Authority under Challenge

LI, HAO
University of British Columbia

October 15, 2025*

Abstract. We consider a game-theoretical model of referee decision-making in sports matches. A referee draws an informative but imperfect signal about a play and chooses whether or not to make a foul call to penalize the player, who wants to get away with a foul play but prefers a legal play if anticipating a foul call. Both a call and a no-call decisions by the referee may be reviewed by an independent monitoring technology that is informationally superior to the referee's signal but remains imperfect. The review adds concerns about having their initial decision overturned to the referee's interests in avoiding making a foul call on a legal play and letting a foul play go unpenalized. A harsh equilibrium, where the referee always makes a foul call when their signal indicates a foul play but mixes between a call and a no-call otherwise, coexists with a lenient equilibrium, where the referee alway makes no call when their signal indicates no foul play but mixes otherwise. An increasing use of the review makes the referee call for a foul more often in a harsh equilibrium and less often in a lenient equilibrium, and can induce the player to commit fouls more often. The theoretical findings are consistent with evidence from European football leagues.

*I would like to thank Hui Zhang for research assistance.

1 Introduction

Economic models of decision-making authority emphasize the tension between a principal's need to delegate to an agent who has better information about decisions to be made, and misalignment of interests of the agent with those of the principal (Holmstrom, 1984; Aghion Tirole, 1997). In practice, however, authority in practice comes with not just the right, but also the obligation to make decisions. A principal saying "Buck stops with me" means that they often have to confront real-time challenges to their authority.

An example of authority under challenge is refereeing decisions in European football after the introduction of the Video Assistant Referee (VAR). Currently in all top-flight European leagues, including the Premier League in England since the 2019-2020 season and LaLiga in Spain since the 2018-2019 season, VAR assists the on-site match referee by reviewing decisions using video footage and providing recommendations to the referee.[1] Upon stopping the match to review the video evidence after being interrupted by VAR, the match referee can either ignore the recommendation of VAR or overturn their initial decision. For some categories under possible review by VAR, such as whether a player has engaged in a violent conduct, video evidence provided by VAR is definitive and leaves the referee no choice but to following the VAR recommendation. However, other categories under possible review by VAR – the decision of whether or not to issue a penalty for a foul inside the penalty area – provide leeway for the match referee to take either action, because the corresponding rules are subject to interpretations even with video evidence. Comparing to how referees made their decisions prior to the introduction of VAR, today's referees may be concerned with having their authority of making on-site decisions chipped away by VAR. These concerns can affect how they officiate the match, and indirectly affect incentives of football players to commit fouls.

In this paper, we present a model of authority under challenge in the context of ref-

[1] These two leagues are chosen because they are two of the most followed football leagues in the world, and more importantly for this paper, it is widely known that there are significantly fewer fouls called in the Premier League than in LaLiga.

ereeing decision-making in a sports match. A match referee draws an informative but imperfect signal about a play and chooses whether or not to make a foul call to penalize the player. The player will commit a foul if they can get away with no call, but prefers a legal play if they anticipate a foul call. The referee, on the other hand, wants to avoid making a foul call on a legal play and letting a foul play go unpenalized. When the referee's decision is not reviewed, the game between the referee and the player has both a "harsh equilibrium," in which the referee always makes a foul call when their signal indicates a foul play but mixes between a call and a no-call otherwise, and a "lenient equilibrium," in which the referee alway makes no call when their signal indicates no foul play but mixes otherwise. The co-existence of the two equilibria suggests that "culture" and "tradition" can shape how the game is played and officiated. The referee makes the foul call more often and the player makes a foul play more often in a harsh equilibrium than in a lenient equilibrium.

We then introduce a review technology to the strategic game played by the referee and the player. Both a call and a no-call decision by the referee may be reviewed, who produces an independent signal that is more accurate than the referee's signal but remains imperfect. With the review technology, we assume that there is an additional reputation loss to the referee when their initial decision is overturned upon review. We consider both when the final decision is automatically made by the review (because video evidence of VAR is "objective"), and when the referee makes the final decision upon review (because video evidence is "subjective"). In the case where the review makes the final decision, we show that both a harsh equilibrium and a lenient equilibrium continue to exist when the referee has similar concern for having their foul call overturned as their concern for having their no-call overturned. By assuming that the referee's concerns for having their foul calls and no-call overturned are both sufficiently small, and the review signal is sufficiently accurate, we show that the same harsh and lenient equilibrium are replicated in the case where the review makes a recommendation and the final decision lies with the referee.

An increase in the use of the review means that the final decision is made more often by the review instead of the referee, but has opposite effects on how often the referee makes the foul call in the two equilibria. In the harsh equilibrium, where a foul call is made with a positive probability even when the referee receives a signal indicating no foul play, the player faces a greater incentive to commit a foul as the review makes the final decision of no call whenever the review signal indicates no foul play. The referee thus has to increase the probability of making a foul call to keep the player's incentive to foul in check. In contrast, in the lenient equilibrium the referee has to decrease the probability of making a foul call when their own signal indicates a foul play, in order to compensate for the loss of incentives for the player to commit a foul. An increase in the use of the review has however a generally ambiguous effect on how often a foul is committed by the player, but the effect is the same in the harsh equilibrium and in the lenient equilibrium. This is because in both equilibria the change in the incentives of the referee to make a foul call is caused by the same shift from the decision losses from wrongfully penalizing the player and letting a foul play go unpunished to concerns about having an initial foul call overturned and an initial no-call overturned. If the referee cares relatively more about having their initial foul call overturned than about having their initial no-call overturned, an increase in the use of the review reduces the referee's incentive to make a foul call, causing the player to foul more often in both the harsh and the lenient equilibria.

The evidence from the Premier League and LaLiga suggests that, prior to the introduction of VAR, a lenient equilibrium was played in the Premier League while a harsh equilibrium was played in LaLiga.[2] We focus on the number of match-altering fouls called per game, which in the model corresponds to the expected product of the probability that players commit fouls and the probability that referees make foul calls given their signals. In particular, the average number of penalties awarded per game was around 0.25 in the Premier League in the five years prior to the introduction of VAR in the 2019-2020 season,

[2]The numbers used here are compiled from public websites https://www.whoscored.com/ and https://www.transfermarkt.co.uk/.

and around 0.30 in LaLiga in the five years prior to the introduction of VAR in the 2018-2019 season.[3] This significant and consistent difference can be explained in the model, if referees in the two leagues have similar interests in avoiding wrongful foul calls and no-calls, since players are less likely to foul and referees are less likely to make foul calls in a lenient equilibrium than in a harsh equilibrium. Following the introduction of VAR, the average number of penalties awarded per game went up slightly to around 0.27 in the Premier League in the five years since the 2019-2020 season, and went up significantly to around 0.35 in LaLiga in the five years since the 2018-2019 season. These changes can be explained in the model, if matches in the Premier League continue to be played in a lenient equilibrium, and those in LaLiga in a harsh equilibrium, and if referees in the two leagues have similar concerns about having their initial foul calls and no-calls overturned. In the model, after VAR referees make foul calls less often in a lenient equilibrium and more often in a harsh equilibrium. When referees in the two leagues care relatively more about having their initial foul call overturned than about having their initial no-call overturned, players foul more often in both the harsh and the lenient equilibria. The effects of VAR go in the same direction in a harsh equilibrium, which explains the significant increase in the number of penalties in LaLiga, and they go opposite directions in a lenient equilibrium, which accommodates only a slight increase in the number of penalties in the Premier League. Although in both the Premier League and in LaLiga, VAR may have made referee decisions more accurate, the theoretical findings suggest that matches have not become cleaner in either league, and in LaLiga, are now disrupted with more foul calls by match referees.

[3]Yellow cards and red cards are also match-altering. Unlike penalties, they are sometimes awarded for disrespecting match officials. From the perspective of the model in this paper, it is unclear how this reason for yellow and red cards is affected by the introduction of VAR. For the corresponding time periods, the number of yellow cards issued per game was 3.29 in the Premier League and 5.08 in LaLiga, and the number of red cards per game was 0.19 in the Premier League and 0.25 in LaLiga. These numbers did not change much in either league in the five years after VAR was introduced.

2 Refereeing without Review

A player (P) in a football match decides between a rough but legally clean play (action C) and an illegal foul play (action F). After P has made the choice, a referee (R) makes their observation of whether a foul has been committed, and decides between blowing the whistle (action W) and keeping the whistle in their pocket (action K). If P has committed a foul, R correctly detects the infraction with probability σ_F. If P has not, R wrongly detects an infraction with probability σ_C. To be precise, denote R's signal space as $\{S, X\}$, with signal S representing R detecting an infraction and X representing him not detecting an infraction. Then we have $\sigma_F = \Pr(S|F)$ and $\sigma_C = \Pr(S|C)$. Assume

$$1 > \sigma_F > \frac{1}{2} > \sigma_C > 0.$$

When R gets their decision right – blow the whistle after a foul, or keep the whistle in their pocket after no foul – R's realized "decision loss" is normalized to 0. When R blows the whistle after no foul, R's decision loss is $\lambda_W > 0$; and when R keeps the whistle in the pocket after a foul, R's decision loss is $\lambda_K > 0$. When P gets away with a foul – when they choose F and R chooses K – P's payoff ν_{FK} is the highest; when they are wrongly whistled – when they choose C and R chooses W – their payoff ν_{CW} is the lowest. In between ν_{FK} and ν_{CW}, P's payoff ν_{CK} when they choose C and R chooses K is higher than their payoff ν_{FW} when they choose F and R chooses W. That is, we assume

$$\nu_{FK} > \nu_{CK} > \nu_{FW} > \nu_{CW}.$$

Throughout we assume that both P and R care only about their respective expected payoff. For example, if P chooses F and R chooses W when a foul is detected and K otherwise, then P's expected payoff is $\sigma_F \cdot \nu_{FW} + (1 - \sigma_F) \cdot \nu_{FK}$, and R's decision loss is $\sigma_F \cdot 0 + (1 - \sigma_F) \cdot \lambda_K$.

Comments on the model follow.

- We have assumed that the player has perfect control of their actions, which is arguably not true in reality. A player intending to make a rough play may end up committing a foul, and conversely, they may not always be able to execute a foul play. This assumption matters to the analysis below, as in equilibrium the referee "knows" the strategy of the player. This allows referee to make the call irrespective of their own signal.

- The player is assumed to prefer an unpunished foul play to a rough play so long as it is not mistaken for a foul by the referee. That is, we assume that the player is known to be "dirty" by the referee. If the player is instead known to be fair, the analysis is straightforward but uninteresting.

- In the model the referee cares only about making the right call. In practice referees are trained professionals, and it is natural that their interests of avoiding making a wrongful foul call and avoiding letting players get away with fouls are aligned with those of the sport.

We have assumed that $\sigma_F < 1$ and $\sigma_C > 0$. If instead $\sigma_F = 1$ and $\sigma_C = 0$, so that P's action is perfectly observed by Referee, there is a unique subgame perfect equilibrium in this sequential-move game with observed actions. The equilibrium strategy of P is C, and R's strategy is K if they observe C and W otherwise. This is good for the game – it is clean because there is no foul by P, and tight because R keeps their whistle in the pocket.

The outcome of the "good equilibrium" – P choosing C and R choosing K – does not exist because we have assumed that R does not have perfect competence. The best response for R to P never committing a foul is never blowing the whistle regardless of R's signal, but the best response to that for P is always committing the foul. This striking result illustrates that a slight imperfection human abilities can cause a drastic change in the outcome in a strategic situation.

Instead of the good equilibrium, a "bad equilibrium" for the game – P commits a foul

whenever an opportunity arises and R blows a whistle without relying on their own signal – always exists. The best response for R to P always committing a foul is always blowing the whistle regardless R's signal, and the best response to that for P is always committing the foul. That the bad equilibrium always exists further illustrates the fragility of the good outcome supported a perfect referee.

Now suppose that R adopts the "straightforward strategy" of keeping the whistle in their pocket when they do not detect an infraction and blowing the whistle when they do. P's best response to the straightforward strategy is F if

$$\Delta \equiv \sigma_C \nu_{CW} + (1 - \sigma_C)\nu_{CK} - (\sigma_F \nu_{FW} + (1 - \sigma_F)\nu_{FK})$$

is strictly negative. Indeed, if $\Delta < 0$, the bad equilibrium is the unique equilibrium. Let β be the belief of R that P has chosen F. R weakly prefers W to K if

$$(1 - \beta)\lambda_W \leq \beta \lambda_K. \tag{1}$$

Since $\sigma_F > \sigma_C$, R believes that P is more likely to have chosen F after signal S than after X. As a result, if R weakly prefers W to K after X, then they strictly prefer W after S; conversely, if R weakly prefers K to W at X, then they strictly prefer K after S.

Lemma 1. *The bad equilibrium always exists in Refereeing without Review. Further, there is no other equilibrium outcome if* $\Delta < 0$.

Proof. We only need to show that there does not exist an equilibrium in which P mixes between F and C. Since $\Delta < 0$ and $\nu_{FW} > \nu_{CW}$, P will never be indifferent between F and W for any probability of R choosing W at signal X when R chooses W with probability one at signal S. Similarly, since $\Delta < 0$ and $\nu_{FK} > \nu_{CK}$, if R always chooses K at signal X then as R's probability of choosing W at signal S decreases from 1 to 0, P will never be indifferent between F and W. □

For $\Delta < 0$, R's signal needs to be sufficiently inaccurate both in detecting a foul and in exonerating a rough play. Since $\nu_{FW} < \nu_{CK}$, if σ_F is close to 1 and σ_C is close to 0, condition $\Delta < 0$ fails. To make things interesting, we maintain the following assumption throughout the paper.

Assumption 1. $\Delta > 0$.

Under Assumption 1, there are two equilibria in which P randomizes between F and C. In one of them, R adopts a "harsh strategy" of choosing W at signal S and randomizing between W and K at signal X; in the other, R adopts the "lenient strategy" of choosing K at signal X and randomizing between W and K at signal S. We refer to the first one as the "harsh equilibrium" and the second as the "lenient equilibrium." These will be the two equilibria we focus on throughout the paper.

Proposition 1. *Under Assumption 1, the harsh equilibrium and the lenient equilibrium coexist in Refereeing without Review.*

Proof. For the harsh equilibrium, since $\nu_{FW} > \nu_{CW}$, let $W_h \in (0,1)$ be given by

$$W_h = \frac{\Delta}{\Delta + \nu_{FW} - \nu_{CW}}.$$

The equilibrium probability of R choosing W at signal X is W_h, which ensures that P is indifferent between F and C. Since $\lambda_W, \lambda_K > 0$, there is a unique $F_h \in (0,1)$ such that

$$F_h(1-\sigma_F)\lambda_K = (1-F_h)(1-\sigma_C)\lambda_W.$$

The equilibrium probability of P choosing F is F_h, which ensures that R is indifferent between W and K at signal X. Similarly, for the lenient equilibrium, let $W_l \in (0,1)$ be given by

$$W_l = \frac{\nu_{FK} - \nu_{CK}}{\nu_{FK} - \nu_{CK} + \Delta},$$

and let $F_l \in (0,1)$ be the unique value satisfying

$$F_l \sigma_F \lambda_K = (1 - F_l)\sigma_C \lambda_W.$$

In the lenient equilibrium, R chooses W with probability W_l at signal S, and P chooses F with probability F_l. The proposition follows immediately. □

Now, we compare the two equilibria. To begin, consider how happens to the two equilibria when σ_F increases and σ_C decreases – R's signal becomes more accurate – starting from the knife-edge case of $\Delta = 0$. At the starting point, the equilibrium strategy of R is the same straightforward strategy, with $W_h = 0$ and $W_l = 1$ respectively.

For the harsh equilibrium, as σ_F increases and σ_C decreases, for any probability that R chooses K at signal X, P's payoff from F decreases and their payoff from C increases. As a result, W_h increases – R becomes harsher – to restore P's indifference condition between F and C. At the same time, for any probability of P choosing F, as σ_F increases and σ_C decreases, for R W is more likely to be wrong and K is more likely to be correct at signal X. This implies that F_h increases – P fouls more often – to restore R's indifference between W and K so as to randomize between them. As σ_F goes to 1 and σ_C goes to 0, F_h goes to 1 and W_h becomes irrelevant. The harsh equilibrium outcome converges to the bad equilibrium of P always fouls and R always blows the whistle.

In contrast, for the lenient equilibrium, as σ_F increases and σ_C decreases, R becomes more lenient as the probability W_l that R chooses W at signal S increases, and P fouls less often as the probability F_l that P chooses C decreases. As σ_F goes to 1 and σ_C goes to 0, F_l goes to 0 and W_l becomes irrelevant. The lenient equilibrium outcome converges to the good equilibrium of P always chooses C and R always keeps the whistle in their pocket.

We can make the above discussion more precise. Recall that R is indifferent between W and K at signal X in the harsh equilibrium while they are indifferent at signal S in the lenient equilibrium. This is possible only if P chooses F more often in the harsh equi-

librium than in the lenient equilibrium. To see this, note that from the two indifference conditions that define F_h and F_l we have

$$\frac{F_h}{1-F_h}\frac{1-\sigma_F}{1-\sigma_C} = \frac{\lambda_W}{\lambda_K} = \frac{F_l}{1-F_l}\frac{\sigma_F}{\sigma_C}.$$

Since $\sigma_F > \sigma_C$, we immediately have $F_h > F_l$. At the same time, R chooses W more often in the harsh equilibrium than in the lenient equilibrium. That is,

$$1-(F_h(1-\sigma_F)+(1-F_h)(1-\sigma_C))(1-W_h) > (F_l\sigma_F+(1-F_l)\sigma_C)W_l.$$

To show this, note that since $W_h > 0$, we have

$$(F_h(1-\sigma_F)+(1-F_h)(1-\sigma_C))(1-W_h) < (1-F_h)(\sigma_F-\sigma_C)+1-\sigma_F.$$

Similarly, since $W_l < 1$ we have

$$(F_l\sigma_F+(1-F_l)\sigma_C)W_l < F_l(\sigma_F-\sigma_C)+\sigma_C.$$

Since $\sigma_F > \sigma_C$, by the above two inequalities, the claim that R chooses W more often in the harsh equilibrium follows from $F_h > F_l$ immediately. We summarize the above discussion in the corollary below.

Corollary 1. *Under Assumption 1, P commits fouls and R whistles more often in the harsh equilibrium than in the lenient equilibrium.*

Corollary 1 does not necessarily establish that the harsh equilibrium is "worse" for the game than the lenient equilibrium. To make the comparison clear, we need to compare the loss resulting of whistling a rough play and the loss of letting a foul play go unpunished. To the extent that R is chosen to represent the interests of the game, we can use λ_W and λ_K to make the comparison of the two losses. This leads to comparing R's decision loss

between the two equilibria.

In the harsh equilibrium, R's indifference condition between W and K at signal X means that their expected decision loss can be compute by always choosing W. This gives R's decision loss in the harsh equilibrium as:

$$(1 - F_h)\lambda_W = \frac{(1 - \sigma_F)\lambda_K \lambda_W}{(1 - \sigma_F)\lambda_K + (1 - \sigma_C)\lambda_W},$$

where the equality follows again from their indifference condition between W and K at signal X. Similarly, R's decision loss at the lenient equilibrium is

$$F_l \lambda_K = \frac{\sigma_C \lambda_W \lambda_K}{\sigma_C \lambda_W + \sigma_F \lambda_K}.$$

Comparing the above two expressions, we have the following.

Corollary 2. *Under Assumption 1, the expected decision loss is greater in the harsh equilibrium than in the lenient equilibrium if and only if*

$$\sigma_C \lambda_W > (1 - \sigma_F)\lambda_K. \tag{2}$$

The left-hand side of the above condition (2) is the expected decision loss to R who adopts the straightforward strategy against P choosing C, while the right-hand side is the expected loss against P choosing F. The indifference between W and K at signal X in the harsh equilibrium and at signal S in the lenient equilibrium imply that we can use the straightforward strategy to compare the expected decision losses between the harsh equilibrium and the lenient equilibrium.

Combining Corollary 2 with Corollary 1, we get a clear sense that from the game's perspective the harsh equilibrium is "worse" than the lenient equilibrium. The expected decision loss in the harsh equilibrium can be higher or lower than that in the lenient equilibrium depending on whether (2) holds or not, but P commits a foul more often and R

makes a foul call more often. Indeed, we have observed that when R's signals become very accurate, the harsh equilibrium converges to the bad equilibrium and the lenient equilibrium converges to the good equilibrium. Of course in both the bad equilibrium and the good equilibrium, the decision loss to R is zero. Although in this paper we do not explicitly model the perspective of sports or spectators, to the extent that the bad equilibrium is "worse" than the good equilibrium, we can conclude that the harsh equilibrium is worse than the lenient equilibrium.

3 Refereeing with Review

Suppose that the review technology provides binary evidence about whether P has committed a foul or not. We use the signal space $\{S^*, X^*\}$ for the review technology – S^* is review's signal for detecting a foul and X^* is the review's signal for no foul – and assume that the review signal and R's signal are conditionally independent. Denote $\Pr(S^*|F) = \sigma_F^*$ and $\Pr(S^*|C) = \sigma_C^*$, and assume

$$1 > \sigma_F^* > \sigma_F > \frac{1}{2} > \sigma_C > \sigma_C^* > 0.$$

The review technology is superior to the ability of Referee, but is not flawless.

When R's call is not reviewed, it is the final decision. The payoff R is as given in the previous section. When R's call is reviewed, we consider two scenarios regarding the final decision: in "Refereeing with Review Decision," the final decision is made according to the review signal – W if S^* and K if X^* – and in "Refereeing with Review Recommendation," the final decision is made by Refereeing after re-evaluation of their decision in light of the review signal. In either scenario, we append their decision loss with "reputation loss" of having their decision being overturned by the review. Let $\mu_W > 0$ be the additional loss to R when their chooses W and the final decision is W, and $\mu_K \geq 0$ be the reputation loss when they choose K and the review signal is S^*. The reputation loss is normalized to 0 when R's initial decision coincides with the final decision. For example,

when P chooses F and R chooses W, the latter's total payoff loss is 0 if either there is no review or there is a review but the final decision remains W, and is $\lambda_K + \mu_W$ if there is a review and the final decision is K; when P chooses C and R chooses W, R's realized total payoff loss is λ_W if either there is no review or there is a review but the final decision remains W, and is μ_W if there is a review and the final decision is K.

We continue to assume that P cares only about the final decision. Regardless of whether R is operating under a review technology, and regardless of whether the final decision is made by R after a review or by the review signal, P's payoff depends only their choice between F and C, and the final decision W or K, as in the benchmark case of Refereeing without Review.

We assume that review is "automatic" in that a fixed fraction $\gamma \in (0, 1)$ of all decisions by Referee, whether it is W or K, is reviewed.

Comments on the model follow.

- In the model we have used γ to represent the probability of the review intervention. In practice, VAR interventions can be disruptive to the match, and are restricted to a few categories where the final decision is important to the match outcome. The value of γ should be interpreted as an average across these categories.

- The model of Refereeing with Review Decision applies to categories of VAR interventions where video evidence is incontrovertible, while the model of Refereeing with Review Recommendation applies to categories where video evidence is open to interpretations.

- We have labeled the reputation losses μ_W and μ_K when the match referee's initial decisions of W and K are overturned as reputation losses. This is a reduced-form way of capturing the idea that challenges to decision-making authority are costly to the principal. In the context of sports matches, when initial decisions are frequently overturned upon review, the match referee can appear to be incompetent. A match

referee can have decision losses λ_W and λ_K that are representative of the best interests of the sport, but if their decisions are repeatedly overturned, they can lose control of the match.

3.1 Refereeing with Review Decision

First, we consider the case of Refereeing with Review Decision.

To examine R's choice between W and K, let β be R's belief that P has chosen F. Since R's decision is automatically reviewed, their expected loss from choosing W is

$$\beta\gamma(1-\sigma_F^*)(\lambda_K+\mu_W)+(1-\beta)((1-\gamma)\lambda_W+\gamma(\sigma_C^*\lambda_W+(1-\sigma_C^*)\mu_W)),$$

and the expected loss from choosing K is

$$\beta((1-\gamma)\lambda_K+\gamma(\sigma_F^*\mu_K+(1-\sigma_F^*)\lambda_K))+(1-\beta)\gamma\sigma_C^*(\lambda_W+\mu_K).$$

For notational brevity, define

$$\Delta_C^* \equiv (1-\sigma_C^*)\mu_W-\sigma_C^*\mu_K>0,$$

and

$$\Delta_F^* \equiv \sigma_F^*\mu_K-(1-\sigma_F^*)\mu_W.$$

R weakly prefers W to K if

$$(1-\beta)((1-\gamma)\lambda_W+\gamma\Delta_C^*) \leq \beta((1-\gamma)\lambda_K+\gamma\Delta_F^*). \tag{3}$$

If $\gamma=0$, then the above condition coincides what we would have in the benchmark case of Referring without Review. If $\gamma=1$, condition (3) does not depend on R's decision losses λ_W and λ_K, because the final decision is determined by the review signal regardless of R's decision. In this case, condition (3) depends on their reputation loss from having their

decision overturned by the review: Δ_C^* is the expected reputation loss of having the wrong initial decision overturned the review signal relative to the loss of having the correct initial decision overturned when P has chosen C, while Δ_F^* is the expected relative reputation loss when P has chosen F. We assume that R's concerns for having their decision overturned by the review are sufficiently balanced relative to accuracy of the review signal, so that the two expected relative expected losses are positive.

Assumption 2. $\Delta_C^* > 0$ *and* $\Delta_F^* > 0$.

Assumption 2 is trivially satisfied if $\mu_W = \mu_K$. Under Assumption 2, the comparison between R's choice between W and F under signal S versus under X is qualitatively similar to the condition in the benchmark case of Referring without VAR regardless of the value of γ.

Now, we examine P's choice between F and C. Suppose that R adopts the straightforward strategy of choosing W at signal S and K at signal X. Since R's decision is automatically reviewed and the review signal is final, P's expected payoff from F is

$$(1-\gamma)(\sigma_F \nu_{FW} + (1-\sigma_F)\nu_{FK}) + \gamma(\sigma_F^* \nu_{FW} + (1-\sigma_F^*)\nu_{FK}),$$

and the expected payoff from choosing C is

$$(1-\gamma)(\sigma_C \nu_{CW} + (1-\sigma_C)\nu_{CK}) + \gamma(\sigma_C^* \nu_{CW} + (1-\sigma_C^*)\nu_{CK}).$$

For notational brevity, define

$$\Delta^* \equiv \sigma_C^* \nu_{CW} + (1-\sigma_C^*)\nu_{CK} - (\sigma_F^* \nu_{FW} + (1-\sigma_F^*)\nu_{FK})$$

as P's payoff difference between choosing C and choosing F against the review decision.

Since $\sigma_F < \sigma_F^*$ and $\nu_{FW} < \nu_{FK}$, and since $\sigma_C > \sigma_C^*$ and $\nu_{CW} < \nu_{CK}$, we have

$$\Delta^* > \Delta,$$

which is strictly positive by Assumption 1. Thus, C is the best response by P to R's straightforward strategy.

We continue to refer to a harsh equilibrium where R adopts a harsh strategy of choosing W at signal S and mixing between W and K at signal X, a lenient equilibrium where R adopts a lenient strategy of choosing K at signal X and mixing between K and W at signal S. We have the following result.

Proposition 2. *Under Assumption 1 and Assumption 2, in Refereeing with Review Decision, there is a harsh equilibrium if and only if*

$$(1-\gamma)(\nu_{FW} - \nu_{CW}) > \gamma\Delta^*, \tag{4}$$

and there is a lenient equilibrium if and only if

$$(1-\gamma)(\nu_{FK} - \nu_{CK}) > \gamma\Delta^*. \tag{5}$$

Proof. Consider the harsh equilibrium first. Let $F_h^* \in (0,1)$ be the unique probability that P chooses F that satisfies

$$\frac{F_h^*}{1-F_h^*}\,\frac{1-\sigma_F}{1-\sigma_C} = \frac{(1-\gamma)\lambda_W + \gamma\Delta_C^*}{(1-\gamma)\lambda_K + \gamma\Delta_F^*},$$

so that (3) holds as an equality for the likelihood ratio $\beta/(1-\beta)$ after R's signal X. When (4) is satisfied, there is a unique probability W_h^* that R chooses W at signal X that makes P indifferent, given by

$$W_h^* = \frac{(1-\gamma)\Delta + \gamma\Delta^*}{(1-\gamma)(\Delta + \nu_{FW} - \nu_{CW})}.$$

In the lenient equilibrium, let $F_l^* \in (0,1)$ be the unique probability that P chooses F that satisfies

$$\frac{F_l^*}{1-F_l^*}\,\frac{\sigma_F}{\sigma_C} = \frac{(1-\gamma)\lambda_W + \gamma\Delta_C^*}{(1-\gamma)\lambda_K + \gamma\Delta_F^*}.$$

When (5) is satisfied, there is a unique probability W_l^* that R chooses W at signal S that makes P indifferent, given by

$$W_l^* = \frac{(1-\gamma)(\nu_{FK} - \nu_{CK}) - \gamma\Delta^*}{(1-\gamma)(\nu_{FK} - \nu_{CK} + \Delta)}.$$

The proposition follows immediately. □

Conditions (4) and (5) are both satisfied if the value of γ is sufficient low, that is, if the review rarely intervenes with R's initial decision.

Now we examine how the review affects the harsh equilibrium and the lenient equilibrium. This is just comparative statics with respect to an increase in γ. From the indifference conditions of R that define the harsh equilibrium and lenient equilibrium, we see immediately that an increase in γ makes the harsh equilibrium harsher, by increasing W_h^*, and the lenient equilibrium more lenient, by decreasing W_l^*. The intuition is straightforward. Take the harsh equilibrium for example. An increase in γ means that P faces a greater chance that R's call is reviewed. In a harsh equilibrium R chooses K with a positive probability only when their signal is X, while in contrast no foul is called by the review whenever the signal is X^*. For any probability that R chooses W at signal X, P faces less incentive to choose F as opposed to C. To restore P's indifference, R has to choose W with a greater probability at signal X.

Next, we consider the effect on the probability of P choosing W. Straightforward calculations show that an increase in γ increases F_h^* (decreases F_l^*) if and only if

$$\frac{\Delta_C^*}{\Delta_F^*} > \frac{\lambda_W}{\lambda_K}. \tag{6}$$

To understand the above condition, note it is R's indifference condition between W and K that determine P's choice between F and C. From the indifference condition that determine F_h^* and F_l^* respectively, as γ increases, R's incentive to choose W as opposed to K, shifts from the ratio of decision losses λ_W/λ_K – wrongfully whistling a rough but legal play to letting a foul play unpunished – to the ratio of expected differences in reputation losses Δ_C^*/Δ_F^* – having a wrong call for foul overturned relative to having a correct no-call over turned to having a wrong no-call overturned relative to having a correct foul call overturned. In Refereeing without Review – $\gamma = 0$ – a greater ratio of decision losses λ_W/λ_K means that R is more concerned with a wrong foul call than with a wrong no-call, and therefore has less incentive to choose W, requiring P to choose F more often to restore R's indifference in both the harsh and the lenient equilibrium. If all final decisions are made by the review – $\gamma = 1$ – a greater ratio of expected differences in reputation losses Δ_C^*/Δ_F^* also reduces R's incentive to choose W, and thus similarly requiring P to choose F more often, but through a mechanism. For any fixed distribution of the review signal, the ratio of expected differences in reputation losses Δ_C^*/Δ_F^* increases with the ratio of reputation losses μ_W/μ_K, because R is relatively more concerned with having their foul call overturned by the review. For fixed ratio of reputation losses μ_W/μ_K, the ratio of expected differences in reputation losses Δ_C^*/Δ_F^* increases as the review signal becomes more accurate in exonerating a rough but legal play P, that is, σ_C^* decreases, and decreases as the review signal becomes more accurate in detecting an illegal foul play, that is, σ_F^* increases.

Corollary 3. *Under Assumption 1 and Assumption 2, increasing adoption of the review technology makes R choose W more often in the harsh equilibrium and less often in the lenient equilibrium, and makes P choose F more often in both harsh equilibrium and the lenient equilibrium if and only if (6) holds.*

Unlike in the benchmark case of Refereeing without VAR, the bad equilibrium may not exist and the good equilibrium may exist.

Corollary 4. *In Refereeing with Review Decision, the bad equilibrium exists if and only if (4) holds, and the good equilibrium exists if and only if (5) fails.*

When γ is close to 1, both (4) and (5) fail. By Proposition 2 and Corollary 4, the good equilibrium exists and there is no other equilibrium. In this case, R's initial decision becomes irrelevant to the final decision. Unlike R's decision, which is strategic and responds to changes in P's choice between F and C, the review decision is automatic and does not respond to P's choice. This explains why the good equilibrium exists in Refereeing with Review Decision, but not in Refereeing without Review.

3.2 Refereeing with Review Recommendation

To begin, we first consider R's re-evaluation of their decision in light the review signal. Let β' be R's belief that P has chosen F after receiving the review signal. If R's initial decision was W, then it is optimal to switch to K if

$$\beta' \lambda_K + \mu_W \leq (1 - \beta')\lambda_W. \tag{7}$$

If R's initial decision was K, then it is optimal to switch to W if

$$(1 - \beta')\lambda_W + \mu_K \leq \beta' \lambda_K. \tag{8}$$

Without the reputation losses, i.e., if $\mu_W = \mu_K = 0$, the optimal decision for R upon review is W if the likelihood ratio $\beta'/(1-\beta')$ is greater than λ_W/λ_K and K if otherwise. In the extreme case where R's reputation losses from being overturned by the review are greater than the corresponding decision losses, the review technology becomes useless. The following result follows from conditions (7) and (8) immediately.

Lemma 2. *In Refereeing with Review Recommendation, the harsh equilibrium and the lenient equilibrium in Refereeing without Review remain equilibria with R never switching their decision upon review, if $\mu_W \geq \lambda_W$ and $\mu_K \geq \lambda_K$.*

From now on, we make the following assumption to make things more interesting.

Assumption 3. *$\mu_W < \lambda_W$ and $\mu_K < \lambda_K$.*

Under Assumption 3, reputation losses create an "authority inertia." When the likelihood ratio $\beta'/(1-\beta')$ lies between $(\lambda_W - \mu_W)/(\lambda_K + \mu_W)$ and $(\lambda_W + \mu_K)/(\lambda_K - \mu_K)$, R sticks to their wrong initial decision if it is W and $\beta'/(1-\beta') < \lambda_W/\lambda_K$, or if it is K and $\beta'/(1-\beta') > \lambda_W/\lambda_K$.

Let β be R's belief that P has chosen F after receiving their own signal. Since review signal S^* upgrades β to a higher belief that P has chosen F, and X^* downgrades it, if R weakly prefers to switch from W to K when the review signal is S^*, then they strictly prefer to switch after X^*. As a result, after choosing W initially, R anticipates three (pure) evaluation rules: stick to W regardless of the review signal; stick to W if the review signal is S^* and switches to K if X^*; switch to K regardless of the review signal. Similarly, after choosing K initially, R will stick to K regardless of the review signal, or stick to K if the review signal is X^* and switches to W if S^*; switch to W regardless of the review signal. We first rule out the re-evaluation rule of switching the initial decision regardless of the review signal.

Lemma 3. *There is no equilibrium in Refereeing with Review Recommendation in which R switches their initial decision regardless of the review signal.*

Proof. Suppose that R initially chooses W when they believe that with probability β P has chosen F. The expected loss from choosing W and then switching to K regardless of the review signal is

$$\beta\gamma(\lambda_K + \mu_W) + (1-\beta)((1-\gamma)\lambda_W + \gamma\mu_W).$$

The expected loss from choosing K and then stick to K regardless of the review signal is instead

$$\beta((1-\gamma)\lambda_K + \gamma\lambda_K).$$

Since R finds it optimal to switch to K after the review signal S^*, and since the updated belief β' after S^* that P has chosen F is greater than β, from (7) we have

$$(1-\beta)\lambda_W > \beta\lambda_K.$$

Thus, choosing W and then switching to K regardless of the review signal is strictly dominated by choosing K and sticking to K regardless of the review signal. The same argument rules out choosing K and then switching to W regardless of the review signal. □

In Refereeing with Review Recommendation, the possibility of re-evaluating their initial decision in light of the review signal complicates R's initial choice between W and K. We assume that the review signals are sufficiently accurate so that R will never stick to their initial decision no matter what that is and no matter what the review signal is.

Assumption 4.

$$\frac{\sigma_C^*}{\sigma_F^*}\,\frac{\lambda_W+\mu_K}{\lambda_K-\mu_K} < \frac{1-\sigma_C^*}{1-\sigma_F^*}\,\frac{\lambda_W-\mu_W}{\lambda_K+\mu_W}.$$

Under Assumption 4, when the likelihood ratio $\beta/(1-\beta)$ at R's initial decision falls on the interval given in the assumption, R's re-evaluation rule involves "double switching" – switch to K after choosing W when the review signal is X^*, and switch to W after choosing K when the review signal is S^*. When R's likelihood ratio $\beta/(1-\beta)$ at the initial decision is outside the interval, instead there is "single switching:" when it is above the interval, R switches to W after choosing K at the review signal S^* and sticks to W after choosing W regardless of the review signal; when the likelihood ratio is below the interval, R switches to K after choosing W at X^* and sticks to K after choosing K regardless of the review signal. If Assumption 4 fails, there is an interval of the likelihood ratio $\beta/(1-\beta)$ at R's initial decision under which R sticks to their initial decision regardless of the review signal. By ruling out such scenario, Assumption 4 strengthens Assumption 3.

Using Assumption 4, we can now examine R's initial decision between W and K. Let β

be the belief of R that P has chosen F at the initial decision. Suppose that R weakly prefers W to K at β. Under Assumption 4, we distinguish two polar opposite cases according to the likelihood ratio $\beta/(1-\beta)$. In the first case, $\beta/(1-\beta)$ is below the interval defined by Assumption 4, and R's re-evaluation rule involves a single switching to K after choosing W. Since R weakly prefers W to K at β,

$$\beta\gamma(1-\sigma_F^*)(\lambda_K+\mu_W)+(1-\beta)((1-\gamma)\lambda_W+\gamma(\sigma_C^*\lambda_W+(1-\sigma_C^*)\mu_W))\leq\beta\lambda_K.$$

We can rewrite the above as

$$\frac{\beta}{1-\beta}\geq\frac{(1-\gamma)\lambda_W+\gamma((1-\sigma_C^*)\mu_W+\sigma_C^*\lambda_W)}{(1-\gamma)\lambda_K+\gamma(\sigma_F^*\lambda_K-(1-\sigma_F^*)\mu_W)}, \tag{9}$$

because $\lambda_K > \mu_K$ by Assumption 3 and $\Delta_F^* > 0$ by Assumption 2 together ensure the denominator on the right-hand side of (9) is strictly positive regardless of the value of γ. In the second polar case, $\beta/(1-\beta)$ lies above the interval defined by Assumption 4. As in the first case, under Assumption 3 and Assumption 2, R weakly prefers W to K if and only if

$$\frac{\beta}{1-\beta}\geq\frac{(1-\gamma)\lambda_W+\gamma((1-\sigma_C^*)\lambda_W-\sigma_C^*\mu_K)}{(1-\gamma)\lambda_K+\gamma(\sigma_F^*\mu_K+(1-\sigma_F^*)\lambda_K)}. \tag{10}$$

We now make assumptions to rule out these two polar cases where R's re-evaluation rule involves single switching.

Assumption 5.

$$\frac{\sigma_C^*}{\sigma_F^*}\,\frac{\lambda_W+\mu_K}{\lambda_K-\mu_K}<\frac{(1-\gamma)\lambda_W+\gamma((1-\sigma_C^*)\mu_W+\sigma_C^*\lambda_W)}{(1-\gamma)\lambda_K+\gamma(\sigma_F^*\lambda_K-(1-\sigma_F^*)\mu_W)},$$

and

$$\frac{1-\sigma_C^*}{1-\sigma_F^*}\,\frac{\lambda_W-\mu_W}{\lambda_K+\mu_W}>\frac{(1-\gamma)\lambda_W+\gamma((1-\sigma_C^*)\lambda_W-\sigma_C^*\mu_K)}{(1-\gamma)\lambda_K+\gamma(\sigma_F^*\mu_K+(1-\sigma_F^*)\lambda_K)}.$$

Under Assumption 2, the two conditions in Assumption 5 are satisfied when σ_C^* is suf-

ficiently close to 0 and σ_F^* is sufficiently close to 1. By (9), the first condition in Assumption 5 implies when the likelihood ratio $\beta/(1-\beta)$ at R's initial decision is below the interval defined by Assumption 4, R strictly prefers K to W; by (10), the second condition implies when the likelihood ratio $\beta/(1-\beta)$ at R's initial decision is above the interval defined by Assumption 4, R strictly prefers W to K.

Now we consider the case where R's likelihood ratio $\beta/(1-\beta)$ at the initial decision lies in the interval defined by Assumption 4. R weakly prefers W to K at β if and only if (3) holds. There two subcases, one that corresponds a harsh equilibrium where (3) holds as an equality for the likelihood ratio $\beta/(1-\beta)$ after R's signal X, and the other that corresponds to a lenient equilibrium where (3) holds as an equality for the likelihood ratio $\beta/(1-\beta)$ after S. In the harsh equilibrium, the likelihood ratio $\beta/(1-\beta)$ after S either still lies in the interval defined by Assumption 4, in which case R strictly prefers W to K – anticipating switching after either W or K – by condition (3), or it lies above the interval, in which case R also prefers W to K – anticipating switching only after K – by condition (10) and Assumption 5. Symmetrically, in the lenient equilibrium, R strictly prefers K to W after signal X.

We can now provide conditions for the harsh and the lenient equilibrium in Refereeing with Review Recommendation to replicate their counterparts in Refereeing with Review Decision.

Proposition 3. *Under Assumptions 1 to 5, in Refereeing with Review Recommendation, the same harsh equilibrium as in Refereeing with Review Decision exists when condition (4) holds, and*

$$\frac{\sigma_F}{\sigma_C}\,\frac{1-\sigma_C}{1-\sigma_F}\,\frac{(1-\gamma)\lambda_W+\gamma\Delta_C^*}{(1-\gamma)\lambda_K+\gamma\Delta_F^*} < \frac{1-\sigma_C^*}{1-\sigma_F^*}\,\frac{\lambda_W-\mu_W}{\lambda_K+\mu_W}, \tag{11}$$

and the same lenient equilibrium as in Refereeing with Review Decision exists when condition (5) holds, and

$$\frac{1-\sigma_F}{1-\sigma_C}\,\frac{\sigma_C}{\sigma_F}\,\frac{(1-\gamma)\lambda_W+\gamma\Delta_C^*}{(1-\gamma)\lambda_K+\gamma\Delta_F^*} > \frac{\sigma_C^*}{\sigma_F^*}\,\frac{\lambda_W+\mu_K}{\lambda_K-\mu_K}. \tag{12}$$

Proof. Let $F_h^* \in (0,1)$ be the same probability that P chooses F in the harsh equilibrium in Refereeing with Review Decision. Then,

$$\frac{F_h^*}{1-F_h^*}\,\frac{\sigma_F}{\sigma_C} = \frac{\sigma_F}{\sigma_C}\,\frac{1-\sigma_C}{1-\sigma_F}\,\frac{(1-\gamma)\lambda_W + \gamma\Delta_C^*}{(1-\gamma)\lambda_K + \gamma\Delta_F^*}.$$

Compare the above to the upper-bound of the interval defined by Assumption 4. In the lenient equilibrium, let $F_l^* \in (0,1)$ be the same probability that P chooses F in the lenient equilibrium in Refereeing with Review Decision. Then,

$$\frac{F_l^*}{1-F_l^*}\,\frac{1-\sigma_F}{1-\sigma_C} = \frac{1-\sigma_F}{1-\sigma_C}\,\frac{\sigma_C}{\sigma_F}\,\frac{(1-\gamma)\lambda_W + \gamma\Delta_C^*}{(1-\gamma)\lambda_K + \gamma\Delta_F^*}.$$

Compare the above to the lower-bound of the interval defined by Assumption 4. The proposition follows immediately. □

The conditions for (11) and (12) are both satisfied, regardless of the values of σ_C and σ_F, and regardless of the value of γ, if σ_C^* is sufficiently close to 0 and σ_F^* close to 1. In this case, Refereeing with Review Recommendation and Refereeing with Review Decision are the same observationally in so far the harsh equilibrium and the lenient equilibrium are concerned. However, there is one critical difference. Instead of Corollary 4 in Refereeing with Review Decision, we have the following result:

Corollary 5. *Under Assumptions 1 to 5, in Refereeing with Review Recommendation, there does not exist the good equilibrium, and the bad equilibrium always exists.*

In contrast to Refereeing with Review Decision, the good equilibrium does not exist because when the probability that P chooses F is sufficiently close to 0, the likelihood ratio $\beta/(1-\beta)$ is below the interval defined by Assumption 4 even after R's signal S. This means that R strictly prefers K to W regardless of their own signal, and sticks to it regardless of the review signal after choosing K. But then P's best response is to choose F instead. The existence of the bad equilibrium follows from a similar logic, except that P's

best response to Refereeing always choosing W and sticking to W regardless of the review signal is F. In Refereeing without Review, the good equilibrium does not exist and the bad equilibrium always exists. Corollaries 5 and 4 further illustrate that challenge to decision-making authority can change what decisions are made in a strategic environment.

4 Concluding Remarks

From perspectives of sports, the ideal outcome is that players never commit fouls and match referees never have to interrupt the match to discipline foul plays. This is however inconsistent with incentives of players and referees, so long as referees do not have perfect observations of play. A review technology is naturally intended to help match referees make better decisions. However, when referees care about their reputation of being competent, a review technology presents a real-time challenge to their decision-making authority. Review can make matters worse from perspectives of sports, by simultaneously inducing match referees to make more foul calls and players to commit more fouls. Unless the decision-making authority is stripped from match referees, even a near perfect review technology would not be able to bring match outcome close to the ideal.

Our model of authority under challenge is static. We have taken a reduced-form approach by directly assuming that a principal with decision-making authority suffers from a reputation loss whenever their decision is overturned by independent review. Such reputation losses can arise "endogenously" in a dynamic model where having an initial decision overturned upon review today causes more challenges to the principal in the future. We leave dynamic analysis of authority under challenge for future research.

References

[1] P. Aghion and J. Tirole, Formal and real authority in organizations, The Journal of Political Economy, 105(1), pp. 1-29, 1997.

[2] B. Holmstrom, On the theory of delegation, in: M. Boyer, R. Kihlstrom (Eds.), Bayesian Models in Economic Theory, North-Holland, New York, 1984.