How sensemaking by people and artificial intelligence might involve different frames

Hebah Bubakr and Chris Baber

University of Birmingham, UK

ABSTRACT

Sensemaking can involve selecting an appropriate frame to explain a given set of data. The selection of the frame (and the definition of its appropriateness) can depend on the prior experience of the sensemaker as much as on the availability of data. Moreover, artificial intelligence and machine learning systems are dependent on knowledge elicited from human experts, yet, if we trained these systems to perform and think in the same way as a human, most of the tools will be unacceptable to be used as criterion because people consider many personal parameters that a machine should not use. In this paper, we consider how an artificial intelligence system that can be used to filter curriculum vitae (or résumés) might apply frames that result in socially unacceptable decisions.

KEYWORDS

Sensemaking, data frame model (DFM), artificial intelligence, human computer interaction.

Introduction

Sensemaking generally means to understand, comprehend and provide explanation for complex or uncertain events (Klein et al., 2006a). *"Sense-making is a motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively"* (Klein, 2006a, p71). There are many situations in which the decisions made by artificial intelligence (AI) might not be acceptable to humans (O'Neil, 2016). Taking job applications as a starting point, we consider how AI and humans might frame the decision to select an applicant in different ways. The idea is to develop an approach which could ultimately serve as a pre-mortem on decision models prior to their being applied. We begin by describing a motivating example.

Motivating example

Kyle Behm, a student who was looking for a minimum-wage job, applied for a part time at Kroger after his friend recommended him. Kyle 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, Kyle was not called for an interview and when he asked, he was told that he failed the personality test he answered during the application. These tests look into account individual motivations, preferences and differences between people. This personality test was used along with other factors like experience and interviews in the past but as the process is more automated these tests are used to eliminate applicants in early stages. So unfortunately, Kyle's honest answers to mental health questions always led to him being rejected by the job market.

Bias and personnel selection from curriculum vitae (CV) filtering

An experiment to study race in the labour market was conducted by researchers to investigate how CVs were processed and filtered (Bertrand and Mullainathan, 2004). The researchers sent around five thousand fake CVs for different job ads that were offered by newspapers. The CVs covered specific occupational categories and with each category the quality of the CV was divided into high and low. Then identities were assigned to each CV using a personal name to suggest the race of the applicant, for example, Emily and Brad suggest white names, while Kenya and Jamal would suggest African American names. The researchers found a significant difference between the two race's call backs. White name applicants receive 50 percent more call backs than African American names. Further, the call back rate for white applicants with a higher quality CV was statistically significant different compared to African-Americans with higher-quality CVs. Again, the results shown raises issues with decision making in the labour market, the question is whether algorithms in modern hiring systems have similar bias. Seeking to understand how such a system works and what are the beliefs, values and expectations of stakeholders we applied the process of sensemaking to the example above using the data/frame model (DFM).

DFM and sensemaking in interpreting CVs

To provide a reasonable explanation DFM divides the sensemaking process into seven elements: mapping data and creating an initial frame; elaborating a frame; questioning a frame; preserving a frame; comparing frames; reframing; and finding or seeking a frame. Therefore, the flexibility that this model provides helps with complex data and ambiguous environment. Using DFM we will try to provide an explanation to the hiring system process by explaining the relationships between the data in that environment and present what each stakeholder believes, and their expectation of the system.

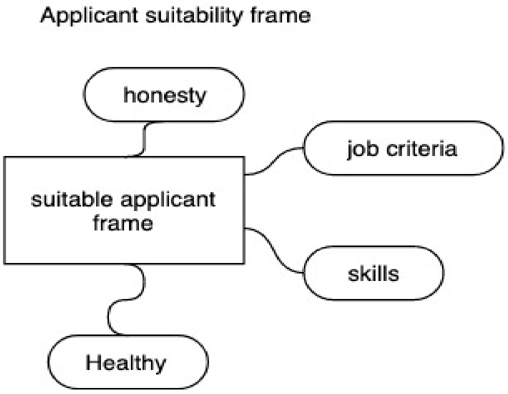

Figure 1: Applicant suitability frame

Applicant frame will address the users' (applicants') suitability for the job based on different elements like job criteria, and the skills they have to perform the job efficiently. Applicants will have an expectation of their suitability for the job based on their beliefs, so if all the data were met the user will expect to be accepted for the job or at least called for an interview, not eliminated completely in early stages. Hence, when Kyle was rejected, he did not accept the output of the system because in his suitability frame (he believes) he has all the needed elements for that specific job. Applicants can revise, change and compare their data elements and the relationships between them to influence the suitability frame.

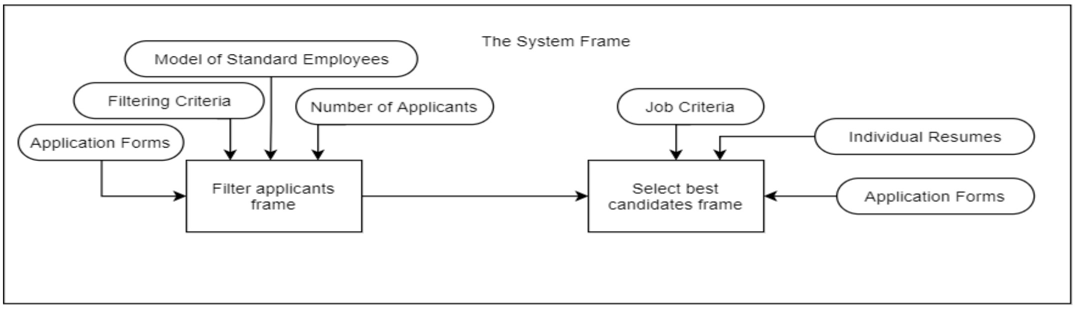

Figure 2: The system suitability system frame

To understand why the system rejected Kyle (and why other systems might eliminate applicants with African American names), we have a look at the system's frame and the data that affect its decision. Unlike the applicant's single frame, the system, in figure 2, has multiple frames. The first subframe is filtering the applicants, when the number of applicants exceed the required number, it is important for the system to filter and reduce the number based on specific criteria. The filtering frame builds decisions based on the data, which are the model of standard employees, the number of applicants, the filtering criteria and of course the application forms. However, these inputs open a wide number of options for filtering, especially when the number of applicants is large.

Additionally, any new applicants that do not resemble the standard employee of the company will be at risk of being rejected because their information does not match the frame's. This explains Kyle's rejection because his mental illness made him an unwanted example compared to thousands of other employees who had clean mental health records. Moreover, in this frame when the number is considered reasonable the system then will select the best candidates. The selecting frame will compare the application forms and the CVs to the job criteria and then rank the candidates based on their qualifications. Having applied these two subframes the system believes that it fulfils the whole frame by providing the best recommendations for the HR department. However, the main problem here is that the system frame of suitability does not match the ones of the applicant or HR. The system believes that Kyle is not suitable because he failed to meet a specific requirement which conflicts with Kyle's frame of suitability. Moreover, when the system also rejected individuals because of their skin colour, foreign name, gender or others that are not related to the job criteria, the system believes that it did a good job by selecting only the good candidates. However, these outcomes are socially unacceptable. The problem in this specific frame can be the bias: examples of successful data that trained the system, false algorithms that handled individual forms unethically or the criterion of judging. The final results do not match any ethical expectation.

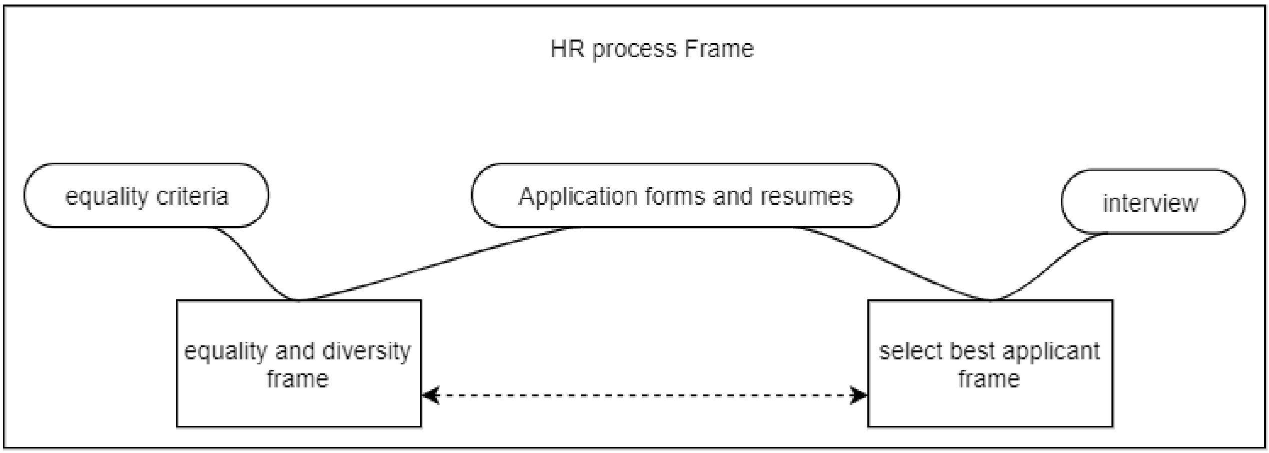
HR process Frame
equality criteria
Application forms and resumes
interview
equality and diversity frame
select best applicant frame

Figure 3: HR suitability frame

Similar to the system frame, the HR frame has multiple sub frames (see Figure 3). Using this frame HR can select the best employee for the offered job. By applying an 'equality and diversity' subframe, HR ensure that they treat all applicants the same. After ensuring equality the HR frame will use another subframe which uses interviews and evaluates individuals to select the best applicant. However, this frame is dependent on the results of the previous frame, so it is not guaranteed that this frame will lead the HR department to the right selection or even the ethical one.

Conclusions

The obvious problem in these frames is that they can only make change locally. Applying different DFM activities like preserving, questioning, changing, comparing and revising will be within the frame itself only. They have no ability to influence each other. Therefore, the result of Kyle's rejection or acceptance, for instance, has a different perspective in each frame. In the applicant frame, he was eligible. In the system frame he was not eligible because he failed the personality test in his application which is an important data element in preforming the frame. In the HR frame he might be eligible but his application was filtered and rejected by the system so he did not have the chance to be considered or interviewed by the HR department. Due to the tools that were used in the filtering process being false and unacceptable and utilising so many parameters that should not be considered. A key recommendation is that companies who develop these systems must teach the machine how to be fair by ensuring that the selected data are fair and equal to all the users in that environment. For example, filtering criteria should only be focused on knowledge, education, experience, not gender, name or race. O'Neil (2016) states *"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"*.

References

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):991-1013.

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

Klein, G., Moon, B., Hoffman, R. R., (2006 a). Making sense of sensemaking 1: Alternative perspectives. IEEE intelligent systems, (4):70-73.

Klein, G., Moon, B., Hoffman, R. R., (2006 b). Making sense of sensemaking 2: A macrocognitive model. IEEE Intelligent systems, 21(5):88-92.

Klein, G., Phillips, J. K., Rall, E. L., Peluso, D. A. (2007). A data-frame theory of sensemaking. In Expertise out of context: Proceedings of the sixth international conference on naturalistic decision-making, 113-155. New York, NY, USA: Lawrence Erlbaum.

Lundgren, H., Kroon, B., 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):198-221.

Duffy, M. "Sensemaking in Classroom Conversations," Openness in Research: The Tension between Self and Other, I. Maso et al., eds., Van Gorcum, 1995, 119-132.

Militello, L., Lipshitz, R., Schraagen, J. M. (2017). Making sense of human behavior: Explaining how police officers assess danger during traffic stops. In Naturalistic Decision Making and

Macrocognition (147-166). CRC Press.

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

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

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