponent models use a fair strategy, and if so, what aspects lead to a fair process of negotiation. This open discussion generates possible areas for future work, which will benefit the investigation of fairness in opponent models, and will help in developing techniques that give rise to a fair process of negotiation, and consequently, to fair outcomes.

References

Baarslag, T. (2016). *Exploring the strategy space of negotiating agents.*

Baarslag, T., Hendrikx, M. J., Hindriks, K. V., & Jonker, C. M. (2012). Measuring the performance of online opponent models in automated bilateral negotiation.

Baarslag, T., Hendrikx, M. J., Hindriks, K. V., & Jonker, C. M. (2013). Predicting the performance of opponent models in automated negotiation.

Baarslag, T., Hendrikx, M. J., Hindriks, K. V., & Jonker, C. M. (2016). A survey of opponent modeling techniques in automated negotiation.

Barslaag, T., Kaisers, M., Jonker, C. M., Gerding, E. H., & Gratch, J. (2020). When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators. , 4684 -4689.

Buffett, S., & Spencer, B. (2007). A bayesian classifier for learning opponents preferences in multi-object automated negotiation. *Electronic Commerce Research and Applications*.

Cerbone, H. (2021). Providing a philosophical critique and guidance of fairness metrics.

Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in machine learning software: Why? How? What to do?

Diekmann, K. A., Soderberg, A. T., & Tenbrunsel, A. E. (2013). Fairness and ethics in bargaining and negotiation.

Doorn, N. (2009). A rawlsian approach to distribute responsibilities in networks.

Ferguson, J. L., Ellen, P. S., & Bearden, W. O. (2013). Procedural and distributive fairness: Determinants of overall price fairness.

Franke, U. (2021). Rawlsâs original position and algorithmic fairness.

Fujita, K., Ito, T., & Klein, M. (2010). A secure and fair protocol that addresses weaknesses of the nash bargaining solution in nonlinear negotiation.

Hindriks, K., Jonker, C. M., & Tykhonov, D. (2009). The benefits of opponent models in negotiation.

Hindriks, K., & Tykhonov, D. (2008). Opponent modelling in automated multi-issue negotiation using bayesian learning.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. *FAccT 21 Virtual Event, Canada*.

Lin, R., Kraus, S., Wilkenfeld, J., & Barry, J. (2006). An automated agent for bilateral negotiations with bounded rational agents with incomplete information.

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

Rawls, J. (1957). I. justice as fairness.

Rawls, J. (1971). *A theory of justice*.

Roselli, D., Matthews, J., & Talagala, N. (2019). Managing bias in ai.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *International Workshop on Software Fairness*.

Wilkerson, J. (2012). Rawls and Nozick on fairness.

Zeng, D., & Sycara, K. (1998). Bayesian learning in negotiation.