AI Oversight and Human Mistakes: Evidence from Centre Court*

David Almog†, Romain Gauriot‡, Lionel Page§, and Daniel Martin¶

February 18, 2024

Abstract

Powered by the increasing predictive capabilities of machine learning algorithms, artificial intelligence (AI) systems have begun to be used to overrule human mistakes in many settings. We provide the first field evidence this AI oversight carries psychological costs that can impact human decision-making. We investigate one of the highest visibility settings in which AI oversight has occurred: the Hawk-Eye review of umpires in top tennis tournaments. We find that umpires lowered their overall mistake rate after the introduction of Hawk-Eye review, in line with rational inattention given psychological costs of being overruled by AI. We also find that umpires increased the rate at which they called balls in, which produced a shift from making Type II errors (calling a ball out when in) to Type I errors (calling a ball in when out). We structurally estimate the psychological costs of being overruled by AI using a model of rational inattentive umpires, and our results suggest that because of these costs, umpires cared twice as much about Type II errors under AI oversight.

*We thank Lucas Lippman and Leshan Xu for excellent research assistance. We also thank Ashesh Rambachan, Mitchell Hoffman, Colin Camerer, Luis Rayo, Mike Powell, George Georgiadis, Devin Pope, Meghan Busse, Yuval Salant, Jörg Spenkuch, Ben Enke, Raphaël Raux, David Argente, Andrew Caplin, and audiences in the Kellogg Strategy Brown Bag Seminar and Booth Behavioral Lab for helpful comments.

†Kellogg School of Management, Northwestern University, david.almog@kellogg.northwestern.edu.

‡Deakin University, romain.gauriot@deakin.edu.au.

§University of Queensland, lionel.page@uq.edu.au.

¶University of California, Santa Barbara and Kellogg School of Management, Northwestern University, daniel@martinonline.org.

1 Introduction

Powered by the increasing predictive capabilities of machine learning algorithms, artificial intelligence (AI) systems are helping humans to reduce the number of mistakes that they make while performing job-related tasks.[1] One way that AI systems are helping humans is by providing guidance, such as a recommendation or risk score. Humans can use this information if they want, but they retain final "decision rights." However, there are also high-stakes settings where AI systems are used to overrule human mistakes, so AI is given the final decision rights instead. This is because a final decision has to be made quickly or because AI guidance is being systematically ignored (as a result of pride, unfamiliarity, overconfidence, or bias). For example, automobiles now use computer vision systems to steer cars back into their lanes or to apply brakes automatically when it appears an object will be hit. Within the larger scope of error reduction, AI has begun to overrule worker decisions as a part of quality control measures in a variety of industries, such as financial trading and energy management. An example from the air traffic control industry is TopSky-ATC, an AI system that can overrule problematic flight paths to increase flow and improve safety.

Although AI is accurate enough to overrule human mistakes, humans are allowed to make the initial decision (kept "in the loop") for social reasons (tradition, comfort, fairness, etc.), labor market concerns (union power, contractual responsibilities, inequality considerations, etc.), or because humans perform better in general since they can incorporate additional information, understand context, handle edge cases, and adapt to changing circumstances. Thus, keeping human decision-making in place while using AI to overrule erroneous human decisions appears to be a straightforward improvement for societal welfare. In fact, recent papers in the economics literature have shown that there are large potential gains from using AI to overrule human mistakes in other high-stakes settings, including law (e.g., Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan 2018; Rambachan 2022) and medicine (e.g., Raghu, Blumer, Corrado, Kleinberg, Obermeyer, and Mullainathan 2019).

However, assessing the full impact of AI oversight requires understanding whether its presence alters human decision-making. While a large amount of research has focused on how humans respond when assisted by AI, very little is known about how humans respond when their decisions might be overruled by AI. Because being overruled can carry psychological costs (e.g., shame and embarrassment of being overruled) and psychological benefits (e.g., relief at having their mistakes fixed), individuals might alter their decision-making under AI oversight. This illustrates how insights from behavioral economics can play a key role in understanding AI-human interactions (Camerer 2019).

To the best of our knowledge, we provide the first field evidence that AI oversight can

[1] See Section 1.1 for an in-depth discussion of related literature.

impact human decision-making by studying one of the highest visibility settings in which AI oversight has occurred: umpiring in top tennis tournaments. The prospect of AI oversight was introduced in top tennis tournaments through a technology known as Hawk-Eye. When challenged, the umpire's decision was overturned (or, in some cases, the point was replayed) if it contradicted the predictions of this highly accurate AI.

We find that umpires lowered their overall mistake rate by 8% (1.1 p.p.), in line with rational inattention given psychological costs to being overruled by AI. However, this improvement is concentrated in less-focused cases (non-serves). For more-focused cases (serves), the mistake rate actually *increased* by 22.9% (7.3 p.p.) for the closest calls (balls within 20 mm of the line). This puzzling result is explained by a behavioral response that occurs with AI oversight. We find that for the closest calls, umpires increased the rate at which they called balls in by 12.6% (6.2 p.p.) after the introduction of Hawk-Eye, which produced a shift from making Type II errors (calling a ball out when in) to Type I errors (calling a ball in when out). This effect is present in all cases, but for more-focused cases, the rate of mistakes is lower, so the dominant effect is the shift in errors. For less-focused cases, the mistake rate is higher, so the dominant effect is a decrease in mistakes.

This behavioral reaction is a sensible outcome of the asymmetric psychological costs of AI oversight in this setting. For umpires, improperly stopping a point (calling a ball out when it was actually in) became a new cause of concern with the introduction of Hawk-Eye. Even if a player successfully challenges such a call, there is no way to return back to where the game was, so the rules dictate that the umpire must decide whether to replay the point or award it to the challenger. Thus, Type II errors (calling a ball out when in) carry two psychological costs: one from the impossibility of implementing the correct outcome (continuing the point) and another from having to implement an arbitrary decision that, in many cases, unleashes the outrage of the players involved and the audience.[2]

We structurally estimate a lower bound on the psychological costs being overruled by AI using a model of rational inattentive umpires. We employ a two-stage approach in which we use decisions before Hawk-Eye was introduced to recover the perceptual costs of making correct calls, and then decisions after Hawk-Eye was introduced to determine the psychological costs of being overruled by AI. The resulting estimates suggest that these psychological costs lead umpires to care twice as much about Type II errors (calling a ball out when in) after the introduction of Hawk-Eye review.

Our results highlight the fact that while introducing an AI device that overrules apparent human mistakes seems promising in principle, especially since it might motivate higher effort

[2] AI oversight could also lead to reputation and career concerns, and these would be asymmetric if ATP Tour organizers punished referees for calls that induce replay decisions (despite their entertainment value). However, we have no evidence, even anecdotal, that this is the case.

from workers, there are two central reasons to approach this with caution. First, there is a *free-riding* motive in which the technology can dissuade decision-makers from contributing effort under the belief that important mistakes will be recognized and wiped out by the technology anyway. This argument holds more weight in contexts where the decision-maker prioritizes the ultimate outcome over their own performance.

The second reason, which is more important for the context we study, is an *incentive misalignment* motive, as the particular implementation guidelines of the oversight mechanism can shift the incentives for the decision-maker in a way that deteriorate the overall quality of their decisions from the perspective of the social planner. As we will argue in Section 4, we do not have reason to believe that before the introduction of Hawk-eye review a particular type of incorrect call was costlier than the other for umpires. Under this assumption, umpires should be aiming to minimize the total number of mistakes when there was no oversight. However, if the umpires' attention has already reached a point where improvement would demand a tremendous amount of additional effort, distorting the relative costs of the two types of errors could result in an increase in total number of mistakes. Umpires may be inclined to reduce the now costlier type of error (e.g., calling a ball out when it was in) at the expense of increasing, by more than a 1:1 ratio, the now relatively less costly type of error (e.g., calling a ball in when it was out). Our analysis suggests that when serves landed very close to the line, umpires were more inclined to call them in, as an attempt to minimize the occurrence of the more costly type of error. It also suggests that in overall terms, the effect of incentive misalignment outweighed the incentives to improve performance through public oversight, which led to an upsurge of incorrect calls during serves.

The rest of the paper proceeds as follows. In Section 1.1, we review related literature on AI and human decision-making and social image concerns. In Section 2, we provide additional details on tennis umpiring, Hawk-Eye review, and the data sources we leverage in our analysis. In Section 3, we examine overall mistake rates, types of calls, and types of errors. We also explore whether the impact of AI oversight differs by tournament stage. In Section 4, we provide a model of perceptually-constrained umpires and use it structurally estimate the psychological costs of AI oversight. Finally, we conclude with a brief discussion in Section 5.

1.1 Related Literature

A recent body of literature has provided evidence that AI can outperform humans in important prediction problems (see Agrawal, Gans, and Goldfarb 2022 for an in-depth discussion of AI capabilities and its impact on society and business.). Some examples of machine learning algorithms eclipsing experts include bail judges predicting pretrial misconduct (Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan 2018), radiologists predicting pneumonia from chest X-rays (Rajpurkar, Irvin, Zhu, B. Yang, Mehta, Duan, Ding, Bagul, Ball, Langlotz, Shpanskaya, Lungren, and Ng 2017; Topol 2019), and workforce professionals predicting productivity for hiring and promotion (Chalfin, Danieli, Hillis, Jelveh, Luca, Jens, and Mullainathan 2016). Despite these results, a prevalent argument in favor of human-driven decisions over AI-driven ones has been the ability of humans to integrate private information and the power of humans to leverage context. From a theoretical perspective, Iakovlev and Liang (2023) demonstrate that in the presence of sufficient uncertainty in the structure of prediction problems, the value of context diminishes as the number of covariates grows large. This finding challenges the extent to which understanding context can actually benefits humans relative to AI.

While AI continues to make strides, a decision-making process that involves both AI and humans has the potential to yield better results than either acting alone. The prospect of introducing AI into the decision-making process opens the door to multiple possibilities. One approach that is being actively studied involves having humans make the final decision after receiving an algorithmic recommendation. Making decisions in environments with algorithmic recommendations introduces the value of exercising discretion as a new skill. Angelova, Dobbie, and C. Yang (2023) shows that in the context of pretrial misconduct, this methodology has potential, and some judges may be able to incorporate algorithmic recommendations and perform better than the algorithm alone. However, it also showcase the complications of discretion, as many judges would be better off by simply following the algorithmic recommendations as a whole.[3] The findings from Hoffman, Kahn, and Li (2018) highlight challenges in exercising discretion in hiring, as managers are observed overruling recommendations due to personal biases rather than private information motives. Exploring how humans adopt algorithmic recommendations, Agarwal, Moehring, Rajpurkar, and Salz (2023) find that providing radiologists with access to AI predictions does not, on average, result in improved performance. Another noteworthy finding from their study is that, radiologists take significantly more time to make decisions when AI information is provided. This observation exposes certain limitations of the algorithmic recommendation approach, indicating its better suitability for scenarios where making contemplative decisions is feasible.

A much more understudied approach to combining AI and human decision-making is what we describe as human decision-making under AI oversight. To the best of our knowledge, our paper provides the first field evidence on this use of AI. With AI oversight, humans are in charge of making decisions, but these decisions are placed under AI scrutiny and potentially overruled. With this form of human-AI interaction, "discretion", a prominent

[3]One potential advantage of our setting for studying human-AI interaction is that, at least relative to the judicial context, human line judges do not have any clear source of private information. We thank Ashesh Rambachan for raising this point.

force in algorithmic recommendation systems, no longer plays a major role. AI oversight, however, introduces new behavioral forces into play, such as shame, pride, embarrassment, and stress. These behavioral factors likely to be particularly strong in the setting we study because in addition to players, the audience is also heavily invested in the outcome of AI oversight.[4]

Because of this connection, we contribute to the literature on social image and peer effects (Bursztyn and Jensen 2017) by showing that the cost of being corrected in public can overpower the potential incentive to free ride on technology. Butera, Metcalfe, Morrison, and Taubinsky (2022) present a novel methodology for measuring the welfare effects of shame and pride in various experimental scenarios. In our work, recover what we interpret to be a lower bound for the psychological costs of due to being overruled by AI, and we feel that shame is a likely component of these costs. To recover these costs, we implement a model of rational inattention adapted for AI oversight and estimate the psychological costs of due to being overruled by AI for both Type I and Type II errors. An interesting feature of our empirical context is that, with the introduction of AI oversight, one type of mistake became more controversial and led to a larger backlash, including complaints from players and fans. This introduced a change in the costs of different error types, which led to a decrease in the controversial mistake at the expense of increasing the less embarrassing one. This result aligns with recent findings by Engelmann, Lebreton, Salem-Garcia, Schwardmann, and van der Weele (2023), who, through lab experiments, demonstrated that anticipatory stress can impact performance in cognitive tasks.

We also add to the literature on monitoring, which has studied human reactions to being monitored across various domains, including auditing and corruption (Olken 2007; Bobonis, Cámara Fuertes, and Schwabe 2016), environmental policy compliance (Gray and Shimshack 2011; Zou 2021), and workforce productivity (Gosnell, List, and Metcalfe 2020). Nagin, Rebitzer, Sanders, and Taylor (2002) find that many call-center workers behave as in the "rational cheater" model, which predicts that shirking behavior will decrease when monitoring increases. In line with this, we show that using AI systems to increase monitoring can improve performance, even though there are additional incentives to shrink from being able to free-ride on AI corrections. It is an open question as to whether humans respond differently to human and AI monitoring, thus we cannot assume that the behavior we observe in our study would be the same if the umpires were monitored by humans. However, AI technology extends the range of scenarios where monitoring becomes operationally or economically feasible, as there are settings where AI monitoring is less costly, quicker, or more effective than human monitoring. We consider our setting to be one such case.

Lastly, we also contribute to the literature on attention and mistakes (see Caplin 2023 and

[4] The crowd sometimes starts clapping with the countdown of a Hawk-Eye reveal!

Maćkowiak, Matějka, and Wiederholt 2023 for reviews).[5] We extend the canonical model of rational inattention with costs that scale linearly with Shannon mutual information (e.g., Caplin and Dean 2013; Matějka and McKay 2015). This extension allows us to incorporate two general features that we will later show fit well into our empirical setting. First, we allow for asymmetric costs of attention for different states. Second, we include behavioral factors in utility to accommodate the outcome of being overruled, which we will later refer to as the *AI oversight penalty*. Bhattacharya and Howard (2022) find that the standard rational inattention model explains the equilibrium behavior of professional baseball players, estimating the linear cost of attention among other things. Our rich data set, containing tournaments from both periods (with and without oversight), allows us to estimate not only the parameters associated with the cost of attention but also the state-dependent utility loss incurred when being overruled by AI. This novel element is a central component of our research question. Furthermore, our research introduces one of the first cognitive economic models that incorporate the influence of behavioral factors on rational attention allocation. Recent work by Bolte and Raymond (2023) and Almog and Martin (2023) suggest the importance of expanding rational attention modeling to account for emotional states.

2 Setting and Data

In March 2006, at the Nasdaq-100 Open in Key Biscayne (currently known as the Miami Open), Hawk-Eye reveiw was officially used for the first time at an ATP Tour event.[6] Later that year, major tennis tournaments, including the US Open, adopted this new technology, allowing players to challenge calls made by umpires. Hawk-Eye uses 6 to 10 computer-linked television cameras positioned around the court to collectively create a three-dimensional representation of the ball's trajectory. Hawk-Eye performs with an average error of just 3.6 mm,[7] so just like the ATP Tour we will consider Hawk-Eye readings to be the ground truth.

We study the decision-making of tennis umpires before and after the introduction of Hawk-Eye review in professional tennis, focusing on umpire's judgment about whether a ball bounced in or out of bounds.[8] This setting has several advantages for studying the

[5] We acknowledge the ongoing debate regarding what to classify as a mistake (Nielsen and Rehbeck 2022). Here we refer to a *mistake* as a decision that is incorrect ex-post relative to an objectively correct answer. It is fair to argue that tennis umpires are not making mistakes, as incorrect calls are just a result of the cost structure of becoming informed.

[6] The ATP Tour is the top tennis tour organized by the Association of Tennis Professionals.

[7] For perspective, the standard diameter of a tennis ball is 67 mm.

[8] Our consolidated data set includes one tournament that was held without Hawk-eye review after the 2006 Nasdaq-100 Open. However, this does not provide enough statistical power for running a difference-in-difference analysis. Hence, we use the short-hand that there is a *before* and *after* period of Hawk-Eye review, even though this strictly only true at the tournament level.

impact of AI oversight on individual decision-making. First, the use of Hawk-Eye review in tennis was one of the pioneering uses of modern AI to conduct oversight in a work setting. Second, it is a setting of economic significance: the global revenue for the ATP Tour was 147.3 million USD in 2022 and the global revenue of Hawk-Eye was 71.6 million Euros in the 12 months to the end of March 2023.[9] Third, there was a testing period during which the technology was used only to record data and no challenges were allowed, which enables us to track the umpire's performance both before and after the formal introduction of Hawk-eye review to a tournament.[10] Fourth, in this setting there is an objectively correct decision and a reliable measure of it, which allows us to identify human mistakes. Finally, this setting offers the simplest possible decision-making problem, as the task is simply to match a binary action (call in or out) to a binary state (ball bounces in or out), which greatly simplifies our empirical and theoretical analysis.[11]

2.1 Professional Tennis, Umpires, and Hawk-Eye

Professional tennis is a racket sport that is played either one-on-one (singles) or two-on-two (doubles). In singles tennis, two players compete against each other on opposite sides of the tennis court. The objective is to score points by hitting the ball "in" (within the bounds of the opponent's side of the court). We will concentrate solely on men's singles matches (singles matches where both players are men) because the analysis of other formats is underpowered in our data.[12] The scoring system for tennis is based on a series of points, games, and sets. Players accumulate points to win a game and accumulate games to win a set. Matches are typically played as the best of three or five sets, with each set requiring a player to win at least six games.

A crew of up to 10 umpires is involved in officiating a match. The roles typically include one chair umpire who oversees the entire match, making decisions on points, game penalties, and overall match control. Additionally, there are usually nine line umpires positioned around the court, each responsible for specific lines, with the sole duty to determine whether the ball bounced in or out when it is relatively close to their respective line.[13] The chair umpire has the authority to overrule the decisions of the line umpires if necessary, so drawing

[9] Sources: https://www.breakingnews.ie/sport/revenues-at-hawk-eye-firm-increase-to-e71-6m-as-higher-costs-hit-profits-1534936.html and https://www.zippia.com/atp-tour-careers-985960/revenue/.

[10] We do not know whether umpires were aware that the Hawk-Eye system was collecting data before the introduction of challenges. However, if they were aware that they were being recorded, then we are likely underestimating the impact of AI oversight on their decision-making.

[11] See Palacios-Huerta (2023) for a comprehensive review of other sports settings that capitalized on favorable empirical situations to address economic questions.

[12] If more data was available on the other formats, it might be of interest to examine whether the impacts of AI oversight vary by gender or team composition.

[13] Figure 7 in the Appendix provides a visual representation of the positioning of each umpire.

inferences on the performance of an individual umpire is complicated, especially since we do not have data on whether the chair umpire overruled an individual line umpire. Thus, in this paper we evaluate performance of the umpire crew as a whole.

The way the Hawk-Eye review protocol works is that players are endowed with 2-3 challenges per set.[14] If a challenge is successful, players do not lose a challenge opportunity. When a player challenges a call, a computerized path and the final landing location of the ball are displayed on a large screen in the stadium for the umpire, players, and the crowd to observe the outcome of the challenge. The public nature of the challenge process adds excitement to the spectator, but simultaneously adds pressure to the umpires, as their decisions are being publicly scrutinized.

An important component of the review system implementation is the asymmetric resolution of incorrect calls, which varies depending on whether the challenged call was initially in or out. If a ball is initially ruled in by the umpires and the challenge successfully overturns the call, then the point ends with the challenger winning the point and the correct outcome being enforced. However, when a ball is initially ruled out by the umpires but the review shows otherwise, enforcing the correct outcome is not always possible because the point was unnecessarily stopped. In this case, the umpire has to make an arbitrary decision on whether the opponent of the challenger had a real chance to return the ball. If the answer is yes, the point has to be replayed from scratch. This situation can be perceived as detrimental to the game due to the inability to implement the correct outcome (continuing the point from where it stopped) and because it forces umpires to make arbitrary decisions that, in multiple instances, have been shown to infuriate one of the players involved.

2.2 Data

While this paper is not the first to utilize tennis data for drawing inferences on decision-making, we have assembled a novel and rich data set that, for the first time, enables us to comprehensively assess the performance of professional tennis umpires. To achieve this goal, we utilized three distinct data sources. Our primary data set, the *Hawk-Eye Base* data set, encompasses precise information on points and, crucially, the location of every bounce for over 100 tournaments.[15] Our second data set, the *Challenge* data set, tracks the outcomes of challenges during a sample of tournaments that took place in the first few years following the implementation of Hawk-Eye review in professional tennis. In the period of time where challenges are used, the first data set only permits drawing conclusions on players' behavior

[14]In practice, players very rarely exhaust their challenges. This is advantageous for our research question, as umpires are almost always under the threat of being challenged.

[15]We excluded from the analysis the 15 clay court tournaments, we elaborate on this decision in the Appendix A.1.

because, when observing a correct call, it is not possible to disentangle whether the umpire made the correct call initially or if a Hawk-Eye review corrected the umpire's incorrect call. In contrast, the second data set allows precise identification of incorrect calls, as every won challenge is, by definition, an admission of the umpire's mistake. Nonetheless, this second data set is incomplete because it only includes points in which players decided to challenge, and this selection is susceptible to the perceptual and behavioral biases of players. Merging the first two data sets was challenging due to systematic inconsistencies in the Challenge data set (e.g., the score was often flipped across players and sometimes missing). To overcome this limitation, we turned to a third data source: manual review of points by replaying match videos. This allowed us to identify the ground truth for challenges in a subset of matches, enabling us to determine the optimal approach for merging the first two data sets.[16]

The consolidated data set produced by merging the Hawk-Eye Base data set and the Challenge data set comprises a total of 698 matches across 35 distinct tournaments, and summary statistics for this data set can be found in Table 1.[17] This data set includes 109 matches from 7 tournaments that were played before Hawk-eye review and 589 matches from 28 tournaments after Hawk-eye review was active. We were able to merge 2,038 out of the 2,108 challenges registered for those 28 tournaments (a 97% merge rate). We will now offer a more comprehensive description of the composition and role of each of the three data sources.

2.2.1 Hawk-Eye Base Data Set

The main source of data for this paper is the Hawk-Eye Base data set, which consists of the official Hawk-Eye data for matches played at the international professional level between March 2005 and March 2009. This data set provides a number of pieces of information for each point, including the position of every bounce captured by the Hawk-Eye system during that point, the serving player, the ongoing score, and the point winner. Altogether, this data set includes information on 1.8 million bounces from more than 1800 men's singles matches, spanning various prestigious tournaments such as Grand Slam tournaments, Masters 1000 tournaments, and International series tournaments. This data set has been used in the past to study important economics questions such as risk management (Ely, Gauriot, and Page 2017), tournament incentives (Gauriot and Page 2019), and mixed strategy play (Gauriot, Page, and Wooders 2023). These papers have focused on the players' perspective, and the Hawk-Eye Base data set is well-suited for this purpose. However, this data set is insufficient to analyze umpires' performance, as it does not permit us to identify whether the final call

[16]The video auditing process was also instrumental in validating the best rules for determining when a call was made incorrectly.

[17]Figure 8 in the Appendix provides information on the month, category, court type, and the number of matches for each tournament.

	Before Hawk-Eye review (PostHK=0)	After Hawk-Eye review (PostHK=1)	Total
A. Players	71	161	174
B. Tournament tier			
ATP 250	2	12	14
ATP 500	0	3	3
ATP 1000	5	13	18
All	7	28	35
C. Matches			
Final	7	27	34
Semifinal	14	53	67
Quarterfinal	19	88	107
Round of 16	14	113	127
Other	55	308	363
All	109	589	698
D.Points	15,439	83,898	99,337
E. Bounces (share)			
< 20 mm from the line	556 (0.7%)	2,760 (0.6%)	3,316 (0.6%)
< 100 mm from the line	2,622 (3.6%)	14,190 (3.5%)	16,812 (3.5%)
Serves	20,942 (29.2%)	111,065 (27.6%)	132,009 (27.8%)
Non-serves	50,696 (70.8%)	291,381 (72.4%)	342,093 (72.2%)
All	71,638 (100%)	402,446 (100%)	474,102 (100%)
F. Average speed in km/h (s.d.)			
Serves	147.7 (24.2)	150 (23.1)	149.6 (23.3)
Non-serves	81.7 (21.4)	83.3 (20.4)	83.1 (20.5)
All	101 (37.4)	101.7 (36.5)	101.6 (36.6)

Table 1: Summary statistics for the consolidated data set.

comes from the umpire or via an overturned challenge.

2.2.2 Challenge Data Set

The Challenge data set was originally obtained from the ATP Tour and encompasses all challenges recorded during 35 tournaments across the initial three years following the introduction of Hawk-Eye (2006-2008). It captures comprehensive information on 2,784 challenges, including details about the players involved, the match itself, the specific point in the match when the challenge was made, and the outcome of each challenge. In order to compile this data set, Abramitzky, Einav, Kolkowitz, and Mill (2012) undertook the arduous effort of collecting and compiling umpire match sheets. Abramitzky, Einav, Kolkowitz, and Mill (2012) acknowledge the selection problem involved in only observing the challenged points (around 2.6% of total points). Nevertheless, they are able to draw inferences on the optimality of decision-making from the players standpoint by relying on the idea that a higher propensity to challenge implies challenging less conservatively. However, without the point-by-point data, it is hard to assess the umpire's performance once Hawk-Eye review was active. By merging the first two data sets, we solve the aforementioned selection problem and gain information on what was originally called by the umpire crew.

2.2.3 Video Auditing

In order to assess the quality of our merging algorithm, we audited full-match video replays using TennisTV, the official ATP streaming service. We identified 43 matches that appeared in both the Hawk-Eye Base and Challenge data sets, providing us with assessment on how well a given merging algorithm would work in the 144 challenges witnessed during these matches.

Using variables such as match, point, distance, and who hit the ball, we developed a merging algorithm with a 96.5% merge rate, for which only 2.1% of the video-audited challenges were merged to an incorrect point. The algorithm consists of 8 iterations, starting with the most stringent merging rule and gradually relaxing criteria for challenges that persist without merging.[18]

As an additional benefit, auditing match replays enabled us to test the effectiveness of the four criteria we implemented to identify incorrect calls (two criteria for each type of mistake). We deem an in call to be incorrect when a player's stroke is recorded as bouncing out in the Hawk-Eye Base data set, yet they still win the point, or if the data set records at least three more strokes after an out bounce. We deem an out call to be incorrect if all the

[18] A detailed explanation of the merging algorithm is provided in Appendix A.2.

strokes of a point are recorded as in, and the player delivering the final hit loses, or if there is a second serve after the first one was recorded as in the Hawk-Eye Base data set.[19]

3 Empirical Results

In this section, we use our consolidated data set to study the impact of AI oversight on umpiring decision-making in professional tennis. We begin by examining changes in the overall mistake rate for all hits and then separately for serves and non-serves. We then study changes in the frequency of each call and the resulting impact on different types of mistakes. In particular, we highlight the influence of distance from the line. We end by considering whether there is important heterogeneity by tournament stage.

3.1 Overall Mistake Rate

Before Hawk-eye review was introduced, the umpire mistake rate was only 0.61%. However, if we look at the 2,622 (3.6%) bounces that fell within 100 mm of the line, the mistake rate jumps to 13.89%. Looking at the 556 (0.78%) of bounces that fell within 20 mm of the line, the mistake rate jumps even more to 32.91%. Given our interest in the impact of AI oversight on human mistakes, our primary focus will be on calls within 100 mm or 20 mm of the line.[20]

While the likelihood of a close call (a ball bouncing within 100 mm or 20 mm of the line) did not change substantially after the introduction of Hawk-Eye and the probability of a close call being in or out did not change substantially either,[21] we find that the mistake rate on close calls did change substantially. We estimate the effect of AI oversight on umpires' performance using the following specification:

$$\mathbb{1}(Incorrect\,Call)_{ipm} = \alpha_0 + \alpha_1 \boldsymbol{PostHK_m} + \beta X_p + \gamma Y_m + \epsilon_{ipm} \tag{1}$$

$\mathbb{1}(Incorrect\,Call)$ is an indicator variable equal to 1 if the call i made by the umpire in point p in match m was incorrect. $\boldsymbol{PostHK}$ is an indicator variable equal to 1 if the match m belongs to a tournament with the Hawk-Eye review active. X is a vector of point characteristics: distance fixed effects (by 20 mm bins), speed (whether is below or above the

[19] Appendix A.3 documents these criteria further.

[20] We also find that over 93% of the challenges in our sample are for calls within 100 mm of the line.

[21] See Table 1 for the likelihood of a close call. The probability that a ball was out when it bounced within 100 mm of the line was 45.3% before Hawk-Eye review and 46.8% after and the corresponding numbers for 20 mm are 48.4% and 48.9%.

	(1)	(2)	(3)	(4)
	Incorrect call			
PostHK	-0.014**	-0.011	-0.011	-0.011
	(0.007)	(0.006)	(0.07)	(0.007)
Point controls		X	X	X
Game controls			X	X
Cluster level				Match
N	16,812	16,812	16,335	16,335
Baseline mean	.139	.139	.136	.136

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

(a) For all hits.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Incorrect call (Serves)				Incorrect call (Non-serves)			
PostHK	-0.006	-0.002	0.000	0.000	-0.024**	-0.021**	-0.023**	-0.023**
	(0.010)	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.011)
Point controls		X	X	X		X	X	X
Game controls			X	X			X	X
Cluster level				Match				Match
N	9,253	9,253	8,990	8,990	7,559	7,559	7,345	7,345
Baseline mean	.139	.139	.137	.137	.138	.138	.136	.136

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

(b) Separately for serves and non-serves.

Table 2: OLS regressions of umpire mistakes for balls bouncing within 100 mm of the line. $PostHK = 1$ if the Hawk-Eye review is active; point controls include distance fixed effects (by 20 mm bins), speed (whether is below or above the median), score, game, set, and an indicator if the point played is in the tie-break stage; and match controls include round and tournament tier.

median), score, game, set and an indicator if the point played is in the tie-break stage.[22] Y is a vector of match characteristics that has round and tournament tier.[23]

Our main coefficient of interest is α_1, which can be interpreted with OLS regression as the effect in percentage points of AI oversight on the probability that an umpire's call is incorrect. Table 2a reports the results of estimating this equation for calls within 100 mm of the line. In our primary specification, the mistake rate is estimated to decrease 8% (1.1 p.p.). This decrease in the mistake rate is consistent with a model of rational inattention in which the psychological costs of being overruled by the AI outweigh any benefits (see Section 4 for a model of this form).

[22] A tie-break is a one-off game held to decide the winner of a set when two players are locked at 6-6

[23] Given that the period without oversight has no 500-tier tournaments, we aggregate the 250 and 500 groups, so we basically control for whether the match is a Masters 1000 tournament or not.

3.1.1 Serves and Non-Serves

There is value in distinguishing between serves and non-serves,[24] as we can think of them as distinct tasks for the umpires. The primary differences from an officiating standpoint, between serves and subsequent hits, lie in speed and anticipation. Serves may be considered more challenging due to their significantly higher speed: the average speed for a serve in our sample is 156 km/h, compared to 82 km/h for non-serves.[25] However, serves have the advantage of anticipation: umpires know the ball is about to be served and will most likely bounce in a very specific region of the court (the serving box). This appears to important: when comparing the slowest quartile of serves (averaging 133 km/h) with the fastest quartile of non-serves (averaging 106 km/h), we find that the error rates for serves are smaller, despite still containing faster strokes. The average incorrect rate during the period without Hawk-Eye review for the first quartile in serves and the fourth quartile in non-serves is 11.6% versus 15.6% for balls within 100 mm of the line, and 27.5% versus 37.1% within 20 mm of the line. Therefore, even though serves are much faster than non-serves, there is compelling evidence to support that the anticipation advantage benefits umpires during serves.

Table 2b reports the results from estimating equation 1 separately for serves and non-serves that bounced within 100 mm of the line. In our main specification we do not find evidence that AI oversight impacted umpire performance during serves. However, we do find evidence that AI oversight corresponds to a 2.3 percentage point decrease in incorrect calls for non-serves. This corresponds to a 17% reduction compared to the baseline mean level of 0.136.

3.1.2 Distance from the Line

Another important component to call difficulty is the distance from the line. If we just consider those calls when the ball bounces between 100 and 20 mm from the line, Table 3a, shows that the there is a statistically significant reduction in mistakes for both serves and non-serves. However, we find something different if we look at the closest calls. In Table 3b, we present the estimates for equation 1 when restricting the sample to bounces within 20 mm. This subset of calls is arguably the most complicated for umpires, and improving performance further would entail excessive cognitive costs. For these calls, the introduction of Hawk-Eye resulted in a 7.3 percentage point increase in incorrect calls for serves, representing a 22.9% increment from the baseline. While this result focuses on a very specific subset of the sample (serves within 20 mm), it provides evidence of which type of situations can lead AI oversight to backfire.

[24] The serve starts off every point in tennis, with players alternating serving each game.

[25] Figure 9 in the Appendix shows the speed distribution for both types.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Incorrect call (Serves)				Incorrect call (Non-serves)			
PostHK	-0.020**	-0.018**	-0.017**	-0.017**	-0.012	-0.017**	-0.020***	-0.020**
	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.011)
Point controls		X	X	X		X	X	X
Game controls			X	X			X	X
Cluster level				Match				Match
N	7,499	7,499	7,238	7,238	6,047	6,047	5,785	5,785
Baseline mean	.092	.092	.091	.091	.082	.082	.080	.080

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

(a) For balls bouncing 20-100 mm away from the line.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Incorrect call (Serves)				Incorrect call (Non-serves)			
PostHK	0.063**	0.063**	0.073**	0.073**	-0.035	-0.037	-0.031	-0.031
	(0.031)	(0.031)	(0.032)	(0.030)	(0.031)	(0.031)	(0.033)	(0.034)
Point controls		X	X	X		X	X	X
Game controls			X	X			X	X
Cluster level				Match				Match
N	1,804	1,804	1,752	1,752	1,512	1,512	1,470	1,470
Baseline mean	.325	.325	.319	.319	.333	.333	.325	.325

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

(b) For balls bouncing within 20 mm of the line.

Table 3: OLS regressions of umpire mistakes (separately for serves and non-serves). $PostHK = 1$ if the Hawk-Eye review system is active; point controls include distance fixed effects (by 20 mm bins), speed (whether is below or above the median), score, game, set, and an indicator if the point played is in the tie-break stage; and match controls include round and tournament tier.

What could explain the surprising result that the number of incorrect calls increases for serves after the introduction of Hawk-Eye review? In the following sections, we show that AI oversight leads to a change in the types of calls that are made, which can be rationalized as a shift in preferences for Type I and Type II errors, and this shift can explain the increase in mistakes.

3.2 Shift in Type I and Type II Errors

Figure 1 illustrates the relationship between mistake rates and the distance to the line for the 16,812 bounces within 100 mm of the line. In this figure, each dot represent the rate at which umpires made an incorrect call for all balls bouncing within one of ten 20 mm bins. The 5 bins to the left of the dashed line correspond to balls that landed out of bounds, while the 5 bins to the right side of that line represent balls that landed inside of the line.[26] The blue dots were calculated using tournaments before the introduction of Hawk-Eye review. The red dots represent tournaments with Hawk-Eye review. While it is expected that umpires' performance decreases as the ball gets closer to the line, we can also observe that the red line lies mostly under the blue line (this holds true for 8 of the 10 bins). This mirrors the results in Table 2a, which shows that overall performance of umpires improves after the introduction of AI oversight, and in Table 3a, which shows that this is particularly true for balls bouncing within 100 and 20 mm from the line. In addition, it is noticeable that the only region of the court where the introduction of Hawk-Eye increased the rate at which umpires make mistakes is the two closest bins on the outside of the line.

In Figure 2, we break down the incorrect call rates by serve and non-serves. This figure shows that the shift in types of errors is present in both serves and non-serves. For serves, the dominant effect is the shift in errors. For non-serves, the dominant effect is a decrease in mistakes. This suggests that non-serves may have a higher potential for wholesale improvement, which could be due to their more manageable speed and the opportunity they provide to rectify mistakes caused by distraction or sparse attention.

To further analyze this change in the type of errors made, we examine how the impact of AI oversight on umpire performance varies with distance to the line. We re-estimate equation 1, but now we interact the $PostHK$ indicator variable with each the indicator variables for each of the ten 20 mm distance bins into which the ball bounced. We report the estimated coefficients of these interaction terms graphically in Figure 3. For each bin, we plot the estimate coefficient as a dot and provide the respective 95% confidence interval around the estimated coefficient. The red dashed line, which separates the in and out parts of the court, shows a sharp discontinuity. The first bin to the left of the red dashed line (balls bouncing

[26] For the rest of the paper, negative distances will be used to denote the distance to the line for balls that bounced outside.

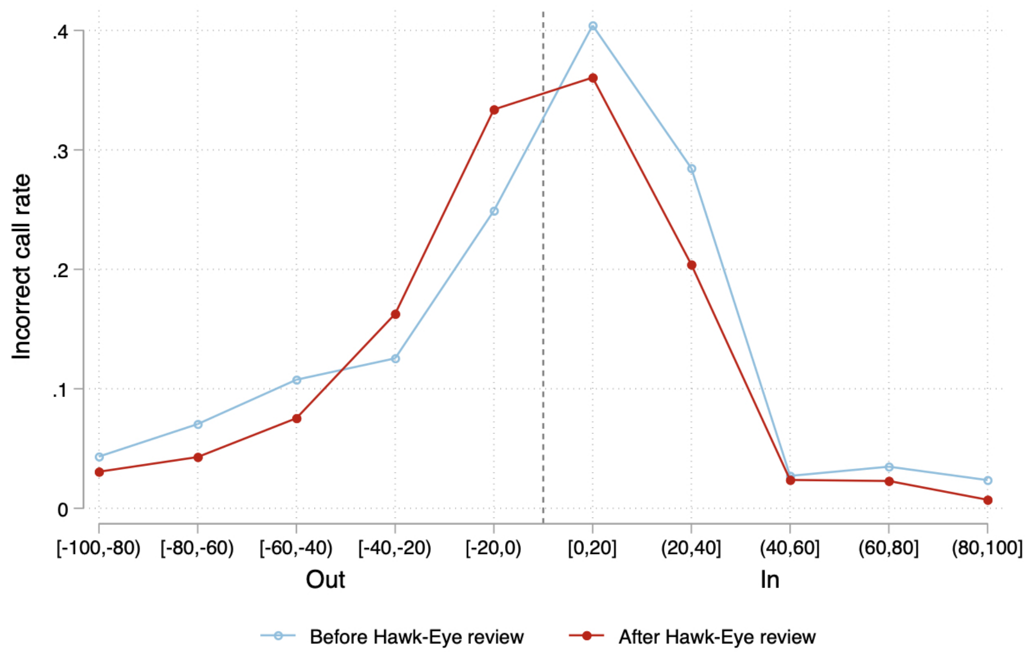
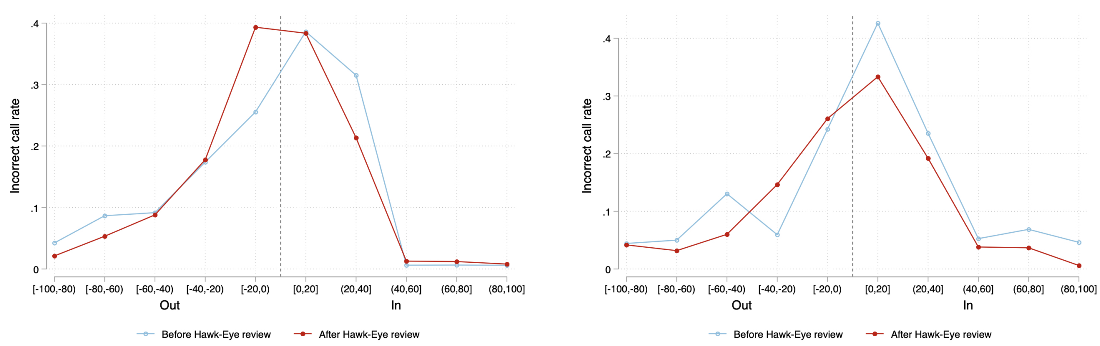


Figure 1: Incorrect call rates by proximity to the line. Each dot is the rate of incorrect calls for a bin of 20 mm. Dots to the left of the dashed line represent bins out of bounds, and the right of the dashed line represents bins in bounds.



(a) Serves.

(b) Non-serves.

Figure 2: Incorrect call rates by proximity to the line (separately for serves and non-serves). Each dot is the rate of incorrect call for a bin of 20 mm, and dots to the left of the dashed line represents bins out of bounds, and the right of the dashed line represents bins in bounds.

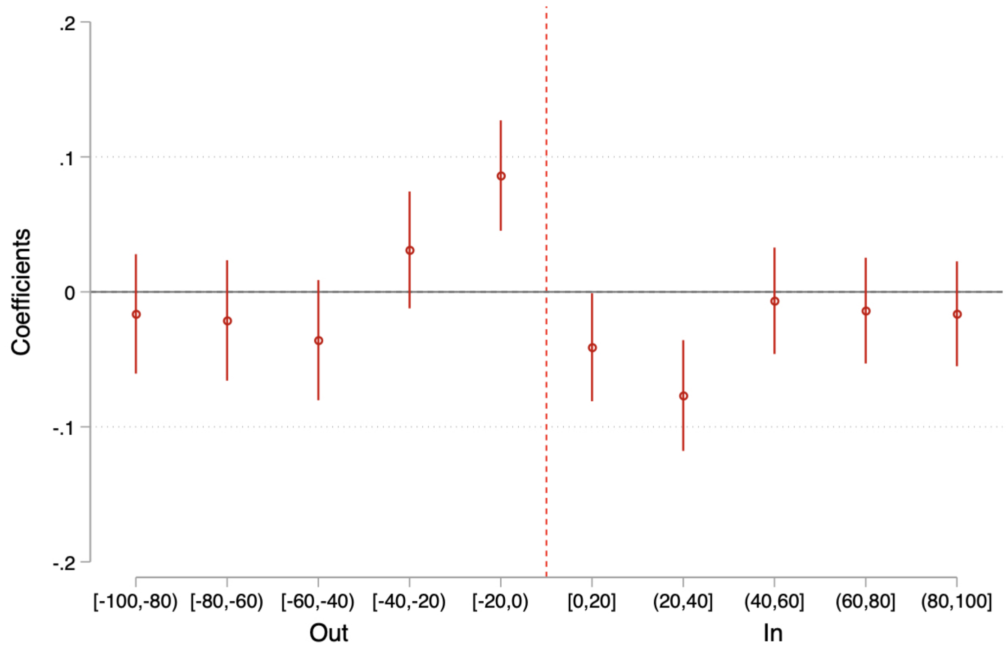


Figure 3: Impact of AI oversight on the incorrect call rate by proximity to the line. Each dot represents the coefficient on the interaction between distance bin and PostHK, the indicator variable that equals 1 if the Hawk-Eye review system is active.

just out) exhibits the most significant positive increase in incorrect calls, with an estimated coefficient of 8.6 percentage points (significant at the 1% level). From that point onward, the coefficients gradually adjust back down to the average treatment effect, just below zero. In the first two bins to the right of the red dashed line (balls bouncing in), we observe the most substantial decrease in incorrect calls, with estimated coefficients of -4.1 (significant at the 5% level) and -7.7 (significant at the 1% level) percentage points. Similarly, as we continue to move to the next bins in that direction, the estimated coefficients gradually increase to the average treatment effect level.

We provide an event-study version of Figure 3 in Figure 4. This figure illustrates how the impact of AI oversight on umpire performance progresses over time by separately examining matches in the first half and second half of the period we study. The first group comprises tournaments from 2006 and the first half of 2007, while the second group includes tournaments from the second half of 2007 and those from 2008. Figure 4 suggests that the shift in umpires' error types was not immediate, but instead took over a year to fully manifest. We chose this division into two time sub-periods to yield a similar number of observations in both groups, and the results are robust to breaking the tournaments into more time sub-periods, though the effects become more noisily estimated.

Clearly, the key shift in errors occurs for the closest calls. What drives this shift? Figure

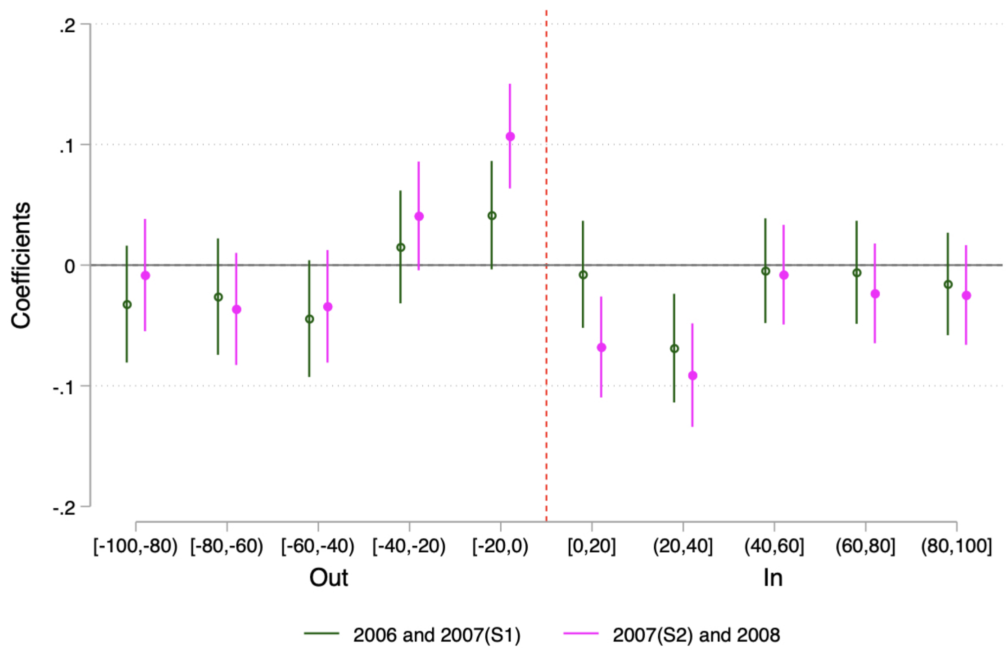


Figure 4: Impact of AI oversight on the incorrect call rate by proximity to the line (by time sub-period). Matches are group into those with Hawk-Eye review in all of 2006 and the first half of 2007 and those with Hawk-Eye review in the second half of 2007 and all of 2008. Each dot represents the coefficient on the interaction between distance bin and PostHK, the indicator variable that equals 1 if the Hawk-Eye review system is active.

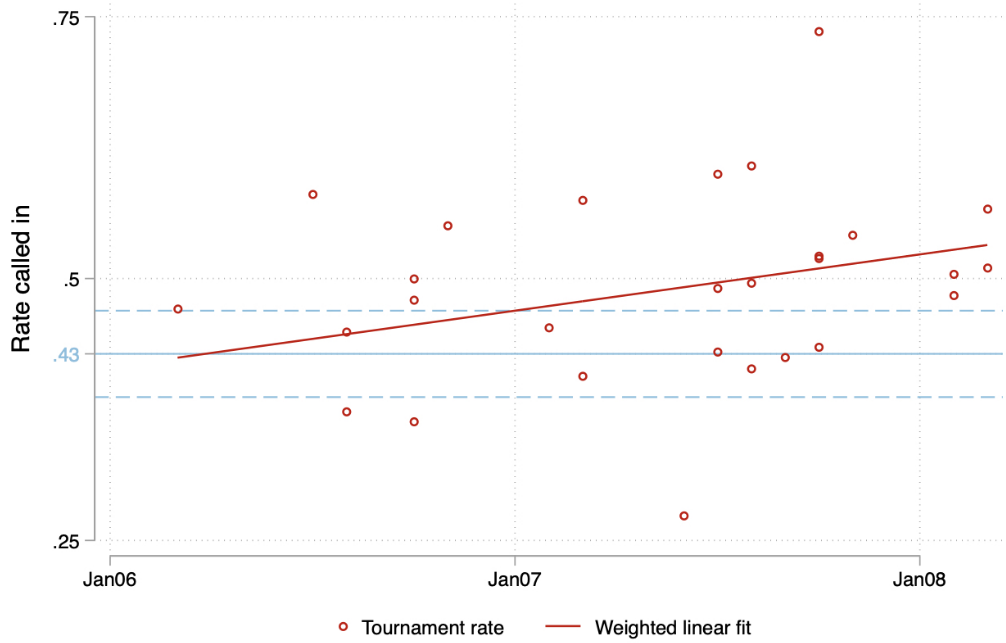


Figure 5: The rate of calling a ball in after the introduction to Hawk-eye review for balls landing <20 mm from the line, regardless of which side of the line they actually bounced on. Each dot represents the rate of calling a ball in for a tournament. The red line is the best linear fit using the dots as observations and weighting them based on the number of calls each tournament contributed. The blue solid line represents the rate of calling a ball in for the 7 tournaments that did not have challenges (the blue dash lines indicate the 95 confidence interval).

5 shows the tournament-specific rates at which umpires called a ball in when it bounced within 20 mm of the line, regardless of which side it bounced. Each of the 28 dots represents the rate at which balls were called in for a particular tournament that allowed challenges, sorted by time. The first observation is that this rate increased from 42.8% to 49% with the introduction of challenges. In 22 out of the 28 tournaments with challenges, umpires called the ball in more frequently than the mean of 42.8% observed in tournaments without challenges. In addition, Figure 5 also suggests a gradual change, with umpires opting to call inside more frequently as time progresses, which could reflect increasing experience with AI oversight. The slope coefficient for the regression in Figure 5 is 0.45 pp per month (p-value=0.11), an effect that would translate to over 5.4 pp in a year.[27]

[27]Figure 11 in the Appendix shows that these results are almost identical when broken down by serves and non-serves.

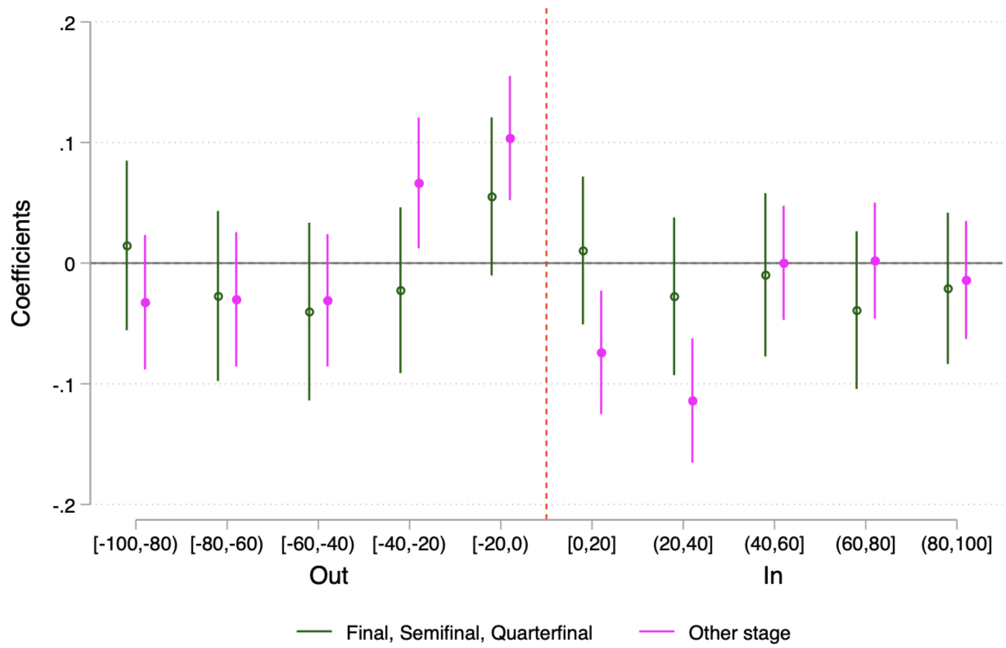


Figure 6: Impact of AI oversight on the incorrect call rate by proximity to the line (by tournament stage). Matches are grouped into those at the final, semifinal, and quarterfinal stage and those at all other (earlier) stages of a tournament, and then analyzed separately. Each dot represents the coefficient on the interaction between distance bin and PostHK, the indicator variable that equals 1 if the Hawk-Eye review system is active.

3.3 Differences by Tournament Stage?

To see if the impact of AI oversight differs by tournament stage, we split our sample into two groups: the first comprising Final, Semifinal, and Quarterfinal matches, and the second containing the remaining matches. In Figure 6, we examine the AI oversight effect separately on the two groups of matches we described. The change in the type of errors in bins closer to the line is more pronounced and only significant for matches in the early stages, as highlighted by the magenta markers. The estimated effects for advanced stage matches (green markers) are not statistically significant. Despite having fewer observations, the estimated coefficients are also smaller in absolute terms in bins around the dashed line.

What could be driving this difference? While we lack specific umpire information, it seems likely that advanced stages feature higher-skill umpires, as they are typically assigned based on performance evaluations. In a study on radiologists, Chan, Gentzkow, and Yu (2022) found that aversion to false negatives tends to be negatively related to radiologist skill. In that context, low-skill radiologists are more prone to falsely diagnose a healthy patient with pneumonia (false positive) because it is less costly than erroneously missing a diagnosis. Applied to our setting, this finding would imply that less-skilled umpires are more likely to shift away from Type II errors (calling a ball in when out).

It is important to note that more advanced matches also involve higher stakes. It is challenging to disentangle whether lower-skilled umpires or lower stakes is what is driving the response in our data, but we can also examine differences in stakes by comparing highly regarded tournaments, such as Master 1000 to lower tournament tiers like 500 and 250.[28] If we were to expect that lower stakes drive the reaction observed in Figure 6, we might expect to see the same in lower-ranked tournaments. Of course, there might also be umpire skill heterogeneity across different tournaments, which would complicate this comparison.

4 Recovering the AI Oversight Penalty

Finally, we structurally estimate the psychological costs of being overruled by AI using a model of rational inattentive umpires.[29] In our model of perceptually-constrained umpires, the set of states is $\omega \in \{\omega^I, \omega^O\}$ for the ball being in (ω^I) or out (ω^O), and the set of actions is $a \in \{a^I, a^O\}$ for calling the ball in (a^I) or out (a^O). The probability of having an incorrect call be challenged is η^I when the ball is in and η^O when the ball is out, and both are equal to zero in tournaments without challenges. When the umpire's call is not challenged, we assume that they receive a normalized utility of 1 when correct and 0 when incorrect. When the umpire's call is challenged, we also assume that they receive a normalized utility of 1 when correct (when their call is upheld). But when incorrect (their call is overturned), we assume that they receive $1 + c^I$ when the ball is in and $1 + c^O$ when it is out. Thus, before taking into account attentional costs, gross expected utility U is

$$U(a,\omega) = \begin{array}{cc} \begin{array}{cc} \omega_I & \omega_O \end{array} & \\ \begin{pmatrix} 1 & \eta^O\left(1 + c^O\right) \\ \eta^I\left(1 + c^I\right) & 1 \end{pmatrix} & \begin{array}{c} a_I \\ a_O \end{array} \end{array}$$

We parameterize the utility of an overturned call in this way so that we can interpret c^I and c^O as the disutility of having been caught making a mistake by the AI system. For shorthand, we refer to these parameters as the *AI oversight penalty* when the ball is in or out. We expect that c^I and c^O are less than or equal to 0 because for the parameters to be negative it would need to be that the umpire is happier having their incorrect call overturned than being correct in the first place. When c^I and c^O are equal to zero (when there is no AI

[28]Figure 12 in the Appendix shows that the discrepant patterns we observed between early and advance-stage matches do not hold when we compare higher and lower-ranked tournaments.

[29]While it could be interesting to consider the dynamic effects of being overruled by AI, we start by considering a static model.

oversight penalty), then the umpire receives a utility of 1 after their call is overturned. This is as if they only care that the correct call is implemented (due to altruism, relief, etc.) and do not not care at all about having been caught making a mistake by the AI system (due to shame, embarrassment, subsequent arguments, etc.). When c^I and c^O are equal to -1, then the umpire receives a utility of 0 after their call is overturned. This is as if any utility gain from having the correct call implemented is perfectly offset by the AI oversight penalty. When c^I and c^O are less than -1, the AI oversight penalty dominates any utility gain from having the correct call implemented.

As is standard in models of attention, we assume that the umpire starts out with prior belief μ, where $\mu(\omega)$ is the probability of state $\omega \in \{\omega^I, \omega^O\}$. After receiving a noisy mental signal of whether the ball is in or out, the umpire forms posterior $\gamma \in \Gamma$, where $\gamma(\omega)$ is the probability of state $\omega \in \{\omega^I, \omega^O\}$. Given a posterior belief γ, the umpire decides what call to make (mixing between actions is allowed), and we assume the umpire maximizes expected utility when making a choice of how often to take each action at that posterior.

Under the assumption that the umpire updates their prior belief using Bayes rule, their attention can be represented by a Bayes-consistent information structure π, which stochastically generates posterior beliefs. Specifically, π is a function that maps the state into $\Delta(\Gamma)$, the set of probability distributions over Γ that have finite support, so that $\pi : \omega \to \Delta(\Gamma)$. Let Π denote the set of all such functions, $\pi(\gamma)$ be the unconditional probability of posterior $\gamma \in \Gamma$, $\pi(\gamma|\omega)$ be the probability of posterior γ given state ω, and $\Gamma(\pi) \subset \Gamma$ denote the support of a given π. We limit the set of information structures to those in $\Pi(\mu) \subset \Pi$ that generate correct posteriors for a given prior belief μ, so that

$$\Pi(\mu) = \left\{ \pi \in \Pi | \forall \gamma \in \Gamma(\pi), \forall \omega \in \Omega, \gamma(\omega) = \frac{\mu(\omega)\pi(\gamma|\omega)}{\sum\limits_{\omega \in \Omega} \mu(\omega)\pi(\gamma|\omega)} \right\} \tag{2}$$

4.1 Initial Bounds on the AI Oversight Penalty

The results of Caplin and Martin (2015) can be applied to show that umpire behavior is consistent with Bayesian expected utility maximization if and only if their choices satisfy the *No Improving Action Switches (NIAS)* condition. The key object of analysis is state-dependent stochastic choice data $P(a, \omega)$, which is the joint probability of action a and state ω. In this setting, $P(a, \omega)$ is the joint probability of making call a (calling in or out) and the ball being in state ω (being in or out).

The NIAS condition requires there should be no utility gain from wholesale action switches. That is, expected utility should not go up if there is either: (1) a wholesale switch to calling in every time that a ball was called out or (2) a wholesale switch to calling out every time that a ball was called in. Formally, this condition is

$$P(a^I,\omega^I)\left(1-\eta^I\left(1+c^I\right)\right) \geq P(a^I,\omega^O)\left(1-\eta^O\left(1+c^O\right)\right) \tag{3}$$
$$P(a^O,\omega^O)\left(1-\eta^O\left(1+c^O\right)\right) \geq P(a^O,\omega^I)\left(1-\eta^I\left(1+c^I\right)\right) \tag{4}$$

When there are no challenges, η^I and η^O are equal to zero, so this condition reduces to a very simple set of inequalities that are easily satisfied in our data. Denoting $P^N(a,\omega)$ as the state-dependent stochastic choice data when no challenges are available, the NIAS condition is simply

$$P^N(a^I,\omega^I) \geq P^N(a^I,\omega^O) \tag{5}$$
$$P^N(a^O,\omega^O) \geq P^N(a^O,\omega^I) \tag{6}$$

When there are challenges, the NIAS condition contains two unobservables, the AI oversight penalty when the ball is in (c^I) and when the ball is out (c^O). In fact, NIAS can be used to produce bounds on one unobservable in terms of the other. Denoting $P^C(a,\omega)$ as the state-dependent stochastic choice data when challenges are available, the bounds on c^I generated by NIAS are

$$c^I \leq \frac{P^C(a^I,\omega^I)(1-\eta^I)-P^C(a^I,\omega^O)(1-\eta^O)}{P^C(a^I,\omega^I)\eta^I}+\frac{P^C(a^I,\omega^O)\eta^O}{P^C(a^I,\omega^I)\eta^I}c^O \tag{7}$$
$$c^I \geq \frac{P^C(a^O,\omega^I)(1-\eta^I)-P^C(a^O,\omega^O)(1-\eta^O)}{P^C(a^O,\omega^I)\eta^I}+\frac{P^C(a^O,\omega^O)\eta^O}{P^C(a^O,\omega^I)\eta^I}c^O \tag{8}$$

To estimate these bounds from umpire decisions under AI oversight, we use both our estimate of *true* probability of each action and state (P^C), which is the *observed* probability of each action and state, and our estimates of the *true* challenge rates (η^I and η^O), which are the *observed* challenge rates. For calls within 20 mm of the line, the estimated bound is

$$0.6207+0.5012^O \geq c^I \geq -1.2115+1.7744c^O \tag{9}$$

If $c^O=-1$, which is when the benefits of a call being overturned are in balance with the AI oversight penalty, we get

$$0.1195 \geq c^I \geq -2.9859 \tag{10}$$

Next, we assume information structures are chosen and each information structure $\pi \in \Pi(\mu)$ has an additively-separable cost $K(\pi)$ in expected utility units. Together with NIAS, the *No Improving Attention Cycles (NIAC)* condition of Caplin and Dean (2015) characterizes this general rational inattention model, which puts no assumptions on the cost of information structures beyond being additively-separable. NIAC sharpens the bounds above by providing an additional condition on the disutility from having been caught making a mistake by the AI system when the initial call was out c^I that is given by

$$c^I \leq \frac{(P^C(a^O,\omega^I) - P^N(a^O,\omega^I))\eta^I + (P^C(a^I,\omega^O) - P^N(a^I,\omega^O))\eta^O}{(P^N(a^O,\omega^I) - P^C(a^O,\omega^I))\eta^I} \tag{11}$$

$$+\frac{(P^C(a^I,\omega^O) - P^N(a^I,\omega^O))\eta^O}{(P^N(a^O,\omega^I) - P^C(a^O,\omega^I))\eta^I} c^O \tag{12}$$

For calls within 20 mm of the line, our estimate of this bound is

$$c^I \geq 1.0105 + 2.0105 c^O \tag{13}$$

Also, whenever $c^O = -1$, NIAC gives that $-1 \geq c^I$. Combined with the NIAS restrictions, we get that $-1 \geq c^I \geq -2.9859$.

4.2 Uniquely Identifying the AI Oversight Penalty

To sharpen identification of these parameters, we assume the cost of information structures scale linearly with the Shannon mutual information between posteriors and the prior (e.g., Matejka and McKay 2015), which we refer to as the "Shannon model." Formally, K is determined by the function

$$K(\pi,\kappa,\mu) = \kappa \left(\left[\sum_{\gamma\in\Gamma(\pi)} \pi(\gamma) \sum_{\omega\in\Omega} [\gamma(\omega)\ln(\gamma(\omega))] \right] - \sum_{\omega\in\Omega} [\mu(\omega)\ln(\mu(\omega))] \right) \tag{14}$$

where $\kappa \in \mathbb{R}_{++}$ is a linear cost parameter that is interpreted as the marginal cost of attention. We expand on the standard Shannon model by allowing for different costs of attention for different states,[30] so that

$$K(\pi,\kappa,\mu) = \kappa^I \left(\left[\sum_{\gamma\in\Gamma(\pi)} \pi(\gamma)\gamma(\omega^I)\ln(\gamma(\omega^I)) \right] - \mu(\omega^I)\ln(\mu(\omega^I) \right) \tag{15}$$

$$+\kappa^O \left(\left[\sum_{\gamma\in\Gamma(\pi)} \pi(\gamma)\gamma(\omega^O)\ln(\gamma(\omega^O)) \right] - \mu(\omega^O)\ln(\mu(\omega^O) \right) \tag{16}$$

where $\kappa^I, \kappa^O \in \mathbb{R}_{++}$.

With the Shannon model, the optimal information structure has one posterior at which each is taken. We denote these posteriors as γ^I when call in and γ^O when call out. By the

[30] Whitney, Wurnitsch, Hontiveros, and Louie (2008) use data from Wimbledon 2007 to show that umpires call more tennis balls as being out (when actually in) than in (when actually out). Their results are consistent with psychometric experiments that document how moving objects generate a perceptual bias and are perceived as being shifted in the direction of their motion. Extending the Shannon model by allowing different costs of attention for different states would help to accommodate these types of biases more broadly.

Invariant Likelihood Ratio (ILR) of Caplin and Dean (2013), these optimal posteriors must obey

$$\frac{\gamma^I(\omega^I)}{\gamma^O(\omega^I)} = e^{\frac{U\left(a^I,\omega^I\right)-U\left(a^O,\omega^I\right)}{\kappa^I}} \tag{17}$$

$$\frac{\gamma^O(\omega^O)}{\gamma^I(\omega^O)} = \frac{1-\gamma^O(\omega^I)}{1-\gamma^I(\omega^I)} = e^{\frac{U\left(a^O,\omega^O\right)-U\left(a^I,\omega^O\right)}{\kappa^O}} \tag{18}$$

For matches where there were are no challenges, the ILR condition, given by (17) and (18), allow us to express the marginal costs of attention κ^I and κ^O cleanly as

$$\kappa^I = \frac{1}{\ln\gamma^I(\omega^I) - \ln\gamma^O(\omega^I)} \tag{19}$$

$$\kappa^O = \frac{1}{\ln\gamma^O(\omega^O) - \ln\gamma^I(\omega^O)} \tag{20}$$

When there are challenges, (17) and (18) allow us to solve for the AI oversight penalty, which is the umpire's disutility from having been caught making a mistake by the AI system. Those values are given by

$$c^I = \frac{1-\kappa^I\left(\ln\gamma^I(\omega^I) - \ln\gamma^O(\omega^I)\right)}{\eta^I} - 1 \tag{21}$$

$$c^O = \frac{1-\kappa^O\left(\ln\gamma^O(\omega^O) - \ln\gamma^I(\omega^O)\right)}{\eta^O} - 1 \tag{22}$$

4.3 Structural Estimates

To structurally estimate the AI oversight penalty, we undertake two steps. First, we estimate the marginal costs of attention when there are no challenges. When there are no challenges, these marginal costs are fully determined by the optimal posteriors $\gamma^I(\omega^I)$ and $\gamma^O(\omega^O)$. Because there is a single posterior for each action in our model, the optimal posteriors are equal to the *revealed posterior*, which is $P(\omega|a)$, the probability of each state conditional on the action taken. Our estimate of the *true* revealed posterior is the *observed* probability of each state conditional on the action taken.

As shown in Table 4, our estimate of the revealed posterior for the ball being in when calling in is 0.599 and our estimate of revealed posterior for the ball being out when calling out is 0.751 for all calls within 20 mm of the line before the introduction of AI oversight. Based on (19) and (20), the marginal costs of attention that rationalize these values are $\kappa^I = 2.492$ and $\kappa^O = 0.906$.

As a second step, we use these estimated costs of attention and the observed challenge rates (our estimates of the true challenge rates η^I and η^O) to estimate the AI oversight

Parameter	Recovery equation	<20 mm	20-100 mm
$\gamma^I(\omega^I)$	N/A	0.599	0.912
$\gamma^O(\omega^O)$	N/A	0.751	0.913
κ^I	19	2.492	0.428
κ^O	20	0.906	0.425

Table 4: Estimated optimal posteriors and costs of attention before the introduction of AI oversight.

penalties. As shown in Table 5, for calls within 20 mm of the line, the observed challenge rates when balls were in is 0.427 and when balls were out is 0.410. Given the values of κ^I and κ^O determined without challenges, the rationalizing AI oversight penalties are $c^I = -2.003$ and $c^O = -0.088$. It is important to highlight that the AI penalty in our case can be seen as a lower bound to the costs of shaming. Under the scenario that umpires do not care at all about the final outcome being correctly implemented, then the cost of shaming equals the AI oversight penalty, but otherwise, the shaming effect is bigger. These results are consistent with anecdotal evidence regarding the increased controversy surrounding this type of error.

Parameter	Recovery equation	<20 mm	20-100 mm
η^I	N/A	0.427	0.479
η^O	N/A	0.410	0.421
c^I	21	-2.003	-1.337
c^O	22	-0.088	-1.079

Table 5: Estimated challenge rates, AI oversight penalties, and mistake utilities.

These estimates suggest that the introduction of AI oversight left the utility of Type I errors $(1 + c^O)$ increases sharply, as there is an increase from 0 before Hawk-Eye review to 0.912 after. This result implies that the umpire cares much more about ensuring the correct call is implemented than being caught making a mistake by the AI system (the AI oversight penalty). On the other hand, the utility of Type II errors $(1 + c^1)$ reduces sharply, from 0 before Hawk-Eye review to -1.003 after. Here the AI oversight penalty dominates any utility boost from having the correct call implemented.

These changes are hard to interpret in fractional terms, so we consider a re-normalization of utility so that the utility of being correct is 0 and the utility of being incorrect is -1. Under this re-normalization, the first stage estimates of the marginal cost of attention do not change, and the second stage estimates of the AI penalty decrease by 1 to $c^I = -2.003$ and $c^O = -0.088$, which means that the (dis)utility of Type II errors (calling a ball out when in) decreases from -1 to -2.003. Thus, these estimates suggest that umpires care twice as much about Type II errors after the introduction of Hawk-Eye review.

5 Conclusion

We document that the introduction of AI oversight in professional tennis corresponded with an improvement in umpires' overall performance. However, when we zoom in on the closest calls, we find evidence that performance in serves deteriorated. We can view the calls on these serves as a cautionary tale illustrating the dangers of distorting the cost of errors, especially in a context or task where the cost of improved performance is already very steep. Reacting to a cost change of this kind may increase the overall number of incorrect calls, which could have negative welfare implications, as it introduces a misalignment of incentives between the decision-maker (umpire) and the other parties involved (ATP, audience, players, etc.).

We view our paper as an initial building block for understanding the implications of AI oversight on humans. Future research could be conducted in experimental settings to complement our findings, gaining more control over AI and human costs, and delving deeper into the underlying mechanisms.

References

Abramitzky, Ran, Liran Einav, Shimon Kolkowitz, and Roy Mill (2012). "On the Optimality of Line Call Challenges in Professional Tennis". *International Economic Review* 53.3, pp. 939–964.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz (2023). "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology". *Working Paper 31422, National Bureau of Economic Research.*

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2022). *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*. Harvard Business Press.

Almog, David and Daniel Martin (2023). "Rational Inattention in Games: Experimental Evidence". *Working paper.*

Angelova, Victoria, Will Dobbie, and Crystal Yang (2023). "Algorithmic Recommendations and Human Discretion". *Working paper.*

Bhattacharya, Vivek and Greg Howard (2022). "Rational Inattention in the Infield". *American Economic Journal: Microeconomics* 14.4, pp. 348–93.

Bobonis, Gustavo J, Luis R. Cámara Fuertes, and Rainer Schwabe (2016). "Monitoring Corruptible Politicians". *American Economic Review* 106.8, pp. 2371–2405.

Bolte, Lukas and Collin Raymond (2023). "Emotional Inattention". *Working paper.*

Bursztyn, Leonardo and Robert Jensen (2017). "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure". *Annual Review of Economics* 9.1, pp. 131–53.

Butera, Luigi, Robert D. Metcalfe, William Morrison, and Dmitry Taubinsky (2022). "Measuring the Welfare Effects of Shame and Pride". *American Economic Review* 112.1, pp. 122–168.

Camerer, Colin F. (2019). "Artificial Intelligence and Behavioral Economics". *In The Economics of Artificial Intelligence: An Agenda. , ed. Ajay Agrawal, Joshua Gans and Avi Goldfarb*, pp. 587–608.

Caplin, Andrew (2023). *Science Of Mistakes, The: Lecture Notes On Economic Data Engineering*. Vol. 16. World Scientific.

Caplin, Andrew and Mark Dean (2013). *Behavioral implications of rational inattention with shannon entropy*. Tech. rep. National Bureau of Economic Research.

— (2015). "Revealed Preference, Rational Inattention, and Costly Information Acquisition". *American Economic Review* 105.7, pp. 2183–2203.

Caplin, Andrew and Daniel Martin (2015). "A Testable Theory of Imperfect Perception". *The Economic Journal* 125.February, pp. 184–202.

Chalfin, Aaron, Oren Danieli, Aandrew Hillis, Zubin Jelveh, Michael Luca, Ludwig Jens, and Sendhil Mullainathan (2016). "Productivity and Selection of Human Capital with

Machine Learning". *American Economic Review: Papers & Proceedings* 106.5, pp. 124–127.

Chan, David C., Matthew Gentzkow, and Chuan Yu (2022). "Selection with Variation in Diagnostic Skill: Evidence from Radiologists". *The Quarterly Journal of Economics* 137.2, pp. 729–783.

Ely, Jeffrey, Romain Gauriot, and Lionel Page (2017). "Do Agents Maximise? Risk Taking on First and Second Serves in Tennis". *Journal of Economic Psychology* 53, pp. 135–142.

Engelmann, Jan B., Maël Lebreton, Nahuel A. Salem-Garcia, Peter Schwardmann, and Joël J. van der Weele (2023). "Anticipatory Anxiety and Wishful Thinking". *American Economic Review* Forthcoming.

Gauriot, Romain and Lionel Page (2019). "Does Success Breed Success? a Quasi-Experiment on Strategic Momentum in Dynamic Contests". *The Economic Journal* 129.624, pp. 3107–3136.

Gauriot, Romain, Lionel Page, and John Wooders (2023). "Expertise, Gender, and Equilibrium Play". *Quantitative Economics* 14.3, pp. 981–1020.

Gosnell, Greer K., John A. List, and Robert D. Metcalfe (2020). "The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains". *Journal of Political Economy* 128.4, pp. 1195–1626.

Gray, Wayne B. and Jay P. Shimshack (2011). "The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence". *Review of Environmental Economics and Policy* 5.1, pp. 3–24.

Hoffman, Mitch, Lisa Kahn, and Danielle Li (2018). "Discretion in Hiring". *The Quarterly Journal of Economics* 133.2, pp. 65–800.

Iakovlev, Andrei and Annie Liang (2023). "The Value of Context: Human versus Black Box Evaluators". *Working paper.*

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018). "Human Decisions and Machine Predictions". *The Quarterly Journal of Economics* 133.1, pp. 237–293.

Maćkowiak, Bartosz, Filip Matějka, and Mirko Wiederholt (2023). "Rational inattention: A review". *Journal of Economic Literature* 61.1, pp. 226–273.

Matějka, Filip and Alisdair McKay (2015). "Rational inattention to discrete choices: A new foundation for the multinomial logit model". *American Economic Review* 105.1, pp. 272–298.

Nagin, Daniel S., James B. Rebitzer, Seth Sanders, and Lowell J. Taylor (2002). "Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment". *American Economic Review* 92.4, pp. 850–873.

Nielsen, Kirby and John Rehbeck (2022). "When Choices Are Mistakes". *American Economic Review* 112.7, pp. 2237–68.

Olken, Benjamin A. (2007). “Monitoring Corruption: Evidence from a Field Experiment in Indonesia”. *Journal of Political Economy* 115.2, pp. 200–249.

Palacios-Huerta, Ignacio (2023). “The Beautiful Dataset”. *Working paper*.

Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan (2019). “The Algorithmic Automation Problem: Prediction, Triage, and Human Effort”. *Working paper*.

Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng (2017). “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”. *Working paper*.

Rambachan, Ashesh (2022). “Identifying Prediction Mistakes in Observational Data”. *Working paper*.

Topol, Eric (2019). “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”. *Nature Medicine* 25, pp. 44–56.

Whitney, David, Nicole Wurnitsch, Byron Hontiveros, and Elizabeth Louie (2008). “Perceptual Mislocalization of Bouncing Balls by Professional Tennis Referees”. *Current Biology* 18, R947–9.

Zou, Eric Yongchen (2021). “Unwatched Pollution: The Effect of Intermittent Monitoring on Air Quality”. *American Economic Review* 111.7, pp. 2101–2126.

A Data Management

A.1 Exclusion of Clay Tournaments

The officiating dynamics in clay tournaments are different. Players are allowed to request the chair umpire to review the ball bounce marks on the surface, and changing the call based on that is permitted. The clay surface presents some particular difficulties; as the surface is continuously changing during the game, the accuracy of Hawk-Eye decreases and requires constant re-calibration to operate at its best.[31]

Tournaments on clay surfaces rejected the idea of incorporating Hawk-Eye challenges. They argued that they already had a mechanism in place to review incorrect calls (ball marks) and expressed reluctance, citing concerns that Hawk-Eye might not be precise enough on this surface. Another reason for rejecting the use of Hawk-Eye on clay is to avoid controversies where the ball's mark and Hawk-Eye might disagree. As noted in The Guardian, Lars Graf, one of the ATP's most experienced officials, explained the decision not to use Hawk-Eye on clay: "We decided not to use Hawk-Eye on clay because it might not agree with the mark the umpire is pointing at".[32]

At first glance, clay tournaments may appear as an ideal control group for comparing umpire reactions in tournaments with Hawk-Eye challenges versus those on clay. However, several complications arise, leading us to the decision of excluding clay tournaments. Firstly, before the introduction of Hawk-Eye, umpires in clay tournaments had different incentives. They could delegate some responsibility by calling more bounces as *in* and allowing the receiving player to request a mark review if there was evidence suggesting it was *out*. Since players cannot cross to the other side of the court, the receiving player has a direct view of the ball mark, making it easier for umpires to delegate in that direction. Secondly, although clay tournaments did not allow player challenges, the camera system was installed, and Hawk-Eye predictions were shown on TV broadcasts. This could have an effect on umpires, but it should differ from corrections made in the stadium. A third reason for avoiding the use of clay data is our inability to determine from our dataset which calls were corrected after examining the clay marks.

In summary, we excluded clay tournaments due to the unique incentive scheme based on ball marks, the differing treatment they received, and the complication in observability they present.

[31] https://www.perfect-tennis.com/why-there-is-no-hawk-eye-on-clay/

[32] https://www.theguardian.com/lifeandstyle/2009/jun/27/tennis-hawk-eye

A.2 Merging Algorithm

Our merging algorithm was designed to identify the highest number of the 144 challenges (derived from the 43 matches with video replays) within the Hawk-Eye Base data set, while keeping the number of false positives (challenges matched into the wrong point) as low as possible. We used standard variables, such as set, game, score, distance difference, player hitting the ball, and whether it was a tie-break, in order to minimize over-fitting.

Our final merging algorithm comprises eight iterations, where we start using stricter criteria and gradually relax them in subsequent iterations. Some iterations, like 5 and 6 focus on very specific situations of the game (tie-breaks), and while in the testing sample do not seem useful, they might prove useful in a bigger sample.

The algorithm merged 143 out of the 144 challenges, achieving a merge rate of 99.3%. Out of these, 136 were merged to the correct point, as validated through video auditing replays, resulting in an accuracy rate of 94.4%. It would be hard to improve this efficacy as there is many challenges with missing variables like set and game. We will now elaborate on the specific criteria used in each of the eight iterations, followed by Table 6 that summarizes the number of challenges successfully merged (and false positives) achieved in each iteration.

Merging criteria for each iteration:

Iteration 1. Same set, game, score, and player (w/ distance difference <35 mm)

Iteration 2. Same set, game, and player (w/ distance difference <15 mm)

Iteration 3. Same set, score, and player (w/ distance difference <15 mm)

Iteration 4. Same game, score, and player (w/ distance difference <10 mm)

Iteration 5. For tie-breaks: Same game and score (w/ distance difference <35 mm)

Iteration 6. For tie-breaks: Same game. If multiple, pick closest in distance (w/ distance difference <15 mm)

Iteration 7. Same set, game and player. If multiple, pick closest point in score and then in distance (w/ distance difference <35 mm)

Iteration 8. Same set and player. If multiple, pick closest point in distance (w/ distance difference <35 mm)

Iteration	Correct Merge	False Positives	Efficacy	Challenges Left
1	98	2	98%	44
2	25	2	93%	17
3	5	0	100%	12
4	1	0	100%	11
5	0	0	–	11
6	0	0	–	11
7	5	0	100%	6
8	2	3	40%	1

Table 6: Performance for each iteration of the merging algorithm.

A.3 Identification of Incorrect Calls

The Hawk-Eye Base data set contains the precise location of every ball bounce and identifies the winner of each point. However, it lacks direct information on the umpires' calls. Therefore, we are tasked with identifying the original umpire calls, which we do through the application of four criteria. The first two determine points where the umpire incorrectly called something in, and the final two identify the other type of incorrect call (umpire calling something incorrectly out). In the period involving challenges, we first restore the umpire's original calls following a successful challenge. Then, we assess whether any of the four criteria are satisfied. We are confident that these four criteria comprehensively identify potential mistakes in our dataset.

Criterion 1. A player stroke is out and wins the point

We selected strokes where the first player to hit a ball out in a given point also won the point. We identified 55 points that satisfy this criterion, and 54 of them were confirmed by the video replays to be mistakes, resulting in a 98% accuracy rate.

Criterion 2. The point continued after an out

To complement Criterion 1, we must also identify points where a player hit a ball out, and despite the error, the point continued, but the same player ultimately lost the point. To achieve this, we examine instances where, after a player hits the ball out, there are at least three more strokes registered. Out of the 25 points satisfying this criterion, 24 were umpire mistakes, yielding an accuracy rate of 96%.

We also tested a more relaxed criterion that identifies points recording at least 2 more strokes after an out. However, this criterion proved to be too lenient, encompassing multiple cases where the point was correctly called, and players continued hitting the ball afterward. When the criterion is relaxed to require at least 2 more strokes, the accuracy rate drops to

52%. This adjustment results in the identification of one new mistake at the expense of 24 incorrectly identified instances. This is why we adhere to the 3+ rule.

Criterion 3. All the strokes of a point are in, and the player hitting last loses

We identify strokes where the last player to hit the ball in, did not win the point. Out of the 25 points identified in the auditing matches, 23 were confirmed to be mistakes, resulting in a 92% accuracy rate.

Criterion 4. Observing a second serve after the first serve was in

This criterion helps identify two types of instances where there is no winner in a point as direct consequence from an umpire mistake. First, it recognizes those first serves that bounced in but were incorrectly ruled out, resulting in no winner for the point, which then moves to a second serve. Secondly, it detects points that lasted multiple strokes but had to be replayed because the umpire stopped the point by incorrectly calling a ball out. The accuracy rate for this criterion is 86%, with 36 of the 43 selected points identified as umpire mistakes. Some inaccurately selected points by this criterion include lets (serves that hit the top of the net and the point is replayed) and other unusual events (e.g., a server touching the line while serving, and the serve not counting). The occurrence and our identification of these events should not change with the introduction of Hawkeye

For this criterion, we limit the analysis to strokes within 40mm of the line, as beyond this threshold, the criterion becomes widely inaccurate. For balls bouncing between 40-100mm away from the line, this criterion has a 30% accuracy rate, as only 7 of the 23 points audited were actual umpire mistakes.

B Additional Tables and Figures

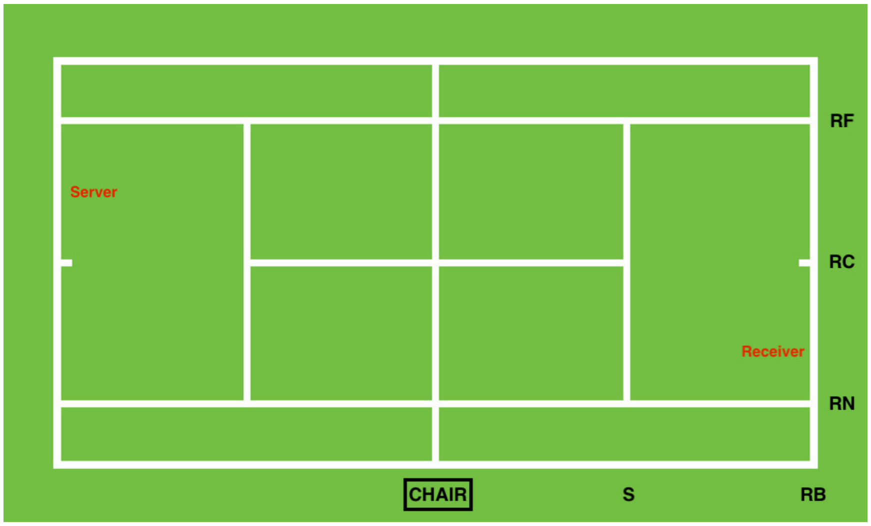


Figure 7: The location of the umpiring crew: S = the serve-line umpire, RB = the right baseline umpire, RN = the right near long-line umpire, RC = the right center-line umpire, RF = the right far-side long-line umpire. The left side of the court corresponding line umpires, except for the serve-line umpire who moves to the left side when the right side player has the serve.

Tournament	Start Month	Category	Court Type	Number of Games (Pre HK & Post HK)			
				2005	2006	2007	2008
Marseille	Feb	International (250)	Hard				25
Rotterdam	Feb	International Gold (500)	Hard			15	26
Dubai	Mar	International Gold (500)	Hard				19
LasVegas	Mar	International (250)	Hard				19
IndianWells	Mar	Masters (1000)	Hard	17		27	
Miami	Mar	Masters (1000)	Hard	15	18	26	
Queens	Jun	International (250)	Grass	13	19	18	
UCLA	Jul	International (250)	Hard			23	
Indianapolis	Jul	International (250)	Hard		7	23	
Washington	Jul	International (250)	Hard			21	
Montreal	Aug	Masters (1000)	Hard	20		27	
Toronto	Aug	Masters (1000)	Hard		25		
Cincinnati	Aug	Masters (1000)	Hard	11	19	24	
NewHaven	Aug	International (250)	Hard			18	
Beijing	Sept	International (250)	Hard			20	
KremlinCup	Oct	International (250)	Carpet06 Hard07		8	15	
Madrid	Oct	Masters (1000)	Hard		30	30	
Basel	Oct	International (250)	Hard			12	
Paris Masters	Oct	Masters (1000)	Carpet06 Hard07		31	33	
Shangahi	Nov	Masters (1000)	Carpet05 Hard06-07	14	15	15	

Figure 8: Information on the 35 tournaments in the consolidated dataset. The numbers in each cell indicate the matches available in our dataset. The color represents the group of tournaments, with blue indicating Pre-HK and red indicating Post-HK.

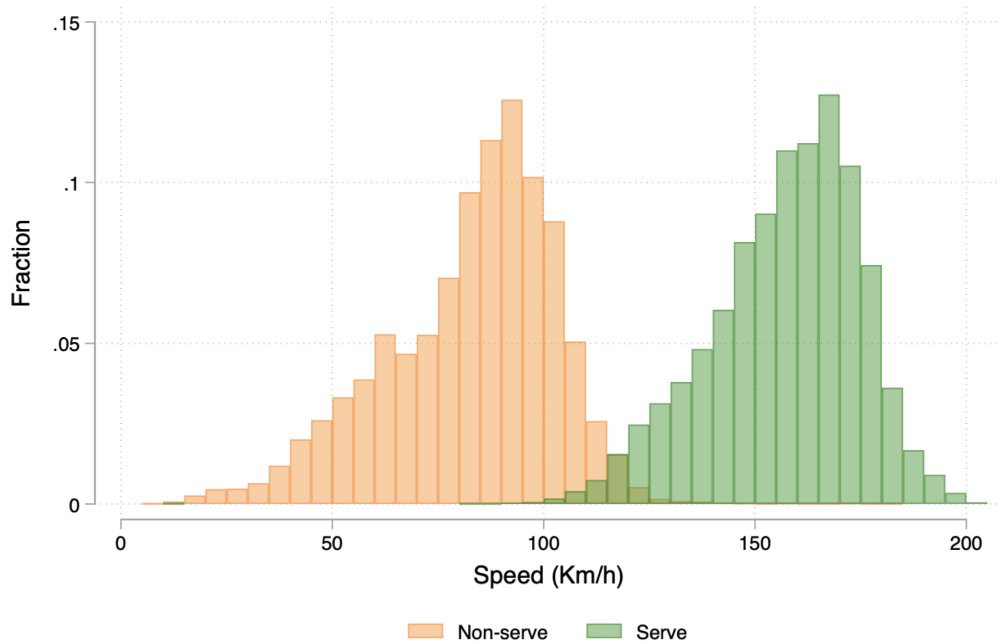
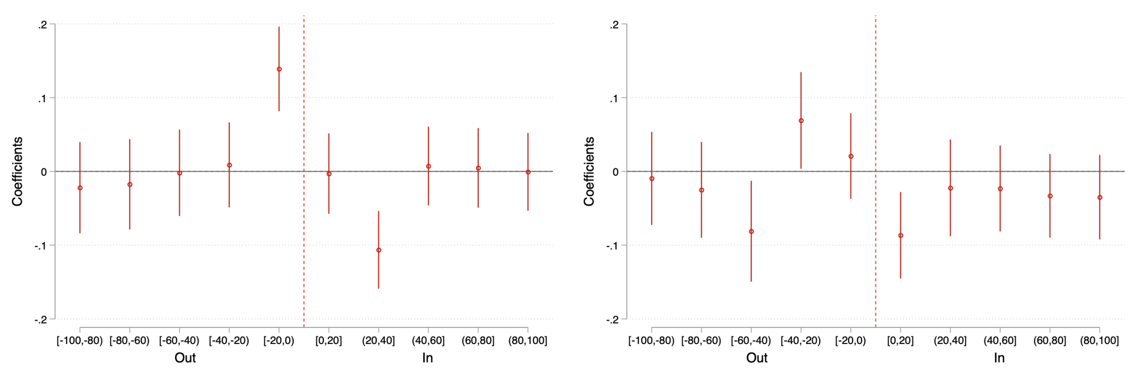


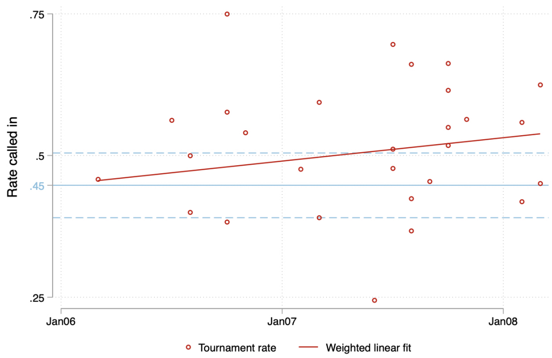
Figure 9: Speed distribution for all the balls that bounced within 100 mm of the line, separated by non-serves and serves.



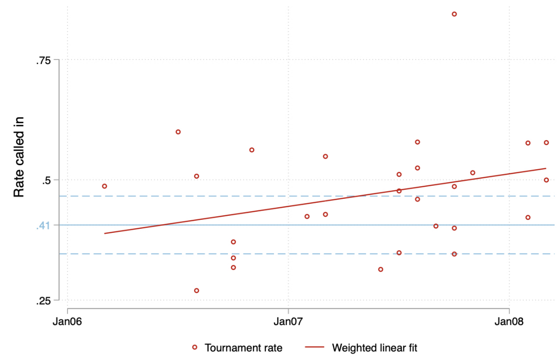
(a) Serves.

(b) Non-serves.

Figure 10: An analog of Figure 3, separating serves and non-serves.



(a) Serves.



(b) Non-serves.

Figure 11: An analog of Figure 5, separating serves and non-serves.

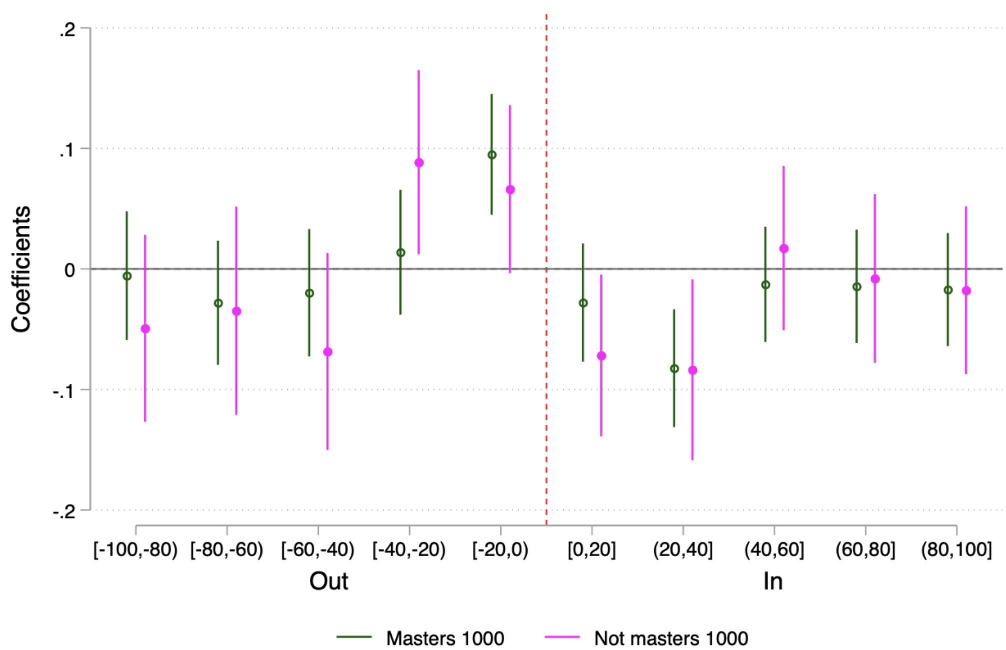


Figure 12: Heterogeneous Impact of AI Oversight on Umpires' Performance Across 20 mm Distance Bins by Tournament Category. We conducted Other separate analyses, categorizing matches from higher ranked tournaments (Masters 1000) vs lower ranked (International 500 and 250).