Probing Athletes' Perceptions Towards Electronic Judging Systems – A Case Study in Gymnastics

Elena Mazurova
Aalto University School of Business
elena.mazurova@aalto.fi

Esko Penttinen
Aalto University School of Business
esko.penttinen@aalto.fi

Abstract

We study athletes' perceptions towards the transition to electronic judging systems. Using purposive sampling, we select an area of sports that is undergoing a somewhat disruptive change in the way athletes are evaluated: gymnastics. We draw on interviews conducted with gymnasts to probe their perceptions of electronic judging systems. We find that gymnasts are quite positive towards the implementation of these systems, although they expressed some uncertainties (i.e. how these systems influence the artistic side of gymnastics) and risks (i.e. technical problems) of the technology. The positive side of the transition to electronic judging systems mainly relates to the deficiencies of the human-based judging, it being vulnerable to biases, human error, human fatigue, judges' personal preferences, and inherent lack of explanation. Our informants expressed that electronic judging systems contain affordances that could efficiently mitigate the said challenges associated with human-based judging.

1. Introduction

Many sports are undergoing a transition from human-based judging to electronic judging systems. In tennis, the Hawk Eye system helps the chief umpire to determine whether the ball was in or out [16]. In soccer, the Goal-line technology gives the umpire a notification when the ball crosses the goal line [5]. In gymnastics, an electronic judging system is being developed which captures the gymnast's movements with 3D laser sensors and gives suggestions on scores [8]. Often, this development towards electronic judging systems is due to athletes becoming stronger, faster, and better, thus making it difficult for the human eye to accurately make judgments.

Most electronic judging systems are simple rule-based systems coupled with sensor technology. In some cases, however, the electronic judging system is developed using sophisticated artificial intelligence (AI) tools. This is the case in gymnastics where Fujitsu is using machine learning to teach the electronic system how to distinguish between pure and impure performances[1]. While the athletes would most probably have no problems in accepting the decisions of rule-based systems, judging based on AI might be exposed to the phenomenon called explainable AI, where users of the system would require explanations on the outcomes of AI.

Motivated by the recent surge in research interest towards explainable AI and the introduction of electronic judging systems in many sports, we ask: *"How do athletes perceive the introduction of electronic judging systems?"*. We are especially interested in probing how trust is established so that gymnasts feel comfortable being judged by the electronic system.

To be able to respond to the research question outlined above, we collaborated with the Finnish Gymnastics Association and conducted interviews with gymnasts, coaches and directors to get a holistic understanding of how the electronic judging system is perceived by different stakeholders, most notably by the gymnasts themselves.

We proceed as follows. After this introduction, in Section Two, we discuss the notion of explainable AI and link it to electronic judging systems. In the third section, we outline our methodological choices regarding the use of the grounded theory approach. In the fourth section, we present the findings of the empirical study and in the remaining sections, we draw conclusions and offer avenues for further research.

[1] See the video on Fujitsu's system at: https://www.youtube.com/watch?v=jHRQxtbh3uw and announcement of the system at: https://medium.com/syncedreview/meet-fujitsus-ai-gymnastics-judges-8cb52613b2a.

URI: https://hdl.handle.net/10125/64285
978-0-9981331-3-3

2. Literature review

2.1. Performance evaluation in sports in general and gymnastics in particular

Performance evaluation in sports containing an artistic side is notoriously difficult. Prior literature informs us that, in comparison with other types of gymnastics, judging of artistic gymnastics suffers from a lack of quality, accuracy, fairness, validity, and reliability [10]. Athletes suffer from the biases and human errors of the panel of human judges. Judges tend to award higher scores to athletes from their own country, give the same scores as other members of the jury, and assess gymnasts based on their previous training capabilities. Judges' accuracy of visual perception can be also affected by their sitting position or the angle of observation. The judges' personal preferences and many other factors affect judges' decision-making on deductions and total scores [15]. Further considering the biases of the judging process, prior literature has identified two groups of factors negatively influencing on judges' decision-making: subjective personal preferences and technical human error, which is caused by, for example, tiredness, blinking or distraction [6, 12].

One of the most influential factors of personal preference is national bias, which often exists in the judging process, especially at large international competitions [10]. Judges' over-scoring or under-scoring of the gymnasts, based on the national attribute, influence not only on the quality of judging but also on the gymnasts' overall ranking in the general standing [10]. Despite the fact that according to the current Code of Points in the judging system, judges cannot judge the athletes of the same nationality in apparatus finals, national biases in favor, for example, neighboring countries, countries with the same or similar political, ethical or religious structure may still exist [6]. Prior research identifies one more important personal factor influencing human judges' scores and deductions: a judge's comparison of the current athlete's performance with the previous one [9]. The performance of the previous athlete and his/her score was found to influence the judge's perception of the following athlete's performance.

Many researchers assume that it is almost impossible to eliminate biases without supporting systems, as the factors affecting them are inherent human features. As a result, electronic judging systems have been conceived to mitigate these inherent challenges associated with human-based judging systems.

2.1.1. Electronic judging systems. To increase such indicators of judging system of artistic gymnastics as quality, accuracy, fairness, validity, and reliability, several researchers recommend using supporting electronic systems [2, 15]. They assume that the factors negatively influencing the judges' objectivity, accuracy, and impartiality in artistic gymnastics, can be if not eliminated then at least smoothed by using smart computer technologies and electronic judging systems during the competitions [9]. Such functions, available for the judges in real time, as video-recording, video replay, slow-replaying, and time-lapse play as well as different kinds of sensors, catching athlete's fulminant movements and measuring the execution time, may significantly improve reliability of the judging system and reduce conformity bias and arithmetic errors in the scoring of athletes' performance [15]. The use of electronic judging systems increases the level of fairness and impartiality in competitions [2].

Electronic judging systems were introduced in sports in mid-1990s. In various sports, the use of new electronic judging systems was triggered primarily due to the inclusion of these types of sport in the official program of the Olympic Games [20]. After the Olympic games in Beijing, a new intensive wave of development and improvement started, as many disadvantages of electronic judging systems prevented them from fitting the strict requirements of the judging process at the Olympic games [20]. Nowadays, the application and a constant process of development and improvement of electronic judging systems have become even more intensive [2, 12]. Research in different fields of Olympic sports shows that the comprehensive use of electronic judging systems will be soon an integral part of all international championships [2].

The use of electronic systems for assessment of the athlete's progress is extensively studied, resulting in promising conclusions about their efficiency, accuracy, and fairness in terms of judging criteria [15]. Such attributes of electronic judging systems as high-speed digital image acquisition devices catching the athletes' body movements in three-dimensional space, and image recognition processing software allow to elevate the judging process to a completely new level and significantly decrease the scope of human error that, in turn, decreases the number of complaints and inquiries submitted by coaches [2, 15]. The use of electronic judging systems at the international competitions may partly decrease the influence of human factors on the judgment process and improve the quality of competitions, making them more demonstrative and exciting for the audience [4]. Thus, the use of electronic judging systems in sports increases the objectivity and clarity of judging and has a positive effect on the technical development of the performance evaluation in sports in general and gymnastics in particular [2, 4, 15, 20].

2.1.2. Explainability of electronic judging systems. In our paper, we focus on electronic judging systems that are "intelligent", referring to systems that employ machine learning and thus differ from rules-based judging systems. Given that a formal, commonly-agreed upon definition of the term "explainable AI" has remained elusive [18] it is important to define explainability. The main source for the difficulties in defining explainability lies in its non-monolithic nature [11]; explainability refers to more than one concept. Explainability needs to be conceptually separated from a related concept: interpretability. *Interpretability* refers to the degree of being understandable to an observer, often a technically versed person [1, 13]. *Explainability*, in contrast, is an outward-oriented and social concept by nature because it entails explanations between two or more (human- or machine-based) agents, called explainer and explainee [13, 19]. As a result, creating a shared meaning between these agents is important when pursuing a higher degree of explainability. To further conceptually separate between interpretability and explainability, we define interpretability as a necessary condition to explainability. In other words, in order to be able to explain the operations and outcomes of AI between agents, one first needs to build an interpretation, a translation of things to understandable format. Existing literature makes a distinction between ante-hoc explainability and post-hoc explainability [7, 11, 14]. *Ante-hoc explainability* denotes approaches geared towards improving the transparency of the mechanisms by which AI systems work. By definition, ante-hoc explainability occurs before the event in question, for example, by incorporating explainability directly into the structure of an AI-model [7]. *Post-hoc explainability*, on the other hand, refers to interpretations that might explain predictions without elucidating the mechanisms by which models work [11]. Post-hoc explainability occurs after the event in question, for example, by explaining what the model predicts in terms of what is readily interpretable [7].

3. Methodology

Since our study aims to improve understanding of human perceptions related to technological change, we chose to conduct a qualitative case study following the Grounded Theory Methodology (GTM) which has often been selected for the study of technological change in emerging research domains in information systems research [21]. It provides strategies and systematic procedures for conducting rigorous qualitative research that requires shaping and handling of qualitative materials [3].

3.1. Case selection and data collection

To select the empirical setting and informants, we employed purposeful sampling to find a relevant, information-rich empirical setting [17]. To identify such a case and informants, the following criteria were used. First, the setting would need to be such that it would be in the process of undergoing a transition from human-based judging to electronic judging systems. Second, the informants would need to be athletes that will be affected by the transition to electronic judging systems (e.g. participation in international competitions at the senior level in World Championships). Based on these criteria, we decided to focus on gymnastics, which is currently undergoing a transition to employing AI in judging systems. To collect empirical qualitative data, the study focused on conducting semi-structured theme interviews. The data collected included nine interviews with gymnasts, directors, and coaches (see Table 1 below). All interviews were conducted in April-May 2019. When developing the interview questions, our primary aim was to initiate an intensive sharing of the participants' opinion about an existing judging system and their perceptions and expectations from the new electronic system as well as comparison of them both. Thus, our interview included open questions which were slightly revised after the first interview. Due to space limitations, the interview questionnaire is available from authors upon request.

Table 1. List of interviews

Interviewee	Role	Familiarity w/ electronic judging system
James	Gymnast	Average
John	Gymnast	Average
David	Gymnast	Expert
Thomas	Gymnast	Average
Mark	Gymnast	Novice
Steven	Director	Expert
Mary	Director	Expert
Paul	Coach	Novice
Kevin	Coach	Novice

3.2. Data analysis

All interviews were tape-recorded and transcribed to enable efficient analysis through Atlas.ti, a software for qualitative data analysis. The data were analyzed using three coding techniques: 1) open coding, 2) axial coding and 3) selective coding. First, open coding was used to identify common ideas, opinions and patterns among different participants' interviews. This was done via cross-reading and comparison of the interview transcripts and identification of common ideas and opinions among them. Thus, we identified several main opinions and perceptional patterns, which were similar among different participants and based on them we formed different groups according to the main concepts and patterns. Those opinions, which did not fit any of the formed groups were analyzed separately. Second, in order to identify the relationships between different groups and patterns we used axial coding. Identification of the similarities and patterns was done among the groups as well as verification and confirmation of these patterns were done within the groups. Third, to integrate our detected concepts and the theory and to build theoretical propositions of our study we used selective coding. This part of the analysis includes both identified patterns corresponding to the existing theory as well as some new ideas, which were formed based on interview results.

4. Findings

In gymnastics, a panel of judges usually consists of 6-8 judges. The judges are divided into difficulty and execution judges. The *Difficulty score* (D) evaluates the content of the exercise on three criteria: difficulty value, composition requirements and connection value. The *Execution score* (E) evaluates the performance according to the execution and the artistic impression of the routine. The base execution score is equal to 10.0. During the routine, the judges take away points (make deductions) from this base score for small (0.1 deduction), medium (0.3 deduction) and large (0.5 deduction) errors in artistry, execution, technique, and composition. For falling off the apparatus the deduction is 1.0. The D- and E-scores are summed-up at the end of the routine and form the gymnast's final score on each apparatus.

Next, we turn to presenting the results of the analysis of our empirical data on three fronts: perceptions of challenges with human-based judging, perceptions of the AI-based judging system, and perceptions of trust in AI-based judging system.

4.1. Perceptions of challenges (biases, errors, lack of explanation) with human-based judging

One of the disadvantages of the current judging system, mentioned by all participants of our interviews, was the big variance in the deductions and final scores that judges make.

Steven: "*It's becoming more and more complex. You have to be really precise with the deductions and the execution. And there are so many different things that you can take points off from execution and as a judge you have to remember so many different things and you have to see and register so many things during maybe one second period or two seconds period. And the gymnasts are better, faster, stronger every year."*

This variance exists due to a variety of different "human" factors affecting the panel of human judges that, according to the opinion of our respondents, lead to the unfair judging system. These factors are human error, human fatigue, judges' personal preferences, overly critical fault-driven approach to judging, and lack of explanations on deductions.

4.1.1. Human error. According to the gymnasts, all routines are performed fast, and the judges have to observe the athlete, judge him/her and write down the deductions, all at the same time. Thus, it is hard for them to see and notice each particular detail each particular second of the routine. As a result, sometimes the gymnasts feel that judges can make mistakes. However, as this problem of human error has existed in the artistic gymnastics judging already for many years, many athletes accept this problem just as an inherent feature of gymnastics.

David: *"In my personal experience, one of the worse aspects of human panel judges is that it's very subjective. You're doing a routine but one judge is going to give this score, another judge is going to give you another score. There is no absolute definite agreed upon judging criteria. There's always a human error aspect."*

Steven: *"It's almost impossible for a human eye to register all those mistakes and to write them down because at the competition the routine continues immediately. So, in that case, in some aspects, it's too complicated for humans."*

4.1.2. Human fatigue. At international competitions, judges have to sit and judge the athlete's performances for eight hours continuously. Informants cited human fatigue as something that influences the judges' capabilities and the value of deductions that they take.

James: *"Judges get tired during the whole day and the whole competition. In the morning they are fresher and in the evening till the last routine, they're more tired"*

David: *"Even if you are going to a World championship, European championships, they are very experienced judges, they know what they're doing, but they're operating very long hours, they have many days of competition. I think there's a point of exhaustion sometimes by the end of the day."*

Thus, according to our respondents, there is a difference with regard to the final score for each routine, in what time of the day to compete. Overall, our informants perceived that if you compete in the morning, you will get a lower score, and if you compete in the evening, you will get a higher score.

John: *"It's always like this: if you compete in the morning, judges are harder on you, they easily take away much more points. They want to be good, strict and do their job properly. And in the evening, they get tired, seeing the same thing over and over again. Thus, if you compete in the morning, they can take a bigger execution or deduction, and in the evening, if you do exactly the same mistake they're will not take so much from your total score."*

James: *"If you compete earlier in the morning when judges are more awake, and in the evening they've been sitting there for ten hours and they started to get tired, it's a little bit easier to score higher points, when they're tired."*

The competing time is usually chosen by the representatives of the Federation of all countries, participating in the competition at the closed session, organized by FIG (International Federation of Gymnastics). However, one unspoken rule exists:

John: *"In the morning sub-division, the worst countries are going, in the middle sub-division, better countries, and in the evening, the best countries are going. So, it's always better to be in the end. Finland is always either in the first or the second sub-division. Not very often in the last one."*

However, a gymnast can really do nothing to change his/her competing time and get into the evening sub-division.

4.1.3. Judges' personal preferences. Another aspect that may often affect the judges work is their "human" personal preferences, according to the opinion of our respondents. These preferences may vary depending on judges' knowledge of some gymnasts and knowing his/her personal progress outside of the particular competition within the country. As the gymnastics community in Finland is not so big, many judges are former coaches, who are used to train different gymnasts in different clubs before they became judges.

Mark: *"If the judges are from the capital city and the gymnasts is also from the same city, maybe they will be giving a higher score to "their" guy."*

However, at international competitions, there are also similar kinds of human judges' preferences, but at more "global level":

Thomas: *"Of course, if you have a friendly country, there's always such thing as, ok, I'm not taking that much from him, he's a friendly guy."*

In the panel of judges at either domestic or international competitions, there exists always a representative of a competing country, which may make some additional scores for one gymnast, but as well as, some additional deductions for others.

John: *"Of course, judges from the same country are trying to help their own athletes. If they see that you're fighting for a final, they can make less deductions."*

4.1.4. Overly critical, fault-finding approach. Another matter in terms of the competing time is unofficial judgment requirement that the judges should follow, such as, for example, not giving too high scores at the beginning of the competing day, as they have to follow some average value of the scores and "to save" some higher points for the evening sub-division.

Mark: *"There are always too many differences in deductions that are given in the morning and in the evening. Judges have a certain average from a morning competition, and they need to keep this average between the morning and the evening score. So, they are afraid to give high scores from the start, as it will be harder for others to get a higher score in the evening. Thus, they need to keep this average between the morning and the evening score. Thus, they don't give too good scores in the morning, and the better scores are coming in the evening."*

Another participant, James, also named one more unofficial requirement that judges follow and that he noticed at the competitions: *"Human judges, even if they see something perfect, like a perfect routine, they can't leave the papers empty, they need to find something in the routine to fill in the papers. That's why it's so hard to get 10.0 nowadays."*

4.1.5. Lack of explanation. Currently, at any kind of competition, domestic or international, gymnasts do not get any explanation or clarification of either their final score or deductions. Neither athletes themselves or their coaches are allowed to talk to the judges during the competition and ask for any feedback on their performance.

John: *"We know how much deduction and execution scores are. We just need to figure out what was not perfect. The judges don't tell us what exactly we get the deduction for."*

Mark: *"Usually I know myself what I did wrong, what I did right."*

It can be done privately, after the competition, based on friendly relations between the coaches and the judges. However, according to our informants, gymnasts have the possibility to meet some of the domestic judges in the training camps, which are organized by the Federation several times per year, and ask the questions that they are interested in, although it does not happen in regards to the competition.

And the only option that the gymnasts have after the competition is to submit an inquiry in case if the gymnast does not agree with a final score. The inquiry should be submitted right after completion of the routine, before the next athlete starts his/her own performance, in case if the previous gymnast and his/her coach do not agree with the final score. And again, this inquiry does not provide any additional information on the possible deductions and faults done during the routine, but gives an opportunity to change the final score, if the judges find it reasonable.

James: *"I asked for an inquiry, and then you have to wait if it's submitted or rejected. In my case, it was rejected, so I didn't get anything for this. I didn't hear anything about my score or how I did."*

The price for the submission of an inquiry varies from 250 to 1000 dollars. The first inquiry that an athlete makes is 250 dollars. If s/he does it again during the same competition, then the price will rise up to 500 euros. And, for the third time, it will cost 1000 euros. In case the inquiry is accepted by the judges, the final score can be changed, according to the inquiry, and this fee is not paid. However, if the judges do not consider the inquiry reasonable and, thus, do not change the score, the price of 500 euros should be paid by the Federation.

Steven: *"You really have to think if you really want to do it, because you're using the Federation money for that."*

Mark: *"The score is as it is. So, I would not complain about it. It's really hard to get any explanation from the judges about what I did wrong and to complain that it's not fair. It is what it is."*

Thus, on the gymnasts' opinion, it makes sense to submit this kind of inquiry only in rare cases, and this practice is more popular among the competitors from big countries such as Russia, Japan, China or the United States at the international championships, as they are fighting for the medals and even a small change in the final score can play a crucial role.

Thus, coaches and gymnasts use their own opportunities to get to know how successful their performance was, for example, by video recording their performance. This enables, afterward, the analysis of the whole routine and potential mistakes in detail.

Mark: *"My coach records everything that I do, and then I will see what did I do wrong."*

It should be noted that all possible mistakes are noted by the athletes and coaches only approximately, as after the competition they do not get any clarifications for the deductions, but only the final score. However, other athletes say that the explanation for the final results is not so much important for them as they do not have so much trust for the judges' opinion. Thus, they fully rely on their own experience in the assessment of their possible faults made during the competition, as well as on their coaches' opinions.

Mark: *"About the clarification of the score, I get it from my coaches afterward, but not from the judges. I always know how I did and my coaches know it as well. Judges always have their own opinions and it's always different from mine or from what my coach says. My coaches know better always."*

4.2. Perceptions on AI-based judging

4.2.1. Overall perception of AI-based judging. Overall, the informants had a somewhat positive perception about the implementation of the new AI judging system in gymnastics competitions. Our respondents felt excited about it, assuming that the system can resolve some of the existing problems of human panel of judges, discussed above. So, our respondents expect the system to be fairer, to be more accurate, equally judge everybody, not to have preferences of different gymnasts, not to get tired by the end of the day, to be able to run the competition and provide the scores faster, and provide some clarification/explanation of the final scores that can be used in the further training process.

John: *"It is better because it's much fairer. It's going to be the same for everyone. It will have the same rules for everyone. It sounds that it's much better for the gymnasts. It's accurate and the scores are more accurate every time. It is how gymnastics is supposed to be."*

Thomas: *"I think I'm excited. I think I really want to see how it works. I think if we have AI, making all calculations, then it would be more equal and much* *fairer for all gymnasts. I think that we're going to use it anyway, it's the future. Why Would not we use it if we have a chance to use it? Just to make the competition fairer."*

James: *"I feel 50/50 right now, equally. I would like to see how it works. If there's a machine there, it would be fairer, it can see all the details probably better, how much deductions exactly. And if it works, then I would most probably like it."*

David: *"AI judging system? I'm in full support of it. We definitely benefit from that because there would be more agreed upon reasons the staff on performance."*

Steven: *"It eliminates human error. And it's also more precise regarding some angle deductions or holds for length."*

4.2.2. Uncertainties about AI-based judging. Despite an overall positive perception that the AI system evoked among the stakeholders, they expressed some concerns and uncertainties about it. These concerns are mostly related to the technical characteristics of the new system, its judging capabilities and the need for participation of a human being in the process of technical support of the system.

John: *"Everything should be done with no errors. Of course, everybody is competing, everybody is human beings, and you make a small mistake, and the computer can also make mistakes and ruin everything. However, better not to. It's a little bit scary."*

James: *"Right now, it's a bit scary, you don't know what it can do and maybe it can do something wrong. You never know, maybe there's something wrong with the system, and suddenly, it stops to work. What to do then?"*

Steven: *"I'm interested and excited. But I also have my doubts at the same time. I'm a little bit skeptical about it because there are so many variations of every skill and every error. Basically, there are as many variations of the skills, as there are gymnasts. Nobody looks exactly the same. So, I'm not sure if a machine can learn all the different variations. Besides, I'm not 100% sure if the machine is correct every single time because gymnastics is so complicated. So, I think it could be good to have human judges and then to have a machine for a check-up, back-up."*

Time in gymnastics competitions can play an important role. At the moment, the waiting time for the scores after passing each apparatus varies from 1 to 5 min. And in some special cases, this process is taking longer, when, for example, *"if you did really bad and the judges need to discuss what score you deserve"* (John).

So, if using the new judging system leads to a decreased time of, for example, waiting for the score and faster move to a new apparatus after the previous routine, an athlete may not have a chance to concentrate and mentally prepare for the next routine.

John: *"I think gymnasts need to sit after the routine, and take it easy and focus. That's why I mean that we need our time before or after the routine, to focus up and get our mind together. We have to sit and rest and think about the next routine. It's good that we have a little bit of time to focus on the next apparatus, but not to run from one directly to another one. It would be really hard for us."*

However, if, in contrast, implementation of the new judging system increases the waiting time between different routines and, as a result, the overall waiting time, it can also negatively reflect on the athletes' physical form due to muscle cooling and longer nervous tension.

John: *"Of course, it's a bit hard to sit for a long time and compete. It's harder for the gymnasts as our body is getting tired."*

4.2.3. Explanations on AI-based judging. The lack of any explanation or clarification of the results is perceived as crucial by our informants, as it negatively influences their performance. So, if the technical capabilities of the system afford to get some explanation or clarification of how the judgment was done by the system, most of our respondents would be glad to get it. In their opinion, it would benefit their training process and would have a positive impact on their future results.

John: *"I think it would be very good for the gymnasts and everybody in the world, in the whole gymnastics, if you could get the clarification or explanation of deductions and execution. I actually never got any explanation or clarification of my results at the competitions. And at the international competition, it's very different. The judges don't really interact with the gymnasts. Thus, I think, it would be really good, if after the competition you can see your small and bigger mistakes that you made and what you have to do better next time. It's always good to know what I did wrong so that I can prepare for the next championship and do a better routine."*

However, here time matters as well. One concern about the time in terms of providing the explanation is a possible increase in the overall time of competition. As the increased time of the competition can negatively influence on the physical form of athletes and, as a result on their overall performance, then in their opinion it is better not to have any clarification of the results at all.

James: *"If AI can provide some explanation, but the competition will take longer, then, I think I would probably leave it because I don't want the competition to be longer as you're getting tired. Then, it's probably better without the explanation."*

Also, the opinions regarding when exactly the explanation should be provided, during or after the competition, were different:

John: *"Maybe it's better to provide an explanation, not during the competition, because you have to focus on the next apparatus, but after the competition, when you're sitting in the hotel, doing nothing, you can look at it."*

Thomas: *"I think that the feedback should come up somehow right after your performance because, after the competition, it has already past two hours, you can't remember precisely what you did on wrong."*

Additionally, the gymnasts have different opinions about which form the explanation should be done in, such as, for example, visual or oral explanation, or the list of the deductions, or video recording of their performance.

James: *"Just a list, for example, where you can see all deductions and in what skills, that would be fairer, I think. Maybe a video explanation would be also good. I think it would be nice to see every skill again, and then the judges can explain how much deductions you got."*

4.3. Trust in AI-based judging

In terms of the trust, the overall opinion of all respondents can be summed-up: if the new judging system proves that it can work properly without any interruption and errors in the judging process, gymnasts feel positive about its implementation and are willing to trust it.

James: *"If I know that it works for 100 %, then I would trust this new technology more, than human beings, because it doesn't get tired, it can see every small detail in my performance, probably, better. Sometimes it goes so fast in gymnastics, when you do some twisting and turning, that it's hard to notice by a human eye."*

Thomas: *"I think in the near future, there's always going to be someone backing it up, but maybe in ten years, or 15 or 20, if we see that the AI system works, it's reliable and it's fair for all gymnasts, then we can trust it and then we can just give the computer the whole power and rights. But of course, there can be some black boxes and some problems with the system, but if you get it working, then, I think, it would be perfect. Maybe, in the beginning, there have to be some people to back-up the system, of course. Because it's new and you can't fully rely on it. But, me, personally, yes, I'm going to trust it."*

Mark: *"I think I will trust it more than human judges because it's the same judging system for all of us."*

4.3.1. AI as substitute of human-based judging. Despite the overall very positive opinion of almost all our respondents about the newly introduced system, most of them still doubt whether the system can fully replace human judges in the future. All respondents say that an AI system can be a very useful complement to the judges, but they do not believe that the system can fully replace the panel of judges. According to their opinion, first, the system should provide proven efficiency, it should be able to work independently with a low number of errors and a high level of accuracy and fairness in order to evoke human trust. However, all respondents confessed that if the decision to fully replace all human judges is taken at the level of FIG or the Olympic Committee, they will accept it as a fact.

John: *"Maybe, it's not the best idea. I think, sometimes, you need humans also to support the work of the computer. Of course, I know that computers are really accurate and much better than the human mind. But it sounds a bit scary to know that there will be no judges, sitting there."*

James: *"I would like to see it first at some smaller competitions, and maybe not introduce it to the Olympics straight away. And maybe in the future, it could be both, maybe one human being judge and this system, and you can see the scores coming from both of them separately so that you can see the difference between what score each of them has given."*

4.3.2. Human interaction. Many gymnasts, as well as judges, think that an important part of each gymnastic performance is a human interaction between the gymnasts and judges before, during, and after the routine. It makes each competition more exciting both for the gymnasts and judges and for the fans as well. There are some long-standing traditions concerning, for example, the greeting of the judges with your lifted-up hand, and asking for permission to start the routine. Also, most of the gymnasts say that they feel this human interaction during the routine. They try to show their routine to the judges and to the fans so that all of them can enjoy it. Thus, they assume that if AI takes the leading role in the process of judging, the lack of human interaction may make them feel uncomfortable during the competition.

John: *"There's always some interaction between me and the judges. We used to, before the routine, to lift the hand up and look at the judges. And if there is only AI and no human judges at all, who are we going to greet with our hand then? They can say something or show the hand to let you go or stand. I think this interaction is important at the competitions. It feels more human if it's not just a big machine, working, which doesn't do anything for you, not waving, not winking."*

James: *"I used to see people, human beings, sitting there, judging. It would be weird to show your routine to some machine. It, kind of, feels weird to lift your hand to some robot, because you actually want to show it to some people. I still prefer some human interaction. Because it's artistic gymnastics, and if in the future there is only some machine that takes this score, maybe the artistic part of gymnastics will not be there anymore, it will go away. So, I think there's still a need for human judges to be there, to see how artistic you're, as artistic gymnastics is also about the feelings, the music, etc., especially, for the women's side."*

Additionally, some of the respondents doubt that artistic gymnastics can be well judged by the AI system without the participation of any human judges. Most of our respondents are concerned about the artistic part of gymnastics, thinking that only human beings can actually judge it correctly. Considering the gymnastics' fans who come to see the competitions, our participants expect them to be less interested and excited in gymnastics without any human interaction between the gymnasts and the judges.

Mary: *"Gymnastic is the gymnasts, the coach, and the judge. And all are human beings. It is cooperation, this is a contact with someone, with a person. What makes it artistic gymnastics it's the feelings and human interaction. When we speak about artistic gymnastics, the replacement of human judges by AI is not possible. This is what I think. Otherwise, we don't talk about artistic gymnastics anymore, not about the feelings, only the technic. I don't think it's interesting anymore. It's not a real sport anymore."*

However, not everybody thinks in the same way. Some gymnasts are so much concentrated on their routine and are so goal-oriented to compete that this "human interaction" part of the competition is not really crucial for them.

Mark: *"I don't really care if there are human judges there or AI, actually. I'm so concentrating on myself and on what I'm doing that I don't notice anybody. They're doing their job and I'm doing mine. However, I still think there have to be human judges as well, not only AI because it has to be some human opinion not only a machine. Human judges can actually see how nicely you're doing the routine."*

5. Discussion and conclusions

Based on our empirical results, we put forward three propositions. Our findings indicate that athletes expect the electronic judging system to be fairer and to be able to equally assess the routines. Also, athletes feel excited about the technical capabilities of the new judging system, which will be able to more accurately and quickly assess such skills as static positions, holds, and angles, which represent the cornerstones of making a pure technical performance in gymnastics. Additionally, athletes felt that such problems as human error, human fatigue and personal preferences of human judges can be, at least, partly resolved by a new, more technically perfected and unbiased electronic judging system. Thus, we propose:

Proposition 1. Athletes perceive electronic judging systems as a way of addressing many of the challenges related to biases and human errors in current judging systems.

Our second finding relates to the explainability of AI systems in that the informants felt that, via the new technology represented by the AI-based judging system, the lack of explainability and interpretability of the current judging system can be, at least partly, resolved. This finding is somewhat contradictory to earlier literature highlighting the challenges associated with AI black-boxes and their inherent lack of explainability and interpretability. However, in the case of electronic judging systems, the explainability of the decisions taken by the AI can be improved via such additional functions as video recording, video replay and stored data of the sensors. These simple functions potentially allow to record the data about the scores, routines and the judges' deductions during the competitions and send them out to the gymnasts after the competition. Despite the fact that the function of video recording is sometimes currently applied in artistic gymnastics, today, it does not have any systematic structure. Thus, the AI judging system was perceived as a way to allow the provision of explanations regarding scores in a more organized manner. Thus, we propose:

Proposition 2. Electronic judging systems have the potential to afford explainability features that would positively influence the athletes' perception of electronic judging systems.

Our third finding on the athletes' perceptions towards electronic judging systems juxtaposes the technical and artistic dimensions of judging. Most of the interviewees reminded that artistic gymnastics is mainly based on the human interaction between gymnasts, coaches, judges, and fans. Both gymnasts and coaches expressed their concerns regarding an AI-based judging system's capabilities to take into account such features of artistic gymnastics as feelings, emotions, music, beauty, and artistry of the routine. Thus, the risk that a new AI judging system may eliminate or decrease these components raised concerns among all stakeholders about the future of artistic gymnastics. Thus, we propose:

Proposition 3. While athletes are somewhat comfortable with electronic judging systems in evaluating the technical aspects of the routine, they are somewhat uncertain how the artistic aspects will be taken into account by the system.

5.1. Limitations and further research

Like most empirical studies, ours is not without its limitations. First, we acknowledge that the sample size was relatively small as we had nine informants, and, most notably, only five gymnasts. Further research should examine the perceptions of gymnasts with larger sample sizes. Second, our study is heavily focused on one specific context: gymnastics. Further research could probe the perceptions of athletes in different sports, employing different types of electronic judging systems (e.g. rules-based vs. AI-based, individual sports vs. team sports). Third, we were limited to using interview data. Further research could investigate the issue by conducting real-life field studies where gymnasts would be studied in their natural setting.

6. References

[1] Biran, O. and Cotton, C. Explanation and justification in machine learning: A survey. *IJCAI-17 Workshop on Explainable AI (XAI)*, (2017), 8.

[2] Can, H., Lu, M., and Gan, L. The research on application of information technology in sports stadiums. *Physics Procedia*, (2011), 604–609.

[3] Charmaz, K. and Belgrave, L.L. Qualitative interviewing and grounded theory analysis. In *The SAGE Handbook of Interview Research: The Complexity of the Craft*. 2012.

[4] Ferger, K. and Hackbarth, M. New way of determining horizontal displacement in competitive trampolining. *Science of Gymnastics Journal 9*, 3 (2017), 303–310.

[5] Gibbs, S. World Cup goalline technology: how does it work? *The Guardian*, 2014.

[6] Heiniger, S. and Mercier, H. *National Bias of International Gymnastics Judges during the 2013-2016 Olympic Cycle*. arXiv preprint arXiv:1807.10033, 2018.

[7] Holzinger, A., Biemann, C., Pattichis, C.S., and Kell, D.B. What do we need to build explainable AI systems for the medical domain? *arXiv preprint*, (2017), 1–28.

[8] Houser, K. AI Will Help Judges Score Gymnastics Events at the 2020 Olympics. *Futurism*, 2018.

[9] Kramer, R.S.S. Sequential effects in olympic synchronized diving scores. *Royal Society Open Science 4*, 1 (2017), 1–9.

[10] Leskoŝek, B., Cuk, I., Pajek, J., Forbes, W., and Bucar-Pajek, M. Bias of judging in men's artistic gymnastics at the european championship 2011. *Biology of Sport 29*, 2 (2012), 107–113.

[11] Lipton, Z.C. The Mythos of Model Interpretability. White paper (2016), 1–27.

[12] Mercier, H. and Klahn, C. Judging the Judges: Evaluating the Performance of International Gymnastics Judges. *MIT SLOAN, Sport Analyttics Conference, Hynes Convention Center, March 3-4, 2017*, (2017), 1–18.

[13] Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence 267*, (2019), 1–38.

[14] Montavon, G., Samek, W., and Müller, K.R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal 73*, (2018), 1–15.

[15] Omorczyk, J., Nosiadek, L., Ambroży, T., and Nosiadek, A. High-frequency video capture and a computer program with frame-by-frame angle determination functionality as tools that support judging in artistic gymnastics. *Acta of bioengineering and biomechanics 17*, 3 (2015), 85–93.

[16] Owens, N., Harris, C., and Stennett, C. Hawk-Eye tennis system. *International Conference on Visual Information Engineering VIE 2003*, (2003).

[17] Patton, M. *Qualitative Evaluation and Research Methods*. Sage Publications, Thousand Oaks, CA, US, 1990.

[18] Preece, A. Asking "Why" in AI: Explainability of intelligent systems – perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management 25*, 2 (2018), 63–72.

[19] Roth-Berghofer, T. and Richter, M. On Explanation. *Künstliche Intelligenz 22*, 2 (2008), 5–7.

[20] Taymazov, V., Bakulev, S., Pavlenko, A., Simakov, A., and Chistyakov, V. To a question of electronic refereeing systems application in taekwondo (VTF). *Uchenye zapiski universiteta imeni P.F. Lesgafta*, (2013), 155–160.

[21] Wiesche, M., Jurisch, M.C., Yetton, P.W., and Krcmar, H. Grounded Theory Methodology in Information Systems Research. *MIS Quarterly 41*, 3 (2017), 685–701.