Human Responses to AI Oversight: Evidence from Centre Court*

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

March 4, 2025

Abstract

Powered by the increasing predictive capabilities of machine learning algorithms, artificial intelligence (AI) systems have the potential to overrule human mistakes in many settings. We provide the first field evidence that the use of AI oversight can impact human decision-making. We investigate one of the highest visibility settings where AI oversight has occurred: 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, but also that umpires increased the rate at which they called balls in, producing 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 attention-constrained umpires, and our results suggest that because of these costs, umpires cared 37% more about Type II errors under AI oversight.

*We thank Lucas Lippman, Anastasiia Morozova, and Leshan Xu for their excellent research assistance. We are very grateful to the executives at Hawk-Eye Innovations Ltd and a leading tennis official at the time of Hawk-Eye's introduction for providing valuable institutional information. We also thank Ashesh Rambachan, Mitchell Hoffman, Colin Camerer, Antonio Rangel, Devin Pope, John List, Paul Raschky, Christopher Mills, Luis Rayo, Michael Powell, George Georgiadis, Alex Imas, Meghan Busse, Yuval Salant, Jörg Spenkuch, Ben Enke, Raphaël Raux, David Argente, Andrew Wooders, Andrew Caplin, and audiences at EC 2024, BDRM, AI and the Future of Work at Wharton, ESIF Economics and AI+ML Meeting, Machine Learning in Economics Summer Conference, Zurich Workshop in AI+Economics, WEBEAS, Monash, UT Dallas, Kellogg Strategy Brown Bag Seminar, and Booth Behavioral Lab for many 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, danielmartin@ucsb.edu.

1 Introduction

In the coming years, firms will have the option to use Artificial intelligence (AI) systems to correct worker mistakes across a wide array of settings. This is due to the confluence of two forces. First, machine learning algorithms are becoming increasingly good at predicting human mistakes. This includes important decisions such as bail judges predicting pretrial misconduct (Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan 2018), radiologists predicting pneumonia from chest X-rays (Rajpurkar, Irvin, Zhu, Yang, Mehta, Duan, Ding, Bagul, Ball, Langlotz, et al. 2017; Topol 2019), and workforce professionals predicting productivity for hiring and promotion (Chalfin, Danieli, Hillis, Jelveh, Luca, Jens, and Mullainathan 2016). Second, there has been a substantial drop in the cost of monitoring worker behavior, brought about by the rise in digitization.[1]

For a concrete example of the potential for AI oversight, consider Zalando, a leading e-commerce company specializing in fashion retail. Zalando allows category managers to suggest discounts based on their private information about fashion trends. Huelden, Jascisens, Roemheld, and Werner (2024) show that while human interventions appear promising, the negative impact of poor interventions (stemming from overconfidence and other biases) offsets the benefits of good interventions (based on private information). However, they show that using algorithmic tools to predict and block undesirable human interventions has the potential to recoup these gains. 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 (Rambachan 2024) and medicine (Raghu, Blumer, Corrado, Kleinberg, Obermeyer, and Mullainathan 2019).

Using AI to overrule human errors appears to be a straightforward improvement for societal welfare. 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 (e.g., Hoffman, Kahn, and D. Li 2018, Grimon and Mills 2025, L. Raymond 2024), 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.[2]

However, determining whether AI oversight impacts human decision-making is challenging for several reasons. First, it is necessary to have observations on decision-making both

[1] The strong impact of these complementary forces – digitization and AI – was demonstrated by L. Raymond (2024) in the real estate market.

[2] This illustrates how insights from behavioral economics can play a key role in understanding AI-human interactions (Camerer 2019).

with and without AI oversight, but firms may be resistant to share such data given the legal and public relations risks involved. Second, to determine the types of errors people make, we need a clear "ground truth" (an objectively correct answer). In some cases, the best thing you can do is aggregating many expert opinions, such as in diagnostic radiology (Agarwal, Moehring, Rajpurkar, and Salz 2023). Third, to get a full picture of the types of errors people make, we need to observe or estimate counterfactual outcomes, as not to run afoul of incomplete data due to "selective labels" (Rambachan 2024). Fourth, it is necessary to control or otherwise account for the private information of decision-makers, as this may introduce confounds in analysis. Fifth, we need a setting where AI is very good at replacing human mistakes, because AI oversight is only likely to be implemented in such cases. Additionally, if we are interested in studying behavioral responses related to being observed, we need a setting where there is an audience that the decision-maker would care about.

Given these challenges, we identified the introduction of AI oversight in umpiring at top tennis tournaments as an ideal field setting for studying the humans response to AI oversight. AI oversight was introduced in top tennis tournaments through an AI technology known as Hawk-Eye. Due to the very small distances involved, camera imaging alone is not sufficient for determining if a ball lands in or out of bounds, so Hawk-Eye leverages state-of-the-art machine learning algorithms to use the path and spin of the ball to predict if a ball landed in or out.[3] Starting in 2006, players were given the option to challenge a line call, and if the call contradicted the highly accurate predictions of Hawk-Eye, the umpire's decision was overturned, or in some cases, the point was replayed.

While this is a very specific case of AI oversight, it adheres to the external validity (EV) "Litmus Test" proposed by List (2020), as it is very hard to find better setting for studying whether AI oversight impacts human decision-making.[4] Our setting differs from others because AI is nearly perfect, but this unique feature is very advantageous for studying the short-term impacts of AI oversight because the strong performance of AI allows us to sidestep common problems in studying human and AI interaction: uncertainty about ground truth, non-observability of counterfactual outcomes, and agents' private information.[5] Another advantage of this setting is that we have Hawk-Eye data from the period immediately before the introduction of AI review, which allows for a direct comparison before and after the introduction of AI review. This comparison is especially clean because many important factors stayed the same over the period we study: there were no substantial changes to the

[3]For similar reasons, Hawk-Eye review is being considered for sports like American football, soccer, and baseball, where camera imaging alone is insufficient to provide a definitive assessment often.

[4]In Section 2.3, we provide a more detailed assessment of EV, including a discussion of the transparency conditions outlined in the *Onius Probandi* of List (2020).

[5]A potentially interesting follow-up question is whether human reactions to AI oversight differ by the ability of the AI to replace human mistakes, but it is unlikely that firms will want to replace human mistakes with lower-quality AI, so the right comparison would be between nearly-perfect and excellent AI.

training and performance assessments of umpires, the pool of umpires, or the positioning and instructions to line judges.[6]

By combining our proprietary Hawk-Eye data with publicly available data on player challenges, we uncover strong evidence that human decision-making changes under AI oversight. We find that after the introduction of AI review, umpires lowered their overall mistake rate by 8% (1.1 p.p.) for close calls (balls within 100 mm of the line). However, this aggregate decrease masks two important sources of heterogeneity: the distance of the ball from the line and whether the ball was in or out. For balls just outside of the line (within 20 mm), the mistake rate actually *increased* by 34% (8.5 p.p.). 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 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 resume the point where it stopped, 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.[7]

Our reduced-form estimates suggest that the psychological costs of being overruled by AI led umpires to shift the rate of Type I and II errors they made, but do not indicate the magnitude of these psychological costs. To get a sense for this, we structurally estimate a lower bound on the psychological costs of being overruled by AI using a model of attention-constrained umpires. In this model, putting in more attention is effortful (cognitively costly) but results in more accurate perception and predictions. The umpire trades off these cognitive costs with the psychological costs of incorrect calls and being overruled by AI. The umpire has two levers for achieving this balance, and we find supportive evidence of both channels: the umpire can increase their attentional effort and vary the threshold belief at which they call a ball in or out, shifting the fraction of Type I and Type II errors.

We employ a two-stage approach to estimate the parameters of this model. We first use decisions before Hawk-Eye was introduced to recover the perceptual costs of making

[6] These institution details were verified by a leading tennis official at the time of Hawk-Eye's introduction.

[7] AI oversight could also lead to career concerns, and these would be asymmetric if ATP Tour organizers punished referees for calls that induce replay decisions (despite their entertainment value). However, the ATP Tour did not use Hawk-Eye to assess umpires during the period that we study.

correct calls, and we then use 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 37% more about Type II errors (calling a ball out when in) after the introduction of Hawk-Eye review.

We view these findings as more than just insights into sports officiating, but as a unique opportunity to observe the behavior of professional workers in a natural setting with distinct empirical advantages. In Section 2.3, we argue that our results are likely to extend to contexts where the outcome and overruling mechanism are visible, as these are the key parameters in the theory guiding our structural model. Using a nearly perfect technology like Hawkeye provides a rare opportunity to observe the ground truth while opening up discussion on whether the behavioral responses we detect generalize to contexts with only highly accurate overruling mechanisms. However, this does not mean that our results are limited to contexts with perfect technology. What matters is that when technology is used for corrections, it is sufficiently accurate. The former distinction may seem restrictive, but the latter aligns with a wide range of widely adopted technologies. For example, while fully autonomous self-driving cars have not yet taken over the streets, many vehicles already include lane-keeping assistance systems that intervene when a driver drifts. These systems, though imperfect, operate with a high degree of confidence in the moments they are engaged, similar to other decision-support technologies in professional settings.

The economics literature on AI and human decision-making, which we review in Section 1.1, has focused almost exclusively on the paradigm in which humans are provided AI recommendations. One reason for focusing on this paradigm is that giving humans final decision rights is more socially acceptable than handing these rights over to AI. We take this literature in a new direction by considering a paradigm where AI systems are used to overrule human mistakes. Despite the negative social desirability of this paradigm, firms will consider using AI to overrule human mistakes because AI recommendations can fail to improve choices. Documented factors for this include poor skills at incorporating AI recommendations into decisions (which requires metacognition and time) or biased beliefs (Agarwal, Moehring, Rajpurkar, and Salz 2023; Caplin, Deming, S. Li, Martin, Marx, Weidmann, and Ye 2024).

To the best of our knowledge, we provide the first field evidence that AI oversight can influence human decision-making. While this field evidence comes from one specific setting, our results and theoretical framework provide a broader lesson on the impact of AI oversight. While introducing an AI system to overrule apparent human mistakes seems promising in principle, especially since it might motivate higher effort from decision-makers, there is a strong reason to approach this with caution: incentive misalignment. The particular implementation guidelines of the oversight system can shift the incentives for the decision-maker. This becomes problematic when the incentive shift reduces the overall quality of their decisions from the perspective of the social planner. As we 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, at least from the perspective of the umpire. 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 balls 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, which led to an upsurge of incorrect calls during balls that were just out. This shift in the types of errors can be highly detrimental in medical and judicial contexts, where Type I and Type II errors have widely differing implications.

In this regard, our study also provides additional perspective to the debate surrounding when formal decision-making authority should be given to humans versus AI (Ide and Talamas 2025; Athey, Bryan, and Gans 2020). This is nearly certain to be an issue that policymakers, regulators, and legal authorities are forced to grapple with in the coming years.

The rest of the paper proceeds as follows. 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. In Section 4, we provide a model of perceptually-constrained umpires and use it to structurally estimate the psychological costs of AI oversight. In Section 5, we explore additional sources of heterogeneity. Finally, we conclude with a brief discussion in Section 6.

1.1 Related Literature

While AI is increasingly good at prediction, humans are kept in the decision process (kept "in the loop") because of social reasons (tradition, comfort, fairness, etc.), due to labor market concerns (union power, contractual responsibilities, inequality considerations, etc.), because of the costs of implementing AI systems, and because humans can sometimes perform better by incorporating additional information, understanding context, handling edge cases, and adapting to changing circumstances. One form of joint AI and human decision-making that has been actively studied is when AI provides recommendations to human decision-makers. Grimon and Mills (2025) show that Child Protective Services workers are able to reduce child hospitalizations when provided with an algorithmic risk score. However, unexpectedly, the gains did not seem to come from following the algorithmic predictions but rather from

reallocating their attention to different features of the allegations. L. Raymond (2024) finds that the availability of algorithmic tools in the housing market generates responses from non-adopter investors as well, shifting their focus to market sectors where human investors have comparative advantages over algorithms. Cho (2023) uses a quasi-experimental sports setting to examine how baseball umpires are affected when they work with AI assistance, focusing on their skill development.[8]

Making decisions in environments with algorithmic recommendations elevates the need for humans to exercise proper discretion, which can lower the effectiveness of these recommendations. The findings from Hoffman, Kahn, and D. 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 reach a decision when AI information is provided. Kreitmeir and Raschky (2024) use a ban on ChatGPT in Italy to show that high-productivity programmers have worse output in the presence of AI assistance.

Because giving AI final decision rights is potentially problematic, why would firms not just give their workers AI recommendations? The results above highlight two potential reasons why. First, humans can have difficulty exercising appropriate discretion when given AI recommendations, either by taking AI recommendations when they should not or by ignoring them when they should not. For example, AI guidance can be systematically ignored due to overconfidence or adopted due to under-confidence (Caplin, Deming, S. Li, Martin, Marx, Weidmann, and Ye 2024). The second reason why firms might consider AI oversight instead is because a final decision has to be made quickly, and incorporating AI recommendations into decision-making can be time consuming.

With AI oversight, discretion, a prominent force in algorithmic recommendation systems, no longer plays a major role. However, it introduces new behavioral forces into play, such as shame, pride, embarrassment, and stress. 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, we estimate a lower bound for the psychological costs of being overruled by AI, and we feel that shame and embarrassment are likely components of these costs. An interesting

[8]One potential advantage of our setting for studying human-AI interaction is that, like the baseball setting, human line judges in tennis do not have any clear source of private information. We thank Ashesh Rambachan for raising this point.

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 number of controversial mistakes at the expense of an increase in the number of less embarrassing ones. It is worth noting that Hawk-Eye decisions were available for television broadcasts both before and after the introduction of AI review, so this aspect of embarrassment does not change over the period we study.

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 when introducing new incentives to shirk by free-riding 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, there are reasons to believe that these forms of oversight might lead to differential responses. For example, Avery, Leibbrandt, and Vecci (2023) provide field evidence that using AI recruitment review induces a response from the supply side because women increase their application completion rates. From a theoretical perspective, Iakovlev and Liang (2023) consider the differences between AI and human evaluation, primarily the impact of context on humans, as AI evaluation is based on fixed covariates. However, regardless of whether there is a behavioral difference between AI and human oversight, 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 Maćkowiak, Matějka, and Wiederholt 2023 for reviews).[9] 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

[9]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.

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*.[10] Bhattacharya and Howard (2022) find that the standard rational inattention model explains the equilibrium behavior of professional baseball players, and they estimate the linear cost of attention in this setting. 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 works by Bolte and C. Raymond (2023) and Almog and Martin (2024) suggest the importance of expanding rational attention modeling to account for emotional states. Our paper also adds to the literature on attention by considering the impact of AI tools on attention, a connection suggested by the results of Grimon and Mills (2025).

2 Setting and Data

In March 2006, at the Nasdaq-100 Open in Key Biscayne (currently known as the Miami Open), Hawk-Eye review was officially used for the first time at an ATP Tour event.[11] Later that year, many tennis tournaments that use non-clay surfaces, including the US Open, adopted this new technology, allowing players to challenge calls.[12] Hawk-Eye uses six to ten 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,[13] 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 the umpire's judgment about whether a ball bounced in or out of bounds.[14] This setting has several advantages for studying the

[10] Other papers have examined how algorithmic recommendations directly influence decision-makers' preferences. McLaughlin and Spiess (2024) propose a theoretical framework in which algorithmic recommendations not only shape beliefs but also establish a reference point. More specifically related to our work, Albright (2024) exploits a natural experiment in bail courts to isolate the causal effect of recommendations, independent of the prediction component, demonstrating that recommendations influence judges' preferences by reducing the cost of making lenient decisions.

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

[12] Players have a fixed number of challenges that they can make per match, but successful challenges do not count against this limit. Players were not allowed to challenge chair umpire decisions on non-clay courts before this system was put in place.

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

[14] Our consolidated data set includes one tournament that was held without Hawk-Eye review after the

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.[15] Third, there was a testing period during which the technology was used solely for broadcast and data recording purposes, without a challenge system in place, allowing us to track umpire performance both before and after the formal introduction of Hawk-Eye review in a tournament.[16] 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.

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.[17] 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 ten umpires is involved in officiating a match. The roles typically include one chair umpire who oversees the entire match, making decisions on points, 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

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 an *after* period of Hawk-Eye review, even though this is strictly only true at the tournament level.

[15]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/.

[16]Our officiating source noted the following about the broadcast period: *"Even though these replays are not shown on the big screen, they inform the media and TV commentators, which can lead to widespread comments about the umpire being wrong."* This suggests that we are likely underestimating the impact of AI oversight on umpire decision-making.

[17]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.

bounced in or out when it is relatively close to their respective line.[18] The chair umpire has the authority to overrule the decisions of the line umpires if necessary, so drawing 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 the performance of the umpire crew as a whole. International chair umpires are certified with Gold, Silver, or Bronze badges, while line umpires are graded according to national or other systems. Certification and grading methods remained unchanged during the study period, and Hawk-Eye metrics were not incorporated into these evaluations for the first couple of years. A leading tennis official who was involved in testing and introducing Hawk-Eye provided valuable institutional insights that will be used throughout the paper.

The Hawk-Eye review protocol works by endowing players with 2-3 challenges per set.[19] 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 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,

[18] Figure B.1 provides a visual representation of the positioning of each umpire.

[19] 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.

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.[20] 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 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 that players decided to challenge, and this selection is susceptible to players' perceptual and behavioral biases. 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: a 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 an effective approach for merging the first two data sets.[21]

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.[22] This data set includes 109 matches from seven 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

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

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

[22] Table B.1 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.

set includes information on 1.8 million bounces from more than 1,800 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 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. Of the 35 tournaments, we use the 28 that also appear in our Hawk-Eye Base dataset. 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 an assessment of 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 99.3% merge rate, for which only 4.9% of the video-audited challenges were merged to an incorrect point. The algorithm consists of eight iterations,

starting with the most stringent merging rule and gradually relaxing criteria for challenges that persist without merging.[23]

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 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.[24]

2.3 External Validity

Given that our paper provides the first field evidence on how AI oversight impacts human decision-making, we view our project as a "WAVE1" study in which the first tests of a hypothesis are conducted (List 2020). Specifically, we test the hypothesis that humans will change their decision-making in response to the (potentially asymmetric) psychological costs of AI oversight. Despite the early stage of our study, we use this section to make an assessment of its external validity. To structure this section, we build around the four transparency conditions given as a *Onus Probandi* in List (2020).

2.3.1 Selection

This condition concerns the representativeness of the studied group compared to the underlying population, in particular in "observables that might impact preferences, beliefs, or individual constraints." Professional tennis umpires, especially at the top tennis tournaments we study, have long experience and are highly skilled, so our study is more likely to generalize to other specialist populations, like expert technicians. However, these tennis umpires are not required to undertake years of formal eduction to be qualified for their roles, unlike doctors and judges, so to the extent that this matters for AI oversight, we might be understating the effect for such populations, which could have stronger concerns about AI oversight.

2.3.2 Attrition

This condition asks if "there motivational or incentive differences between subject and target groups to maintain compliance." As noted in the introduction, one of the leading tennis

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

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

officials at the time of Hawk-Eye's introduction confirmed that many important factors remained consistent over the period we study: there were no substantial changes to the training and performance assessments of umpires, the pool of umpires, or the positioning and instructions given to line judges. This is ideal for isolating the impact of psychological factors related to AI oversight, but in practice, there may be other important factors that drive the human response to AI oversight, such as career concerns.

2.3.3 Naturalness

This condition addresses the artificiality of a setting, such as in the choice task. We view the task in our study as having parallels to doctors making diagnosis decisions based on images that are hard to read. However, because it is more of a perceptual task than about general intelligence, this task might not capture ego-relevant concerns, perhaps understating the effect in a general population. A potentially interesting follow-up study is to investigate how humans respond to AI oversight in ego-relevant tasks.

Another aspect of our setting is that there is a live audience of spectators. We view this aspect of our setting as having parallels to many professional decisions that occur under public scrutiny – for example, doctors make diagnostic decisions in front of patients and colleagues, judges determine bail outcomes in court with prosecutors, clerks, and officers present, and HR decisions are often made in team settings. However, having a larger audience might induce additional human responses that do not apply to more private or less intimate settings.

A final factor about our task, which we discussed previously, is that AI is nearly perfect at the task we study. Being overruled by less precise technology may influence the behavioral forces we study in different ways, as humans can shift blame or defer shame. However, this does not mean that our findings only generalize to contexts where the overruling mechanism is highly accurate, like Hawk-Eye. Rather, what matters is that the technology in use is very precise when deployed, allowing for a wider applicability in various settings. In the Zalando category managers case, technology can block a manager's price changes only when the algorithm detects a high probability that the decision is based on non-private information. Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2018) show that judges release the riskiest 1% of defendants (as predicted by the algorithm) at a rate of 48.5%, suggesting a judicial protocol in which defendants identified by the algorithm as among the riskiest are denied bail, regardless of the judge's decision.

2.3.4 Scaling

This condition concerns whether the effects will persist when the intervention are scaled. We have already addressed issues related to horizontal scaling (moving to other populations), but here we address issues of vertical scaling (scaled up to a larger population). For cost reasons, only the top tournaments implemented AI oversight, but all umpires at this level faced this technology, so all of the umpires' professional peers faced the same AI oversight. Thus, our study sheds light on this level of scale. One concern could be that the widespread introduction of AI oversight might dampen the psychological costs of AI oversight. However, we see that the response to AI oversight gets larger with time, so the exposure and visibility that comes with increases in scale might actually increase the effect.

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.[25] We begin by examining changes in the overall mistake rate for all hits. We then highlight the influence of distance from the line and whether a ball was in or out.

3.1 Overall Mistake Rate

Before Hawk-Eye review was introduced, the umpire mistake rate was only 0.61% of all ball bounces. 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%) 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.[26]

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,[27] we find that the mistake rate

[25] Our officiating expert confirmed that the pool of umpires remained stable throughout the study period. This eliminates the possibility that the estimated effects are due to changes in the umpire sample. Additionally, suspensions and terminations are implemented solely for disciplinary reasons rather than performance, which further alleviates this concern.

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

[27] 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)	(0.007)	(0.007)	(0.007)
Point controls		X	X	X
Match 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

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

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 it 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.

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 it is below or above the median), score, game, set and an indicator if the point played is in the tie-break stage.[28] Y is a vector of match characteristics that has round and tournament tier.[29]

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 2 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 by 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, such as attentional free-riding (see Section 4 for a model of this form).

One reason why the estimates of this specification might be noisy is that the impact

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

[29] 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.

of Hawk-Eye might depend on two main features. First, an important component to call difficulty is the distance from the line, so the ability to change attention enough to fix a mistake might depend on the distance from the line. Second, it might matter which side of the line the ball is on, and umpires might have different perceptions of the cost of Type I and Type II errors after the introduction of Hawk-Eye review.

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 represents the rate at which umpires made an incorrect call for all balls bouncing within one of ten 20 mm bins. The five bins to the left of the dashed line correspond to balls that landed out of bounds, while the five bins to the right side of that line represent balls that landed inside of the line. For the rest of the paper, negative distances will be used to denote the distance to the line for balls that bounced outside.

The blue dots in the figure 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 eight of the ten bins). This mirrors the results in Table 2, which shows that the overall performance of umpires improves after the introduction of AI oversight, and in Table 6a, 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.

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 (including point and match controls), but now we interact $PostHK$ with 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 2. For each bin, we plot the estimated 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 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)

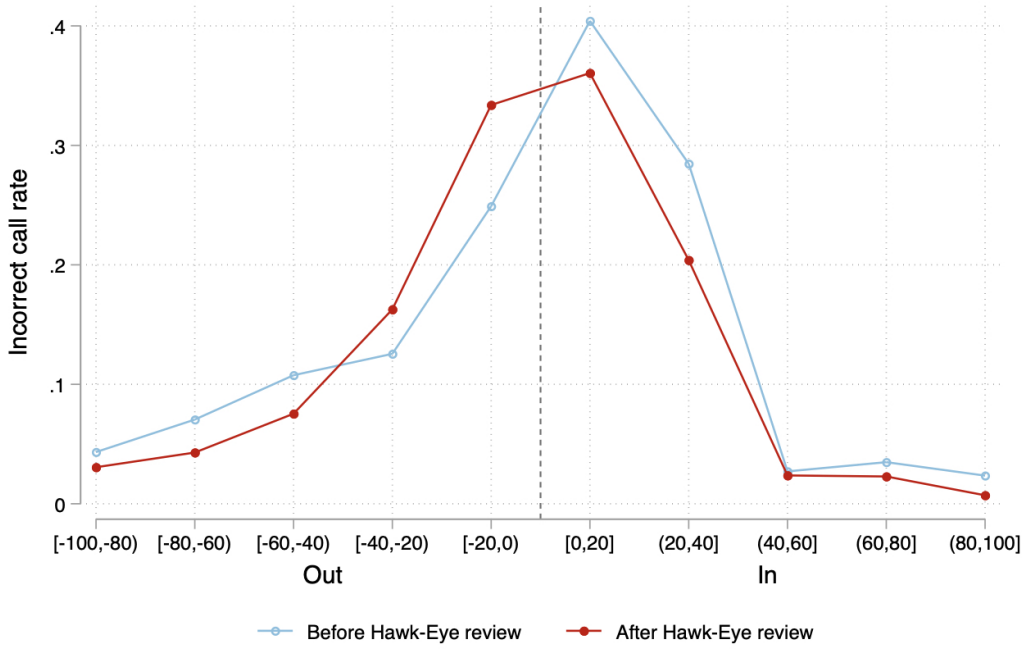


Figure 1: Incorrect call rates by proximity to the line. Each dot is the rate of incorrect calls for a bin of 20 mm. Dots to the left of the dashed line represent bins out of bounds, and the right of the dashed line represents bins in bounds.

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. As highlighted by Goldsmith-Pinkham, Hull, and Kolesár (2024), our estimates from Figure 2 may suffer from "contamination bias" due to the regression of multiple treatments while controlling for observables in an additively separable manner. To address this concern, we implement one-treatment-at-a-time regressions, one of the suggested solutions by Goldsmith-Pinkham, Hull, and Kolesár (2024). Figure B.2 shows that our original estimates do not suffer from this type of bias.

We provide an event-study version of Figure 2 in Figure 3. 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 3 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

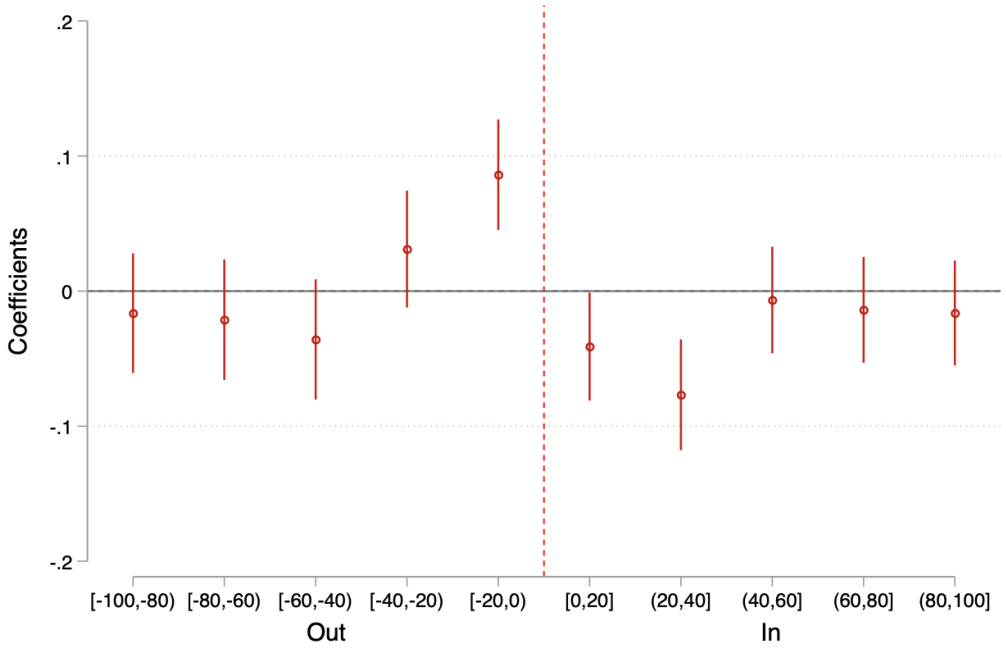


Figure 2: Impact of AI oversight on the incorrect call rate by proximity to the line. Each dot represents the coefficient on the interaction between the distance bin and PostHK, the indicator variable that equals 1 if the Hawk-Eye review system is active. Controls for point and match characteristics are included.

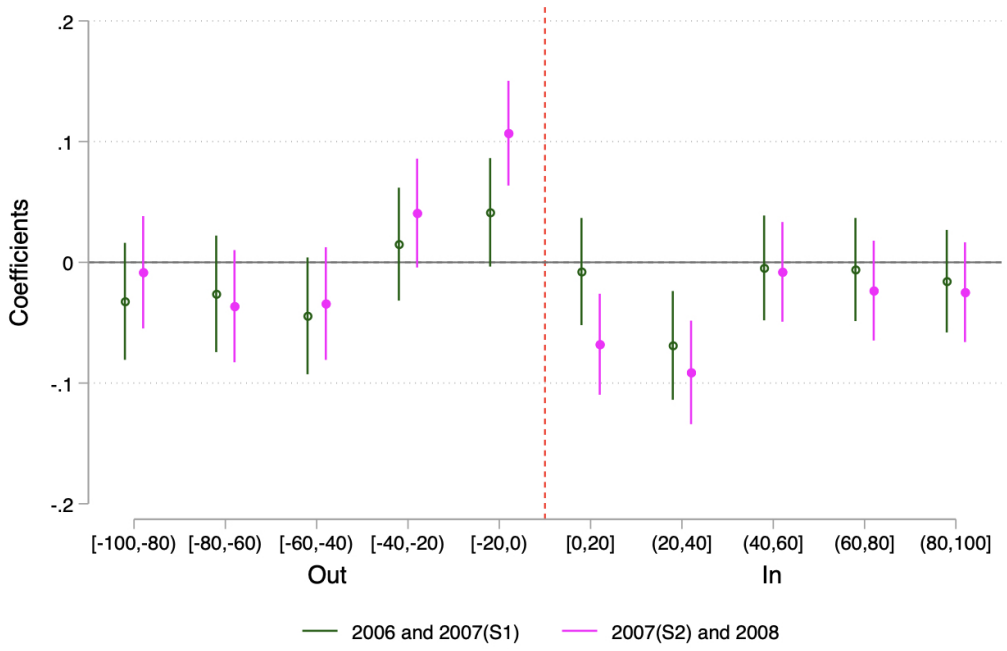


Figure 3: Impact of AI oversight on the incorrect call rate by proximity to the line (by time sub-period). Matches are grouped into those with Hawk-Eye review in all of 2006 and the first half of 2007 and those with Hawk-Eye review in the second half of 2007 and all of 2008. Each dot represents the coefficient on the interaction between the distance bin and PostHK, the indicator variable that equals 1 if the Hawk-Eye review system is active.

4 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 AI review, sorted by time. The first observation is that this rate increased from 42.8% to 49% with the introduction of AI review. In 22 out of the 28 tournaments with AI review, umpires called the ball in more frequently than the mean of 42.8% observed in tournaments without AI review. Figure 4 suggests a gradual increase in the frequency of inside calls by umpires over time, which may reflect their growing adaptation to the psychological forces introduced by AI oversight.[30] The slope coefficient for the regression in Figure 4 is 0.45 pp per month (p-value=0.11), an effect that would translate to over 5.4 pp in a year.[31] Borrowing a common instrument from cognitive science used to study imperfect perception and cognitive limitations, we compare the psychometric curves across periods in Figure B.4. Both psychometric curves have similar slopes in the middle region, ensuring that the response we observe in the closest calls results from a shift in actions, and not from a change in information acquisition.

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.[32] Here, we apply the rational inattention theory introduced by Sims (2003) and characterized by Matějka and McKay (2015) and Caplin and Dean (2013), which we refer to as the "Shannon model." Bhattacharya and Howard (2022) use this model to study attention constraints in baseball, while Brown and Jeon (2024) use it to study consumers choosing among complex health insurance plans.

In our model, 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 AI review. 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

[30] The leading official we contacted, who played a key role in the deployment of Hawk-Eye in the ATP, confirmed that umpires were not explicitly instructed to call more 'ins' for close calls.

[31] Figure B.3 shows that these results are almost identical when broken down by serves and non-serves.

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

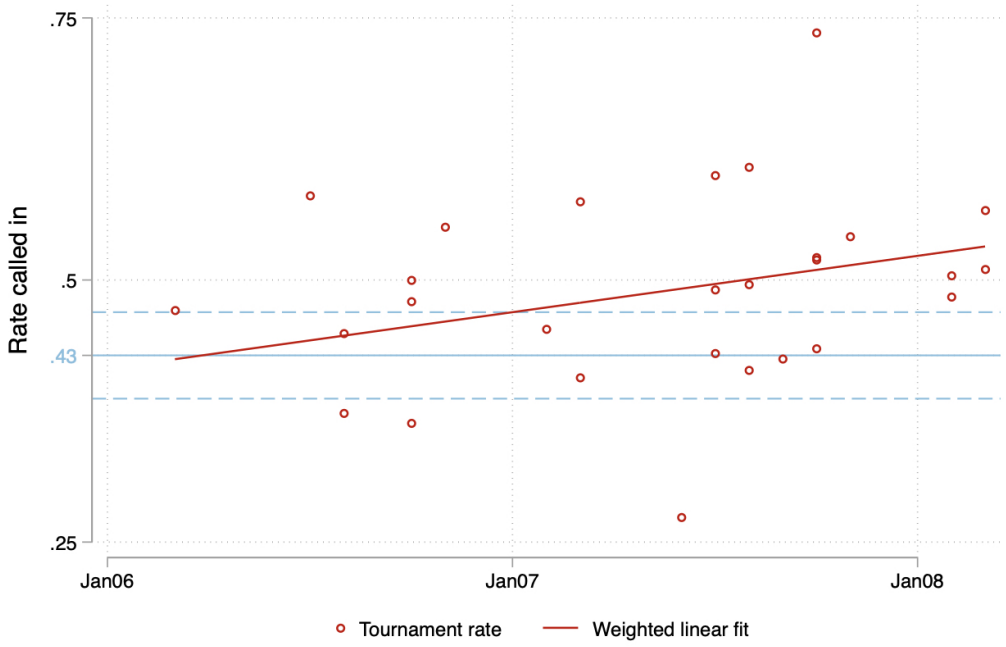


Figure 4: The rate of calling a ball in after the introduction to Hawk-Eye review for balls landing <20 mm from the line, regardless of which side of the line they actually bounced on. Each dot represents the rate of calling a ball in for a tournament. The red line is the best linear fit using the dots 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 seven tournaments that did not have AI review (the blue dash lines indicate the 95 confidence interval).

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

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 the umpire would need to be 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 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 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

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

Next, we assume information structures are chosen, which we interpret as the umpire choosing how much attention to pay and what aspects of the game to pay attention to,

as in Bhattacharya and Howard (2022). Each information structure $\pi \in \Pi(\mu)$ has an additively-separable cost $K(\pi)$ that scales linearly with the Shannon mutual information between posteriors and the prior. Formally, K is determined by the function

$$K(\pi, \kappa, \mu) = \kappa \left(\left[\sum_{\gamma \in \Gamma(\pi)} \pi(\gamma) \sum_{\omega \in \Omega} [\gamma(\omega) \ln(\gamma(\omega))] \right] - \sum_{\omega \in \Omega} [\mu(\omega) \ln(\mu(\omega))] \right) \tag{3}$$

where $\kappa \in \mathbb{R}_{++}$ is a linear cost parameter, which can be interpreted as the marginal cost of attention. We assume that κ is increasing in the difficulty of judging where a ball landed and assume that the difficulty of this perceptual task is unimpacted by the introduction of Hawk-Eye review.

We extend the standard Shannon model by allowing for different costs of attention for different states, so that

$$K(\pi, \kappa, \mu) = \kappa^I \left(\left[\sum_{\gamma \in \Gamma(\pi)} \pi(\gamma) \gamma(\omega^I) \ln(\gamma(\omega^I)) \right] - \mu(\omega^I) \ln(\mu(\omega^I)) \right) \tag{4}$$

$$+ \kappa^O \left(\left[\sum_{\gamma \in \Gamma(\pi)} \pi(\gamma) \gamma(\omega^O) \ln(\gamma(\omega^O)) \right] - \mu(\omega^O) \ln(\mu(\omega^O)) \right) \tag{5}$$

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

With the Shannon model, the optimal information structure has one posterior at which each action is taken. We denote these posteriors as γ^I when call in and γ^O when call out. By the Invariant Likelihood Ratio (ILR) of Caplin and Dean (2013), these optimal posteriors must obey

$$\frac{\gamma^I(\omega^I)}{\gamma^O(\omega^I)} = e^{\frac{U\left(a^I, \omega^I\right) - U\left(a^O, \omega^I\right)}{\kappa^I}} \tag{6}$$

$$\frac{\gamma^O(\omega^O)}{\gamma^I(\omega^O)} = \frac{1 - \gamma^O(\omega^I)}{1 - \gamma^I(\omega^I)} = e^{\frac{U\left(a^O, \omega^O\right) - U\left(a^I, \omega^O\right)}{\kappa^O}} \tag{7}$$

For matches where there was no AI review, the ILR condition, given by (6) and (7), allows us to express the marginal costs of attention κ^I and κ^O cleanly as

$$\kappa^I = \frac{1}{\ln \gamma^I(\omega^I) - \ln \gamma^O(\omega^I)} \tag{8}$$

$$\kappa^O = \frac{1}{\ln \gamma^O(\omega^O) - \ln \gamma^I(\omega^O)} \tag{9}$$

When there is AI review, (6) and (7) 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

$$c^I = \frac{1 - \kappa^I \left(\ln \gamma^I(\omega^I) - \ln \gamma^O(\omega^I)\right)}{\eta^I} - 1 \tag{10}$$

$$c^O = \frac{1 - \kappa^O \left(\ln \gamma^O(\omega^O) - \ln \gamma^I(\omega^O)\right)}{\eta^O} - 1 \tag{11}$$

4.1 Structural Estimates

To structurally estimate the AI oversight penalty, we undertake two steps. First, we estimate the marginal costs of attention when there is no AI review. When there is no AI review, 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. We present the estimates recovered from the data, along with the bootstrap standard errors obtained from drawing 1000 random samples with replacement, each containing the same number of observations as the original dataset.

As shown in Table 3, our estimate of the revealed posterior for the ball being in when calling in is 0.849, and our estimate of the revealed posterior for the ball being out when calling out is 0.876 for all calls within 100 mm of the line before the introduction of AI oversight. Based on (8) and (9), the marginal costs of attention that rationalize these values are $\kappa^I = 0.580$ and $\kappa^O = 0.510$.

Parameter	Recovery equation	Estimate
$\gamma^I(\omega^I)$	N/A	0.849
$\gamma^O(\omega^O)$	N/A	0.876
κ^I	8	0.580
		(0.025)
κ^O	9	0.510
		(0.024)

Standard errors in parentheses

Table 3: Estimated optimal posteriors and costs of attention before the introduction of AI oversight.

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 penalties. As shown in Table 4, the observed challenge rates when balls were in is 0.449 and when balls were out is 0.415. Given the values of κ^I and κ^O determined without AI review, the rationalizing AI oversight penalties are $c^I = -1.374$ and $c^O = -0.903$. 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	Estimate
η^I	N/A	0.449
η^O	N/A	0.415
c^I	10	-1.374
		(0.121)
c^O	11	-0.903
		(0.116)

Standard errors in parentheses

Table 4: 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 $(1+c^O)$ potentially unchanged, as there is an increase from 0 before Hawk-Eye review to 0.097 after, but the difference is not statistically significant. On the other hand, the utility of Type II errors $(1 + c^1)$ reduces sharply, from 0 before Hawk-Eye review to -0.374 after. Here, the AI oversight penalty dominates any utility gain 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 $c^I = -1.374$ and $c^O = -0.903$, which means that the (dis)utility of Type II errors (calling a ball out when in) decreases from -1 to -1.374. Thus, these estimates suggest that umpires care 37% more about Type II errors after the introduction of Hawk-Eye review.

5 Examining Heterogeneity

In this section, we study three main sources of heterogeneity: skill, stakes, and type of perceptual task. Practically, this relates to differences by tournament stage (finals or earlier stage), tournament tier (higher or lower tier), and shot type (serve and non-serve).

5.1 Heterogeneous Response by Umpire Skill

Umpires are rewarded for good performance with assignments to advanced-stage matches, as the pool of officials needed becomes smaller and organizers can be more selective as the tournament progresses. Our officiating source confirmed that not all umpires (particularly line umpires) are professional officials, and you can expect to see more of them in the early stages. These selection dynamics within tournaments create variation in umpiring skill. Using later tournament stages to identify more skilled umpires may confound the analysis by including matches with higher stakes. However, we demonstrate that this is not the case at the end of the section.

To examine whether the impact of AI oversight differs by tournament stage, we split our sample into two groups: the first group includes Final, Semifinal, and Quarterfinal matches, where we expect to find the most skilled umpires, while the second group contains the remaining matches. In Figure 5, 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.

In a study on radiologists, Chan, Gentzkow, and Yu (2022) found that aversion to false negatives tends to be negatively related to radiologist skill. Specifically, lower-skilled radiologists are more likely to falsely diagnose a healthy patient with pneumonia (false positive) because this error is less costly than missing a diagnosis. Our findings align with those of Chan, Gentzkow, and Yu (2022), as less-skilled umpires are more likely to avoid the more costly type of error (calling a ball in when it is actually out).

It is important to note that advanced matches also involve higher stakes. It is challenging to disentangle whether the observed effects are due to lower-skilled umpires or higher stakes. To address this, we can compare differences in stakes by examining highly regarded tournaments, such as Master 1000 events, against lower-tier tournaments like the 500 and 250 series. Figure B.5 shows that the discrepant patterns we observed between early and advance-stage matches do not hold when we compare higher and lower-ranked tournaments.

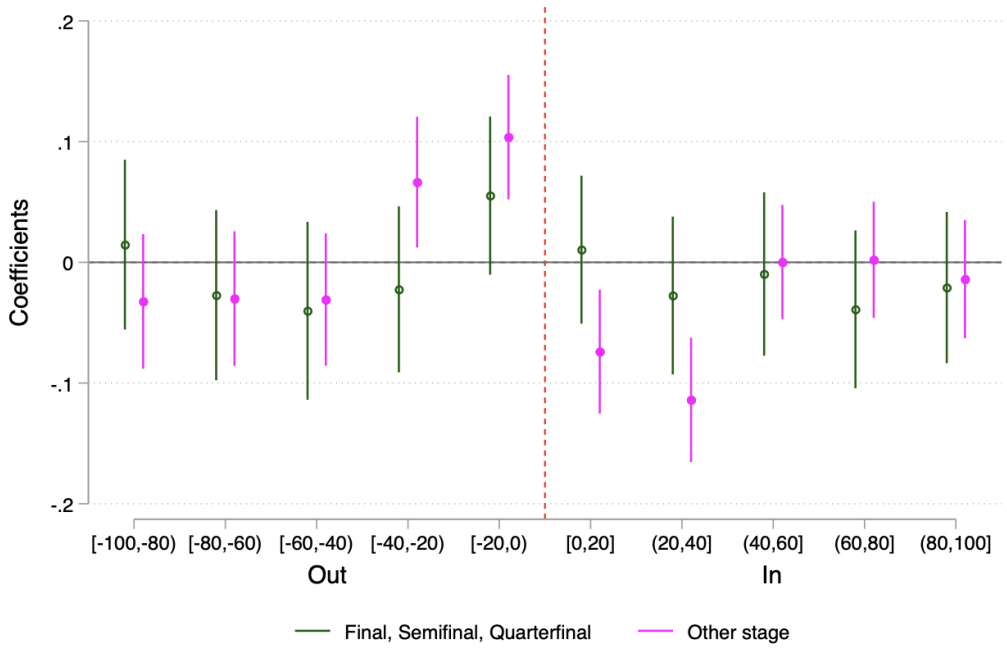


Figure 5: Impact of AI oversight on the incorrect call rate by proximity to the line (by tournament stage). Matches are grouped into those at the final, semifinal, and quarterfinal stages and those at all other (earlier) stages of a tournament and then analyzed separately. Each dot represents the coefficient on the interaction between the distance bin and PostHK, the indicator variable that equals 1 if the Hawk-Eye review system is active.

If the difference in stakes is what drives the heterogeneity observed in Figure 5, we should expect to see the same patterns between lower and higher-ranked tournaments, but we do not. One might be concerned that the pool of umpires is more talented in Master 1000 tournaments. However, our officiating expert confirmed that the quality of the umpire pool in our sample does not necessarily vary by tournament tier. Instead, it is influenced by a range of factors, including geographic location, organizer budget, and the number of courts. Therefore, we can effectively use different tournament tiers to demonstrate the effects of varying stakes while assuming that any variation in umpire skill is uncorrelated.

5.2 Serves and Non-Serves

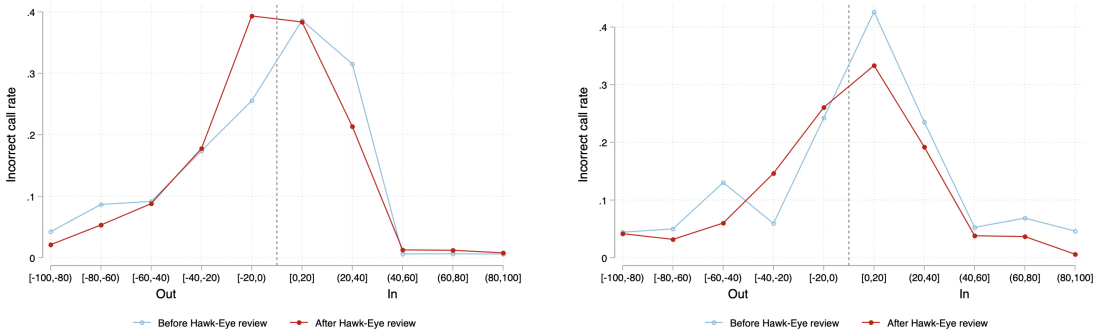
There is value in distinguishing between serves and non-serves,[33] as we can think of them as distinct tasks for the umpires. The primary differences, from an officiating standpoint, between serves and subsequent hits lie in speed and anticipation. Serves may be considered more challenging due to their significantly higher speed: the average speed for a serve in our sample is 156 km/h, compared to 82 km/h for non-serves.[34] 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 be 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 lower, 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.

In Figure 6, 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.

Table 5 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

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

[34] Figure B.6 shows the speed distribution for both types.



(a) Serves.

(b) Non-serves.

Figure 6: 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.

	(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
Match 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

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

Table 5: OLS regressions of umpire mistakes for balls bouncing within 100 mm of the line separately for serves and non-serves. $PostHK = 1$ 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.

0.136.

If we just consider those calls when the ball bounces between 100 and 20 mm from the line, Table 6a, shows that 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 6b, 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.

6 Conclusion

The main goal of this paper is to leverage a unique empirical setting to study what the adoption of AI oversight in tennis can reveal about human behavior. We do not aim to suggest an optimal practice for this application, as it is obvious that, in terms of accuracy, line umpires in tennis should have been replaced a long time ago. In 2020, the ATP employed electronic line judges in some tournaments due to COVID-19 limitations, and by 2025, the ATP aims to fully replace line umpires in top tournaments. This provides an important lesson about labor displacement: if it took 20 years and a pandemic to replace line umpires with a superior (and almost perfect) technology, it is hard to imagine AI replacing doctors or judges anytime soon. Collaboration between AI and humans will be a topic of first-order relevance for years to come, and we believe this paper provides valuable information to guide the design of better AI interventions.

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 exploring in more detail the underlying mechanisms.

	(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
Match 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
Match 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 6: 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 it 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.

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*.

Albright, Alex (2024). "The Hidden Effects of Algorithmic Recommendations". *Working paper*.

Almog, David and Daniel Martin (2024). "Rational Inattention in Games: Experimental Evidence". *Experimental Economics* 27.4, pp. 715–742.

Athey, Susan, Kevin Bryan, and Joshua Gans (2020). "The Allocation of Decision Authority to Human and Artificial Intelligence". *AEA Papers and Proceedings* 110, pp. 80–84.

Avery, Mallory, Andreas Leibbrandt, and Joseph Vecci (2023). "Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech". *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*.

Brown, Zach Y and Jihye Jeon (2024). "Endogenous information and simplifying insurance choice". *Econometrica* 92.3, pp. 881–911.

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.

Caplin, Andrew, David Deming, Shangwen Li, Daniel Martin, Philip Marx, Ben Weidmann, and Kadachi Ye (2024). "The ABC's of Who Benefits from Working with AI: Ability, Beliefs, and Calibration". *Working paper*.

Chalfin, Aaron, Oren Danieli, Andrew 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.

Cho, Sungwoo (2023). "The Effect of Robot Assistance on Skills". *Working paper*.

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.

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.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár (2024). "Contamination Bias in Linear Regressions". *American Economic Review* 114.12, pp. 4015–51.

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.

Grimon, Marie-Pascale and Christopher Mills (2025). "Better Together?: A Field Experiment on Human-Algorithm Interaction in Child Protection". *Working paper*.

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

Huelden, Tobias, Vitalijs Jascisens, Lars Roemheld, and Tobias Werner (2024). "Human-Machine Interactions in Pricing: Evidence from Two Large-Scale Field Experiments". *Working paper*.

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

Ide, Enrique and Eduard Talamas (2025). "Artificial Intelligence in the Knowledge Economy". *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.

Kreitmeir, David and Paul A Raschky (2024). "The Heterogeneous Productivity Effects of Generative AI". *arXiv preprint arXiv:2403.01964*.

List, John A. (2020). "Non est Disputandum de Generalizability? A Glimpse into The External Validity Trial". *NBER Working Paper No. 27535*.

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.

McLaughlin, Bryce and Jann Spiess (2024). "Algorithmic Assistance with Recommendation-Dependent Preferences". *arXiv Preprint arXiv:2208.07626*.

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.

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 (2024). "Identifying Prediction Mistakes in Observational Data". *The Quarterly Journal of Economics*.

Raymond, Lindsey (2024). "The Market Effects of Algorithms". *Working paper*.

Sims, Christopher A (2003). "Implications of Rational Inattention". *Journal of Monetary Economics* 50.3, pp. 665–690.

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 a match, the accuracy of Hawk-Eye decreases and requires constant re-calibration to operate at its best.[35]

Tournaments on clay surfaces rejected the idea of incorporating Hawk-Eye review. 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".[36]

At first glance, clay tournaments may appear as an ideal control group for comparing umpire reactions in tournaments with Hawk-Eye review versus those on clay. However, several complications arise, leading us to the decision to exclude 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.

[35] https://www.perfect-tennis.com/why-there-is-no-hawk-eye-on-clay/
[36] 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 match (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 are 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 A.1 which 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 A.1: Performance for each iteration of the merging algorithm. Out of the 144 challenges audited, only one remained unmerged.

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's 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 two 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 two 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 a direct consequence of 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 Hawk-Eye

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 seven of the 23 points audited were actual umpire mistakes.

B Additional Tables and Figures

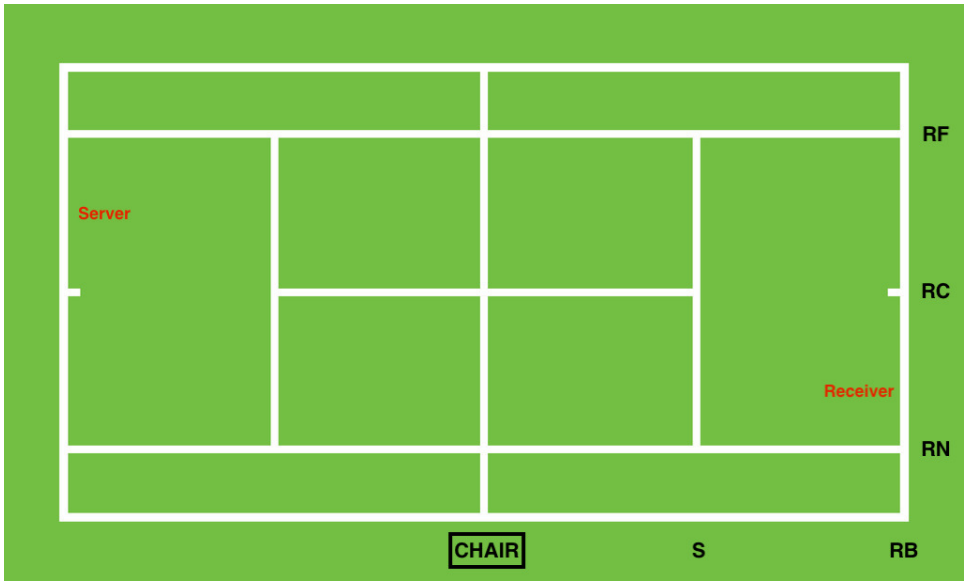


Figure B.1: The location of the umpiring crew: S = the serve-line umpire, RB = the right baseline umpire, RN = the right near long-line umpire, RC = the right center-line umpire, RF = the right far-side long-line umpire. The left side of the court has corresponding line umpires, except for the serve-line umpire, who moves to the left side when the right side player has the serve.

Tournament	Start Month	Category	Court Type	Number of Games			
				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	Sep	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	
Shanghai	Nov	Masters (1000)	Carpet05 Hard06-07	14	15	15	

Table B.1: 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.

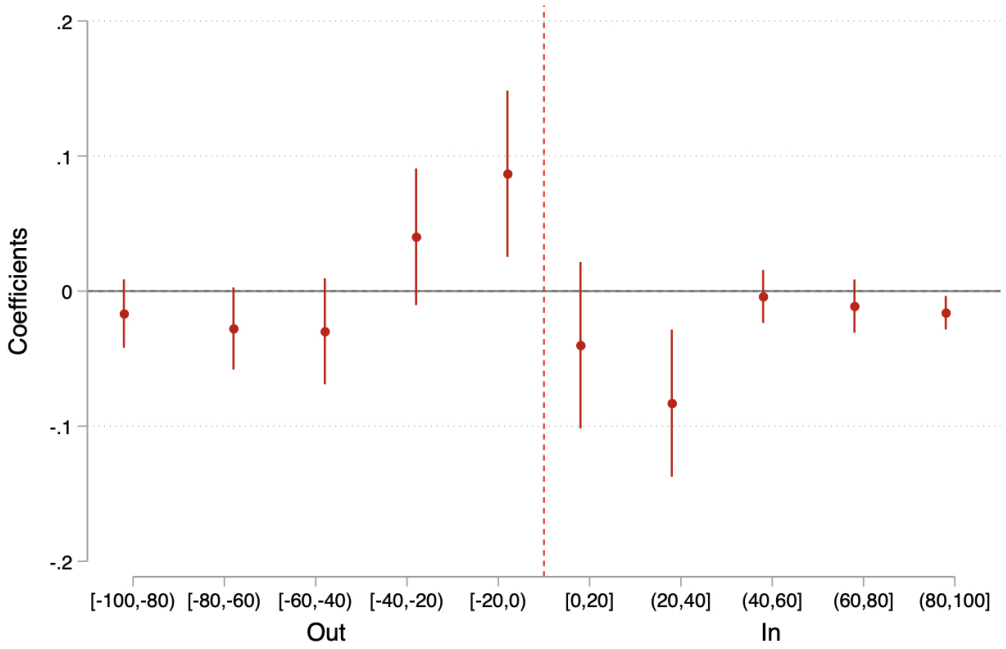
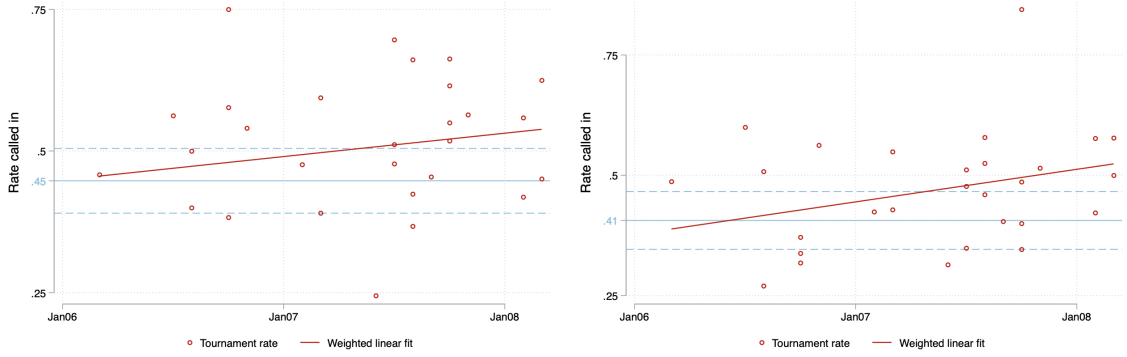


Figure B.2: An analog of Figure 2, using a one-treatment-at-a-time approach.



(a) Serves.

(b) Non-serves.

Figure B.3: An analog of Figure 4, separating serves and non-serves.

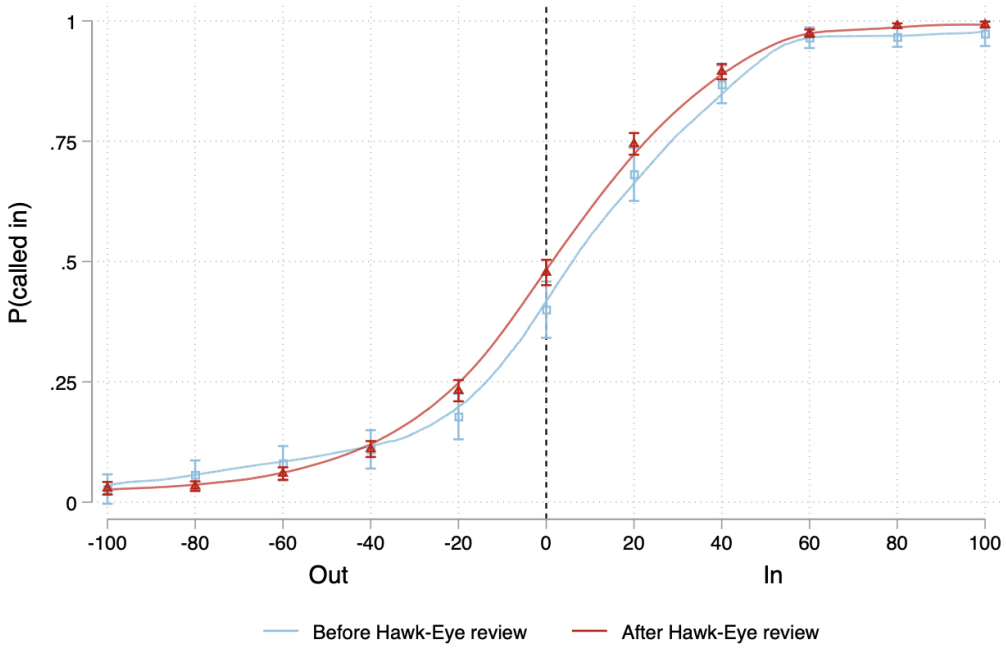


Figure B.4: The dots represent the average rate at which umpires call a ball "in" when it bounces within a 10 mm radius (e.g., at x=0 is the average rate for balls bouncing between -10 mm and +10 mm). The psychometric curves were approximated using locally weighted scatterplot smoothing.

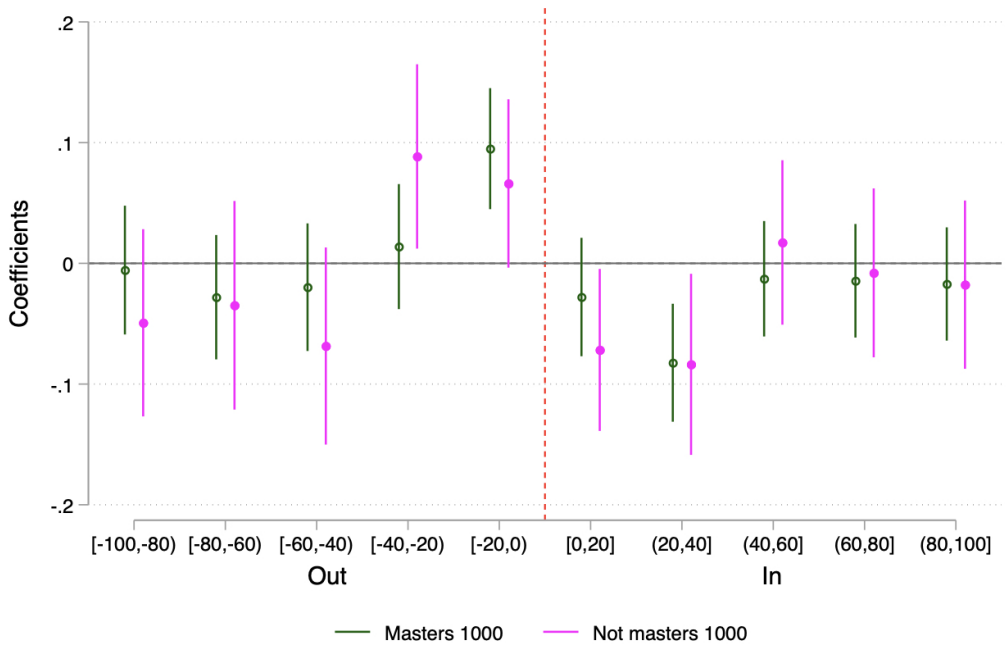


Figure B.5: Heterogeneous Impact of AI Oversight on Umpires' Performance Across 20 mm Distance Bins by Tournament Category. We conducted other separate analyses, categorizing matches from higher ranked tournaments (Masters 1000) vs lower ranked (International 500 and 250).

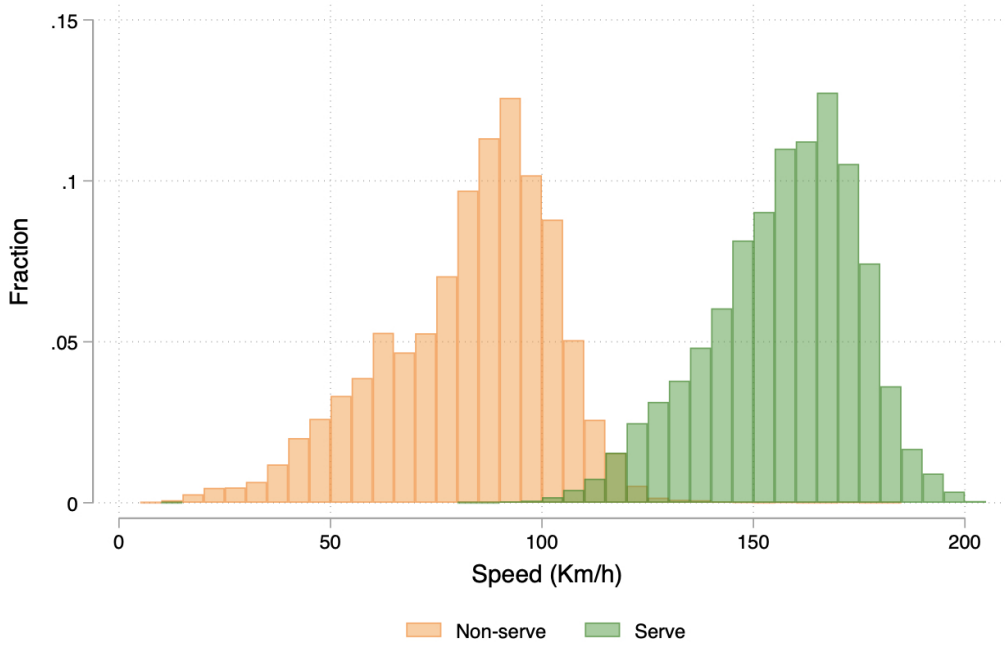


Figure B.6: Speed distribution for all the balls that bounced within 100 mm of the line, separated by non-serves and serves.