The Unintended Consequences Fairness Brings to Automated Negotiation

Nick Ouwerkerk
Supervisor(s): Luciano Cavalcante Siebert, Sietze 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

In this paper, the unintended consequences, also named edge cases in this paper, of integrating fairness into the automated negotiation process are researched. By finding these unintended consequences, we can deal with them accordingly or avoid them, as to not cause any problems with our fairness metric that might make our negotiation process less fair or cause any undesired behaviour. Edge cases are searched for in a small-scale experiment by implementing the difference principle from the 'Justice as Fairness' notion by John Rawls into the negotiation process. The negotiation has two agents, and the behaviour of one of the agents is changed to adhere to the difference principle. By using these agents in automated negotiations with different domains, the behaviour and outcomes of these negotiations between two agents will be checked for any abnormalities that could be considered an unintended consequence. From this, we conclude that agent implementing fairness has a smaller available bidding space which leads to a staler negotiation process. Furthermore, the outcomes show that not always an optimal result is found. However, no unintended consequences directly related to fairness were found. Since finding edge cases is an exhaustive process, which can be compared to finding bugs in a computer program, the research done is not proof that the Rawlsian notion of fairness, or any kind of fairness for that matter, has no other unintended consequences. The research in this paper can be used as inspiration for further research into the edge cases of fairness in automated negotiation, and to have a general idea as to what unintended consequences Rawlsian fairness brings.

1 Introduction

Fairness plays a critical role in negotiation. By having some notion of fairness integrated into the negotiation process we increase the chance of all parties of the negotiation being satisfied with the result, create a healthy negotiation environment for parties of any background and increase the trust between parties for current and future negotiations. Fairness also impacts the negotiation process. Nancy A. Welsh notes that "negotiators rely upon their perceptions of distributive and procedural fairness in making offers and demands, reacting to the offers and demands of others, and deciding whether to reach an agreement or end negotiations" [1, p. 753]. Since fairness influences both the negotiation process and outcome, automated negotiation should use fairness too to imitate this behaviour. Furthermore, T. Baarslag *et al.* argue that one of the major challenges in having autonomous negotiators be able to perform negotiations in real-world applications is user trust [2]. For a user to relinquish control to an autonomous agent, it should properly represent the user, which would include their idea on the definition of fairness. This makes it important that a notion of fairness is properly integrated into these autonomous agents.

Fairness is a concept that has been discussed and studied for a long time, and because of this, we have come to realise there are many interpretations of what it means to be fair. This has resulted in a large library of literature on the theory of fairness in all kinds of scientific fields like philosophy, economics and game theory. Important relevant work to the theory of fairness includes John Rawls' *A Theory of Justice* [3], Robert Nozick's *Anarchy, State, and Utopia* [4] and Hobart Peyton Young's *Equity* [5]. Among other things, these works show that fairness is a multi-dimensional concept, and can be defined and interpreted differently by different individuals.

However, the study on fairness in automated negotiation is quite sparse. As alluded to earlier, research on the role of fairness in 'normal' negotiation shows how the perception of fairness of the negotiators has a major impact on the negotiation process [1], and C.

Albin even suggests that "notions of fairness may create a motivation to resolve a particular problem through negotiation in the first place, and thus have an impact on the positions and expectations which parties bring to the table." [6, p. 223]. With the rise of the digital age we have seen a necessity in researching fairness in relation to the domain of computer science for more fair algorithms and programs, like fairness in classification [7], but also increasing interest and advancements in automated negotiation, autonomous agents and their use-cases [2, 8, 9]. Since fairness can heavily alter a negotiation process, the lack of research on fairness in automated negotiation creates a problem in making and using autonomous agents for negotiation that can truly represent their human counterparts.

When implementing fairness into automated negotiation, we should be aware of the unintended consequences our implementation of fairness brings. An unintended consequence, also referred to as edge cases in this paper, of fairness is any peculiar behaviour in the negotiation process or the outcome of the negotiation when a form of fairness is used to some degree in the negotiation process. These unintended consequences could put the negotiation process in an undesirable situation. In the worst case, these vulnerabilities could lead to exploitation by a malicious party for their personal gain when not taken care of. This would contradict the general idea of implementing fairness since these unintended consequences could lead to unfair results. In this case, by implementing fairness, we have made the negotiation unfair and this should be avoided at all costs. As an example of an unintended consequence, let's look at equality as a notion of fairness, where in the context of (automated) negotiation all participating parties should roughly end up with the same outcome. An unintended consequence could be that all parties ended up with sub-optimal results, meaning that all parties could have gotten more out of the negotiation, even when this would make the outcome minimally less fair.

The purpose of this paper is to search for some of the unintended consequences when integrating fairness in autonomous agents through a small-scale investigation in an automated negotiation process. In particular, the Rawlsian notion of fairness will be used. For any of the undesirable situations found, a solution or way to avoid the problem is explored. Since finding undesirable consequences is an exhaustive process and usually context-specific, it is impossible to find all edge cases or even know that all situations where consequences arise are found. However, the research in this paper could serve as a basis and inspiration for further research.

The structure of this paper is as follows: Firstly, some background information on fairness is given. After this, the methodology and results are given. After this is done, a section will be spent discussing responsible research in relation to the experiments done. Following is a critical discussion on the research done and some headers for further research. Lastly, the conclusion of the research done is given.

2 Background on Fairness

Defining what fairness means and how to apply its multi-dimensional ruling has been on the mind of mankind for a long time and is proven to be burdensome. Although most people have an intrinsic idea of what fairness means and what could be considered fair, defining the term 'fairness' and instantiating rules that would be considered fair in every situation is a huge undertaking. John Rawls wrote about his interpretation and definition of fairness and gave his ideas on what would be considered a just and fair society. His ideas on fairness are going to be used during the experiment, so it is important that we first look at Rawls' ideas on fairness as described in his book *A Theory of Justice* [3], so that we can apply some of

his ideas later on. Rawls thought that fundamentally, justice is connected to fairness and under a fair system, no one should be able to be exploited.

Rawls' notion of fairness is based on two principles. The first principle is called the liberty principle, and the second principle is a combination of two principles named the fair equality of opportunity and the difference principle. These principles are ordered in importance, such that the priority of these principles is clear when there is a situation in which the principles form any conflicts in any practical setting. In this paper, this order is not followed rigorously for convenience's sake, however, it is still worth keeping in mind.

His first principle, the liberty principle, states that "each person has the same and indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all" [10, p. 42]. The liberty principle entails that every person, no matter their inherent characteristics, has the same baseline level of liberties and freedoms. The liberties of certain people or groups could be toned down for the sake of overall liberty, for example when those groups or people use their liberties to impair the liberties of others.

The second principle of Rawls' theory on fairness, containing the fair equality of opportunity and the difference principle, states that "social and economic inequalities are to satisfy two conditions: first, they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least advantaged members of society" [10, pp. 42-43]. In essence, this idea contains the idea of equality according to Rawls. The first part, fair equality of opportunity, means that in general, no inequality is allowed. Any person should have a fair chance to obtain benefits in society. However, Rawls did not deny that inequalities cannot always be avoided. The second part of the second principle, the difference principle states that inequalities, mainly social and economic, should provide the most benefit for the least advantaged. This means that the well-off parties should not be able to get richer at the expense of the lower class.

Rawls' notion of fairness is not the only one out there. Although we will use Rawls' fairness as justice as the basis of fairness in this paper, it is important to know that there are other definitions and concepts of fairness. As an example, justice as fairness originated as a extension of egalitarianism, whose idea of fairness is that every human being is equal and has equal rights, both in a social and legal sense. Another example is Nozick's fairness notion in *Anarchy, State, and Utopia* [4], that emphasises the idea that people are self-owners of their body, spirit and opportunities, and all results, like possessions, that these previous three points yield. When trying to use fairness in an algorithmic sense, it is crucial that we keep in mind that not everyone will agree with the notion of fairness used and that we thoroughly explain why that specific notion of fairness is used, even if not all will agree with it.

3 Method

Now that the groundwork of Rawls' definition of fairness is known, we would like to apply his idea of fairness to automated negotiation to check for any unintended behaviour. After all, negotiation needs fairness, but leaving edge cases untouched might lead to serious problems. This means that we would like to keep in mind the unintended consequences of our Rawlsian fairness implementation, and for this, we would like to know what some of these consequences could be. In the following experiment, the Rawlsian notion of fairness is implemented such that the negotiation process adheres to principles of Rawls' justice as fairness as much as

possible, in particular, the idea that any inequalities should benefit the least advantaged the most.

For this paper, the SAOP (Stacked Alternating Offers Protocol) is used as the negotiation protocol. For this protocol, when an agent receives an offer from the other agent, the receiving agent can either accept the offer, create a counterbid and sent it to the other agent, or end the negotiation leaving both agents empty-handed. For a negotiation process with two agents, most other negotiation protocols are redundant since they are usually more suited for multi-party negotiations like AMOP (Alternating Multiple Offers Protocol) and MOPAC (Multiple Offers Partial Consensus). This makes the SOAP protocol most suited for this paper.

An automated negotiation process is set up with two agents, which we will call 'agent 1' and 'agent 2'. This is done in GeniusWeb, an open architecture for negotiation via the internet. The most important detail that needs to be kept in mind is that we will consider agent 1 as the 'advantaged' group, and agent 2 as the 'disadvantaged' group to mirror the advantaged and least advantaged groups as given in the difference principle. These assumptions are based on external factors from the negotiation, which means that these characteristics of the agents are true before the negotiation begins, and should influence how the agents enter the negotiation. These agents will run multiple negotiations against each other, and both the process of the negotiation and the outcome will be inspected for any odd behaviour. We would like to see that the outcome of the negotiations adheres to the Difference Principle, in other words, the outcome should be optimized for the disadvantaged group, agent 2. However, automated negotiation is usually not deterministic, and having one set result deemed fair before the negotiation begins somewhat defeats the purpose of negotiation if that result would be the only acceptable outcome. This is why the outcome does not need to be exactly the optimal Rawlsian outcome to be considered fair. In general, only behaviour and outcomes that show a large deviation from the normal and fair baseline will be considered unintended consequences. Checking the outcomes will be done through the utility of the accepted bid. This utility value for each agent is a value between 0 and 1 which indicated how desired the outcome is for that agent.

To integrate Rawls' fairness into the negotiation process, the advantaged agent, agent 1, cannot create bids that decrease the utility of the other agent:

$$\max_{available\_bids\_agent\_1}(utility\_agent\_2)$$

This imitates the idea that the advantaged agent cannot create bids that increase their own utility at expense of the other agents, in other words, the advantaged group will not get more advantaged at expense of the disadvantaged group. This also has the effect that agent 1 cannot create bids that decrease the utility for both agents, which means that there are not any sub-optimal bids (where both parties' utilities can be increased) from agent 1 and conveys the idea that inequalities are allowed if benefiting both agents. For agent 2, nothing special is done in terms of fairness implementation. This is done to keep the negotiation as open and dynamic as possible, while still ending up with outcomes that can be considered fair.

For simplicity's sake, both agents know each other's preference profiles. These preference profiles are used to create the bids for both agents and contain the preferences of the issues and issues values for each agent. Usually, these preference profiles are created by the agents through opponent modelling, which is when an agent tries to approximate the issue and issue value weights of the other agents during the negotiation. Creating such an opponent

modelling strategy is outside of the scope of this paper, but the general idea of the experiment would still work if an opponent modelling strategy is used.

Both agents use a trade-off strategy as their bidding strategy. The trade-off strategy starts with a very good bid for the bidding agent, while also trying to optimize the utility of the bid for the other agent. As more rounds get played, the threshold for the minimum utility of the bid for the bidding agent lowers, thus increasing the bidding space (all bids the agent can choose from). When creating the bids, since the agents know each other's preference profiles, the agents can more easily optimise their opponents' utility for the bid they are about to propose.

For the acceptance strategy, the strategy that decides for an agent when to accept an offer, a simple strategy is used that checks if the utility of a bid received by an agent is larger than the utility of the counterbid the agent would make, and accept based on this:

$$Utility(incoming\_bid) > Utility(outgoing\_bid)$$

This should ensure that a compromise is made between the agents with which both agents are content with the results since the acceptance strategy is partly based on its own bidding strategy.

Unintended consequences are searched for in the automated negotiations set up. This will be done through both the results of the negotiation and the negotiation behaviour throughout. The idea of an unintended consequence will be kept broad. By comparing the behaviour of the agents and any peculiar results, any interesting finds will be mentioned. Edge cases can be linked to either the notion of fairness used or any behaviour that seems to be linked to implementation or external factors. When any unintended consequences are found, suggestions will be made that could help avoid or combat the problem.

4 Results

In the research done, unintended consequences are found by looking at both the result of different negotiations, as well as looking at the behaviour of the agents throughout the negotiation process. As explained earlier, the definition of unintended consequences is interpreted quite broadly. Any result or behaviour that seems unusual, which will mostly be judged by comparing the behaviour of both agents, will be considered. Both abstract edge cases that are linked to the notion of fairness itself and more concrete implementation edge cases are considered. This way, a general idea is formed as to where unintended consequences can be found if the concrete notion of fairness described in this paper is used. Furthermore, some ideas and speculations are given as to why these unintended consequences might have occurred.

Firstly, the implementation of the difference principle, the idea that the advantaged should not be able to get more advantaged at the expense of the disadvantaged, has some consequences. The number of bids when setting up a counterbid of the agent implementing this principle, agent 1, generally decreased because the agent cannot pick bids that would lower the utility of the opponent. This leads to less variety during the negotiation process from agent 1, usually offering the same bid over a period of time until it comes across a bid that improves the utility of the other agent. An example of the variety of bids being different for the agents can be seen in figure 1. Notice that agent 2 seems to be more free in its bidding space, leading to more variety during the negotiation process while agent 1 has many reoccurring offers. Furthermore, if one of the earlier bids of agent 1 gives a relatively

high utility to agent 2, the time frame in which agent 1's bids stay the same can be a major part of the total negotiation.

Secondly, there are cases where the implementation of agent 2 leads to negotiation outcomes that do not conform to Rawls' fairness. This is due to the fact that although agent 1 keeps fairness in mind when negotiating, agent 2 does not do this on a surface level. As just discussed, agent 1's bidding space is restricted such that the utility of agent 2 cannot decrease during the negotiation. However, this is not the case for agent 2's bidding space. this means that agent 2 can explore its bidding opportunities more, and since the agent starts with its own best interest for the bids and slowly compromises according to the trade-off bidding strategy, this usually means an 'optimal' outcome (somewhere around the Pareto-front) is found and in accordance to or close to Rawls' fairness. Even so, there are cases where agent 2 picks a sub-optimal outcome for itself and agent 1 accepts this offer. This leads to outcomes where either there are bids left that would improve both parties' utilities, Or the outcome is the result of agent 1 receiving a better utility while agent 2's utility is not optimized, thus in both cases not conforming to the fairness notion of Rawls.

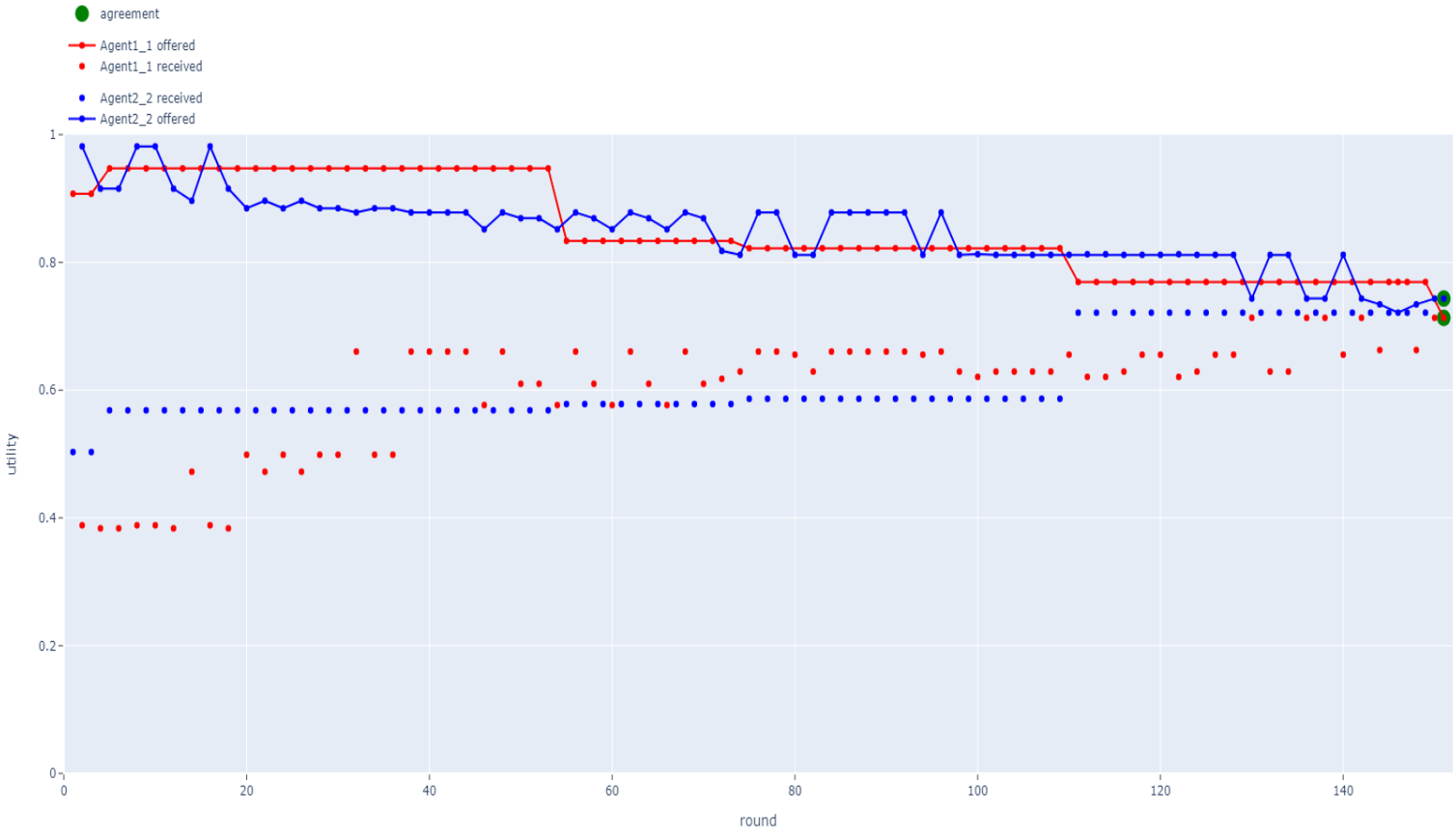

Figure 1: A negotiation process. The red points are the utilities of agent 1 and the blue points are the utilities of agent 2 of the bid for the given round.

Some aspects of the results that were tested did not get impacted by the fairness implementation. If one of the two agents consistently ended up with a higher utility than the other agent at the end of the negotiation, when the profiles of the agents are switched, the utilities of the agents are switched too, such that the other agent now ends up with the higher utility. This suggests that the fairness implementation does not heavily impact which agent gets their preferred bid or is not able to pull the negotiation in their favour.

Moreover, there were no unintended consequences found strictly in relation to the notion of fairness itself. The outcomes of the negotiations show no unusual signs, usually ending up with both agents having similar utility. When comparing agent 1 and agent 2, although the agents' negotiation process is different, they will usually end up with similar outcomes. This indicates that in the tested negotiations, the notion of fairness integrated into the negotiation process does not lead to unintended consequences. However, this could also be caused implementation-wise, but this is something that would need further investigation to be sure of its cause.

5 Responsible Research

This section on responsible research will cover the ethical implications of the research done and also discusses the reproducibility of the methods used in this paper. In terms of ethical implications, on a surface level, the research is quite ethically harmless without much context on its own. No sensitive data is used during the experiment and since the negotiation processes performed are imaginary situations and negotiations, they do not hold much ethical responsibility for any real-life situations, discussions or political discourse. However, looking at the big picture, we will have to keep in mind that by experimenting with fairness, we automatically make ethical implications. Our used notion of fairness affects the outcome of the negotiations, thus determining what each party will get and what their utility is in relation to the outcome. Although in the context of this experiment alone this does not have a big impact, as the topic of fairness in (automated) negotiation is being more researched, we have to be aware that eventually, the way the researcher defines fairness can have ethical implications if the research done is used in real-life applications. This is why we must make sure it is clear to the reader what type of fairness is used, and how this affects the negotiation process.

In general, the negotiation process set up in the experiment in the paper is quite reproducible. Assuming the method given is followed as much as possible, it is reasonable to assume the results will be similar. Still, there is a chance that the reproducibility of the experiments is not perfect. This will be mainly because of technical implementation differences that could lead to a (minor) deviance from the results in this paper. Although the methodology goes as in-depth into how the negotiation process was set up as possible, there will possibly be implementation differences, for example because of a different interpretation of Rawls' fairness, that could lead to differentiations in the results. For example, these might be platform-dependent if the implementation is not done through GeniusWeb, or differences in the implementation of the method. Also, with automated negotiation being nondeterministic, similar experiments might not end up with exactly the same results, however it is to be expected that the same conclusions can be made with similar experiments setups.

6 Discussion

Following the results, some ideas on how to avoid the unintended consequences found will be discussed. Although the fairness implementation in agent 1 has some drawbacks, they do not necessarily lead to unfair results. This means that this unintended consequence does not have to be negative by itself. However, since the bidding space is heavily shrunk, many other interesting bids are not explored which might also be considered reasonably fair. To avoid this behaviour randomness could be used. Instead of picking the bid with the highest utility of the other agent, select a few of the highest utility bids and pick a random one from those. Agent 1 still cannot decrease the other party's utility for their own gain, however, a larger bidding space is explored which has a higher chance to lead to a compromise. One thing we should keep in mind for this is that we still have to make sure the result is fair. A larger bidding space means more room for error in selecting a fair result, so there might be cases where using this change consistently leads to unfair results.

It is also shown that agent 2's implementation (or rather lack of implementation) occasionally leads to results that do not conform to Rawls' notion of fairness. Since agent 2 does not keep optimality in mind, it could give itself a sub-optimal bid that the other agent accepts. An improvement could be for agent 2 to check if the incoming and outgoing bids are fair. Since agent 2 has the most to gain from a fair result since it is the disadvantaged agent, it would make sense for this agent in particular to make sure the result is fair. However, we have to keep in mind that the experiment is done with the assumption that the agents know each other's preference profiles. This is normally not the case, and although usually agents create opponent models during negotiation, the fact that these are quite inaccurate at the beginning of the negotiation and potentially all throughout the negotiation, there will be cases where confidently knowing if a bid is fair for both parties from the perspective of one party will be difficult. At the end of the day, negotiation is not a deterministic procedure. There will be cases where even if both agents try to be as fair as possible, the result is determined to be not fair or not fair enough. In those cases, it could be worth it to redo the negotiation a certain amount of times.

This paper focuses only on a fragment of what could be studied for improving fairness in automated negotiation. In computer programs, finding edge cases can be a iterative and tedious task, and can be quite context-dependent. The same idea applies to the unintended consequences of fairness in automated negotiation. Since finding edge cases and unintended consequences of fairness is an exhaustive process with possibly infinitely many problems and challenges that need to be addressed, we would never be able to cover all situations in this paper, or in any paper for that matter. Finding situations where fairness does not behave as we would like it to is a context-specific issue, and as a consequence, the results found in the research done in this paper might not apply in the use case of a similar but different setting. This is why it would be very worthwhile to research the abstraction of these issues of consequences of fairness, such that we can create some framework that can be applied to a multitude of situations to combat undesirable behaviour of fairness.

7 Concluding Observations

Fairness is a critical concept in negotiation, and since autonomous agents should reflect and represent their users, the importance of fairness also extents to automated negotiation. However, it is important to know if integrating fairness into the negotiation process brings any unintended consequences. In this paper, a small-scale experiment is presented that

integrates Rawls' idea of fairness into the automated negotiation process, runs this process over various domains, and searches for unintended consequences that could occur when using Rawls' notion of fairness.

From the research done in this paper, we can conclude that there are unintended consequences are present when integrating fairness into automated negotiation. The results show that some of the unintended consequences are a smaller bidding space for the agent keeping fairness into account, and the outcome not being optimal for both agents, which contradicts Rawls' notion of fairness. These consequences seem to be related to the implementation. Consequences directly related to the notion of fairness implemented were not found. This does not necessarily means that they do not exist, and are a great pointer for further research. This paper has shown that edge cases for fairness in automated negotiation can be very context-specific, which is why it is important to research a variety of different test setups.

the research in this paper has plenty of room for extension. For example, the reader might not agree or relate to the Rawlsian notion of fairness, or might simply be interested in researching how a different notion of fairness could impact the negotiation process. As society has not settled on one unified definition of fairness (and mostly likely will not for some time, given the historical timeline of research on fairness) testing all different kinds of definitions of what it means to be 'fair' in negotiation would lead to a larger toolset to deal with unintended consequences during automated negotiation.

Furthermore, the acceptance and bidding strategy could be changed to see if this introduces any other unintended consequences with relation to fairness. Acceptance strategies that could be used are time-dependent strategies or combinations of other acceptances strategies. Examples of bidding strategies that could be experimented with are hardliner agents or conceder agents. This paper focuses on quite elementary and simple agents to get the point of unintended consequences across, however these agents can be customised and made more complex to reflect more realistic agents.

In short, this paper fixated a lot of parameters to conduct its experiment. Any of these parameters could be altered, improved or removed and this could lead to new or different edge cases that are worth exploring. Possibly the most worthwhile improvement would be to make the negotiation setting reflect real-life situations more, such that the agents can be placed in context of society and be used to aid realistic automated negotiations.

References

[1] N. A. Welsh, "Perceptions of Fairness in Negotiation," *Marquette Law Review*, vol. 87, no. 4, pp. 753–767, 2004.

[2] T. Baarslag, M. Kaisers, E. H. Gerding, C. M. Jonker, and J. Gratch, "When Will Negotiation Agents Be Able to Represent Us? The Challenges and Opportunities for Autonomous Negotiators," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, (Melbourne, Australia), pp. 4684–4690, International Joint Conferences on Artificial Intelligence Organization, Aug. 2017.

[3] J. Rawls, *A Theory of Justice*. Cambridge, Massachussets: Belknap Press of Harvard University Press, 1 ed., 1971.

[4] R. Nozick, *Anarchy, State, and Utopia*. New York: Basic Books, a member of the Perseus Books Group, 2013.

[5] H. P. Young, *Equity: in Theory and Practice*. Princeton, N.J: Princeton University Press, 1994.

[6] C. Albin, "The Role of Fairness in Negotiation," *Negotiation Journal*, vol. 9, pp. 223–244, July 1993.

[7] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, (Cambridge, Massachusetts), pp. 214–226, ACM Press, 2012.

[8] N. Jennings, P. Faratin, A. Lomuscio, S. Parsons, M. Wooldridge, and C. Sierra, "Automated Negotiation: Prospects, Methods and Challenges," *Group Decision and Negotiation*, vol. 10, no. 2, pp. 199–215, 2001.

[9] G. E. Kersten and H. Lai, "Negotiation Support and E-negotiation Systems: An Overview," *Group Decision and Negotiation*, vol. 16, pp. 553–586, Oct. 2007.

[10] J. Rawls and E. Kelly, "Two Principles of Justice," in *Justice as Fairness: A Restatement*, pp. 42–43, Cambridge, Mass: Harvard University Press, 2001.