Fairness by Discussion: An Alternative View on the Fairness of Protocols in Automated Negotiation

Zohar Cochavi (Delft University of Technology)
Supervisor(s): Sietze Kuilman, Luciano Cavalcante Siebert
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

The field of automated negotiation promises to improve negotiations, thus, a fair outcome and process should also be considered when building these systems. However, issues exist with computational approaches to fairness with which the field of computer science is mainly concerned. To this end, we propose a new approach to fairness based on that of essentially contested concepts to see if argumentation-based negotiation could be used as an extension to the Stacked Alternating Offers Protocol to improve fairness. Looking at fairness as an essentially contested concept shows that discussion between people somehow influenced by the negotiation system is necessary to maintain its fairness. This in turn means that systems that provide accessible context are fairer than systems that would not do so. Thus arguments, if implemented in an accessible manner, add more context to the negotiation, in turn making an SAOP negotiation fairer.

Fairness by Discussion

An Alternative View on the Fairness of Protocols in Automated Negotiation

Introduction

Fairness in computation, especially machine learning is a topic that has gotten increasing attention, and with good reason. COMPAS was a statistical tool that aided some U.S. states in determining how likely an individual was to recommit a crime. The tool, however, could display significant racial bias towards black individuals further reinforcing bias in human decision-making ("Fairness in Machine Learning", 2020).

Another field trying to augment human decision-making using computation is that of automated negotiation. It promises to improve the outcome, and process of negotiations by assisting humans or replacing them altogether (Baarslag et al., 2017). To be clear, there have not been such drastic fairness-related harms in automated negotiation, but that does not mean there never will be. Especially since machine learning tools are also being used to improve the performance of these negotiating agents (the computer program negotiating on behalf of a party) in, for example, opponent modeling[1] (He et al., 2016).

This raises the question: Would there be a way in which we could improve the fairness of automated negotiations? The question in and of itself is too broad and has to be scoped down in order to be meaningfully answered. Starting with fairness and the eternal discussion regarding the subject.

Fairness is Hard

Giving a single good definition of fairness is in no way trivial. Something Gallie (1955) also observed. In his research, he coined the term of an *essentially contested concept*, which tries to answer the questions of why some concepts are to hard to define in

[1] *Opponent modeling* is when one tries to estimate the preference profile of their adversary. This allows for more selective consideration of bids and, by extension, a quicker resolution of the negotiation process (Carmel & Markovitch, 1996).

a general context.

Fairness being an essentially contested concept could imply that on some topics there cannot be a single agreed definition. If that is the case, perhaps creating a system in which definitions are more easily investigated and adapted could be considered fairer.

The relevance of an essentially contested concept is further emphasized by the history of fairness in philosophy. Simply look at the number of opinions on fairness, or most any topic, all taking vastly different angles in an attempt to define the concept (Rawls, 1973; Wolff, 1998).

Even though there is discussion around fairness in computer science, the discussion does not seem to be as diverse as in philosophy. In a lot of research, similar approaches are taken and mostly based on computational or statistical approaches to fairness (Cerbone, 2021; Jacobs & Wallach, 2021).

Although using computational approaches to the topic makes sense in the context of computer science, some researchers have raised concerns about computational approaches to fairness (Jacobs & Wallach, 2021). This further motivates the need for a different approach to fairness in computation and thus automated negotiations.

On Negotiation

One way in which we could influence the fairness of negotiations is by establishing certain rules one has to follow during the negotiation. The set of rules followed during a negotiation is called a *negotiation protocol*.

One example of such a protocol is *SAOP* or the *Stacked Alternating Offers Protocol* (Aydoan et al., 2017). In this protocol parties present proposals in an alternating fashion, with one party initiating the process. At each proposal, the other party (or parties) can choose to accept the offer, or propose a counter-offer , to which the other party can respond again by accepting or proposing a counter-offer, etc. This protocol is one of the most basic ones, pay attention next time you haggle for a new car at the dealer, most probably you will follow the rules of SAOP.

If you decide to rely on a seller's empathy by arguing that you really cannot afford that price, you are employing an *Argumentation Based Negotiation* protocol or *ABN* protocol. While not technically a complete protocol, allowing the usage of arguments is part of a protocol[2]. In these types of negotiations, a party is allowed to provide reasoning behind their proposal in order to inform, or possibly manipulate, an adversary (Rahwan et al., 2004).

These kinds of negotiations are interesting because they contain more information about the motivations of a party than SAOP. Especially in the context of automated negotiations as we will be able to see why an agent makes certain decisions (Rahwan et al., 2004). Furthermore, they can be mathematically proven to be able to allow parties to reach a more satisfying agreement more quickly (Jennings et al., 2001, p.205).

Of course, more properties of negotiations exist. Another elementary example is direct negotiation between two individuals, called a *bilateral negotiation.* This differs from, for example, the negotiation method used when buying a house where a real estate agent acts as a *mediator* through which the parties bid. In addition to these two, there is a more 'general' type of negotiation, namely a *multilateral* negotiation. Simply said, this type of negotiation is one in which more than two parties are involved.

Outline

Given that the usage of arguments provides a great advantage, it is interesting to investigate if it could be considered as an extension to SAOP to improve fairness. The simplicity of SAOP makes it a good candidate for determining how much fairer a negotiation would be *with* arguments instead of *without* since we will be able to focus solely on the impact arguments have on the fairness of a negotiation.

To this end, we consider the fairness of SAOP and that of ABN in the context of fairness as an essentially contested concept. In short, with this definition, we will be able to

[2] For brevity, I will often refer to SAOP and ABN as "the protocols". Even though, as mentioned before, this is not technically correct.

consider fairness for automated negotiations without having to solely rely on computations. We thus take a different angle at the problem than other approaches in computer science (Cerbone, 2021; Dwork et al., 2011; Pessach & Shmueli, 2020), allowing us to resolve concerns some researchers have raised about such approaches (DeBrusk, 2018; Jacobs & Wallach, 2021).

Because of its importance, we will start with a larger discussion on fairness in Fairness by Discussion in which we will cover how we define fairness and why. After that we consider the impact arguments would have on the fairness of SAOP in Accessible Argumentation Drives Discussion. We discuss some topics for future study or discussion in Other Remarks and conclude the argument in A Promising Argument.

Fairness by Discussion

To assess the protocols in terms of *fairness*, the term has to be properly defined. In this section, we will briefly explore current ideas on fairness in computer science and philosophy in A Brief History of Fairness. Following that, we will further explain what an essentially contested concept is and why Fairness is Essentially Contested. We then argue why discussion is necessary for an essentially contested concept in The Necessity of Open Discussion, reflect back on what that means for fairness in Back to Computation, and summarize our findings in Putting it Together.

A Brief History of Fairness

Plenty of work exists on fairness in philosophy. One example is that of Wolff, who considers fairness as follows:

> Fairness is the demand that no one should be advantaged or disadvantaged by arbitrary factors. (Wolff, 1998, p.106)

The question then becomes, what is an *arbitrary factor*? and what does it mean to gain an advantage (or disadvantage) over someone else? While these questions have been answered in numerous ways within the realm of philosophy (Cerbone, 2021; Rawls, 1973;

Wolff, 1998), computer science is rather homogeneous in its opinion of fairness. Most popular answers fall somewhere in between a *Rawlsian* and *egalitarian* view of the distribution of goods or services (Cerbone, 2021; Pessach & Shmueli, 2020).

There has been a growing amount of research into resolving these issues related to computational fairness, mainly through the awareness of bias, both in systems and culture (DeBrusk, 2018; Pessach & Shmueli, 2020). This gives researchers and policymakers great tools to determine where 'unfairness' could originate from but does not solve the question of what explicitly would be unfair.

Good reasons do exist for computer science to use the definitions of fairness it does now. As an example, in Dwork et al. (2011) the definition of fairness is a statistical model of 'similar individuals should be treated similarly'; similar distributions should have similar mappings. Maximizing fairness is then a matter of minimizing the distance between the distributions after the mapping. These models are relatively easy to implement, quantifiable, and therefore easier to analyze than more 'typical' philosophical descriptions of fairness (the theories that take 20 pages to explain, and 20 years to understand).

Problems, however, exist with 'computable' approaches to fairness. There is a potential for some significant fairness-related harms that come with computational systems, as mentioned by Jacobs and Wallach (2021). They argue that fairness cannot be computed without essentially simplifying some parts of it. In turn, leading to possibly unfair scenarios because of simplifications that might be justifiable in all scenarios in which the model is used.

Of course, this does not imply that no good definition of fairness exists in certain contexts, but it does go to show how contested the topic is. Current computational approaches to fairness seem to do exactly as described; they somehow try to compute fairness. Which is the cause of the problems mentioned by Jacobs and Wallach (2021) and Kuhn (1996).

Having a definition that is independent of computation could therefore be a good contributor to the discussion. This brings us back to the essentially contested nature of fairness and how it relates to discussions.

Fairness is Essentially Contested

We have previously given an informal definition of an essentially contested concept, but the definition is more exact and has important implications for our discussion. Formally, Gallie (1955) states four conditions a concept should satisfy to be considered essentially contested:

1. It must be appraisive in the sense that it signifies or accredits some kind of valued achievement (Gallie, 1955, p.171).

2. This achievement must be of an internally complex character, for all that its worth attributed to it as a whole (Gallie, 1955, p.171).

3. Any explanation of its worth must therefore include reference to the respective contributions of its various parts or feature (Gallie, 1955, p.172).

4. The accredited achievement must be of a kind that admits of considerable modification in the light of changing circumstances; and such modification cannot be prescribed or predicted in advance (Gallie, 1955, p.172).

In short, it is a concept that should (1) signify value, (2) be multidimensional, i.e. there are multiple factors that all contribute to something being regarded as the concept, (3) it can only be *properly* defined in context (the definition must therefore refer to its contributions), and (4) be time-dependent. It is a concept that is considered valuable/important, but its definition is volatile concerning context.

(1) Fairness is a valuable feature, there has been a lot of research in *trying* to improve fairness in systems (computational as well as organizational), as covered in the previous section. You will rarely hear someone talk about trying to minimize the fairness of a system unless they are in a particularly sadistic mood.

(2) Often, fairness does depend on more than one factor at once. Take again the definition of fairness in terms of 'arbitrary factors' which, quite literally, depends on (the lack of) multiple factors. A less obvious example would be that of *Rawlsian* fairness which depends on an individual being able to be ignorant of his position in society (Rawls, 1973) to form a fair judgment. This depends on an individual being able to be ignorant to extend that one is not able to 'see' their position in society, but not to such so far such that they are not aware of any other's position in said society, and it depends on those individuals being 'rational' and 'free persons' (Rawls, 1973, p.11), etc.

(3) Furthermore, fairness is only well-defined within context. It is not hard to think of counterexamples of any type of fairness if the context in which it does hold is not part of the definition. Take for example *egalitarianism*, which is the belief that resources should be shared equally[3] (Wolff, 1998). But what if the individuals do not put in the same amount of work to collect those resources? The definition should therefore be: Resources will be shared equally among those who put in a similar amount of work[3]. If the context is not well-defined, the definition will be incomplete.

(4) Furthermore, it is hard to predict what people will find fair in the future. Take, for example, capital punishment. While prevalent throughout European history, most countries have now abolished the practice and no longer consider it a fair form of punishment (Neumayer, 2008).

Naturally, the theories presented here are slightly cherrypicked to prove my point. These are, however, definitely not the only ones and plenty more examples can be given.

[3] I'm oversimplifying here, but that is exactly the point. There are nuances to the term depending on your stance. Even including them, counterarguments are rarely in short supply.

The Necessity of Open Discussion

Assuming that fairness is an essentially contested concept, we can explore the relation between essentially contested concepts[4] and discussions. This relation will show how arguments could contribute to a more fair system.

The lack of a discussion about any concept could indicate a couple of things: (I) either an agreement, (II) an 'agreement on disagreeing', or simply (III) not realizing that there is a disagreement. We will go through each scenario, arguing why it would imply that discussion is eventually necessary if that concept is essentially contested.

(I) If there is an actual agreement, it means some definition is accepted. Considering that points (3) and (4) infer exactly that there is no single definition (since it is so sensitive to context) for an essentially contested concept, this implies that a discussion is necessary for an essentially contested concept[5]. Both because of the simple passage of time that implies that the definition changes (4), and because different people have different backgrounds, i.e. different contexts in which they think about fairness.

(II) Secondly, the moment two parties decide not to discuss the topic, it does not mean that the discussion will never arise. We can argue that because of the valuable nature of an essentially contested concept (1), people will have to argue about it at some point. This will, however, only be the case if the subject is valuable enough.

Fairness could be considered valuable enough to eventually give rise to discussions. If one feels like they are being treated in unfairly, and the situation is *open*, most anyone will say that they do not agree. Here, open, refers to a situation in which if an individual were to voice their opinion on fairness they know that their context will be 'added' to the

[4] In this discussion, I will use the term essentially contested concept and fairness interchangeably for the purposes of readability. Anything said in this section applies to all essentially contested concepts, unless explicitly mentioned that it is not.

[5] Importantly, it does not mean that a discussion is *sufficient* for something to be a essentially contested (i.e. discussion implies essentially contestedness).

definition of fairness. This means that more open discussions contain more context, and therefore fairer, definition of fairness.

(III) In the last case, since the concept is regarded as valued (1), we can again assume that, even if parties have not yet voiced their opinions, they will at some point do so. Thus concluding that discussion is indeed necessary for fairness as an essentially contested concept.

On the other hand, what happens if an individual cannot take part in the discussion? If individuals that are influenced by the given definition of a concept cannot contribute to the definition, i.e. their context is not part of the definition of fairness, they are essentially subject to an incomplete definition of fairness (3). Having to 'use' an incomplete definition of fairness is considered unfair from the perspective of the individual whose context is missing.

Inhibiting discussion, or somehow inhibiting stakeholders from actively participating in that discussion (by for example not having an open situation), would therefore be unfair. Making a discussion more open would improve the fairness of a system or situation.

Back to Computation

We can extend this idea to computer science and its definition on fairness. Considering the premise that computer science has a rather homogeneous opinion on fairness, the field could indeed be better off having more people that can contribute to the definitions of fairness in computational systems. We will avoid a larger topic on the other end of the discussion (having individuals not be part of the discussion might sometimes be a good thing)[6] since it does not apply to the current scenario.

This does depend on the openness of a system. Indeed, plenty of sources talk about

[6] There are, of course, scenarios in which one would be better of excluding certain people from the discussion. Especially if people are uninformed, or worse, *think they are informed* about a certain context of fairness. But this is out of scope of the discussion and not applicable to the current scenario of complex systems, since problem is that there are not enough people that can have an opinion on the matter.

the necessity of ’open systems’ of computation and the different risks that a lack of transparency brings to complex systems such as automated negotiation, especially when machine learning is involved (Hagras, 2018, p.29).

Furthermore, if fairness is regarded as something that can only be well-defined in one context, how can we justifiably implement one definition in a system? Even if, hypothetically, that definition would fit in the context of that particular system and all of its stakeholders, the aforementioned definition would change over time (4). This means that as the implementation ages, the context necessary for the definition to be considered fair is missing.

Mitigating the issues with this scenario would require constant maintenance on behalf of the developers, and be sure that the system will *never* be used outside of its intended context. Humans are notoriously bad at *not doing* certain things when they are told to.

Putting it Together

Having reached the end of our philosophical rabbit hole, we can start putting things together.

The usage of fairness in computer science is rather homogeneous, which is not a problem per sé, but these definitions all rely on computational approaches to fairness. Multiple sources mention a problem with approaching certain problems, including fairness, in a computational matter, suggesting that there could be a better way of defining fairness.

To this end, we consider fairness as an essentially contested concept. It tells us that fairness is a valuable attribute that is so context-sensitive that a general definition is practically impossible to establish, and that any definition will only hold within the given context.

Looking at fairness from this perspective implies that discussion is necessary to call a system fair. Inhibiting discussion about, or excluding individuals that are somehow affected by the system is therefore considered unfair, and the more open the discussion regarding a given definition of fairness, the fairer the definition.

Accessible Argumentation Drives Discussion

Choosing to extend SAOP with ABN essentially means that, besides just proposals, arguments will are included in the negotiation. This can be done in a variety of ways: at every counter-offer, only when a party ’feels like it’, etc. While the concrete usage of arguments could definitely have an effect on the fairness, we will not concern ourselves with this matter in depth.

Before being able to discuss the advantages and disadvantages of different implementations, we first have to assess if arguments even contribute to fairness at all. Therefore, we will now limit ourselves to the general case of how arguments contribute to fairness. In the last section, we will briefly return to the point of implementation.

A Machine’s Motivations

Having arguments included in a negotiation has numerous advantages. Certain advantages, however, are certainly more ’absolute’ than others, which might depend on context if they could be considered advantages.

Furthermore, this information provides insight into the machine’s motivations. It gives people the opportunity to reason *with* the machine, instead of about it. Reasoning about the machine requires knowledge of its inner workings, in turn requiring background knowledge. This limits the number of stakeholders being able to participate in a discussion regarding the system.

Reasoning *with* the machine, however, is possible because the agent shows the reasoning behind its actions. If the machine can explain themselves about the current issue (e.g. their opinion on the state of the negotiation, and their wishes regarding its outcome), this allows a stakeholder *regardless of their background* to have an opinion on the *fairness of the process*[7]. Thereby increasing the ’openness’ of the discussion regarding the system.

[7] This does require a computer to express its motivations and reasoning in natural language. Recent advancements in the field of *eXplainable Artificial Intelligence*, *XAI*, have shown some progress towards

The process of how a computational system arrives at a conclusion contributes a lot more to the discussion since it provides more context. As previously discussed, this context is at the heart of an essentially contested concept and therefore necessary for healthy discussion. This allows stakeholders to have a more contextualized opinion (i.e. an opinion that contains contributions relevant to their definition of fairness (3)) on the fairness of the system.

Therefore the inclusion of arguments would provide a certain contextualization of the given arguments which would, given the importance of context for fairness, improve the fairness of SAOP.

The Necessity of Accessible Arguments

An important assumption was the 'layman' being able to understand the arguments of the machine. While this seems natural, it is definitely not a given.

It should not be necessary to have a computer science degree to understand that an agent took advantage of an adversary's poor position in a negotiation. Whether the action of the agent is fair or not is unimportant. It is about a non-expert having the ability to have an opinion on the matter that is relevant within the context of the negotiation.

This accessibility is a requirement because most stakeholders will not be experts[8]. Accessibility here refers to a non-expert being able to understand and access the arguments that are given. A person simply using the negotiating agent should be able to access the argumentation history of the negotiation and understand it as if two humans were conversing with each other.

Therefore, if non-experts cannot understand or access the arguments given by the agents, the inclusion of arguments does not improve fairness from the perspective of essentially contested concepts. They do not provide the context which would otherwise

this goal (Hagras, 2018).

[8] Here, we refer to an expert as someone with the knowledge required to create such an agent (i.e. a computer scientist).

improve a person's opinion on the fairness of the system, and neither improve the 'openness' of the discussion as it is would be about as useful as looking through the source code.

A Different Perspective

While non-accessible arguments might not improve fairness from the perspective of essentially contested concepts, there are other advantages to using arguments. Jennings et al. (2001) mention that it can be mathematically proven that negotiations containing arguments converge to a more satisfactory agreement in less time. This does not directly impact the *fairness* of the system, but it does seem more respectful towards its stakeholders to consider a system that saves them both time and 'wasted' utility[9].

Furthermore, Wolff (1998) considers fairness to, in some cases, be inferior to respect. He proposes a solution to the issue raised before about egalitarianism and people who do not contribute as much to the gathering of certain resources compared to others (i.e. lazy people) by saying that this is 'disrespectful'.

In that case, even if the arguments are not accessible, ABN could be considered to be 'fairer' than their non-argumentation-based counterparts.

There is no such Thing as Free Lunch

These advantages, however, are not without cost. There is an argument to be made that having a more complex system makes it *less accessible* to the layman. Considering this, would a simpler system not be fairer if more people can be more easily informed?

The costs of creating such a system are significant. Not only is the human required to understand it, but, if we want a truly argumentation-*based* negotiation, the opposing agent also has to be able to parse and use these arguments. Doing this will require the adversary agent to use natural language processing to parse the passed arguments.

Even if the negotiation is not truly negotiation-based and the arguments are only

[9] *Utility* refers to the amount personal value, or reward, of something. In this case, it refers to the value of the outcome of a negotiation. Each party will have their own value attributed to the outcome, so they will attribute different *utility* to it.

provided to improve fairness (as an essentially contested concept), these arguments still have to be created which is nowhere near trivial (as alluded to before[7]).

While we will not discuss the feasibility, it is worth noting that it adds significant complexity over standard SAOP. Although not directly impacting the fairness of the decisions and process of the system, it does limit the number of individuals that will be able to implement such a system which in turn could raise several ethical considerations.

Other Remarks

While I have tried to make this discussion as complete as possible, some topics have been left undiscussed. Here, we will briefly touch upon these topics before moving to the conclusion. This is definitely not an exhaustive list, but they are among the most important ones to consider for future study.

The primary focus has been on *if* arguments could improve fairness in SAOP, not by how much. In the last section we have briefly touched upon this, but the practical potential of this kind of application of arguments heavily depends on how the feasibility compares to the actual improvement of fairness. This would, however, require quantifying fairness which would depend on a computational approach to fairness. As mentioned before, computational approaches to fairness have great advantages. Combining the two might prove especially effective.

Furthermore, while we have tried to argue that discussion is necessary for all essentially contested concepts, we were only successful in doing so for fairness. The assumption was that fairness is valuable enough that people will voice their opinion if the situation allows it (which is a big if in some environments), but this is hard to say for all essentially contested concepts.

Another undiscussed topic is that of liars, specifically, an agent lying in an argument. Lying in this case could either refer to making up arguments to manipulate and gain a 'crafted' advantage over an adversary or by manipulating the individual reading the arguments to construct an opinion on the fairness of the negotiation. This could have an impact on the fairness of the negotiation, but, if a deceptive agent is caught the backlash would be great if the discussion is open enough (assuming most stakeholders do not like being deceived).

A Promising Argument

Systems are going to get more complex, which is unavoidable as our hunger for technological advancement is insatiable. Machine learning and other AI techniques are already being used in automated negotiation in for example opponent modeling[1] (He et al., 2016). These computational models can and have caused significant fairness-related harms (DeBrusk, 2018; "Fairness in Machine Learning", 2020; Jacobs & Wallach, 2021). It is therefore important we avoid similar situations arising in automated negotiations.

If implemented in an accessible way (that is, in a way that non-experts can access and understand), the inclusion of arguments could provide a way to make these systems more understandable. This would mean that stakeholders can create a definition of fairness for themselves with more context, and it would make it easier to participate in the discussions regarding the system.

This improves fairness because we have considered open discussion necessary for a fair system; the more open a discussion and the more information available about a system (since this provides more context), the fairer it is. The necessity of an open discussion follows from the fact that we consider fairness to be an essentially contested concept. Being essentially contested means that the definition of a valued concept only makes sense within the context in which it is defined. Failing to include the context in its definition will lead to an incomplete (and in our case unfair) definition.

Especially where some suggest that computational approaches to fairness seem to cause problems (Jacobs & Wallach, 2021; Tang & Ito, 2018), this approach to fairness could provide a promising argument for the use of arguments in SAOP.

Responsible Research

Having written this argument for a more accessible and open type of negotiation protocol, my hope is that this document can have a positive effect on the fairness and trustworthiness of these systems and the people influenced by it.

Since this document is specifically an analysis about fairness and how to improve it, I will not cover the ethical implications of defining fairness as such, seeing as it has been extensively covered throughout.

It is also appropriate to mention that most assumptions have been based on prior research which has been properly referenced and mentioned where appropriate. Equally, factual statements all refer to their respective sources. Wherever this is not the case, it has been indicated that this is my own opinion, assumption, or intuition.

References

Aydoan, R., Festen, D., Hindriks, K. V., & Jonker, C. M. (2017). Alternating Offers Protocols for Multilateral Negotiation. In K. Fujita, Q. Bai, T. Ito, M. Zhang, F. Ren, R. Aydoan, & R. Hadfi (Eds.), *Modern Approaches to Agent-based Complex Automated Negotiation* (pp. 153–167). Springer International Publishing. https://doi.org/10.1007/978-3-319-51563-2_10

Baarslag, T., Kaisers, M., Gerding, E. H., Jonker, C. M., & Gratch, J. (2017). When Will Negotiation Agents Be Able to Represent Us? The Challenges and Opportunities for Autonomous Negotiators. *2017*, 4684–4690. https://doi.org/10.24963/ijcai.2017/653

Carmel, D., & Markovitch, S. (1996). Opponent modeling in multi-agent systems. In G. WeiSS & S. Sen (Eds.), *Adaption and Learning in Multi-Agent Systems* (pp. 40–52). Springer. https://doi.org/10.1007/3-540-60923-7_18

Cerbone, H. (2021). Providing a Philosophical Critique and Guidance of Fairness Metrics. *arXiv:2111.04417 [cs]*. https://doi.org/10.48550/arXiv.2111.04417

DeBrusk, C. (2018). The Risk of Machine-Learning Bias (and How to Prevent It). *MIT Sloan Management Review*. Retrieved May 18, 2022, from https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. *arXiv:1104.3913 [cs]*. Retrieved April 20, 2022, from http://arxiv.org/abs/1104.3913

Fairness in Machine Learning. (2020). Retrieved June 19, 2022, from https://sitn.hms.harvard.edu/uncategorized/2020/fairness-machine-learning/

Gallie, W. B. (1955). Essentially Contested Concepts. *Proceedings of the Aristotelian Society*, *56*, 167–198. Retrieved May 18, 2022, from http://www.jstor.org/stable/4544562

Hagras, H. (2018). Toward Human-Understandable, Explainable AI. *Computer*, *51*(9), 28–36. https://doi.org/10.1109/MC.2018.3620965

He, H., Boyd-Graber, J., Kwok, K., & Hal Daumé, I. I. I. (2016). Opponent Modeling in Deep Reinforcement Learning. *Proceedings of The 33rd International Conference on Machine Learning*, 1804–1813. Retrieved June 19, 2022, from https://proceedings.mlr.press/v48/he16.html

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. https://doi.org/10.1145/3442188.3445901

Jennings, N., Faratin, P., Lomuscio, A., Parsons, S., Wooldridge, M., & Sierra, C. (2001). Automated Negotiation: Prospects, Methods and Challenges. *Group Decision and Negotiation*, *10*(2), 199–215. https://doi.org/10.1023/A:1008746126376

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed). University of Chicago Press. Retrieved June 8, 2022, from http://catdir.loc.gov/catdir/toc/uchi051/96013195.html

Neumayer, E. (2008). Death Penalty Abolition and the Ratification of the Second Optional Protocol. *The International Journal of Human Rights*, *12*(1), 3–21. https://doi.org/10.1080/13642980701725160

Pessach, D., & Shmueli, E. (2020). Algorithmic Fairness. *arXiv:2001.09784 [cs, stat]*. https://doi.org/10.48550/arXiv.2001.09784

Rahwan, I., Ramchurn, S., Jennings, N., Mcburney, P., & Parsons, S. (2004). Argumentation-Based Negotiation. *The Knowledge Engineering Review*, *18*. https://doi.org/10.1017/S0269888904000098

Rawls, J. (1973). *A theory of justice* (New ed.). Oxford University Press.

Tang, X., & Ito, T. (2018). Metric for Evaluating Negotiation Process in Automated Negotiation. *2018 IEEE International Conference on Agents (ICA)*, 26–29. https://doi.org/10.1109/AGENTS.2018.8460127

Wolff, J. (1998). Fairness, Respect, and the Egalitarian Ethos. *Philosophy & Public Affairs*, *27*(2), 97–122. https://doi.org/10.1111/j.1088-4963.1998.tb00063.x