Stakeholder-dependent views on biases of human- and machine-based judging systems

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

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

Antti Salovaara
Aalto University School of Arts, Design and Architecture, Finland
antti.salovaara@aalto.fi

Abstract

Motivated by recent controversy over biases associated with algorithmic decision-making, we embarked on studying various stakeholders' perceptions related to potential biases in verdicts from human-based and algorithm-based judging. In an empirical study conducted in the domain of gymnastics judging, we found that, while our informants viewed both human- and AI-based judging systems as being subject to biases (of different types), they were quite welcoming of a shift from human-based judging to machine-based judging. Our findings show that the athletes trusted strongly in unknown, "magic" capabilities of AI, thought to be more objective and impartial. This, in turn, encouraged potential acceptance of new technology. While the gymnasts saw AI-based systems in a positive light, judges demonstrated less favorable perceptions overall and less acceptance of AI technology, expressing concern about possible challenges of AI.

1. Introduction

Fueled by the increasing volume, velocity, variety, and apparent veracity of data [13, 26], algorithms developed to apply machine learning (ML) are penetrating a more and more expansive set of activities in individuals' lives and organizations' practices. While many instrumental outcomes of these algorithms, such as greater accuracy and efficiency, are welcomed by most stakeholders, development toward greater reliance on artificial intelligence (AI) entails some potential negative humanistic outcomes, engendering a host of suspicions. One of these, extensively discussed in the popular press, is related to possibilities that bias in the algorithms developed could lead to algorithms' recommendations being skewed [1, 31].

A recently cited alarming example, covered by, among others, *The New York Times* [22], is an algorithm developed by a criminology and statistics professor that is used by the U.S. government to make decisions on granting probation to prisoners. The algorithm and its use have created controversy and heated discussion surrounding the lack of transparency of its recommendations (i.e., it is hard to see whether gender, age, or ZIP code was a deciding factor) and the chances that biases held by developers of the system are "baked into" it (e.g., with regard to race, socioeconomic status, and geography). Such controversy notwithstanding, algorithms are being extended into nearly every corner of our society.

Motivated by recent debate and expressions of concern related to potentially biased AI-powered machinery of society, we asked, "**How do various sorts of stakeholders think about the biases related to human- and algorithm-based decision-making**?"

To address that research question, we sought an empirically rich context in which to study the phenomenon, one where algorithms are exerting disruptive change in the way verdicts and judgments are rendered. Our search led us to artistic gymnastics, for which AI-powered machinery is being developed by Fujitsu to assess the technical purity of gymnasts' movements and routines. It is currently being pilot tested at competitions.[1]

Our findings reveal a multifaceted picture of the perceived biases of human-based and machine-based judging. Many informants in our study were actually quite welcoming of AI-based judging, mostly on account of problems and biases connected with human-based judging. In light of this, we can return

[1] For the information on the features and functions of the electronic judging system developed by Fujitsu, the reader is referred to: https://www.fujitsu.com/global/about/resources/news/press-releases/2019/1002-01.html.

URI: https://hdl.handle.net/10125/71383
978-0-9981331-4-0

to the example of the probation-decision algorithm with a reminder that any discussion of AI-powered systems' biasedness should be conducted in conjunction with examining the positives and negatives of both human-based and machine-based systems.

2. Literature review

Previous research has shown that AI-based judging has the potential to improve such indicators of judging system in sports as quality, accuracy, fairness, impartiality, validity, and reliability [21]. Thus, electronic judging systems allow significant reductions in the amount of human errors and biases in the scoring of athletes' performance.

2.1. Human biases

Research in the domains of social psychology, behavioral theory, economics, and finance, not only information systems, has shown that people are not always rational, and we are influenced by various cognitive biases [18]. These are systematic deviations in behavior, perceptions, and thinking that stem from subjective beliefs, prejudices and stereotypes, emotion-linked factors, misinterpretation, or faulty analysis of the information present [2]. Biases are inherent to human reasoning, and they prejudice the quality of decisions made by considerable numbers of people [30]. Humans tend to create our own "subjective" reality, which can influence decision-making, judgment, and perception of information. Accordingly, cognitive biases are often referred to as decision-making biases or judgment biases [2]. Human judgment biases may be regarded as deviations from rational thinking on the part of an individual or group, alongside the possible consequences of these deviations [2], Importantly, some of them can, in fact, contribute to more effective human decision-making while others limit one's space for objective judgment. In their various forms, cognitive biases are widely discussed in information systems (IS) literature with regard to their influence on the implementation and use of particular information systems [6, 14, 15, 18, 19, 23, 24, 25, 28, 29]. For instance, Arnott [2] has offered a taxonomy of biases, dividing them into memory, statistical, confidence, adjustment, presentation, and situational biases. Though finding precise boundaries between groups of biases can be complicated, since they are "blurry," such a framework can still aid in discussion.

2.1.1. Memory biases. In psychology, memory biases are conceptualized as cognitive biases that influence memory recall and the amount of time it takes [2]. Memory biases, considered the deepest of the cognitive biases, are related to storage and recall of the information and data in our memory [23]. There are many types of memory biases. In **hindsight bias**, things' obviousness "after the fact" increases people's confidence in their ability to make the right decision or forecast, and it simultaneously decreases their ability to learn from past events. Hence, human ability to predict some events and outcomes is usually heavily overestimated [23]. **Recall bias** affects the ease of "recalling" certain significant events from memory. The more repeated, frequent, familiar, or salient the event, the more easily it will be recalled. One result of the influence of recall bias is that a decision-maker may readily make decisions that accord greater weight to less relevant information than to more relevant or new information. **Similarity bias** appears when a decision-maker makes a judgment about some event in reliance on its similarity to an event in a related class: if an event belongs to some particular class, it gets perceived and judged in line with the decision-maker's opinions/stereotypes connected with its apparent class rather than its specific characteristics [25]. **Testimony bias**, in turn, involves inability to recall details of an event, whereby an inaccurate reconstruction of this event is produced in the human memory. Referring to these non-original memories, the decision-maker produces judgments in the belief that objective evaluation of a "real" event has occurred when, in reality, the evidence of the event is no longer ironclad, if it ever was, and the memories are not clear enough to make for accurate evaluation that avoids errors. Usually, testimony bias is evoked in a memory *post factum* because of memory cues.

2.1.2. Confidence biases. A person affected by confidence biases tends to display unwarranted confidence in his or her decision-making. An important feature of these biases in action is avoiding new information that could lead to questioning previous decisions or judgments. **Desire bias** involves letting expectations linked to some results' desirability influence one's perceptions related to those results. For instance, the decision-maker may overestimate the likelihood of a particular outcome because of wishful thinking. Even though the person who is making the decision may have observed (or be able to access) information that shows the desired outcome to be unlikely, expectations or the desire for some particular result may still bias the decision-making in line with hopes [2]. Where desire bias influences mainly perception by the decision-maker, confirmation and selectivity bias influence also how he or she collects and processes

information related to the desired outcome. These two types of bias are interrelated in the ways they influence people's behavior. Both of them are intensively discussed in IS literature in terms of their impact on users' decision-making process and with regard to resistance vs. acceptance of technologies [7, 8, 11, 15, 17, 19, 24]. With **confirmation bias** comes overconfidence in one's personal beliefs, contributing to interpreting ambiguous evidence as information that supports those beliefs despite evidence to the contrary [2]. When confirmation bias is at play, one tends to search for, interpret, or give preference to information that corresponds to and reflects the favored internal beliefs and thoughts while at the same time ignoring information that contradicts those beliefs [15, 17, 19, 24]. **Selectivity bias** reinforces confirmation bias through filtering. The decision-maker dismisses information that is unfamiliar or appears irrelevant. With selectivity bias, we think our perception and evaluation of the given events are objective. It influences our views as to which information is relevant for decision-making processes [7, 8, 11, 23]. Finally, **overconfidence** is a general bias involving much higher subjective confidence in one's actions, judgments, and decisions than their objective accuracy warrantts [29]. This cognitive bias is rooted in subjective overestimation of one's capabilities [2]. One form of it is manifested in overestimating one's control. Overconfidence in one's judgment can appear also when the results of the judgment/decision cannot be tested.

2.1.3. Anchoring and adjustment biases. The usual human approach to judgment is to "begin at the beginning," choosing a starting point and then adjusting one's opinion. **Anchoring** is a tendency to rely on that starting point – the initial information – during the decision-making process. People tend to process information for decisions by working from a suggested reference point, an "anchor," and making **adjustments** to this to reach their goals [2, 17]. This strategy usually yields good and effective results; however, the starting point for judgment may happen to be wrong, there might not be enough adjustment, and further steps in the decision-making process could end up wrong or problematic. Even when the anchor has been chosen at random and people know this, they still fall victim to anchoring and adjustment biases.

2.1.4. Presentation biases. Among the most important biases from the perspective of the decision-making process are presentation biases, which affect how a person perceives information [2]. One example is **order bias**, in which the order of the information or data's presentation exerts a significant influence on the human judgment process. Studies show that we tend to pay more attention to the first and the last item/object or subject shown, so these "bookend" objects may be overemphasized in our judgment or evaluation. The first one in the set is evaluated more accurately since it gets perceived as primary, and the final one displays a recentness effect. Hence, research has revealed that the sequence of information or data's presentation may hold more sway over the decision outcomes than the data or information *per se* [2].

2.1.5. Situational biases. The last set of biases in this category is connected with how a person responds to the general decision situation. These represent the highest level of biases' abstraction [2]. Firstly, **complexity bias** is evoked by various external factors: time pressure, a stressful task setting, information overload, a highly important task, large volumes of data to process, and others. This bias impairs the decision-making process, contributing to incorrect decisions. Secondly, **rule bias** can occur in judgment situations wherein a person may apply predetermined decision-making rules. If there is an error in the rules, that error gets reflected in the decision. Arnott has stressed that even if no other biases and prejudices are present, this one could ruin the whole decision process [2]. There are several factors that can tie in with rule bias and thereby lead to a biased decision: the expected/desired results (one's choice may depend on such elements as what the outcome "should be") the admissibility of compromise in making the decision, and what aspects of matters are to be considered in the decisionss-making (e.g., whether some should be disregarded or all of them must be factored in).

2.2. AI biases

2.2.1. Racial and gender biases of AI. Recent research shows that ML algorithms can demonstrate racial and gender discrimination [3]. When the facial recognition of three artificial-intelligence-based systems was tested for its accuracy, the facial-analysis software demonstrated gender and racial biases, irrespective of the companies'/developers' claims as to the algorithms' potential for neutrality and fairness [3]. The researchers, Buolamwini and Gerbu, found the software to demonstrate a higher error rate in gender identification for women than for men and for darker-skinned as opposed to lighter-skinned people. According to their report, the AI error in face recognition for black women was about 35% while that with white men was below 1%. These biases are due to the datasets used for the relevant neural networks' training, which exert a huge influence on the resulting model [27]. Ascertaining the level of

various biases of face-recognition programs is crucial since systems of this sort are used in many fields – criminal justice (for identifying suspects), medicine and health care, banking, recruitment, etc. [3].

In the United States, use of certain AI systems has already led to biased decisions in risk assessment connected with criminal **convictions**, not just parole. In their decision-making process, judges have relied on an AI-based system's projected probabilities of some prisoners repeating their criminal actions in the future. Historical patterns connected with specific ethnic and racial groups caused the system to state a 77% higher index value for the probability of black criminals being involved in the same crime again than for white criminals. The AI's forecasts were grounds for many decisions not to grant parole. Only later did an independent investigation reveal that the predictions were wrong in 80% of cases [1]. In 2019, research showed that the Optum **health-care algorithm** gave triage preference to white patients over black patients who were more ill when considering predictions related to treatment costs. It emerged that the system prioritized among patient not in light of the risk or seriousness of the disease but on the basis of patients' race and their "profitability" for the health system [12]. Another example is targeted AI-algorithm-based **Google- and Facebook-supplied personalized ads**, widely used on many Web sites. An empirical study examined which Facebook and Google users see announcements of highly paid vacancies in science, technology, engineering, and mathematics (STEM) fields most often [16]. The results showed that, notwithstanding the overall neutrality of the ads themselves, women saw them less often than men. The system was designed to optimize advertisement costs, and its analysis of the input data produced the conclusion that delivering such an advertisement to men is more profitable than showing it to women.

2.2.2. Reasons for AI bias. The biases of AI result from two sets of biases in combination: cognitive and algorithmic [10]. In the process of creating algorithmic systems, their developers transfer their **cognitive biases** to the algorithms, whereupon the system may start demonstrating some prejudices of its developers. The **algorithmic biases** thus created may arise at various stages in framing the problem, collecting the data to be used for the algorithm's training, and preparing those data [9]. Firstly, framing the problem is a hard process: systems applying ML must address highly uncertain aspects, factors, and characteristics of human life, such as a borrower's credit potential, the likelihood of going bankrupt, or recidivism potential. What the system forecasts depends mainly on the developers' initial framing, which, in turn, relies on their personal perceptions of the problem [9]. Secondly, collection and preparation of data for the system's training may bring in two problems. The first is related to initial errors and possible biases. Because AI uses "historical" data to make predictions, input data that favor or prioritize against certain groups of people will get mirrored: the AI will learn how to "discriminate" on the basis of the same attribute. For example, in 2018, Amazon developers realized that their new AI recruiting system "doesn't like women" [4]. The system tended to choose men for highly paid leadership positions because of data on previous hiring/promotion decisions: men held these positions more often [4]. Another reason for the system's bias was that the input data included more examples of white men of average age, attributes that may have gone unnoticed since they matched the demographics of the developers of the algorithms [5]. After efforts to rectify the problem by deleting such terms as "woman" and "female," the company announced that the system was not gender-neutral and would not be launched. The second problem highlights an issue beyond large and non-validated datasets: how the AI system uses the input. In AI software, algorithms and a set of rules aid in identifying patterns for later decision-making. Developers' cognitive biases may taint the choice of the algorithms to be used in further decision-making by the system. The AI is not initially biased; rather, it "adopts" human biases. A further problem arises in that the existence of algorithmic biases is hard to reconcile with a deeply held human belief in AI's neutrality and objectivity [20].

3. Methodology

To address our research question, we chose to conduct an inductive qualitative case study.

3.1. Selection of the case and collection of data

We employed two main criteria in our search for a suitable context for empirical study. Firstly, the case had to represent a setting of transition to electronic judging systems, so as to facilitate gathering relevant views on both human- and machine-based judging. Secondly, the informants had to be persons directly influenced by the implementation, because we wanted to probe (multiple) stakeholders' true perceptions of the biases possible with both sorts of judging system. We found artistic gymnastics to offer a suitable field since it is currently moving over to employing AI in its judging. We selected our case accordingly and collected data

via 21 semi-structured interviews, with various stakeholders affected by the introduction of a new electronic judging-support system. All interviews were tape-recorded and transcribed (after which the one non-English-language interview was translated into English also). The interviewees are characterized in Table 1.

Table 1: The interviews

Role	Pseudonyms
Gymnast	James, John, David, Thomas, Mark
Director	Steven, Mary
Coach	Paul, Kevin
Judge	Abby, Bella, Charlie, Edward, Harry, Lilly, Nick, Norman, Sarah, Ulla
Vendor	Caleb
FIG	Simon

When developing the interview questions, we were guided primarily by a wish to encourage the participants to share their opinions and perceptions of both the human-based judging system and the "e-system." Therefore, the informants were asked open questions about the following issues: the judges' professional experience in gymnastics, the human-based system, the e-judging system, their perceptions of both, explainability, and the training process. For space reasons, the interview protocol is not reproduced here; the authors will provide it upon request.

3.2. Data analysis

We used ATLAS.ti for data analysis, employing three coding techniques: 1) open coding, 2) axial coding, and 3) selective coding. Via open coding, we obtained 96 distinct codes, for particular ideas and opinions expressed by the various informants. Secondly, we used axial coding, forming 17 code groups and searching for inter-group correlations and patterns. Finally, to integrate the concepts uncovered with theory and to build theoretical propositions from our study, we used selective coding. This part of the analysis involved identifying similarities in opinions between groups of informants. Because our interview questions were divided into sets that corresponded across informant roles, we were able to summarize the participants' opinions and perceptions about the systems and their comparison, and we could identify the differences in expectations for the new judging system across informant roles. Proceeding from this analysis, we were able to identify the main challenges linked with the current judging system and possible corresponding challenges and opportunities brought by the new one.

4. Findings

4.1. Informant perception of human judgment's biases

One of the main problems with the human-based judging systems lies in judges' biases. These may stem from several factors: influences of emotions, personal preferences, familiarity with a given athlete or specific routine, others' expectations of particular athletes, personal prejudice attached to a particular country or athlete, to name a few.

4.1.1. Memory biases – recall, testimony, and hindsight. During a gymnastics competition, judges have to evaluate many elements of an athlete's routine in only a few minutes. The level of accuracy is affected by such factors as the angle of visual observation, fatigue, experience, and the judge's attention. In any case, in the opinion of the judges interviewed, it is nearly impossible to notice every detail of a routine, as noted by Charlie ("Sometimes we really can't see it; the human eye can't always capture the exact moment and the picks of the moment") and director Steven ("It's almost impossible for a human eye to register all those mistakes and to write them down because in the competition the routine continues immediately – so, in that case, it's too complicated for humans in some respects"). The judges must heavily rely on their memory when evaluating the routine and giving it a final score. One can assume that recall and testimony bias might come into play. If the judge cannot recall all the details of the routine, his or her memory tries to "reconstruct" the missing parts. High levels of approximation and low levels of accuracy in judgment can be expected. Indeed, Ulla stated that judges perform large amounts of approximation in their evaluation of heights, speeds, and angles in an athlete's performance and that they estimate scores on the basis of this approximation. Also, judges may unconsciously rely on intuition and experience to fill in gaps in the memory. Hindsight bias can rear its head here. Whatever the outcome of the athlete's performance, the judge might refer to a forecast from well beforehand and conclude that "I knew that would happen!" Sarah described this: "When a gymnast runs [up], I can tell you if it's going to be a catastrophe or not. We anticipate. Anticipation helps you sometimes with your judgment. You have not only what you see at this moment; it's much bigger than just what you see."

4.1.2. Similarity, desire, and mere-exposure biases. Judges' expectations lead to other biases. Similarity bias acts such that a given country's leading position may support a higher score for the athlete. According to Ulla, the scores of athletes from some countries therefore can sometimes be unreasonably higher than others' since judges perceive them as stronger by dint of the strength linked to their country. Desire bias, in turn, follows from expecting a certain quality of performance from a certain athlete. Familiarity with an athlete and his or her earlier success or failure may influence judges' expectations, prognoses, and objectivity. Desire bias may lead judges to perceive the routine as better or worse than it actually is. Abby said, "If you're really familiar with the routine, it can influence your judgment positively or negatively. Maybe you don't see a mistake because you see it all the time and you get used to it. Or maybe you see every little mistake that they make more. Familiarity with the routine can move your judgment up or down." Mere-exposure bias too may affect the judges' evaluation of a routine. Informants mentioned that judges' friendly relations with the athlete or coach may bubble over to the personal preferences of some judges and, in turn, get reflected in the scores. Felicity mused, "Do the judges have preferences? Unfortunately, I have to answer 'yes, I think so.'" Several gymnasts echoed this, with Mark saying, "If the judges and the gymnast are from the same city, maybe they will give a higher score to 'their' guy" and John saying, "Of course, judges from the same country are trying to help 'their own' athletes. They may make fewer deductions." Sarah echoed this: "Sometimes judges and coaches […] set a good relationship. Even though the judge here is supposed to be neutral and working for all the countries, she still has a little affiliation with some country."

4.1.3. Order bias. Decision-makers often give undue weight and attention to the first and the last things encountered. Gymnastics competitions are no exception: the athletes' order is very important for an objective judging process. Our informants confirmed the significant influence of order bias on the judging process, commenting that an athlete who competes in the morning will get a lower score and one competing in the evening gets a higher score. Mark said, "There are always too many differences in the deductions in the morning [vs.] the evening." John expanded, "It's always like this: if you compete in the morning, judges are harder on you; they easily take away many more points. They want to be good, strict, and do their job properly. Thus if you compete in the morning, they can make a bigger […] deduction, and in the evening if you do exactly the same mistake, they will not take so much from your total score."

4.1.4. Rule bias. Rule bias can appear in the human judging system when judges follow unofficial preordained requirements such as to maintain a certain "average score over the course of the day" or not give overly high scores to a "perfect" routine. Felicity referred to the former by saying, "That's what we're told when we have the judges' meeting before we have a long competition day: keep the line the whole day," and Mark echoed this: "Judges have a certain average from a morning competition, and they need to keep this average between morning and evening scores. 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, so 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 come in the evening." As for perfect scores, gymnast James said, "Human judges, even if they see something perfect, like a perfect routine, they can't leave the papers empty. They need to find something [wrong] in the routine, to fill in the papers. That's why it's so hard to get 10.0 nowadays."

4.1.5. Complexity bias. Complexity bias is clearly present, due to such factors as information overload, time pressure, human fatigue, lack of accuracy, the need to pay constant attention, and perceived importance of the judges' task and responsibility. Judges must make their decision in just a few minutes, during the routine. Extensive approximation in their judgment arises from the limits of the human brain: it can process only a certain amount of information within a certain time. Norman said, "There are too many decisions to be taken, so for a human brain it is not possible […]. In one second, you have to make maybe 8–10 decisions, and it is almost impossible because it happens all at the same time." Felicity summed up the issue: "We don't want that, but we all make mistakes when we judge." Human fatigue is a particularly strong complicating factor for the judges. At international competitions, judges have to spend many hours in sometimes uncomfortable conditions. Interviewees cited several examples. Charlie: "When you're sitting down and you have six subdivisions in one day and you have it over two days so you're spending 14 hours a day in the gym, yeah, it's really hard to be fresh from the first moment of the first day until the last moment of the last day." Norman: "The concentration from 10 in the morning till 10 in the evening […] is almost impossible work for humans to maintain the same concentration." Sarah: "I was sitting with the light in my eyes. Toward the end of the day, it was stressful with the lights. Of course, there's a human aspect. You compete at the beginning

of the day, in the middle of the day, when the judges might be tired or thirsty or hungry, needing a break. All of these human components can injure."

4.1.6. Anchoring and adjustment biases. Under the influence of adjustment biases, judges may tend to root their judgment partly in previous achievements of the athlete. The initial anchor for the judge might be the athlete's ranking or "usual" performance level. John said, "Especially in your home country, it's usually not fair when the judges know you and saw you so many times during the training so they kind of know already where you will do your mistake. So if you don't do it in the competition, they think like 'oh, he usually makes this mistake, so it will be a mistake now' even if you do it very well. I like to compete more internationally, where judges don't know me. I usually score higher points." However, judges at international competitions may be influenced by another anchoring bias – first impressions. Once made, the impression is very hard to change or adjust.

4.2. Informants' perception of AI

4.2.1. Stakeholders' awareness of the new system. In our study, many of the stakeholders had little awareness of the new electronic judging system. The judges and gymnasts knew that the AI system is going to be used during the judging process at competitions but did not have a deep understanding of how it works. For instance, judge Edward said, "I know very little about it. They just told us that it's going to be used for the difficult decision-making, but we were not really told how it functions." Along similar lines, Abby stated, "I haven't heard. We have not had a notification about when it's going to be implemented," and Norman stressed that "[w]e don't know anything about the system. We need to know more when it's ready." At one of the international competitions, the new system was introduced to all the judges and federations' representatives. The introduction sessions provided some basic information about the technical capabilities of the system and its supporting functions for judging. The informants did not consider this information to give them sufficient understanding, however, with Ulla saying, "We know that this system exists. After the last Olympic games, we were told that they were working on a supporting judging system that can judge more objectively, but we don't know how it functions and how it gives the scores. What is the score difference between those scores that we give and those that the system provides? We also don't understand how it can support the judges' work." Abby expressed similar concerns: "We need to know how it works. If we have to work with it, then we need to know how to make it work. And we need to know what it can tell us, what information it can give us and how it gives the information, and also how quickly it could give this information." As for the athletes and coaches, they were not even involved in the introduction sessions.

Despite their lack of information or knowledge about the new electronic judging-support system, all informants gave a positive evaluation of the system, had high expectations for it, and expressed favorable perceptions at personal level. Judges, gymnasts, and other stakeholders alike expected the electronic judging system to rectify the biases of the existing human-based judging system, thanks to its technical capabilities. That said, they did discuss some challenges that a new system may create for artistic gymnastics.

4.2.2. Informant-perceived advantages of the system. Most informants expressed a belief that an electronic judging system would be more **accurate** than human judges. Stakeholders of all stripes stated that technology is always more accurate than human beings. Simon: "It helps the accuracy. The goal is to be able to help the judges in cases where better accuracy is needed, aid in judges' education, help the coaches and the athletes with the training, and improve safety. The sky is the limit for this system." Charlie: "I think that artificial intelligence can provide an accurate and detailed breakdown." Lilly: "The computer can do better, can see angles better, and it's more precise than a human." Edward: "What a human eye sees is one thing, but what the machine sees is more accurate. I heard that it's very useful." Felicity: "I believe that the electronic judging system can be more accurate than human judges." Sarah: "This technology is a step further, the more detailed show[ing] of the areas and angles and possibly what muscles are working. It's fantastic, and it's amazing technology."

Considering **objectivity**, judges claimed that they try to be as objective as they can but human biases and preferences may indeed influence their judgment. They assumed that the new AI system is not biased and hoped it would prove more objective and neutral in its evaluations. Ulla: "AI doesn't care which country you're from. It evaluates the technical side of the performance. Judges can hear very often from the coaches that we've been biased with their athletes, and if the routine is evaluated by the system, who can you blame for low scores? Nobody. Because AI is unbiased. It's objective." The judges stated also that the system is unbiased and more objective in that it has no "anticipations" or prior expectations for a given athlete's performance. Edward: "I think it does have its benefits for sure. It can take a lot of objective information that can be transposed to giving the score

for the athletes." Bella: "For objective things, maybe. Because we can make some mistakes about an objective thing, but the system can't."

Overall, our informants found that the capability of an AI system to provide **explanations** of the final results would be very useful on both sides. Ulla: "When the e-system can provide some explanation or even a printed list of all deductions and scores, that would be great! Then it will be clear for everybody – for both coaches and gymnasts – how the judgment was done, and everybody will understand everything." Additionally, informants stated that the system's ability to provide the scores and the list of deductions immediately should aid greatly in expediting inquiries during competitions. Nick: "It looks good, and it could be really helpful when there's an inquiry, for example. We had some cases here where [an appeal] was accepted because of the Fujitsu system." Harry: "When they have something to appeal – I mean inquiry for the superior judges and also for the Technical Committee members – they will use this system to help them to evaluate the whole routine again. It could be very useful." Furthermore, informants stated that AI with an explanation capacity may support athletes' training process after the competition, which is important for an athlete's improvement. Edward: "I come from the American continent. Maybe at the American Championships we don't do it, but on the lower level, after the competition is over, we gather all the coaches and gymnasts and tell them what their mistakes were at each apparatus."

4.2.3. Informant-perceived disadvantages of the system. The biggest worry cited with regard to the e-judging system is its perceived inability to **evaluate the artistry** of the gymnasts' performances. One informant reminded us that artistic gymnastics is not for nothing called "artistic." The artistic component of the athlete's performance is crucial. Judges in particular stressed this. Norman: "It's called artistic gymnastics. And artistic is the key part of it, how it looks. I don't think the machine really can take up this part. We have artistry; we have a lot of things. Beautiful things." Harry: "Impossible, because it can't measure the artistry. It can measure only angles, only time. The computers don't understand what is artistic. If in artistic gymnastics judging is completely done by the computer, it's not artistic gymnastics." Felicity: "If you can teach the computer all the tempo and rhythm, can you really teach things that we call artistry?" Charlie: "[G]ymnastics is the sport of emotion. Artificial intelligence has no emotions thus far." In their view, if the AI-based system cannot perceive and evaluate the artistry of a routine (and "runs the show"), this component of the sport might be eliminated. They held that this must not be done, as it would standardize all routines and reduce the standard of artistic gymnastics competitions. Lilly: "In the end, we will have every exercise look the same, and the personal style of athletes will be lost."

Human interaction is always an invisible element in the process of athletes' performance. A slight welcoming nod from the judges when the athlete steps out onto the floor, raising a hand before one starts the routine, the judges flashing a smile when a gymnast did exceptionally well – all of these are important components of the performances. Human interaction instills a friendly environment, and it makes the athletes feel more comfortable during the routine, positively influencing their performance. Our informants stated that the e-judging system cannot provide the same level of human interaction in the gymnastics. This sense could become a stumbling block to implementing the system for artistic gymnastics. Nick: "Gymnasts standing in front of a computer and saying, 'Hi, I'm starting my exercise.' That's kind of weird for me. We're part of the competition, and it should always be a human aspect of judging at the competition." Charlie: "I'm not quite sure how the athletes will feel. When an athlete does a good exercise and looks over to present to the judge and sees the reaction of the judge, I think that's something that is a human emotion that gives that athlete a good feeling's worth. Or if the judge offers a sympathetic look even though the routine was not good, maybe the athlete still knows that there's someone who is cheering about the performance. Well, I'm not sure if artificial intelligence will be able to provide that type of feedback to the athlete."

Exactness is the flip side to high accuracy, according to some of our informants. Despite the fact that all stakeholders perceived the system's high level of accuracy as an undeniable advantage, they stated that excessive exactness in judgment would upset the balance between the judges' evaluation and the athlete's performance – while judges do not judge accurately, gymnasts do not perform accurately either. Accordingly, the informants assumed that gymnasts will not be able to provide high enough accuracy in their performance to match the level of exactness in the evaluations produced by the e-judging system. Edward: "This system is too perfect. My worry is that it is too perfect. It's a big difference: what a human eye sees is one thing, but what the machine sees is more accurate. Right now, we're humans. Gymnasts are humans. We as judges note certain deductions, certain angular deductions. Sometimes 45 degrees is very difficult to recognize for a human eye. But if a camera sees '44.9 degrees,' it does

not accept the exercise; it makes a deduction. But for a human eye, the normal eye, it may pass. The gymnasts will be mad at the judgment with the machines because it's gonna catch every single mistake they make."

5. Discussions

Despite widespread criticism leveled at AI's negative implications in the popular press and end users' numerous suspicious of AI, systems of this nature are being introduced. The one we considered, soon to enter use in gymnastics for scoring athletes' performance, may even replace human judges. Our study, contributing to new research in this domain, was aimed at identifying how the stakeholders perceived AI-based and human judging, what biases are typical of AI and which are common in humans, and how views of these biases affect users' acceptance/resistance with regard to new information systems. Below, a summary of our findings frames our attempt to explain them.

Firstly, we found that the main challenges of the existing, human-based judging system for artistic gymnastics lie in the biases and subjectivity of the judges. At times, judges demonstrate memory, confidence, anchoring, presentation, and situational biases of several kinds in their judgment. These biases arise for various reasons: human emotion, personal preferences for particular gymnasts or countries, familiarity with a routine or athlete, prejudice and "preset" requirements for evaluation (official or not), fatigue and other factors connected with the length of the competitions, the limits of the human eye's ability to detect several micro-movements of gymnasts simultaneously, the generally low accuracy of human evaluation, etc.

Secondly, our findings show low levels of stakeholder awareness of the AI-based judging-support system. Having been provided with little knowledge and understanding of the AI's operations and technical capabilities, judges and athletes alike filled the gaps with suppositions. Despite their lack of information and knowledge about the e-judging system, they demonstrated positive personal perceptions, offered favorable evaluations, and had high expectations for the new system.

Thirdly, we found that when evaluating the system's capabilities in light of their perceptions, the informants demonstrated confirmation and selectivity biases. Both judges and gymnasts employed selective thinking when assessing the AI, taking into account only that information consistent with their prior knowledge. Thereby, they supported their internal beliefs and ruled out information that might conflict with these. Confirmation bias thus encourages strong reliance on the capacity of a "perfect" AI system – even one with largely unknown capabilities – to resolve all possible challenges of biased judging. Gymnasts in particular believed that, thanks to its technical capabilities, the new electronic judging-support system will demonstrate high accuracy, impartiality, and objectivity while also providing sufficient levels of explanation and clarification of the results. One possible explanation for such excessively positive perceptions might lie precisely in the **lack** of information and knowledge about AI in general or this system specifically and an associated perception of it as a magic "black box." The same bias stimulates manifestation of judges' overconfidence in the objectivity of their decisions and constrains their acceptance of the new technology even though they agreed that the new judging system might be more objective, impartial, and accurate. The judges among our informants showed strong concerns about such disadvantages as the technology's inability to evaluate artistry, a lack of human interaction, and excessive exactitude. Overall, they had a more negative perception of the AI system than the athletes did. This tied in with general resistance to the technology's acceptance among the judges and to their unwillingness to adopt another means of judgment.

Thus, we concluded that stakeholders' biases connected with AI couple with their lack of information, knowledge, and understanding of it to produce different perceptions of the given system: in the case of judges in our study, the outcome was relative resistance to the new technology, while the result among athletes was potential acceptance. Finding and providing means of clearer interpretation of AI-based systems' internal structure for each set of stakeholders could lead to better understanding of AI and, thereby, more appropriate acceptance and user trust.

Regarding generalizability, we see that our results could be applicable in sports that contain similar parameters of evaluation of the athletes: technical and artistic components, and where the judging process is likely to transition to an amalgamation of human-based and machine-based decision-making. For further studies, we recommend closer consideration of the differences between two distinct groups of end users, in response to what our case study showed about the biases that differently influence judges' and athletes' perceptions and, hence, the evaluations of new technology. Both these groups are targeted as end users of the electronic judging system, yet their biases affect their willingness to accept or resist the technology quite differently. In-depth study of the reasons for such different effects of end users' biases

on their perception of technologies could yield interesting and valuable insight.

6. References

[1] Angwin, J., J. Larson, S. Mattu, and L. Kirchner, "Machine Bias", *ProPublica*, 2016.

[2] Arnott, D., "Cognitive biases and decision support systems development: a design science approach", *Information Systems Journal 16*(1), 2006, pp. 55–78.

[3] Buolamwini, J., and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", *Proceedings of Machine Learning Research*(81), 2018, pp. 1–15.

[4] Dastin, J., "Amazon scraps secret AI recruiting tool that showed bias against women", *Reuters*, 2018.

[5] Dickson, B., "What is algorithmic bias?", *TechTalks*, 2018.

[6] Gerlach, J.P., P. Buxmann, and T. Dinev, "'They're All the Same!' Stereotypical Thinking and Systematic Errors in Users' Privacy-Related Judgments About Online Services", *Journal of the Association for Information Systems 20*(6), 2019, pp. 787–823.

[7] Geva, H., G. Oestreicher-Singer, and M. Saar-Tsechansky, "Using retweets when shaping our online persona: Topic modeling approach", *MIS Quarterly 43*(2), 2019, pp. 501–524.

[8] Greenstein, S., and F. Zhu, "Do experts or crowd-based models produce more bias? Evidence from encyclopedia britannica and wikipedia", *MIS Quarterly 42*(3), 2018, pp. 945–958.

[9] Hao, K., "This is how AI bias really happens—and why it's so hard to fix", *MIT Technology Review*, 2019.

[10] Harini, V., "A.I. 'bias' could create disastrous results, experts are working out how to fight it", *CNBC*, 2018.

[11] Hu, N., P.A. Pavlou, and J. Zhang, "On Self-Selection Biases in Online Product Reviews", *MIS Ouarterly 41*(2), 2017, pp. 449–471.

[12] Johnson, C.Y., "Racial bias in a medical algorithm favors white patients over sicker black patients", *The Washington Post*, 2019.

[13] Jones, M., "What we talk about when we talk about (big) data", *Journal of Strategic Information Systems 28*(1), 2019, pp. 3–16.

[14] Kim, A., and A.R. Dennis, "Says who? The effects of presentation format and source rating on fake news in social media", *MIS Quarterly 43*(3), 2019, pp. 1025–1039.

[15] Kim, A., P.L. Moravec, and A.R. Dennis, "Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings", *Journal of Management Information Systems 36*(3), 2019, pp. 931–968.

[16] Lambrecht, A., and T. Catherine, "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads", *Management Science 65*(7), 2019, pp. 2966–2981.

[17] Lee, K., and K. Joshi, "Examining the use of status quo bias perspective in IS research: need for re-conceptualizing and incorporating biases", *Information Systems Journal 27*(6), 2017, pp. 733–752.

[18] Lee, S.Y., H. Rui, and A.B. Whinston, "Is best answer really the best answer? The politeness bias", *MIS Quarterly 43*(2), 2019, pp. 579–600.

[19] Legoux, R., P. Leger, J. Robert, and M. Boyer, "Confirmation Biases in the Financial Analysis of IT Investments", *Journal of the Association for Information Systems 15*(1), 2014, pp. 33–52.

[20] Lipton, Z.C., "The Foundations of Algorithmic Bias", *Technical and Social Perspectives on Machine Learning*, 2016.

[21] Mazurova, E., and E. Penttinen, "Probing Athletes' Perceptions Towards Electronic Judging Systems – A Case Study in Gymnastics", *Proceedings of the 53rd Hawaii International Conference on System Sciences*, (2020).

[22] Metz, C., and A. Satariano, "An Algorithm That Grants Freedom, or Takes It Away", *The New York Times*, 2020.

[23] Minas, R., R. Potter, A. Dennis, V. Bartelt, and S. Bae, "Putting on the thinking cap: Using NeuroIS to understand information processing biases in virtual teams", *Journal of Management Information Systems 30*(4), 2014, pp. 49–82.

[24] Moravec, P.L., R.K. Minas, and A.R. Dennis, "Fake news on social media: People believe what they want to believe when it makes no sense at all", *MIS Quarterly 43*(4), 2019, pp. 1343–1360.

[25] Ramachandran, V., and A. Gopal, "Managers' judgments of performance in IT services outsourcing", *Journal of Management Information Systems 26*(4), 2010, pp. 181–218.

[26] Salovaara, A., K. Lyytinen, and E. Penttinen, "High reliability in digital organizing: Mindlessness, the frame problem, and digital operations", *MIS Quarterly 43*(2), 2019, pp. 555–578.

[27] Schwab, K., "Facial Recognition Systems Are Even More Biased Than We Thought", *FastCompany*, 2018.

[28] Shmueli, O., N. Pliskin, and L. Fink, "Can the outside-view approach improve planning decisions in software development projects?", *Information Systems Journal 26*(4), 2016, pp. 395–418.

[29] Wang, J., Y. Li, and H.R. Rao, "Overconfidence in phishing email detection", *Journal of the Association for Information Systems 17*(11), 2016, pp. 759–783.

[30] Xu, X., J.Y.L. Thong, and K.Y. Tam, "Winning Back Technology Disadopters: Testing a Technology Readoption Model in the Context of Mobile Internet Services", *Journal of Management Information Systems 34*(1), 2017, pp. 102–140.

[31] Yapo, A., and J. Weiss, "Ethical Implications of Bias in Machine Learning", *Proceedings of the 51st Hawaii International Conference on System Sciences*, (2018).