Olympians: Part II

A Case Study for Responsible AI

**Disclaimer:**

This use case has been authored by Andreas Theodorou. While there is a significant number of differences, e.g. the inclusion of discussion point and role-playing activities as this is meant as a teaching use case, the overarching narrative (i.e. creation of an AI hiring system in a company with a homogeneous population) is based upon the case study “Hiring by Machine”, developed by University Center for Human Values (UCHV) and the Center for Information Technology Policy (CITP) at Princeton and can be found at: <https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/10/Princeton-AI-Ethics-Case-Study-5.pdf>

Their and this work are licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).Case Study, Part 2

The system (procured—or developed—in the previous part of this use case) was running for 3 years. During those years, Olympians invested a significant amount of money to make it work with high accuracy for the purposes and culture of their company. They retrained their entire model using resumes sourced from their HR department.

It had successfully handled over 5,000 applications when Marikou—a promising and hard-working computer science student from Lefkara in Cyprus—received a rejection email within hours of applying for a graduate job through the company’s website. She was surprised at having been tossed aside so quickly; she has been convinced that she was an ideal candidate for the company. She had strong academic qualifications. Moreover, she had carefully crafted her resume to reflect her civic commitments and experience working with non-profit organizations like the ones Olympians frequently have as clients. Marikou has mostly been involved in charities working with wheelchair users —such as herself. Her ambitions to develop responsible tech solutions to improve the lives of those with disabilities seemed a perfect match for Olympians’ core mission at benefiting the wider society. Adding to her surprise, the email came in the middle of a Saturday night, i.e. off-office hours.

Disappointed at her rejection, Marikou wrote to the company asking for feedback on her application. Human Resources (HR)—knowing that the General Data Protection Regulation (GDPR) requires an explanation for all decisions made by algorithm—asked a recruiter to investigate Marikou’s case. After having thoroughly reviewed Marikou’s application, a recruiter judged her to be on par with Olympians’ very best graduate hires. Based on her resume, the recruiter expected she would make an excellent addition to the company. He could not see a reason for her application being automatically discarded instead of being invited to an interview.

A major concern for HR was that that the system may have used Marikou’s disability status as a reason to deny her application. An internal investigation was initiated. However, the system’s engineers reassured HR that they had explicitly designed the algorithm so that it would not discriminate against protected categories. Their answer was: “Any text related to disabilities is not entered into the model and, hence, not used by the decision-making system.” But if it wasn’t her disability, then what was it that system had found in Marikou’s resume that had caused it to categorize her as a bad fit? What was it that the humans could not see?

After much digging, the developers found the unlikely answer: *sports*. It turned out that there was a strong correlation between participation in athletics and their propensity to excel at the company; the system had learned to connect a history of playing sports with ‘good fit.’ While it was true that many of employees no longer active participated in sports, their resumes reflected a history of having done so. Marikou, on the other hand, had never been interested in sports – nor any history of athletic activities due to her lifelong disability.

In the interest of openness and honesty, Olympians reached out to her and explained how the company had incorporated an AI system into its hiring processes. While the system was generally a success, the company admitted that there were still some bugs that would need to be worked out. In Marikou’s case, the AI system had considered her resume’s lack of references to physically demanding activities to indicate a weak cultural fit for the company. Olympians apologized on behalf of the company, invited her for an interview, and promised that the company was already searching to solve the problem. Marikou was dismayed to learn that Olympians had delegated decision-making in hiring—an area that could have a profound impact on her life prospects—to an autonomous system. Even worse, that system had then wrongly discriminated against her! Frustrated and angry, she published the company’s response on her website.

Marikou ultimately decided to reject Olympians’ offer for an interview and, instead, she filed an official complaint with the company, incorporating these arguments:

1. **Discrimination** against her as a person with physical disabilities. In fact, Marikou argued that given the history of marginalization and the lack of accommodations traditionally made for persons with disabilities, the fairest thing for Olympians to have done would have been to engineer the system to positively discriminate in favour of those with physical disabilities.
2. **Dehumanisation** for those personally affected by the system, the process of being converted to an “input” and assessed in this manner can feel dehumanizing. While humans are not perfect and they import their own biases into reviewing applications, but Marikou pointed out that they can care and empathize with applicants. When human agents reject worthy applicants, they may feel regret. The sense of dehumanization may extend to the Human Resources workers who had a central aspect of their job replaced by a machine – which ended up acting without oversight.
3. **‘Coerced’ consent** topass the data through an automated system is extracted at time of application. There is no option to deny consenting to the system and still successful apply for a position at Olympians. Arguably, at times of economic crisis and shrinking job markets, any job application counts.

**Discussion point 1**

Do you agree with Marikou’s points? If so, which and why?

At the same time, Olympians’ employees were unhappy that their resumes might have been used to train the underlying datasets without their knowledge or permission. They filed a similar complaint with the company. The legal department first investigated the claims related to the use of data. Olympians’ consent form, all applicants must agree to prior to the transmission of their data to the company, makes it clear that their data will be stored by HR and used to improve their recruitment practices. The consent form also makes it clear that they are subject to automatic processing and may be used in automated systems.

Olympians’ employees may have not only have agreed to this when they first applied for their position, but also their employment contracts make it clear that all data stored by HR can be used by the company and its agents. Afterall, comparing CVs of current employees to potential candidates, has been a common recruitment practice from before the AI system was first introduced. Moreover, as data used to train the model was anonymised the legal department believes that there was even less need to alert individual employees. Overall, the legal department was satisfied.

**Discussion point 2**

One issue neither raised in the complaints nor discussed by legal is about *data retrieval and deletion*. GDPR allows individuals to inquiry and request the delete their data. Scholars argue that personal data is any data that could be ‘singled out.’ Even if anonymised, with sufficient data points, any CV could be mapped to the individual. What if that data has already been used for training of a model? Can – and should – we delete selected data from a deployed model? An option might be to retrain the whole system from scratch; is this feasible?

The other major claim that the legal department felt obliged to investigate is related to *discrimination*. Anti-discriminations laws in Europe require both public and private organisations to provide *equal opportunities* by not considering specific protected characteristics as a disadvantage. In Sweden, one of the characteristics employees are not allowed to discriminate against is *physical* *disabilities*.

If the system had been directed to discriminate against applicants based on their disability status, i.e. by directly considering the disability status of Marikou, Olympians would clearly have violated Swedish law. But that was not the case. The system was not intentionally discriminating against resumes based on protected attributes; rather, “redundant encodings” in Olympians data had allowed the system to infer such attributes from other, seemingly innocuous data. After all, lack of participation in sports is not a protected characteristic. Thus, Olympians’ lawyers believed they could prove the company was legally in the clear.

**Discussion point 3**

Legal compliance does not guarantee ethical compliance –nor vice versa. Even if there was not a *legal* case of discrimination, was the company still morally culpable for its actions? Do you trust the company more given the legality of their actions?

The legal department during its investigation came across Article 22(1) of the GDPR:

“The data subject shall have the right not to be subject to a decision based *solely* on automated processing, including profiling, which produces *legal effects concerning him or her or similarly affects him or her*.”

In the context of GDPR, *solely* implies a decision-making process that is totally automated and excludes any human influence on the outcome. Olympians’ recruitment system decides and applies its decision without human intervention, i.e. it decides to filter out candidates and alerts them of their job application outcome without any human approval. The fact that HR decided to investigate and then altered the decision has very little bearing as Marikou *contested the decision* *after it was made*. If all profiling algorithms required human confirmation, then advertisements, movie recommendations, etc would simply not work. Therefore, there is a clause in this article of the GDPR: “legal effects concerning him or her or *similarly affects him or her*.” While Olympians’ hiring process does not have legal effects, based on Marikou, it may have *similar effects* – that is an effect that has an equivalent impact on an individual’s circumstances, behaviour, or choices. The legal question might be if getting hiring is produced or not such an effect. This *still* needs to be tested in court.

An additional question raised by the legal department relates to compliance with the upcoming AI Act. They realised that even if the system complies with existing legislation, it may not with the AI Act. It is unclear to them if it is a ‘high-risk system’ or not to be covered by the act.

Even if Marikou did not pursue her claim further, Olympians’ senior management decided to restore both the publics and their employers’ trust in the company. They tasked their procurement team to investigate the company’s response and come up with new recommendations.

**Group Task 2 — Company’s response**

*Task 2.1*

Answer the following binary questions:

1. Do you think Olympians should be held liable?
2. Do you think Olympians acted unethical?
3. Do you think Olympians should take down their tool?
4. If fixed, should the tool be rereleased?

*Task 2.2*

Do you think Olympians acted right by disclosing the insidious discrimination to Marikou? Or should they simply ‘cover it up’ by telling Marikou that they reconsidered their decision and want her for an interview?

*Task 2.3*

Do you think, even if legally within their rights, the company should do something for its disgruntled employees? If so, what?

**Group Task 3 — Prevention**

*Task 3.1*

Answer the following questions:

1. Did your ranking change of the 7 Requirements from Group Task 1? If so, what is your new ranking?
2. Do you think that any assessment process could have prevented Helena’s case? If so, provide example assessment criteria.
3. Can you recommend any other means of preventing such incidents?

*Task 3.2*

The type of discrimination practiced by the recruitment system might not seem obvious and shows how the emerging behaviours of a system are difficult to predict. This behaviour may end up amplifying biases and resulting to discrimination. Relying on only checking “what goes into a model” might not be sufficient.

Discuss the following 2 questions and come up with an answer:

1. Do you think that such ‘extreme’ cases (in this case, 1 out of over 5000) are an ‘acceptable risk’ or warrant turning off the system if they can’t be fixed?

2. How can we address this kind of insidious discrimination, which is, by definition, difficult to spot?