Using the Toulmin Model of Argumentation to Explore the Differences in Human and Automated Hiring Decisions

Hebah Bubakr and Chris Baber
The Department of Electronic, Electrical and Systems Engineering, University of Birmingham, Birmingham, U.K.

Keywords: Toulmin Model of Argumentation, Artificial Intelligence, Human Computer Interaction.

Abstract: Amazon developed an experimental hiring tool, using AI to review job applicants' résumés, with the goal of automating the search for the best talent. However, the team found that their software was biased against women because the models were trained on résumés submitted to the company for the previous 10 years and most of these were submitted by men, reflecting male dominance in the tech business. As a result, the models learned that males were preferable, and it excluded résumés that could be inferred to come from female applicants. Gender bias was not the only issue. As well rejecting plausible candidates, problems with the data lead the models to recommend unqualified candidates for jobs. To understand the conflict in this, and similar examples, we apply Toulmin model of argumentation. By considering how arguments are constructed by a human and how a contrasting argument might be constructed by AI, we can conduct pre-mortems of potential conflict in system operation.

1 INTRODUCTION

Argumentation technology can help Artificial intelligence (AI) agents enhance human reasoning (Lawrence et al, 2017; Oguego et al, 2018; Modgil and Prakken, 2013). In this context, an 'argument' has a 'scheme' (which relates to a specific form of reasoning, often inspired by human behaviour) in which there is a structure to how information leads to a conclusion; a 'format' (which encodes the information); and a 'representation' (which visualises the scheme) and for which diagrams are commonly used (Reed, Walton and Macagno, 2007). Information in an argument diagram can be either a proposition or an inference (Bench-Capon and Dunne, 2007; Freeman, 1991). Toulmin diagrams are a well-known form of argument representation (Bench-Capon and Dunne, 2007). In this paper, Toulmin diagrams represent arguments to consider bias in hiring decisions.

For jobs with many applicants, an initial sift of résumés could be performed to reduce the number of applications to consider. However, there is a risk that such a sift could be subject to bias. Bertrand and Mullainathan (2004) sent five thousand fake résumés for different job adverts in the Boston Globe and Chicago Tribune newspapers. The adverts were covered different occupational categories and with each category the quality of the fake résumés was either high and low. Identity was assigned to each résumés using a personal name to suggest the race of the applicant, for example, Emily, Anne and Brad suggested white names; Kenya, Lakisha and Jamal suggested African-American names. There was a statistically significant difference in call-backs (Bertrand and Mullainathan, 2004): résumés for 'white' applicants received 50% more call-backs than those applicants with African American names (even though the content of the résumés was identical). Further, the call-back rate for 'white' applicants with a high-quality résumé was 27% higher than for 'white' applicants with low quality résumés. In contrast, call-back for résumés with African-American names with high quality résumés was only 8% higher than for low quality résumés.

O'Neil (2016) suggests that the desire to bring analytical science to human resources is to make selection fairer (by removing potential for human bias) but are increasingly being used to filter applicants. This raises the question of whether they are optimised for the objective for selection (to aid picking the most appropriate person for the role) or rejecting applicants (to aid in filtering applications down to a manageable set). We propose that these objectives represent different forms of argumentation

Bubakr, H. and Baber, C.
Using the Toulmin Model of Argumentation to Explore the Differences in Human and Automated Hiring Decisions.
DOI: 10.5220/0009129102110216
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 2: HUCAPP*, pages 211-216
ISBN: 978-989-758-402-2; ISSN: 2184-4321

that place emphasis on different aspects of the resume and test scores.

The main purpose of using an automatic system is to ensure efficiency and fairness, and to cut the thousand applications to a reasonable number for HR employees to take to the interviewing process. Moreover, the machine ought to remain unaffected by bias or prejudices, and each applicant should be judged by the criteria outlined in the job specification. However, AI systems can be trained using records of employment data from many years which could mean that bias has been institutionalized in favour of specific employee groups. Gender bias is an issue in jobs that reflect a male dominance like the tech business for example. Amazon developed an experimental hiring tool, using AI to review job applicants' résumés, with the goal of automating the search for the best talent. This would give candidates a ranking from one to five stars. However, the team found that their software was biased against women because the models were trained on résumés submitted to the company for the previous 10 years and most of these were submitted by men. As a result, the models learned that males were preferable, and it excluded résumés that appeared to come from female applicants. Gender bias was not the only issue. As well rejecting plausible candidates, problems with the data lead the models to recommend unqualified candidates for jobs.

In this paper, we ask whether representing an argument in the form of a diagram could help diagnose possible problems or biases in the output of an algorithm. The main concern in this paper is to understand the reason why a machine produces such unacceptable results. Whether the problem was a mistake, biased data, or an unethical developer, it will be always be unacceptable to reject applicants for these reasons. Thus, people's beliefs and cultural values affect how we interact with AI system and accept its results. Users of the system will have expectations based on their prior knowledge. This knowledge is formed by the beliefs and the values of the users. For instance, the HR department might expect the AI systems to choose applicants with higher education and length of experience for administrative positions. Therefore, when a hiring system eliminate suitable candidates because of their name, gender or zip-code, the results will be prejudicial to them because these outputs conflict with people's ethical values and do not match the expectation they had of the system. Seeking to understand how the system works and what are the beliefs, values and expectations of stakeholders we apply Toulmin's model of argumentation. We note, at this point, that other argumentation representations could be equally appropriate to this aim.

2 MOTIVATING EXAMPLE

Kareem Bader, a student who was looking for a minimum-wage job, applied for a part time at super-market chain store after his friend recommended him. Kareem had a history of having bipolar disorder but at the time of sending the application he was a productive, high-achieving student and healthy enough to practise any type of work. However, Kareem was not called for an interview and when he asked, he was told that he failed the personality test he answered during the application. The most commonly used personality test is the Five Factor Model (Lundgren, Kroon, & Poell, 2017), and this is the one which Kareem took. Kareem gave honest answers to mental health questions and real information about his personal background, and this led to him being rejected.

3 TOULMIN MODEL OF ARGUMENTATION

This paper will use Toulmin model of argumentation (Toulmin, 1958). Toulmin provides an argumentation technique to show that there are other arguments than formal ones, which offers more logic and further explanation. His argumentation model distinguished six different kinds of elements: Data, Claim, Qualifier, Warrant, Backing, and Rebuttal (Verheij, 2009). The example below illustrates Toulmin's classic example of Harry, who may or may not, be a British subject.

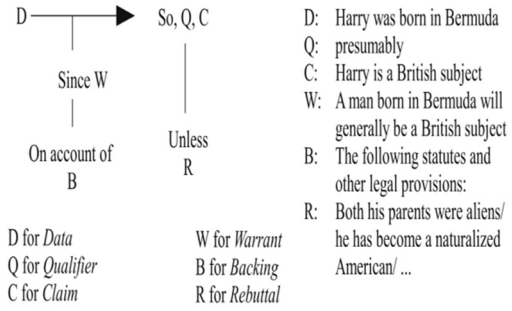

Figure 1: Toulmin's layout of arguments with an example (Toulmin 1958, p104, 105).

In Figure 1, D represents the datum "Harry was born in Bermuda" which is held to be the foundation

of the claim "Harry is a British subject". What does this claim stand on? We need to use a warrant; in our example, this is 'A man born in Bermuda will generally be a British subject.' A point to keep in mind is that the warrant is not universal. We are saying that a man born in Bermuda will *more likely* be a British subject. The warrant serves as a hypothetical statement which acts as a bridge between the Ground (Data that provide facts) and the Conclusion of an argument (claim). The warrant can be considered as an inference, and can be challenged. As a result, a claim needs to be qualified (Verheij, 2009). In our example, the qualifier is to presume that Harry is a British Subject. As the elements in this argument became more explicit, the Warrant needs further support so that it can provide the connection between the data and the claim. In Toulmin's example this is called the backing and refers to statutes and other legal provisions without the need to define them. For this paper, this idea of 'backing' helps to illustrate the hidden values and beliefs that might inform models. The final element, in Toulmin's example, is the Rebuttal which indicates any arguments against the claim or any exceptions to it. According to Toulmin (2003), the rebuttal indicates 'circumstances in which the general authority of the warrant would have to be set aside' or 'exceptional circumstances which might be capable of defeating or rebutting the warranted conclusion'.

Table 1: Translate the example text to Toulmin elements.

Toulmin element	Text from Motivating example
Datum	Kareem gave honest answers to mental health questions. Productive, high-achieving student. Healthy enough to practise any type of work.
Claim	Kareem Baderis suitable for this job
Warrant	Productive, high-achieving, Healthy, was recommended
Backing	Minimum-wage job which is part time at super-market, usually does not have hard to meet criteria.
Rebuttal	Unless he is not suitable

Applying this model to the processing of hiring which include selecting and filtering applicant depending on their resumes and test results, we can construct the argumentation scheme shown below. First to apply Toulmin model of argumentation on the motivation example we need to interpret the terms. We convert the text from the example into Toulmin element's term as shown in table1.

Using table 1, we consider the applicant's argumentation. When a job seeker applies for a specific job they will have some expectation regarding their suitability. When an applicant claims to be suitable for the job, we out to ask why? Their answers will be based on their educational background, skills, heath condition and other information that support their claim. In the Toulmin model, that supportive information represents the datum. However, to claim suitability for the job, providing simple facts is not enough. The Warrant, in the Toulmin model, bridges between data and claim. In figure 2, the warrant states that if applicant met the requirement, they probably will be suitable. That rule is supported by the backing which is the job criteria generated from the employers' job specification. Rebuttal happens when the rule is not met or does not apply to the data. So, when Kareem claims he was suitable for the job, he did not understand why his suitability got challenged.

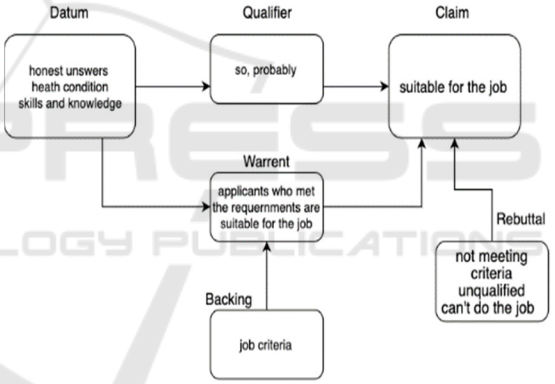

Figure 2: Applicant's argumentation.

Automated systems help HR department hire, fire, and promote employees efficiently. The hiring system might be used when there are too many people applying for the job, and the HR department can only interview a certain number of applicants. Hence, when the number of applicants is too high (Claim) the system must reduce the number by filtering application. The warrant is the limited time for interviewing. This rule is supported by (Backing) which states that this filtering process will save the HR time and effort. However, in the filtering process the system might use a model based on current profiles of employees, job criteria, to judge the applicants' resumes and forms. An exception to this claim is when the system filters good and qualified people.

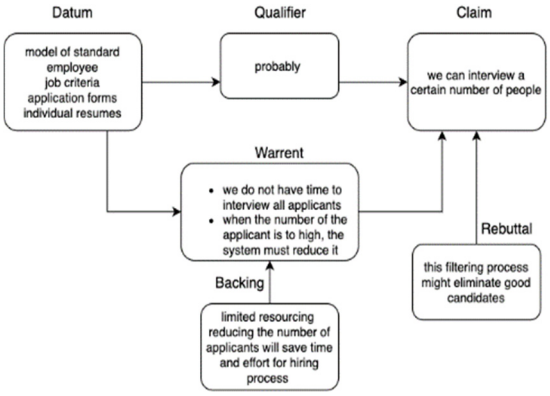

Figure 3: Filtering argumentation.

After reducing the number of applicants to a reasonable amount, the system could select a set of best candidates among the applicants by matching the job criteria with the individual resumes and forms for the applicants then ranking them based on their match to the 'model of employees', see figure 4.

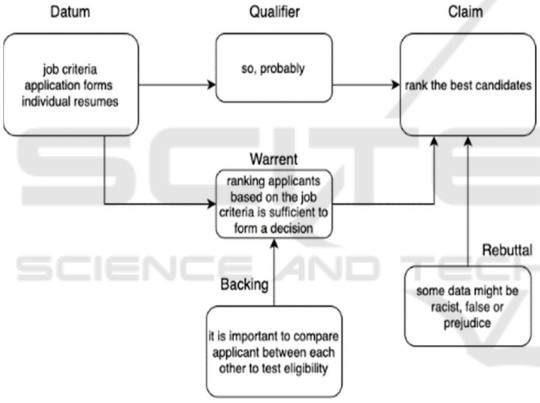

Figure 4: Selection argumentation.

The last stage in the hiring process is in the HR department's hands. As figure 4 shows, the system, which the HR uses to ensure the filtering efficiency, will provide a number of 'best candidates' for the HR to select from. The HR will define the best applicant after reviewing each résumé. This information is supported by the belief that the system is reliable enough to trust its results.

The HR department assumes that this process ensures equality and diversity in its hiring process. It depends on data from the hiring system and specific equality criteria applied to all candidates. However, if the model is inherently biased, e.g., because it is derived from a homogeneous employee population that does not reflect diversity, then the process will have prejudice embedded in it.

As we can see, from figures 2-4, there are conflicts between different claims in different situations from different stakeholders. Each claim represents the values and beliefs of its own claimant. Moreover, each claim was formed on the available data in that situation and carries no authorization to interfere with other claims. Even though the filtering and selecting processes are dependent on each other, each had their own criteria. For instance, in the applicant suitability diagram, figure 2, Kareem claimed that he was suitable for the job which was supported by the data and the warrant that he was good enough for the job. In both filtering and selecting argumentation's diagrams, figures 3 and 4, the system claimed that Kyle was not suitable because he failed one on the essential tests in his application form. Therefore, his application was not transferred to the selecting process nor to the HR department which claimed nothing about Kareem's stability, except that his name was not listed as a possible candidate. So based on the data, warrant and backing, he was not defined as a suitable employee. Now, whether Kareem was filtered from the start or not selected at the end, the system clearly did not see him as fit for the position which shows conflict between values and beliefs' of the applicant and the system. This example illustrates that each argumentation has its own elements that do not interact with elements in other argumentations. So, the claim that was formed by the system cannot use the data from either the applicant or the HR argumentation. Moreover, the same data were treated differently in the filtering and selecting stages. From this example, we conclude that representing argumentation is a way of illustrating how different criteria support an argument can lead to a bias decision.

So far, the argumentation has been hand-crafted and applies Toulmin's criteria. This is presented as a 'pre-mortem' of the process as way of explaining potential bias and conflict. The question is how this could be automatic?

4 ARGUMENT INTERCHANGE FORMAT

Argumentation technology can be supported by computational models. Argument interchange format (AIF) supports representation and exchange of data between argumentation tools (Chesnevar et al. 2006). As shown in figure 5, AIF uses an argument network (AN) which contains two types of nodes, information nodes (I-nodes) and scheme nodes (S-nodes). I-nodes

represent the contents like data and claims, and S-nodes represent the applications of schemes like rules of inference. There are three different types of scheme-nodes, a rule of inference application node (RA-node), a preference called a preference application node (PA-node), and a conflict application node (CA-node) (Chesnevar et al. 2006). Moreover, AN has two types of edge, scheme edges and data edges. Scheme edges originate from S-nodes and are meant to support conclusions. These conclusions may either be I-nodes or S-nodes. Data edges originate from I-nodes, and they must end in S-nodes, and supply data or information to scheme applications. I-to-I edges are not an option because I-nodes needs always an explanation to be attached. I-nodes can have zero incoming edges, but S-nodes link two or more elements, i.e., "for RA-nodes, at least one antecedent is used to support at least one conclusion; for PA-nodes, at least one alternative is preferred to at least one other; and for CA-nodes, at least one claim is in conflict with at least one other" (Chesnevar et al. 2006).

Using AIF, our Toulmin models can be expressed in a standard XML-based syntax. Chesnevar et al. (2006) show how Toulmin elements can be expressed as I-Nodes: claim, data, backing, rebuttal, and qualifier (figure 5). The warrant was presented in this diagram as (S-Node) specifically RA-Node, as it holds both data nodes and the claim together. Likewise, an RA-node links rebuttal nodes to claims, and Qualifier-Application nodes link qualifier nodes to claims. The result of his ontology is represented in Figure 5.

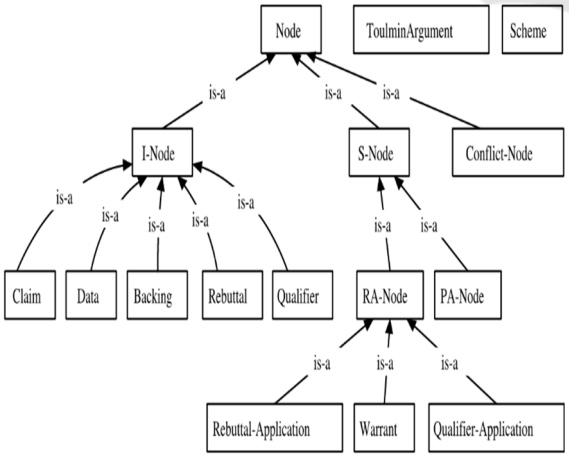

Figure 5: Toulmin argument class hierarchy in RDFS (Chesnevar et al. 2006. P312).

The mark-up language can provide a formalisable definition of Toulmin's elements, but this is not different than what table 1 provides. Indeed, defining and managing mark-up for argumentation could still involve a high degree of human intervention (because, unlike other forms of Natural Language Processing, it is not always obvious whether a given word of phrase is being expressed as data, or claim, or warrant). Hence, an abiding question is how to automatically generate the argument labels from the text.

5 CONCLUSIONS

If Kareem was not filtered, does that mean he was suitable for the job? Not necessarily, because the system decision could be right. The main issue concerns the way the system formed its conclusion. In the above example, when the system is facing a huge number of applicants, an obvious problem in the filtering process is the choice of parameter and use of evidence to accept one of these parameters. A solution is to conduct sensitivity tests of these parameters. i.e., remove some parameters and explore the impact of the resulting decision or consider how the parameter interacts with be personality test, name, gender, age etc. of applicants. For example, in the filtering process all the parameter should be about skills, education and experience not about name, age, gender and mental health. O'Neil (2016) advises to, "…build a digital version of a blind audition eliminating proxies such as geography, gender, race, or name to focus only on data relevant to the job position. The key is to analyse the skills each candidate brings to the company, not to judge him or her by comparison with people who seem and whether or not the output of these systems make sense to them or not similar." (p.177)

We conclude by asking whether it is wrong for job applicants who understand what the system is looking for to *craft* their application with these features in mind, or provide answers to trick the software which can reduce the chance of early filtering. If we allow people to challenge the AI hiring systems, how do we know that they are not going to provide false information to manipulating it?

Even if we ensure that people are honest and responsible when they are dealing with the AI, we cannot ensure the AI is honest and responsible when dealing people. This manipulation also violates the High-Level Expert Group on Artificial Intelligence (2019) that states "AI systems can contribute to achieving a fair society, by helping to increase citizens' health and well-being in ways that foster

equality in the distribution of economic, social and political opportunity".

REFERENCES

Allport, G.W. (1961), *Pattern and Growth in Personality*, Holt, Rinehart & Winston, New York, NY.

Bench-Capon, T.J. and Dunne, P.E., 2007. Argumentation in artificial intelligence. *Artificial intelligence*, *171*(10-15), pp.619-641.

Barrick, M.R., Mount, M.K. and Judge, T.A. (2001), "Personality and performance at the beginning of the new millennium: what do we know and where do we go next?", International Journal of Selection and Assessment, Vol. 9 Nos 1/2, pp. 9-30.

Bertrand, M. and Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, *94*(4), pp.991-1013.

Bex, F., Modgil, S., Prakken, H. and Reed, C., 2013. On logical specifications of the argument interchange format. *Journal of Logic and Computation*, *23*(5), pp.951-989.

Chesnevar, C., Modgil, S., Rahwan, I., Reed, C., Simari, G., South, M., Vreeswijk, G. and Willmott, S., 2006. Towards an argument interchange format. *The knowledge engineering review*, *21*(4), pp.293-316.

Dastin, J., 2018. Amazon scraps secret AI recruiting tool that showed bias against women. San Fransico, CA: Reuters. Retrieved on October, 9, p.2018.

Drucker, K., 2016. Avoiding Discrimination and Filtering of Qualified Candidates by ATS Software.

High-Level Expert Group on Artificial Intelligence (2019) 'High-Level Expert Group on Artificial Intelligence Set Up By the European Commission Ethics Guidelines for Trustworthy Ai', European Commission. Available at: https://ec.europa.eu/digital.

Kirschner, PA et al. (eds.), 2003, Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making. Berlin: Springer.

Lawrence, J., Snaith, M., Konat, B., Budzynska, K. and Reed, C., 2017. Debating technology for dialogical argument: Sensemaking, engagement, and analytics. ACM Transactions on Internet Technology (TOIT), 17(3), p.24.

Lundgren, H., Kroon, B. and Poell, R.F., 2017. Personality testing and workplace training: Exploring stakeholders, products and purpose in Western Europe. European Journal of Training and Development, 41(3), pp.198-221.

Militello, L., Lipshitz, R. and Schraagen, J.M., 2017. Making Sense of Human Behavior: Explaining How Police Officers Assess Danger During Traffic Stops. In Naturalistic Decision Making and Macrocognition (pp. 147-166). CRC Press.

Modgil, S. and Prakken, H., 2013. A general account of argumentation with preferences. Artificial Intelligence, 195, pp.361-397.

Oguego, C. L., Augusto, J. C., Muñoz, A., & Springett, M. (2018). Using argumentation to manage users' preferences. Future Generation Computer Systems, 81, 235-243.

O'Neil, C. (2016) Weapons of Math Destruction. First edit. New York.

Prakken, H., 2010. An abstract framework for argumentation with structured arguments. Argument and Computation, 1(2), pp.93-124.

Rahwan, I. and McBurney, P., 2007. Argumentation technology. IEEE Intelligent Systems, 22(6), pp.21-23.

Reed, C. and Rowe, G., 2004. Araucaria: Software for argument analysis, diagramming and representation. International Journal on Artificial Intelligence Tools, 13(04), pp.961-979.

Reed, C., Walton, D. and Macagno, F., 2007. Argument diagramming in logic, law and artificial intelligence. The Knowledge Engineering Review, 22(1), pp.87-109.

Thomas, J.B., Clark, S.M. and Gioia, D.A., 1993. Strategic sensemaking and organizational performance: Linkages among scanning, interpretation, action, and outcomes. Academy of Management journal, 36(2), pp.239-270.

Verheij, B. (2009) 'The Toulmin Argument Model in Artificial Intelligence Or : how semi-formal , defeasible argumentation schemes creep into logic', pp. 219–238. doi: 10.1007/978-0-387-98197-0.

Toulmin, S.E., 1958. The use of argument. Cambridge University Press.

Toulmin, S.E., 2003. The uses of argument. Cambridge university press.

Weber, L. and Dwoskin, E., 2014. Are workplace personality tests fair. Wall Street Journal, (September 29)