Event report: A Risk-based Approach to AI Services: Risk Chain Models and Recruitment AI

Introduction.

On July 15, 2021, the Institute for Future Initiatives at the University of Tokyo hosted an online event titled "A Risk-based Approach to AI Services: Risk Chain Models and Recruitment AI". While social implementation of artificial intelligence (AI) -based services and products has been boosted in recent years, issues correlated to the trustworthiness and transparency of AI have become a challenge. Various approaches have been developed by both public and private sectors, including the establishment of ethics guidelines and the development of risk prevention tools. However, newly designed risk management approaches have been encountering issues such as different core risks among AI services, managing AI models alone may not adequately remedy the risks, and human components including users may become a risk factor. The so-called Risk Chain Model (RC Model) invented by a research group at the University of Tokyo was introduced at this event to solve the abovementioned concerns. The panelists were invited to explain which risks are critical for AI-based services and products, who is responsible for associated risks, and how to consider risk countermeasures (e.g., tool selection) using recruitment AI as a case study.

Keynote Presentations

Introduction to RC Model and Case Study - Mr. Takashi Matsumoto

Mr. Matsumoto, a visiting researcher at the University of Tokyo Institute for Future Initiatives (IFI) and a member of Deloitte Tohmatsu Risk Services, is engaged in research developing the concept of a risk chain model (RC Model) which focuses on the linkage of various latent risks in the AI development to implementation cycle. He explained the risk chain model along with a case study based on artificial intelligence utilized in the corporate recruitment process. The RC Model provided an opportunity to consider implementing policies and principles to promote the appropriate use of AI in the industry.

In order to control the downside of AI, we must identify and manage risks from various perspectives. In other words, controlling risks in both technical and non-technical areas. Mr. Matsumoto claims the risk chain model (RC Model) serves two purposes. First, the RC Model should clarify the key risks of AI services unique to each organization and specify the value and purpose of using AI, then map the risk elements that may prevent it from achieving those goals. Secondly, we must reconsider whether inspecting AI models alone could fundamentally resolve the risks that have been identified. Unintended negative consequences can occur due to inherent biases in the data utilized to train the model, as well

as changes in the implementation environment including shifts in data formats and updates to the AI application process. As described above, the risk chain model aspires for a comprehensive approach to evaluating risks in the application of AI.

Risk Chain Model (RCModel) Structure

The RC model consists of three layers: the technical layer of AI systems, the providers/engineers of AI services, and the users of AI services. The technical aspects of AI and the provision of services are important perspectives for comprehending the risks.

There are four important technical aspects of AI models including the capability to perform sufficient prediction accuracy, high generalization ability, resistance to noise, and the ability to understand the decisions made by AI. Furthermore, the data required to build AI services must be highly accurate and free of bias.

The requirements for service providers consist of three elements: code of conduct, structured AI operations, and communication. The code of conduct must put the users of the service first, ensure fairness and privacy, and disclose necessary information to users. It also requires, as part of service management, the scalability and sustainability of the organization as a service provider, the ability to respond quickly to AI deficiencies, ensuring the security of the service, managing access rights to the service, and creating a system that allows third parties to verify the AI.

Panel Discussion

The panel discussion on recruitment AI was facilitated by Professor Arisa Ema and Mr. Matsumoto. Panelists were invited from a variety of professional backgrounds, including those with experience in technical development, recruitment AI services, and AI risk governance research to discuss the topic from various viewpoints.

Explanation of "R001: Appropriate Evaluation - R008: New Positions" by Mr. Matsumoto

Before opening the discussion to the panelists, Mr. Matsumoto explained the primary risk factors regarding the real-world application of AI. As a basic framework of the RC Model, there are four values and objectives along with 16 risk scenarios in the utilization of AI. The core four values and objectives are defined as "maintaining the predictive performance of AI," "collaboration between AI services and users," "responding to trend shifts in the business environment," and "maintaining corporate social responsibility through adherence to ethics and compliance". In below, 16 risk scenarios will be explained based on the use of

recruitment AI in the hiring process.

First, there are four main risk scenarios to overcome for achieving and maintaining high predictive performance. AI will not be able to perform an "R001: proper evaluation" of an applicant unless appropriate expectations are set for the position for which the applicant is being hired. There is also the difficulty of "R002: maintaining predictive performance," the hiring quality or candidate selection quality may decline unless acknowledging the declining predictive performance of AI. There are concerns about "R003: noise effects" when AI is processing the motivation letters, resume, or required essays. The algorithm could score the same contents differently due to the use or position of punctuation marks or periods. Finally, there is the risk of "R004: False Application," in which an applicant may be hired even though the information submitted by the applicant is incorrect.

Diverse risks also exist between AI services and user collaboration. “R005: Excessive AI dependence” addresses the concern in which recruiters place too much trust in AI and overlook the wrong decisions made by AI models. In addition, "R006: False Feedback" occurs when the AI is trained on new data to improve its performance and is caused by mislabeling of the teacher data.

Thirdly, recruitment AIs need to reconsider the description of an ideal candidate adapting to transitions in the business environment. If the AI prioritizes recommending the same candidate types continuously ignoring the shift in demands, there is a risk that the AI will not be able to adjust to the changing “R007: human resource trend”. In addition, it will be difficult for AI to make appropriate decisions for "R008: new job types" for which there is no available teaching data necessary for AI to build a recommendation model.

Comments from Yosuke Motohashi, Senior Data Analyst, AI and Analytics Division

As the creator of the system, the risk we are most aware of is the changing “R007: human resource trend”. AI models are trained based on previous data, hence one concern is their inability to adapt to shifting hiring conditions. However, pursuing AI that follows current trends in a cutting-edge-oriented manner may make it difficult to recruit personnel who have inherited the company's traditions and culture. Another risk associated with the recent use of deep learning in recruitment AI is the issue of black boxing: it is impossible to understand why one candidate is preferred by the algorithm over one candidate through complex analysis.

Comments from Ms. Shizuka Shimizu, President of Funleash Co.

From the standpoint of a person in charge of corporate human resources, we recognize the importance of addressing the problems caused by the combined risks of R001 to R0004 (appropriate evaluation, maintaining predictive performance, noise effects, and false application) as explained by Mr. Matsumoto. Certain aspects should not be overlooked from a human resource perspective while the use of AI is meaningful in improving recruiting efficiency and reducing costs in selecting appropriate personnel from a large pool of applicants.

First, AI-driven recruitment methods may encounter difficulties considering three major recruitment factors as candidates are selected through algorithmic labeling. The first aspect is competency: candidates' experience and abilities, which is critical in the hiring process. The second is the job description: candidates' responsibilities and expectations upon joining the organization, which is considered to determine the fit and suitability of each applicant. The third factor is the candidate's motivation, personality, and vision for the future which are particularly valued in the current job market. These three concerns raise the question of whether the uniqueness of the individual can be properly labeled in the candidate evaluation process to ensure a superior hiring process.

Secondly, the mystified or black-box nature of the candidate selection made by the AI is a shared issue among several parties who are using recruitment AI based on previous interview results. We cannot be accountable recruiters without a clear understanding of the algorithm's decision criteria, meanwhile, we also have a problem that the general public may have difficulty comprehending the AI-based recruiting process when explained to them.

Comments from Hideaki Shiroyama, Professor at IFI The University of Tokyo

Mr. Motohashi and Ms. Shimizu raised concerns about AI, however, we must admit that humans may share a common dilemma that AI-based recruiting is facing. Therefore, I think it will be important to identify relative strengths and weaknesses in AI recruitment.

It is important to set reasonable expectations for candidates for an "R001 appropriate evaluation". Although, recruiters, as well as artificial intelligence, may not be able to verbalize the criteria to describe what is optimal for the role. Recruitment is not all about hiring similar types of people simply because of their suitability; there is a need for diversity within the workforce. There are also unspoken nuances behind recruiting, such as the idea of hiring not only the right person for the job but also a candidate who can carry on the

company's traditions to preserve the organizational DNA. These nuances are not always clearly defined as each organization is aware of them differently.

Explanation of "R009 Cost Overruns" - "R016 Privacy Protection" by Mr. Matsumoto

There are risk factors prohibiting from achieving recruitment AI responsive to the changing environment. For instance, using AI services to maintain and improve staffing levels could easily be more costly than before: "R009: cost overruns", when a company aimed to implement an AI system for reducing the cost associated with recruitment. Aside from transitions in business requirements, recruitment strategies must be tailored to the different characteristics of applicants in different regions. Without training AI models specific to each region, there is a risk of not being able to "R010:adapt to local communities". Developing an artificial intelligence capable of catching up with business and local conditions could place a heavy burden on the development system leading to "R011: inadequate development speed".

The fourth value of the risk chain model is compliance with corporate social responsibility. Though there is concern that ethics and compliance will not be prioritized as the AI model is bound to mechanically evaluate the information provided by applicants according to an algorithm. Therefore keywords and information that can be used to gain an unfair advantage in recruitment could potentially be used against the AI model by "R012: selling decision-making basis information for illicit profit". There is also concern about the problem of "R013: Fairness," in which unfair judgments are made against applicants who belong to certain attributes such as country, region, race, gender, and religion. Furthermore, there could be "R014: use of prediction results for other unintended purposes" in which the evaluation made by the algorithm in the candidate selection process is used by HR against their intentions in career development upon joining the company. Finally, there is a risk of "R015 reputational damage" to individuals and "R016 privacy protection" if the information on applicants is leaked externally.

Comments from Mr. Motohashi

Improvements can be made regarding "R012: selling decision-making basis information for illicit profit" and "R013: fairness" from the engineering perspective. In particular, incorporating the practice of different recruitment strategies adjusted to each region and candidate pools to AI models is essential regarding "R010: adapting to local communities". While AI is good at making decisions based on averages values of an existing dataset, it is often considered not good at processing new strains of data with idiosyncrasies. Therefore, it will be difficult to accommodate candidate diversity. Engineers are now required to

reproduce the actual circumstances where hiring decisions are made based on complex judgments from various perspectives, rather than simply hiring people with high positive deviation values derived by machine-made algorithms.

Comment from Ms. Shimizu

There is significant concern over candidates' consent regarding the usage of their data during the AI-drive recruitment process. We must have clear communication with candidates on how their data is intended to be used for decision-making. Moreover, we need well-organized data management for remedying occasional human errors and providing continuous feedback by humans on candidate selections made by AI under the premise that artificial intelligence cannot always make the correct decision. It is also important to attain diverse hiring through the use of recruitment AI by monitoring candidates falling into specific categories to ensure that they are not disadvantaged.

Comments from Prof. Shiroyama

We believe that a recruitment AI with a function that aims to support the hiring of people with various characteristics or backgrounds, rather than homogeneous hiring by HR would contribute to ensuring "R013: fairness" in corporate recruitment.

Case Discussion on Recruitment AI

- Risk Chain Consideration for "R001: Appropriate Evaluation"

As a possible scenario for discussing recruitment AI in terms of appropriate evaluation, the discussion was based on the premise that AI's contribution cannot be properly evaluated without setting appropriate expectations for each hiring position. Five perspectives were introduced that are important to achieve proper evaluation. First, the recruitment AI service provider has the accountability to set appropriate target values for each job category according to the hiring strategy. Secondly, multiple AI systems should be used for a single position to offer scalability for analyzing candidates from multiple perspectives. Thirdly, ensuring sufficient technical capacity to run systems adequately since multiple models may be used at the same time. Fourthly, it is essential to maintain data balance by securing new training data. Finally, accuracy is highly influential to examine whether the goals set for each job opening have been achieved. Traceability is another important aspect to accumulate user feedback, audibility is necessary to verify the performance of the AI model for each job category and company, and sustainability must be achieved by re-training is necessary to maintain the level of recruitment.

Comments from Mr. Motohashi

I think it is very difficult to define a well-performing recruitment AI, although it is necessary to verify its performance to maintain accountability. Recruitment AI is part of the recommendation AI family, yet we must keep in mind that the algorithmic model is not performing at the same level of complexity required to facilitate product purchases: performance is simply determined by customer purchase success rate.
Mr. Motohashi argued that feedback on hiring results will be important to improve AI's performance, and we should evaluate the match rate between the candidates recommended by AI and those hired, and collect direct feedback from candidates after the hiring.

Comment from Ms. Shimizu

Hiring is not the primary goal of human resources among many other responsibilities. Finding the candidate that makes contributions to the organization over the long term is significantly important. Even a very good candidate will be considered a failure in terms of hiring if he or she quits within a few months or in a year. To improve recruitment AI, we should establish a system that provides feedback throughout the entire human resource value-chain for AI performance improvement, from hiring to placement, evaluation, and employee development.

Comments from Prof. Shiroyama

We feel that it will be difficult for the foreseeable future to secure the necessary training data for each type of job. However, Prof. Shiroyama pointed out that we will not run out of study data in the long term if we start accumulating recruitment data now. It is vital to be aware of aspects that are important for post-hiring analysis and that we should not simply analyze the relationship between post-hiring performance and application documents, but should also focus on training, placement, career paths, etc.

- Regarding "R006: False Feedback."

One risk that should be addressed from the user side of AI is incorrect feedback since inaccurate pass/fail labels can degrade the AI's performance. There are five important aspects to consider when examining this risk. First, the effectiveness in terms of users' understanding that incorrect feedback can degrade the model's performance. Secondly, the controllability over the utilization of the feedback collected from each company to improve the performance of the trained model. Thirdly, data must be properly organized to provide accurate feedback as well as maintain data quality to verify the accuracy of the data in the AI system. Fourth, it is important to calculate the prediction accuracy at the stage of training

and re-training the AI model before implementation and verify whether there are any abnormal trends hidden in the user's feedback. For instance, we must detect the abnormal generation of teacher data. Fifth, sustainability is indispensable for the effective use of AI as it includes pursuing the causes of degraded performance of AI models by reverting to previous models if necessary, and cleansing teacher data if there is inappropriate data.

Comment from Ms. Shimizu

We must acknowledge that the user who creates the feedback has an attributional bias for providing appropriate feedback to the AI model. I believe that we can make the most of AI's attractiveness in recruiting by understanding and minimizing the risks that humans have and providing appropriate feedback since artificial intelligence could only be a product of our learning. Hence it is up to the user to unlock the potential of AI in the recruitment process.

Comments from Mr. Motohashi

There are systematic and non-systematic approaches to dealing with erroneous feedback. One systematic approach would be to identify the characteristics of the evaluators who create the feedback and have the AI recognize each of their tendencies to reduce bias. Another possible human-based solution would be to have multiple evaluators create feedback for each candidate so that if an incorrect evaluation is made, the opinions of other evaluators can be taken into account to ensure that no mistakes are made in the total evaluation.

Comments from Prof. Shiroyama

In general, the method of the feedback given to the AI model seems to be labeled as either pass or fail. However, I am curious about the possibility of converting the training data into flexible evaluations for each candidate, such as "will be an immediate asset" or "has potential for long-term growth". For being conscious of the diversity in recruitment, Prof. Shiroyama argued that the creation of a feedback system learning the characteristics of candidates in more detail is ideal. Furthermore, it is important to design a system that is expecting a wide spectrum of evaluation of candidates, such as personnel after being hired, to provide appropriate feedback to AI.

- **R013 Fairness**

Recruitment should not utilize AI models that produce biased decisions for candidates associated with specific countries, regions, races, genders, or ages. While different countries

and regions have different standpoints on fairness, there are three major points to keep in mind to control the risk related to fairness. First, it is important to be aware of data balance and to avoid bias in the training data. Secondly, it is important to prevent the AI model from unintentionally excluding certain candidate groups from the recruitment process when generalizing the data. For example, if a company that has not actively recruited women in the past implements a recruitment AI based on historical data, the AI is likely to make decisions that favor men based on historical statistical trends. Similarly, when a candidate's ethnicity or faith is mentioned in the candidate profile or cover letter, the algorithm should be prevented from making unfavorable decisions based on keywords that it is not accustomed to processing. Finally, the fairness of the AI model ought to be examined to assure biases local to each organization or region are not corrupting the AI's decision-making. For instance, there is a high possibility that the AI will not be able to make appropriate decisions due to unforeseen training data of foreign applicants when recruiting foreigners in Japan. Whenever there is a risk involved with the utilization of AI, it is necessary to elucidate the issue for transparency and build consensus among users regarding the inability to make appropriate predictions. Recruiters must understand the imperfection of AI models to make accurate judgments, and consider flexible measures such as relying on human judgment depending on the situation.

Comments from Mr. Motohashi

Mr. Motohashi claimed that AI will always evaluate candidates through their attributes and labeling to distinguish certain candidates as better. Therefore, it will be difficult to implement the definition into a recruitment AI unless we start with a transparent definition of fairness. He emphasized the importance of evaluating the AI model apparently practicing the definition of fairness retrospectively. For example, part of the evaluation should be to ensure that there is no discrimination among candidates based on gender, faith, etc., and it is worthwhile to consider how unavoidable bias can be rectified.

We believe that the unfair bias created by AI can be remedied through a review of individual evaluations and by considering the entire applicant pool as a group. If the AI scored with bias due to gender differences, an algorithm could be implemented to force the scoring to be corrected for each applicant. On the other hand, our natural solution in hiring is to adjust the entire applicant pool as a group so that the final gender ratio is close to equal. We believe that human judgment should be present in the hiring process to make hiring AI fair to applicants.

Comment from Ms. Shimizu

The issue of fairness in hiring is not limited to AI, yet unfairness can still occur even when recruiting is primarily carried out by humans. Fairness is still difficult to guarantee even if an organization acknowledges and clarifies the concept of fairness in hiring. Therefore, it is necessary to communicate to employees that recruiters will make maximum efforts to recruit and assign personnel fairly to the greatest extent possible. In case anyone feels that they have been disadvantaged, it is necessary to design a system that allows them to file a claim and seek clarification on where the responsibility lies. Recruitment and staffing decisions are based on the context within the organization, so we may dare to hire people of a younger age to lower the average age in the team, or actively recruit people of a certain race to provide racial diversity. We feel it is also important to explain the background of such hiring. Ms. Shimizu expressed the importance to be aware that there is no perfect recruitment process. Hence willingness to attain an ideal recruitment process and to create supporting systems that supplement the inadequacies is essential in utilizing recruitment AI.

Comments from Prof. Shiroyama

Prof. Shiroyama argued that the definition of fairness is very vague and that the concept of fairness is fluid in recruitment. A system that is capable of capturing the necessary change in the definition of fairness tailored to each organization would be highly effective for a fair recruitment. Therefore, we recognize the importance of combining the open-door policy that Ms. Shimizu introduced as well as the voice-collecting system that can provide input related to recruitment and human resources as a complementary system along with recruitment AI.

Summary: Risk Chain Model (RC Model) and future development

We expect that the risk chain model introduced in this event report will contribute to building consensus among relevant stakeholders to control risks. We should expect the use of AI in organizations will become more active as the technological capabilities of recruitment AI continue to increase, especially in the area of human resources. Since it is up to us to maximize the potential of AI, we believe that we should not simply use it to manage and select candidates, but rather pursue ways of using it that will make all relevant stakeholders happy. We should also collect data on various risk chain cases surrounding artificial intelligence in order to formulate policies and regulations in a data-driven manner to guide the appropriate use of such advanced technology.