Implicit Institutional Incentives and Individual Decisions: Causal Inference with Deep Learning Models

By

Stefano Cabras[a] and J.D. Tena[b]

Abstract

Reward schemes guide choices. However, they are not necessarily presented as a collection of written incentive mechanisms but as complex and implicit cues. This paper proposes a methodology to identify tacit organizational incentives based on direct observations of institutional reactions to operational decisions. Football data provides a laboratory for this analysis as referee decisions, and their consequences are subject to public scrutiny. This allows estimating the length of time between referee appointments in Spanish football as a function of referee decisions in the most recent match. A deep learning model is instrumental in this analysis as it allows controlling for many potential confounders. Our results are consistent with the presence of institutional incentives for the referee to take gradual (instead of drastic) decisions to send off home team players and deliver the game's expected outcome. Finally, we discuss the implications of these findings in organizations.

Keywords: Institutional incentives, Cognitive bias, Causal analysis; Deep-learning model; Causal machine learning.

[a] Universidad Carlos III de Madrid, Department of Statistics, C/ Madrid, 126 -28903 Getafe, Spain. Email: stefano.cabras@uc3m.es.

[b] Corresponding author. University of Liverpool, Management School, Liverpool, UK and University of Sassari and CRENoS, Department of Economics, Sassari, Italy. Chatham St, Liverpool L69 7ZH, UK. Email: jtena@liverpool.ac.uk. Phone: 07490181331.

1. Introduction

Unlike neoclassical economics, which conceive firms as perfect profit maximizers, neo-institutional theory focuses on the presence of persistent forces that shape organizational culture. Therefore (non-necessarily optimal) managerial decisions can be made with a routine-based logic of behavior (Gavetti, 2005) which is likely to perpetuate despite their high operational costs (Park & Patterson, 2021).

However, many of the unwritten rules could not be necessarily perceived by individuals. Moreover, even if some workers could become aware of implicit institutional pressures, it is challenging to identify the implications of not following them, given that organizational rewards or punishments occur in conjunction with many other decisions and events. Thus, at least two practical problems prevent the empirical identification of implicit institutional incentives and the estimation of their effects on workers. First, the "gold standard" of scientific research (Varian, 2016b), randomized control trials, is not feasible in settings where the object of study is the institution rather than the individual. Furthermore, it is challenging to find actual examples for a given organization where one can observe the consequences of different decisions for a sufficiently large number of cases to make the estimation possible. Second, even if this observational sample exists for causal analysis, we require an analytical tool that considers the complex interactions of different factors driving each operation.

Decisions made by soccer referees provide a laboratory for this type of analysis. Thus, unlike participants in experiments, soccer referees are professionals in high-stakes situations whose operations are recorded and scrutinized. Moreover, soccer referees have a high degree of

discretion in decision-making, such as whether particular actions are fouls. As association football is a low-scoring game, any decision can substantially impact the match's final score.

In this paper, we explore the incentive scheme of referees in the top tier of the Spanish football league (Primera División) by estimating how a referee's decision in a match could affect the number of rounds a referee must wait to be appointed for the next game. Of course, being prevented from refereeing again for a long time is not the only punishment that a referee may suffer. Still, it is the only one that can be consistently and publicly-observed through all our analysis periods. Traditionally, 'la nevera', or 'the fridge' in English, is an expression in the Spanish vernacular to apply to a referee punished by not working for several rounds because of serious mistakes in his last match. The institution responsible for this decision, the Spanish Football Federation, does not report which referees are in 'the fridge'. Still, its existence itself and how it is affected by different referees' decisions is an empirical question. We analyze how decisions on penalties and the number of sendings off due to yellow, red cards, and penalties affect the next referee appointment's length of time. Moreover, we study if their consequences are significantly different when they favor the home or the away team and incentivize them to deliver expected results. This is relevant to get evidence on whether the incentive scheme offered prevents or encourages irrational decisions.

Our empirical approach relies on comparing the observed length of time between referee appointments with the one predicted under a hypothetical situation if the referee had made a different decision. This prediction is obtained using a deep-learning (DL) model, a subset of machine learning, see Schmidhuber (2015). The last years have witnessed remarkable progress in using machine learning techniques for causal inference (Athey & Imbens, 2015, Zhao & Hastie, 2019). In particular, Zhao and Hastie (2019) indicate that controlling for all factors that impact treatment and response variables allow for causal claims. DL models are

particularly well suited for this purpose. They set a higher bar for the possible omitted regressors by allowing treatment and confounder variables to interact at different layers, which can be used to construct meaningful counterfactual scenarios. The use of DL models for causal analysis has been motivated by Luo et al. (2020). Moreover, this approach has been increasingly employed in empirical applications, see Magazzino *et al.* (2021) and Liu *et al.* (2021), among many others. However, we also use another methodological design, Causal Forest (Athey & Wager, 2019; Athey et al., 2019), finding qualitatively similar results.

The little requirement for human intervention of machine learning, particularly DL, is pertinent in this research. Thus, although variables in the model and the number of layers are analyst decisions, the algorithm determines the nature of the interactions between covariates. Silberzahn and Uhlmann (2015) report the results of a crowdsourcing analysis where different researchers were supplied with the same dataset, asking them to provide an empirical estimation for a specific answer on the racial bias for football referees finding substantial differences in their responses. Here, the machine learning algorithm decides DL specifications using an objective model specification approach. However, this does not imply the analyst does not play a significant role in the research process. Shrestha et al. (2021) highlight that researchers, not data, generate theory by interpreting the outcomes of machine learning algorithms. This consideration affects the act of reasoning in the present study. Thus, rather than deducting theoretical applications from known axioms, we follow an inductive approach by exploring possible explanations for estimated data patterns (Shrestha *et al.*, 2021; Choudhury *et al.*, 2021)

This research is related to three different strands of literature. First, it contributes to management research by estimating implicit incentives faced by decision-makers. Employment practices such as, for example, reward management and promotion constitute fundamental elements of management strategy. This paper shows that referees are nudged to

follow different forms of anchoring bias that punish them for taking drastic decisions to send off players or favor the underdog team. Second, this research is also linked to other social science attempts to model systematic policy decisions across different policy stance periods, such as Taylor rules (Zhang *et al.*, 2022) and fiscal policies (Larch *et al.*, 2021). We estimate the factors underlying political decisions in a more unconventional setting using an estimation technique that takes into account, among many other factors, time-varying (seasons) and individual heterogeneity of decision-makers (referees) and games as well as the specific context of these decisions. Finally, this paper is also connected to recent contributions about the use of machine learning in management studies (Shrestha *et al.*, 2021, Choudhury *et al.*, 2021). While machine learning models are increasingly used for causal analysis in disciplines such as medicine (Liu *et al.*, 2021) or economics (Beloni *et al.*, 2013, Varian, 2016a), applications in management are still scarce.

The remainder of this paper is organized as follows. The following two sections discuss the contribution of the present analysis to the existing literature on reward systems and cognitive biases. Section 4 provides theoretical interpretations of expected results. Section 5 presents our data and the empirical approach used in the analysis. Estimation results are shown and discussed in the context of organization theory in Sections 6 and 7, respectively. Finally, Section 8 suggests some future lines of research.

2. Determinants and consequences of rewards in organizations

Setting a reward system that incentivizes employees to feel more committed to work and increase productivity is a central strategic element in most organizations. Many studies focus on how reward systems influence the behavior of organizations. For example, Baumann and Stieglitz (2014) employ an agent-based simulation model to show that firms can improve performance by offering low-powered rewards to employees to select and implement new

ideas. High-powered incentives could generate an excessive number of good ideas. Han et al. (2012) indicate that the effectiveness of incentive compensation to motivate managerial behavior depends positively on executives' core evaluation and firm performance. In a more recent contribution, Mitsuhashi and Nakamura (2022) employed a difference-in-difference design to study how incentive redesign triggers network changes in Japanese firms.

A different strand of the management literature studies the determinants of reward systems in organizations. A general approach is that the optimal reward design is contingent on the firm's industrial context and strategic orientation. Thus, early papers have already linked incentive pay plans with, for example, high technology (Balkin and Gomez-Mejia, 1987) and more diversified firms (Napier and Smith, 1987). Rajagopalan and Finkelstein (1992) use a sample of 50 utility firms over ten years to show that senior management reward system depends on strategic orientation and environmental change. Kroll et al. (1997) study how CEO rewards depend on acquiring firms' form of control, and Boyd and Salamin (2001) link the strategic reward system to the firm's strategy.

Despite the previous arguments, strategic decisions on rewards are not always objective and logical but may be based on heuristic arguments. For example, Zorn et al. (2019) analyze how directors are biased toward CEOs they hire by offering higher payment and job security. They tested this hypothesis using responses to a survey of MBA alumni from two universities. The study of Nair et al. (2021) illustrates how cognitive bias can influence the evaluations of CEO quality. They used a longitudinal sample of 112 male CEOs across 82 FTSE 100 firms to find a positive impact of CEO vocal masculinity on their compensation. In another relevant contribution, Shin and You (2022) find that the alignment of multiple attributes of individual directors increased the ability to decide on CEO dismissal when the

firm is underperforming relative to expectations in a longitudinal database of S&P 500 firms. This paper is also related to identifying the reward system in a company. However, using the unconventional professional sports setting, we contribute to this literature by providing empirical evidence of implicit institutional pressure guiding reward decisions. Our approach is based on an estimation of the observed reaction of an organization to individual operations. Another particular characteristic of our study is that we consider the whole organization's history. This is relevant as our interest is to estimate persistent institutional pressures instead of those associated with a particular event or business cycle.

3. Institutional cognitive bias and professional sport

Institutional strategy is often conceived as a dual process (Gavetti & Levinthal, 2000, Gavetti & Rivkin, 2007). One part occurs in the world of cognition, managers' mind, and comprises the mental process that constructs particular theories about the firm and its environment. The other regards the world of action, i.e., what the company does. The evolutionary model considers managers to be bounded rational (Simon, 1955). Limitations include the complexity of the problem, cognitive capability constraints, and time restrictions. As a result, implicit and repetitive rules rather than optimal solutions guide managers' behavior (Cohen et al., 1996). Organizational studies have identified cognition problems, for example, in the asymmetrical effect of positive and negative external evaluations (DesJardine & Bansal, 2019), adaptation to technological change (Eggers & Kaplan, 2019), or restrictions imposed by safe routines (Oliver et al., 2017). Moreover, organizational factors and managers' attention can moderate the bounded rationality problem (Gavetti, 2005, Eggers & Kaplan, 2019).

Unsurprisingly, the unconventional setting of professional sport, which requires quick decisions in high-stakes situations, is particularly well-adapted to identifying cognitive biases. Garicano *et al*. (2005) is the seminal analysis on football referee bias; they found

evidence of a tendency for referees in the Spanish football league to increase stoppage time in close games when the home team is trailing compared to when the home team is leading. There also exists evidence of fewer disciplinary sanctions (in terms of red and yellow cards) for home teams in the English Premier League (Dawson *et al*., 2007), the top tier of the Bundesliga and the English Premier League (Buraimo *et al*., 2010) and in European Cup matches (Dawson & Dobson, 2010).

The analysis of referee bias is not restricted to home advantage. Price and Wolfers (2010) find evidence in the American National Basketball Association League of referee preferences for players whose ethnicity is the same as the majority of the referee team, while results in Price *et al*. (2012) explore other types of biases such as referee predilection for close games and loser teams. The third example, Gallo *et al.* (2013), considers implicit discrimination against black African players in the English Premier League via the incidence of disciplinary measures.

A common thread in the presence of referee bias is the role that social pressure exerts on decisions. In the case of home advantage, pressure can be a function of attendance. For example, Garicano *et al*. (2005) and Pettersson-Lidbom and Priks (2010) find that a significant amount of home bias in the top tier of the Spanish and the Italian League respectively are influenced by the ratio of attendance to stadium capacity, while Buraimo *et al*. (2010) find it can be explained by the absence of running tracks in stadia, which dictates the proximity of spectators to the football pitch.

However, pressure is not solely a function of attendance. Social attention can also affect the decision-making process. Pope *et al*. (2018) replicate the analysis on racial bias by Price and Wolfers (2010) using more recent data finding that the effects are no longer significant when they consider the 2007-2010 period. The authors' interpretation is that increased awareness of racial discrimination in NBA refereeing was sufficient to eliminate racial discrimination.

Bryson *et al*. (2011) studied the impact of salary contracts on referees' decisions in English football's top two tiers. Unlike us, their focus is not on explaining incentives as a function of decisions but the other way around. They analyze the determinants of yellow and red cards per game, finding that they are more frequent earlier in the season and in big games. Their finding is particularly relevant to our paper. It suggests that round and information about the contenders should be included in our analysis to control for unobserved heterogeneity in the number of yellow and red cards and other referee's decisions.

However, there is a shortage of research on estimating the specific incentives referees face when making decisions. Organizational context is obtained from the aggregation of individual behavior and, therefore, is also likely to be affected by cognitive biases (Christensen et al., 2022). Price *et al*. (2012) discuss this issue in the specific context of professional sport. They hypothesize that referee preferences for the home team increase consumer satisfaction. A similar situation occurs with close games. They empirically tested this idea by estimating the impact of referee preferences for home teams, close games, and differences in winning percentages between home and away teams (Match-up Coefficients) on the probability that a referee is assigned to a playoff game which can be considered as a prominent and visible form of compensation. Only Match-up Coefficients turn out to be significant in that regression. This was interpreted as indirect evidence of the existence of incentives for bias.

Boeri and Severgnini (2011) study the referees' incentive scheme during the Calciopoli scandal affecting Italian football. They explain referees' allocation to different types of matches using a probit model using referee characteristics and their interaction with previous referees' involvement in match rigging as covariates. They find that past participation in rigged matches increases the probability of being allocated to important games and that career concerns are significant for match-fixing. Boeri and Severgnini's study is focused on

one particular episode in Italian footballing history. In contrast, rather than a specific episode, our interest concerns the entire history of Spanish football when investigating the incentive structures in place for referees. By doing that, we attempt to identify persistent and (probably unconscious) implicit organizational reward schemes instead of occasional and voluntary corruption participation.

4. Theoretical interpretation of expected results

The objective function of the Spanish Football League organizer could be contingent on factors such as referee identity, home and away teams identities (match), season, round, other referee decisions, etc. Moreover, each of these variables could interact in complex ways to affect the organizer's objective function. The central assumption in our analysis is that controlling for all these factors, a referee's decision on, for example, yellow or red cards and penalties should not be systematically penalized or rewarded unless there is an intention to guide these decisions in a particular direction. Therefore, we pay specific attention to the presence of biased organizational penalties associated with observed referee decisions. In this sense, this section does not provide general axioms as it typical in deductive research but some interpretation of the possible data patterns found by machine learning methods that can be replicable by other analysts (Shrestha et al., 2021, Choudhury et al., 2021)

An interesting question concerns the referee's preference for yellow and red cards. A red card is an abrupt decision that sends the player off the pitch for the rest of the match, while a yellow card is a warning that this decision can be made. Tversky and Kahneman (1992) indicate that individuals influenced by the anchoring heuristic will insufficiently and sluggishly adjust away from the anchor (anchoring-as-adjustment). Therefore, organizers could be affected by this bias under the following expected result:

- Expected result 1 (E1): A Spanish Federation affected by the anchoring bias would

incentivize referees to warn players with yellow cards instead of sending them off the pitch in an abrupt decision (red card).

As discussed in the previous section, there is consistent evidence (particularly for the Spanish League) of home referee bias regarding these decisions (Garicano *et al*., 2005). Therefore, other things equal, organizers trying to counteract this type of prejudice must penalize more severely favorable referee decisions to the home team than the away team. However, on the other hand, organizers themselves could also be affected by the pressure generated by home supporters that could bias their decisions. Thus, there is no clear theory on whether the Spanish Federation will punish more (less) decisions favoring the home (away) team. Therefore, we anticipate the following:

- Expected result 2 (E2): A Spanish Federation concerned with referee home bias will punish relatively more penalties, yellow cards, and red cards that favor the home team.

Another relevant assumption concerns the evaluation of the cost of making decisions. Ritov and Baron (1990) were the first to provide evidence of omission bias showing a vaccination scenario where many participants irrationally opted not to give vaccination shots to their children despite being far less risky than the disease itself. Evidence of propensity for inaction in other social contexts can also be found, for example, in Schweitzer (1994) and Baron and Ritov (2004). According to this, it is plausible that organizers could be more indulgent with referees who do not take action. We cannot estimate the impact of any referee inactions, but it is still possible to measure the effect of decisions regarding the model's variables. Accordingly

- Expected result 3 (E3): Compared to the option of doing nothing, referee actions regarding cards or penalties will be penalized.

A separate issue of concern relates to the incentives that referees can face to deliver an unexpected result. We can only speculate in this respect. Following arguments similar to those in Price *et al*. (2012), it is possible is that the Spanish Federation may encourage a certain amount of surprise in the outcomes of games as this maintains an interest in the competition. However, it is more plausible that rational organizers do not blame the referee if the game's outcome is consistent with previous expectations. Based on this, we await the following:

- Expected result 4 (E4): Referees will be penalized by surprising results that can call organizers' attention.

5. Data and Empirical Strategy

5.1. Data

The Spanish Primera Division is the top tier of Spanish football. The first edition of the two top tiers Spanish leagues took place in season 1928-29, and the tournament has continued through to the present day, except between 1936 and 1939, because of the Spanish Civil War. The competition has worked as a round-robin tournament where clubs are promoted and relegated based on performance.

Real Federacion Española de Futbol (RFEF) was the organiser of the Spanish Football League until 1984 when la Liga de Futbol Professional (LFS) was created and took over as the organiser of the competition. The LFP is part of the RFEF, although it is a separate legal entity. The body responsible for appointing referees to games is the Comité Tecnico de Arbitros (CTA), also known as Comité Nacional de Arbitros, a sub-division of the RFEF. Like football clubs, referees can be demoted to a lower division at the end of each season. Here, we focus our analysis on top-tier referees.

Throughout history, there have been 31 different presidents of the RFEF. The last one in our sample, Jose Maria Villar, has been the longest-lasting president in charge from 1988 to the end of our analysis period. The way to allocate referees has been affected by different policy stance periods but, with the exceptions of seasons 1953/54-1956/7, 1971/72 to 1975/76 and 1996/7 to 2004/05 when referees were randomly appointed, there has been some degree of discretion in these decisions. As it will become clear later, removing these years from the data sample under the DL model is unnecessary. The estimation strategy learns from the data and will not use variables (seasons in this case) if they do not contribute to the prediction of the response variable.

We collected match-level data for the whole history (from 1929 to 2017) of the top tier of the Spanish League from the database BDFUTBOL at the url: https://www.bdfutbol.com. For each game, the variable whose response we want to analyze is the number of *rounds* a football referee must wait until he referees the next football match (*time*). We are interested in cases with a larger number of rounds as small rounds between assignments are the norm. Therefore, we consider instances with more than two rounds that a referee must wait. Two essential features must be mentioned about this variable. Firstly, it is measured in terms of rounds rather than actual calendar weeks. Secondly, because of referees being demoted or retiring, this variable has missing values amounting to around 2% of the sample.

[INSERT FIGURE 1 AROUND HERE]

We use 1988, when J.M. Villar took over as president of the RFEF, to split the database into two periods that we denote as recent and earlier. Figure 1 shows the distribution of the response variable for the earlier and recent periods. In both cases, the mass of the distribution is concentrated in the first few rounds suggesting that it is typically expected that a referee will not have to wait for more than 5 rounds to be appointed again after a match. Waiting for

many rounds to be appointed again is a rare event. There are also significant differences between the two periods. The response variable's median value changes significantly from 4 rounds in the earlier period to 3 rounds in the recent period.

The covariates are the number of players sent off with two yellow cards for the home and away teams, *home2yellow* and *away2yellow*, respectively; similar variables are defined for the number of sent-off players with a red card, *homered* and *awayred*; the number of penalties in favor of the home and away teams, *homepen* and *awaypen*. We also include dummy variables to indicate the home and away team, the referee, the number of scored goals for the home and the away team, the outcome of the game, Recent period, round, and season. Finally, we also consider the Brier Score of the match. This was obtained by using the Elo ratings of the teams to specify ordered probit models estimated with an estimation window of 5 seasons. This model was used to obtain probabilities of home victories, draws, and away wins that were considered for computing the Brier Score of the match.

There is a total of 19,636 observations for 22 original variables (descriptive statistics are shown in Table 1), some of which are categorical with many levels. For example, the referee identity variable has 661 levels referring to an equal number of referees. In the DL model, each possible level of each original variable is separately considered. This amounts to 1,152 variables, which are allowed to interact in the DL model freely.

[INSERT TABLE 1 AROUND HERE]

5.2. Rationality behind our estimation analysis.

This study aims to estimate the causal effect for a referee in previous game i with respect to its action D_i on the time he has to wait for refereeing the next match Y_i, such that the Average Treatment Effect on for a referee in game i is defined as $ATE_i = \mathbf{E}_{\pi(Y_i|X_i=x_i,Data)}(y_i(D_i = d_o) - Y_i(D_i = d_c)|X_i = x_i)$, where $D_i = d_o$ is the observed action in previous game i and

$D_i = d_c$ represents the action he could have taken. For instance, $D_i = d_o$ could be that we have observed zero red cards, while $D_i = d_c$ represents the (*what-if*) counterfactual situation that must be predicted, e. g. what if on the previous match i, the referee would have shown two red cards. Furthermore, ATE is defined upon the expectation of the random variable Y_i conditional on the information set.

We identify the causal effect just described under the strong ignorability of treatment assignment (Rosenbaum & Rubin, 1983). Strong ignorabilty requires the fulfillment of two assumptions. First, the unconfoundness assumption requires that conditional on a set of observable variables, potential outcomes are independent of treatment assignment. Second, the overlap, or common support, assumption requires each observation to have a positive probability of receiving each treatment level. Thus, estimating these causal effects requires a model that permits the evaluation of the response variable for a rich collection of interactions between treatment and observable variables. Causal estimation is difficult if confounding variables do not affect the response variable in a linear way or if their distribution is very different across treatment groups.

The concept of strong ignorability mirrors the "backdoor criterion," which implies that adjusting for all factors that influence both treatment and response variables allows for causal interpretation (Zhao & Hastie, 2019). Thus, in our setting, this implies controlling for two groups of variables: (1) omitted variables and (2) dynamic endogeneity (Abdallah *et al.*, 2015). DL models deal with the first issue by accounting for all possible collected confounding variables (i. e., elements of vector X_i) along with their interactions in the predictive model for Y_i instead of including a subset of specific match or referee characteristics. Regarding dynamics effects, Abdallah *et al.* (2015) advise the inclusion in the model of the lagged level of the dependent variable. However, a lagged dependent variable is

a restrictive feature of all the interactions between rounds' and referees' identities already allowed in the DL model.

Like Hill (2011) and Hill and Su (2013), we do not follow the propensity score methodology by estimating two models, one for the assignment mechanism and one for the response surface. In particular, rather than conducting causal analysis based on a previous transformation of the data (Imbens, 2000) or a restricted matched sample (Rosenbaum & Rubin, 1983), we assess the response function by employing a statistical framework (DL) that allows for flexible interactions between the treatment variable and each covariance level. This translates the causal effect estimation problem into a problem of estimating a response surface. Therefore our estimation requires an adequate prediction model rather than an interpretable one, which would restrict the analyst to specify beforehand the relation between the response variable Y_i and the possible causes X_i (see Shmueli, 2010 for a discussion on this point).

5.3. The deep-learning predictive model

A DL model is a neural network with many layers of neurons (Schmidhuber 2015). DL refers mainly to an algorithmic approach rather than a specific probabilistic model, although both components are present in DL (see Breiman, 2001 for the merits of including both elements). Each neuron is a deterministic function such that two connected neurons correspond to a function of a function along with an associated weight w. Figure 2 shows the used neural network architecture. From left to right, we have the first input node made of 1151 columns regarding the match information, then layers of 50 all connected nodes in the following order: a dense layer, a normalization batch layer, a dense layer, a dropout layer of 40% and the output node made of 1 node which is the length of time to next match.

[INSERT FIGURE 2 AROUND HERE]

Essentially, for a response variable Y_i for referee i and a predictor variable X_i (or an entry of the design matrix X) we have to estimate $Y_i = w_1 f_1\left(w_2 f_2\left(\cdots\left(w_k f_k(X_i)\right)\right)\right)$. The larger the k is, the deeper the network with many stacked layers of neurons connected (a.k.a. dense layers). Therefore, it is possible to capture high non-linearities and all interactions among variables. The approach to model estimation underpinned by a DL model is that of compositional function against that of additive function underpinned by the usual regression techniques, including the most modern ones (i.e. $Y_i = w_1 f_1 + w_2 f_2 + \cdots + w_k f_k(X_i)$). See Schmidhuber (2015) for more details. In this setting Y is the scalar random variable of times (in rounds) and X is a vector of dimension 1152.

Estimating a DL model consists of estimating the vectors $w_1, \cdots, w_k$. There are different optimization algorithms to estimate w_s and we used the Adaptive Subgradient Methods (ADAGRAD) (Duchi *et al*., 2011) to minimize the squared loss function, i.e., w_s are estimated to minimize $\sum_{i=1}^{N=19636}(y_i - \hat{y}_i)^2$ the quadratic differences between Y_i and the prediction $\hat{Y}_i = \hat{w}_1 f_1\big(\hat{w}_2 f_2(\cdots \hat{w}_k f_k(X_i)\ )\big)$.

The model structure consists of twenty dense layers, separated by a normalization batch layer and a dropout layer at 40% to avoid overfitting and achieve model parsimony. We have around 234,000 parameters (i.e., weights) to be updated. In a couple of non-reported exercises, we increased and decreased the number of layers, but the decision always resulted in a lower R^2 statistic with a negligible impact on the estimation outcome.

Of course, some weights will be zero to prevent overfitting as they do not contribute to the gradient of the quadratic loss function. Furthermore, to achieve stability in estimation, we introduced a normalization batch between the two hidden layers (Ioffe & Szegedy, 2015). A normalization batch is a standardization (i.e., mean zero and variance one) applied to weights connecting two sets (layers) of all connected neurons. Ioffe and Szegedy (2015) show that

this operation allows for better stability in the gradient of the whole function $Y|X$ estimated with the DL model.

The following graph shows the result of the optimization procedure, iterated for one hundred steps. The loss in the training set (a sample subset randomly defined at a given step and used in the gradient) is practically monotone decreasing, meaning that the model is learning from the data. On the other hand, the loss in the validation set (a subset of the training sample not used for fitting at that particular epoch (optimization step)) is almost always below that in the training set (used to calculate the weights). These two facts indicate that the model does not overfit the data.

[INSERT FIGURE 3 AROUND HERE]

The estimated model can predict 50% of the variability of the response variable. This indicates that the length of time between referee appointments is not purely random, as might be expected, but instead can be forecasted somehow.

5.4. Simulation Study

In order to further validate the proposed approach, and because this is not usual in applications of causal inference in the economic literature, we perform the same simulation study as in Cattaneo et al. (2019). The experiment consists in estimating the causal effect of a treatment in which only 5 out of *k* predictors are relevant, and *k* ranges from 5 to 200 predictors. There are 500 replications for sample sizes *n* of 1000 and 2000. In their example, the true ATE in 0.5. Cattaneo et al. (2019) showed that standard propensity score models produce biased ATE estimates and the bias increases with the number of irrelevant predictors (*k*). They also proposed a bootstrap post-bias-correction. Table 2 shows the simulation results when the ATE is estimated with DL and the approach proposed by Cattaneo et al. (2019). In particular, it reports the bias, the rooted MSE (rmse), the coverage of the corresponding

confidence intervals at a nominal 95% and the corresponding length. Confidence intervals have been obtained using a normal approximation around the point estimators and the empirical standard deviation of the ATE (over all n individual effect estimators). Comparing results in Table 2 with those in Cattaneo et al. (2019), we can clearly see that, even without the need to apply any bias correction, DL provides a more precise ATE estimate. Moreover, our estimation results are not affected by the number of covariates and it improves with the sample size.

The results reported in Table 2 are expected according to the methodological motivation previously discussed. Therefore, the model capability allows for much more covariates than just k=200 at each iteration, and the Dropout effect removes unnecessary covariates from the model.

[INSERT TABLE 2 AROUND HERE]

6. Empirical results

This section estimates the effects of referee decisions on the length of time a referee must wait to be appointed for his next match. More specifically, we assess the impact of disciplinary decisions regarding yellow and red cards, penalties, and an indicator of how surprising the last game's outcome was, measured by Brier Score, on the length of time a referee must wait. For illustration, we start the analysis with a simple linear OLS regression of the response variable against all referee decisions and home and away club and referee dummies. Consistently with the discussion in the data section, the regression also includes a dummy variable to account for the most recent policy stance period. Estimation results shown in Table 3 indicate that this naïve regression explains slightly less than 25% of the response variable (about half of the proportion explained by the DL model). Moreover, a non-reported ANOVA analysis indicates that the 660 referee dummies already explain almost all of this

proportion (23%). Season trend and the 1988 structural change are the only significant variables, while no referee decision seems to affect the length of time for a referee appointment. However, this regression cannot be interpreted as a causal estimation as it does not account for the fact that treatment and confounding factors could interact in very complex ways to answer "what if" questions.

[INSERT TABLE 3 AROUND HERE]

Our core analysis is based on a DL model, including all predictors with the intervention variable D changed to calculate the effect induced by a specific variable. This estimation requires a counterfactual assessment for each referee decision obtained by the fitted DL model. Formally, let $\tilde{\mathrm{X}}$ be the matrix of confounding variables and let D be the intervention variable representing the counterfactual situation, i.e. $\tilde{\mathrm{X}}$ does not have the intervention variable. The range of factual values is set to be $\mathrm{D} = \{0,1,2,3,4\}$ while its assigned counterfactual values are $\mathrm{D} = \{4,3,2,1,0\}$. This implies that the effects are estimated for variations in the counterfactual situation (regarding the factual) of magnitudes $\mathrm{Z} = \{4,2,0,-2,-4\}$ in the intervention variable. To illustrate this estimation, let's consider the decision to show 4 yellow cards to the home team with respect to no showing any yellow card. The first estimation (for $\mathrm{Z} = 4$) requires estimate with DL the length of time up to the next appointment for the same referee at the same match who shows zero yellow cards to the home team and subtracting the expected length of time had this referee shown 4 yellow cards to the home team in that match. This difference estimates the individual treatment effect, to obtain the average treatment effect we average over all differences with Z=4 for all matches (time lengths are estimated with DL in each match). This gives the estimation of the causal effect for Z=4. The same process repeated for the different Z magnitudes.

We evaluate the effect for a given referee on the length of time to be appointed again of changing decisions about *home2yellow*, *away2yellow*, *homered*, *awayred*, *homepen*, *awaypen* and Brier Score. Given that our database corresponds to a highly long historical period, a relevant question to answer is whether referees face different incentive schemes now and in the past. Following the change in policy stance already discussed, we estimate the effects before and after 1988. As explained in the Data section, this period is denoted with the name Recent. However, this distinction was not made in the case of *home2yellow* and *away2yellow* because yellow cards were only introduced in football after 1970.

First, we estimate the expected penalization that a referee suffers due to decisions regarding the number of two yellow cards, red cards, and penalties. Figures 4 to 6 show these effects. As our intervention variables are quantitative, results are always represented by smoothing curves (which connect points on the horizontal axis). Such curves along with the 95% confidence intervals are obtained using GAM models (Wood *et al.*, 2016).

[INSERT FIGURE 4 AROUND HERE]

[INSERT FIGURE 5 AROUND HERE]

[INSERT FIGURE 6 AROUND HERE]

Increasing the number of second yellow cards for both the home and away teams reduces the number of rounds that a referee must wait to be appointed again. On the other hand, increasing the number of red cards produces a penalization in terms of waiting rounds. This effect is significant but of small magnitude. Consistently with E1, a referee's disciplinary sanctions are better appreciated when they are gradually taken rather than in an abrupt decision.

Regarding variations in the number of penalties, results suggest that an increase in penalties awarded to the home side does not increase punishment for referees. There is no punishment

for home penalties in the recent period. However, now there are incentives to give more penalties to the away team, which is not consistent with E2.

There is also some evidence in Figure 5 of an incentive scheme favoring inaction (E3) as penalization increases by increasing the number of red cards. However, no penalization is observed by increasing the number of yellow cards and penalties.

When comparing the two different analysis periods, there is a higher penalization for both an increased number of home and away red cards in the recent period. However, the evidence is mixed when we turn our attention to penalty kicks. While the punishment for a high number of penalties favoring the home team has been reduced, there is a similar incentive to award penalties to the away team in both periods.

To better analyze the possible presence of incentives for home referee bias decisions, we estimate the effect of differences in second yellow cards, red cards, and penalties between the home and the away team for the most recent period. Figure 7 shows these estimations. Results suggest that referees have incentives to show relatively more yellow cards to the home team and more red cards to the away team. Regarding the impact of penalty kicks, our findings suggest that referees are incentivized to not award a disproportionally unbalanced proportion of penalties to the home and away team. Overall, we do not find definitive evidence of an incentive scheme, especially concerning red cards, to counteract the presence of home bias in referee decisions.

[INSERT FIGURE 7 AROUND HERE]

Now we turn our attention to studying how referees are penalized for delivering unexpected results. To explore this, we consider Brier Score as an indicator of how surprising the game's outcome is compared to what was expected. Counterfactual effects for Brier Scores in the two reference periods are shown in Figure 8. Consistent with intuition, organizers try not to

affect the status quo by incentivizing to deliver expected results (E4). Thus, everything else equal; they had to wait more time to be appointed again, the more unexpected their last match's score was.

[INSERT FIGURE 8 AROUND HERE]

To further explore the robustness of our estimation results to the methodological design, we estimated the impact of the different referee actions using an alternative approach: Causal Machine Learning (CML). CML is the name of a subfield of machine learning devoted to estimating causal effects. Among the many methods, Causal Forest (CF) is described in Athey and Wager (2019) and Athey et al. (2019) as a Random Forest-based approach to estimating causal effects. An essential characteristic of this method is that, instead of targeting the minimization of the response variance, it splits variables and nodes to maximize the impact difference between treated and untreated units. Thus, the treatment effect estimation is the average differences over all these terminals for a tree and all trees. Furthermore, the sample is randomly split. One part is used for deciding on splitting variables and separating points, and the other is to estimate the mean of treated and untreated observations. This way of evaluating the causal effect is also known as "honest trees".

Another critical difference with DL, is that CF considers an additive prediction function. The mean effect is given by the sum of the impact in each tree (models). However, as already stated, the prediction with DL is through a compositive function (functions of functions). Comparing these approaches in terms of their R^2, CF only explains 1.6% of the variance of the dependent variable. This value is sensibly inferior to the 50% obtained under the DL model. Therefore, both approaches are very different in prediction, which is a fundamental property of statistical models used for causal analysis (Zhao & Hastie, 2019). Scientific research needs good predictions to get reliable counterfactuals and causal effects.

Despite the differences discussed in the previous paragraphs, CF estimates of marginal effects, reported in Figure 9, are reasonably consistent with DL results. In particular, they suggest the negative (positive) impact of yellow (red) cards on the length of time a referee has to wait to be appointed again. Home and away penalties show a close to zero effect. Moreover, the negative impact of most Briar Score values and the negligible impact of home and away penalties are also consistent with the DL analysis. Causal effects conditional to each specific decision are not reported for the sake of brevity but are available from the authors upon request. However, CF estimates are also qualitatively consistent with DL results.

[INSERT FIGURE 9 AROUND HERE]

7. Lessons and limitations for organizations

The analysis in the previous section suggests the presence of institutional pressures affecting referees' decisions in subtle and complex ways. More specifically, referees are incentivized to make gradual (instead of drastic) decisions to send a home player off the pitch (anchoring-as-adjustment) and to deliver results that are consistent with expectations (status quo bias). However, we did not find conclusive evidence of institutional incentives to favor the home team (home team bias) and not make decisions (inaction bias). Overall, results suggest that organizers incentivize some potential referee bias.

Our study complements the previous literature discussed in Section 2 about organizational culture affecting employment practices. The approach is not deductive and therefore does not set general hypotheses on the nature and motivation of the identified pressures. Instead, our aim is more modest: identifying institutional responses to observed referee actions. However, the estimated effects can be interpreted as institutional biases. Similar types of underlying forces (anchor biases) may exist in other institutional settings. For example, without receiving

written orders, school teachers may be persuaded to pass a proportion of students that do not differ much from a pre-established number, or managers may be incentivized to follow a routine of actions before making a decision. The context of professional sports allows the identification of the reward scheme associated with implicit pressures (Urda & Loch, 2013, Amaral & Tsay,2009). Based on this evidence, individual decisions should be analyzed in a more general framework where self-interest profit maximization also includes institutional biases.

A limitation of our research is that it is specific to the particular context of Spanish soccer. However, while identifying these pressures in standard industries is complex, football provides a laboratory where information about operations, the context where decisions are made, and their consequences are publicly observed. Giambatista et al. (2005) argue that non-sport papers are also concentrated in particular sectors, and whether their results can be generalized elsewhere is unclear. Thus, rather than relying on experiments or questionnaires, we base our analysis on observed institutional reactions to high-stakes decisions.

Another limitation concerns using a model that (unlike parametric specifications) does not describe the role of different confounders in explaining the response variable. However, considering a DL model is instrumental in conducting this type of analysis. It facilitates the adoption of the strong ignorability hypothesis by allowing us to control for many variables and how they interact in a fully flexible way (a total of 230,000 factors) without imposing subjective decisions in parametric specifications. These factors embed variables considered in more standard methods, such as other specific referees and match characteristics. Moreover, it is more difficult to alter results by subjective changes in the model specification or the variables involved.

8. Future lines of research

An exciting venue for future research would be to explore the impact of institutional biases found in our study in more general settings. Thus, unlike nudge theory which focuses on correcting discriminatory behavior by changing the different factors that influence perceived choice (Thaler & Sunstein, 2008), little research has been devoted to how institutions can generate rather than correct biased behavior in an implicit way. For example, some remarkable exceptions are Panagopoulos (2014) and Gomez and Wapman (2017). They study the effect of implied social forces on voting and contraceptive decisions by young black and Latina women, respectively. Still, an essential difference with the analysis in the present paper is that we focus on the presence of cognitive bias faced by workers inside a specific organization rather than social pressure at the country level.

A second possible line of research would be to estimate the presence of institutional pressure by putting together the use of questionnaires with machine learning techniques. Qualitative methods have been criticized in causal analysis as they do not allow the analyst to control for all the variables affecting the outcome of interest (Antonakis et al., 2010). However, while we cannot identify causal relationships by only observing a conjunction of events, qualitative analysis has the advantage that, unlike machine learning, is conducted as a reflexive process at every stage of the project (Maxwell, 2012). Therefore, a combined analysis could reveal the potential bias and limitations of both methods.

Technical innovation provides another possible direction of research. In particular, the implementation of Video Assistant Referees (VAR) in Spanish soccer from season 2018/19 will help referees reduce the amount of uncertainty they face when making decisions and should influence the incentives identified in our study. Moreover, visibility is fundamental in explaining institutional response to external pressure (Okhmatovskiy & David, 2012). In this context, the impact of VAR on organizational pressures faced by soccer referees can be deemed an appealing analysis issue. Other interesting questions to explore could be, for

example, to study the influence of referee decisions on different types of incentives such as salaries or referee relegation and to extend this estimation to cross-sectional rather than longitudinal settings to analyze potential correlations among institutional biases.

References

Abdallah , W., Goergen, M. & O'Sullivan, N. (2015). Endogeneity: how to failure to correct for it can cause wrong inferences and some remedies. *British Journal of Management*, 26(4), 791-804.

Amaral, J., & Tsay, A.A. (2009). How to win "spend" and influence partners: Lessons in behavioral operations from the outsourcing game. *Production and Operations Management*, 18(6), 621–634.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(6), 1086–1120.

Athey, S., & Imbens, G.W. (2015). Machine learning methods for estimating heterogeneous causal effects. Stat, 1050(5), 1–26.

Athey, S., Tibshirani, J. & Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics* 47(2), 1148–78.

Athey, S., & Wager, S. (2019). Estimating Treatment Effects with Causal Forests: An Application. *Observational Studies* 5(2), 37–51.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, 94(2), 74-85.

Belloni, A., Chernozhukov, V., & Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), 608–650.

Boeri, T., & Severgnini, B. (2011). Match rigging and the career concern of referees. *Labour Economics*, 18(3), 349-359.

Breiman, L. (2001). "Statistical Modeling: The Two Cultures" (with Comments and a Rejoinder by the Author). *Statistical Science* 16. Institute of Mathematical Statistics, 199–231.

Bryson, A., Buraimo, B., & Simmons, R. (2011). Do salaries improve worker performance?. *Labour Economics* 18(4), 424-433.

Buraimo, B., Forrest, D., & Simmons, R. (2010). The twelfth man? Refereeing bias in English and German soccer. *Journal of the Royal Statisticial Society: Series A (Statistics in Society)*, 173(2), 431-449.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. (2019) Two-step estimation and inference with possibly many included covariates. *The Review of Economic Studies*, 86(3), 1095-1122.

Choudhury, P., Allen, R.T., & Endres, M.G. (2021). Machine Learning for Pattern Discovery in Management Research. *Strategic Management Journal*, 42(1), pp. 30-57.

Christensen, M., Dahl, Ch. M., Knudsen, T., & Warglien, M. (2022) "Context and aggregation: An experimental study of bias and discrimination in organizational decisions". *Organization Science*, forthcoming.

Cohen, M. D., Burkhart, R., Dosi, G., Egidi, M., Marengo, L., Warglien, M., & Winter, S. G. (1996). Routines and other recurring action patterns of organizations: Contemporary research issues. *Industrial Corporate Change* 5(3), 653–698.

Dawson, P., & Dobson, S. (2010). The influence of social pressure and nationality on individual decisions: Evidence from the behaviour of referees. *Journal of Economic Psychology*, 31(2), 181-191.

Dawson, P., Dobson, S., Goddard, J., & Wilson, J. (2007). Are football referees really biased and inconsistent? Evidence on the incidence of disciplinary sanction in the English Premier League. *Journal of the Royal Statisticial Society: Series A (Statistics in Society)* 170(1), 231-250.

DesJardine, M., & Bansal, P. (2019). One step forward, two steps back: How negative external evaluations can shorten organizational time horizons. *Organization Science* 30(4), pp. 761-780.

Duchi, J., Hazan, E. & Singer, Y. (2011) Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12, 2121–2159.

Eggers, J.P., & Kaplan, S. (2009). "Cognition and renewal: Comparing CEO and organizational effects on incumbent adaptation to technical change". *Organization Science*, 20(2), 461-477.

Gallo, E., Grund, T. & Reade, J. (2013). Punishing the foreigner: implicit discrimination in the Premier League based on oppositional identity. *Oxford Bulletin of Economics and Statistics* 75(1), 136-156.

Garicano, L., Palacios-Huerta, I., & Prendergast, C. (2005). Favoritism under social pressure. *Review of Economics and Statistics*, 87(2), 208-216.

Gavetti, G. (2005). Cognition and hierarchy: Rethinking the microfoundations of capabilities' development. *Organization Science* 16(6), 599-617.

Gavetti, G., & Levinthal, D. (2000). Looking forward and looking backward: Cognitive and experiential search. *Administration Science Quarterly* 45(1), 113-137.

Gavetti, G. & Rivkin, J.W. (2007). On the origin of strategy: Action and cognition over time. *Organization Science* 18(3), 420-439.

Giambatista, R. C., Rowe, W. G., & Riaz, S. (2005). Nothing succeeds like succession: A critical review of leader succession literature since 1994. *The Leadership Quarterly* 16(6), 963–991.

Gomez A.M., & Wapman, M. (2017). Under (implicit) pressure: young Black and Latina women's perceptions of contraceptive care. *Contraception* 96(4), 221-226.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 217–240.

Hill, J.L., & Su, Y.S. (2013). Assessing lack of common support in causal inference using bayesian nonparametrics: implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, 7(3), 1386-1420.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* 87(3), 706–710.

Ioffe, S. & Szegedy, Ch. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In International conference on machine learning, 448-456. Proceedings of the 32nd International Conference on Machine Learning

Larch, M., Orseau, E. & van der Wielen, W. (2021). Do E.U. fiscal rules support or hinder counter-cyclical fiscal policy?. *Journal of International Money and Finance*, 112, 1-21.

Liu, R., L. Wei, L., & Zhang, P. (2021). A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nature Machine Intelligence* 3(1), 68-75.

Luo, Y., Peng, J., & Ma, J. (2020) When causal inference meets deep learning. *Nature Machine Intelligence* 2, 426–427.

Magazzino, C., Mele, M., & Sarkodie, S.A., (2021). The nexus between COVID-19 deaths, air pollution and economic growth in New York state: Evidence from Deep Machine Learning. *Journal of Environmental Management*, 286, 1-12.

Maxwell, J. A. (2012). The importance of qualitative research for causal explanation in education. *Qualitative Inquiry* 18(8), 655-661.

Okhmatovskiy, I., & David, R. J. (2012). Setting Your Own Standards: Internal Corporate Governance Codes as a Response to Institutional Pressure. *Organization Science* 23(1), 155-176.

Oliver, N., Calvard, Th. & Potocnik, K. (2017). Cognition, technology, and organizational limits: Lessons from the Air France 447 disaster. *Organization Science*, 28(4), 729-743.

Panagopoulos, C. (2014). I've Got My Eyes on You: Implicit Social-Pressure Cues and Prosocial Behavior. *Political Psychology* 35(1), pp. 23-33.

Park, S.H., & Patterson, K. (2021). Being counted and remaining accountable: Maintenance of quarterly earnings guidance by U.S. public companies. *Organization Science* 32(3), 544-567.

Pettersson-Lidbom, P. & Priks, M. (2010). Behavior under social pressure: Empty Italian stadiums and referee bias". *Economics Letters*, 108(2), 212-214.

Pope, D.G., Price, J., & Wolfers, J. (2018). Awareness Reduces Racial Bias. *Management Science* 64 (11), 4988-4995.

Price, J., Remer, M., & Stone, D.F. (2012). Subperfect game: Profitable biases and NBA preferences. *Journal of Economics and Management Strategy* 21(1), 271-300.

Price, J., & Wolfers, J. (2010). Racial discrimination among NBA referees. *Quarterly Journal of Economics* 125(4), 1859-1887.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making* 3(4), 263-277.

Rosenbaum, P. R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55.

Schweitzer, M. (1994) Disentabling status quo and omission effects. An experimental analysis. *Organizational Behavior and Human Decision Processes* 58(3), 457-476.

Shmueli, G. (2010). To explain or to predict?. *Statistical science* 25(3), 289-310.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks*, 61, 85–117.

Shrestha, Y.R., Vivianna, F.H., Puranam, P.,& von Krogh, G. (2021). Algorithm Supported Induction for Building Theory: How Can We Use Prediction Models to Theorize?. *Organization Science* 32(3), 856-880.

Silberzahn, R., & Uhlmann, E.L. (2015). Many hands make tight work. *Nature* 526, 189-191.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics* 69(1), 99–118.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Yale University Press.

Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5(4), 297–323.

Urda, J., & Loch, Dh. H. (2013). Social preferences and emotions as regulators of behavior in processes. *Journal of Operations Management* 31(1-2), 6-23.

Varian, HR. (2016a) How to build an economic model in your spare time. *The American Economist* 61(1), 81–90.

Varian, HR. (2016b). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences* 113(27), 7310–7315.

Wood, S.N, Pya, N., & Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association* 111(516), 1548–1563.

Zhang, R., Martínez-García, E., Wynne, M.A., & Grossman, V. (2022). Ties that bind: Estimating the natural rate of interest for small open economies. *Journal of International Money and Finance*, forthcoming.

Zhao, Q., & Hastie, T. (2019). Causal interpretations of black-box models. *Journal of Business and Economic Statistics*, 39(1), 272-281.

Tables

Table 1. *Descriptive Statistics*

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Round	19,636	18.498	10.503	1	10	27	44
Homeyellow	19,636	1.811	1.754	0	0	3	11
home2yellow	19,636	0.057	0.248	0	0	0	3
Homered	19,636	0.065	0.262	0	0	0	4
Homepen	19,636	0.129	0.355	0	0	0	3
Awayyellow	19,636	2.061	1.861	0	0	3	12
away2yellow	19,636	0.075	0.279	0	0	0	3
Awayred	19,636	0.080	0.293	0	0	0	5
Awaypen	19,636	0.074	0.270	0	0	0	2
Homegoals	19,636	1.648	1.398	0	1	2	12
Awaygoals	19,636	0.977	1.039	0	0	2	8
brier.score	19,636	0.111	0.120	0	0	0.2	1
Time	19,636	4.021	2.618	3	3	4	44

Factors variables		# of levels	Levels –(Frequency)
Season	19,636	87	
Referee IF	19,636	661	

Outcome	19,636	3	0 (4273)	0.5 (5039)	1 (10324)
TeamID - Home	19,636	231			
TeamID – Away	19,636	236			
Recent period	19,636	2	TRUE (11774)	FALSE (7862)	

Table 2. *Simulation: DL estimates of the average treatment effect. Bias-Corrected results in Cattaneo et al. (2019) are reported for comparison.*

Panel (a) n=1000										
	DL					Bias-Corrected				
k	bias	sd	rmse	coverage	length	bias	sd	rmse	coverage	Length
5	0.16	0.02	0.09	1.00	1.64	-0.21	4.93	4.93	0.93	18.28
20	0.15	0.02	0.10	1.00	1.67	0.18	5.26	5.27	0.94	19.81
40	0.14	0.02	0.09	1.00	1.70	1.03	5.11	5.22	0.94	19.67
60	0.13	0.02	0.10	1.00	1.71	1.75	5.02	5.32	0.93	19.27
80	0.13	0.02	0.11	1.00	1.72	2.28	4.91	5.41	0.92	18.67
100	0.13	0.02	0.10	1.00	1.73	2.65	4.78	5.46	0.90	18.28
120	0.12	0.02	0.11	1.00	1.74	2.96	4.66	5.51	0.89	17.80
140	0.12	0.02	0.11	1.00	1.75	3.24	4.57	5.60	0.87	17.46
160	0.12	0.02	0.11	1.00	1.75	3.46	4.43	5.62	0.86	17.15
180	0.11	0.02	0.11	1.00	1.76	3.58	4.35	5.63	0.86	16.97
200	0.11	0.02	0.11	1.00	1.76	3.81	4.22	5.69	0.84	16.75
Panel (b) n=2000										
	DL					Bias-Corrected				
k	bias	sd	rmse	coverage	length	bias	sd	rmse	coverage	Length
5	0.19	0.01	0.01	1.00	1.63	-0.12	4.95	4.95	0.93	18.21
20	0.16	0.01	0.02	1.00	1.67	0.06	5.16	5.16	0.94	19.31
40	0.15	0.01	0.02	1.00	1.75	0.54	5.35	5.38	0.94	19.72
60	0.13	0.01	0.02	1.00	1.83	1.18	5.44	5.57	0.93	19.75
80	0.13	0.01	0.03	1.00	1.88	1.82	5.43	5.73	0.91	19.75

100	0.12	0.01	0.03	1.00	1.92	2.33	5.37	5.86	0.90	19.31
120	0.11	0.01	0.03	1.00	1.95	2.74	5.27	5.94	0.90	19.04
140	0.11	0.01	0.03	1.00	1.97	3.21	5.11	6.04	0.88	18.85
160	0.11	0.01	0.03	1.00	1.98	3.53	5.05	6.16	0.87	18.66
180	0.11	0.01	0.03	1.00	2.00	3.87	4.95	6.28	0.85	18.40
200	0.11	0.01	0.03	1.00	2.00	4.13	4.84	6.36	0.85	18.22

Details of this simulation can be found in Cattaneo et al. (2019)

Table 3. *Determinants of length of time for the next referee appointment. OLS regression*

Variable	***Estimate***
homeyellow	0.012 (0.015)
home2yellow	-0.110 (0.078)
awayyellow	-0.015 (0.014)
away2yellow	-0.052 (0.070)
homered	0.073 (0.068)
awayred	0.042 (0.065)
homepen	0.025 (0.049)
awaypen	0.042 (0.065)
Briar.score	-0.042 (0.065)
Season trend	-0.036 (0.007) **
After 1988	0.571 (0.110) **
R-squared	0.245
Adjusted R-squared	0.199

F-statistic	5.31 **

Home and away clubs and referee dummies were also included in the estimation; standard errors in parenthesis, * p<0.05 ** p<0.01

Figures

Figure 1. *Response variable: distribution of length of time between top tier referee's appointments*

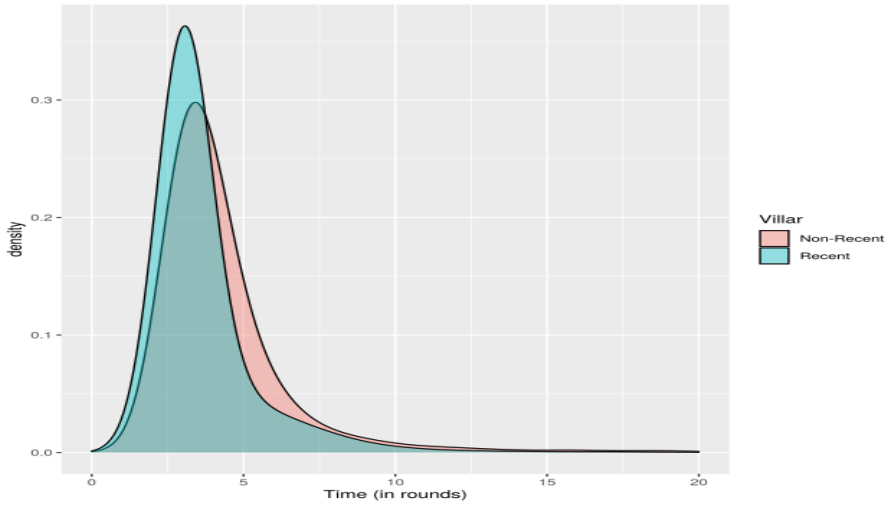

Figure 2. *Neural network architecture*

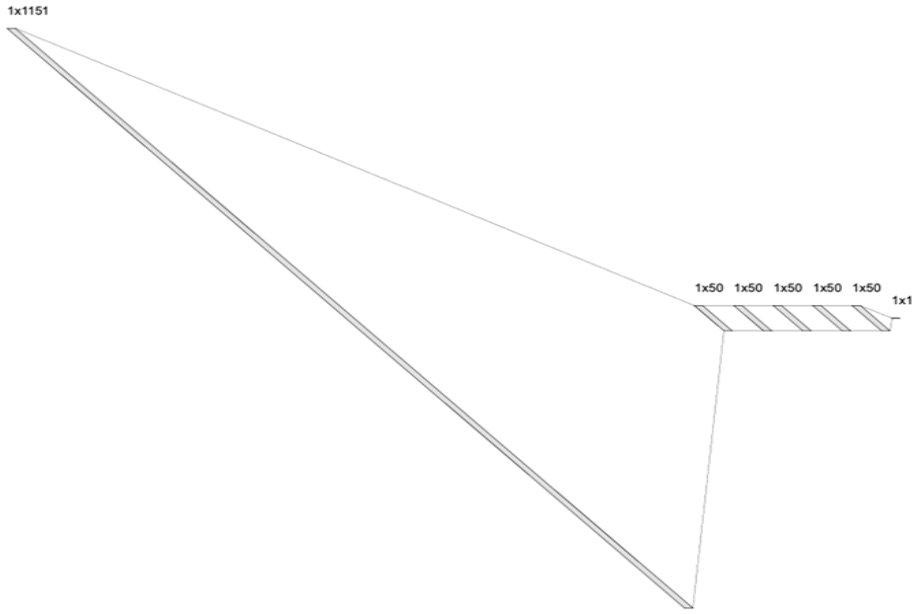

Figure 3. *Results of the Optimization Procedure*

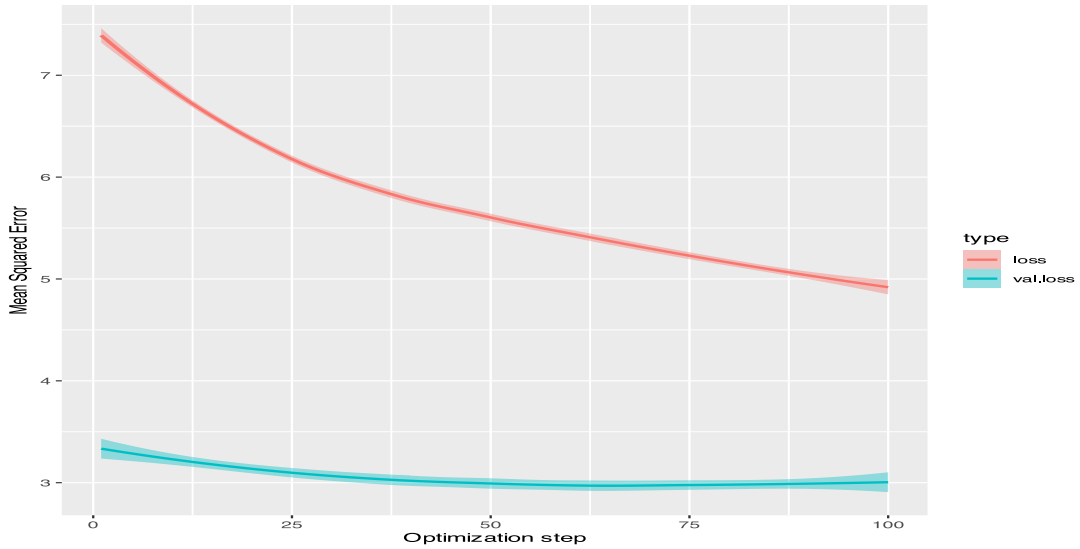

Figure 4. *Causal effect of variations in two yellow cards*

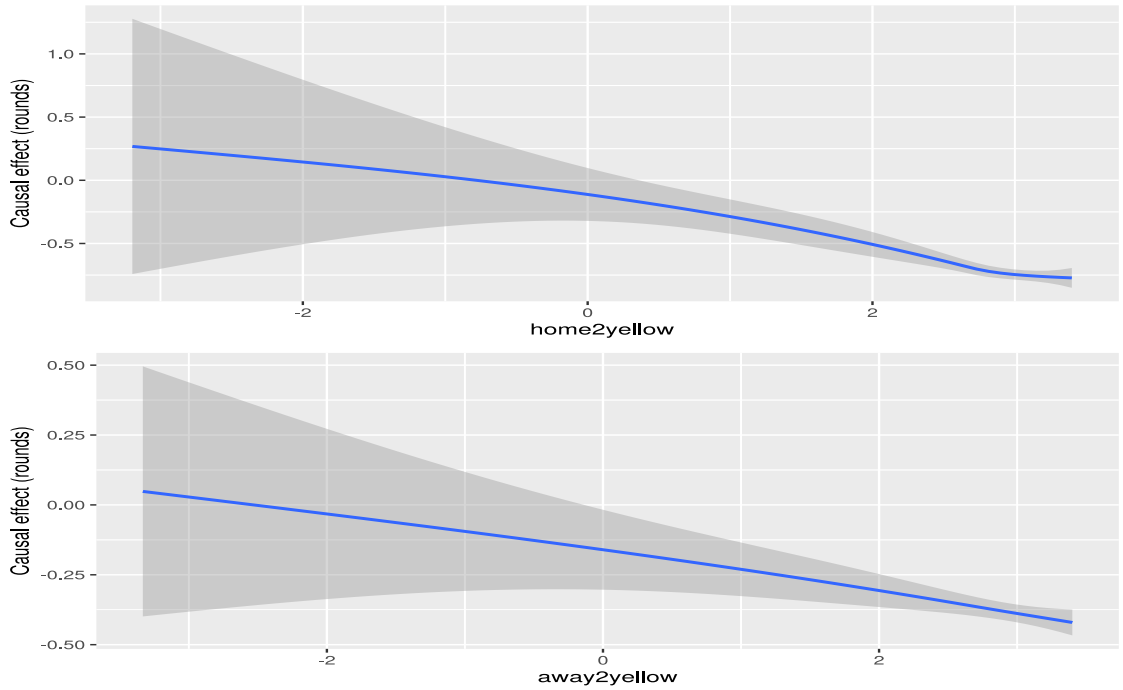

Figure 5. *Causal effect of variations in red cards*

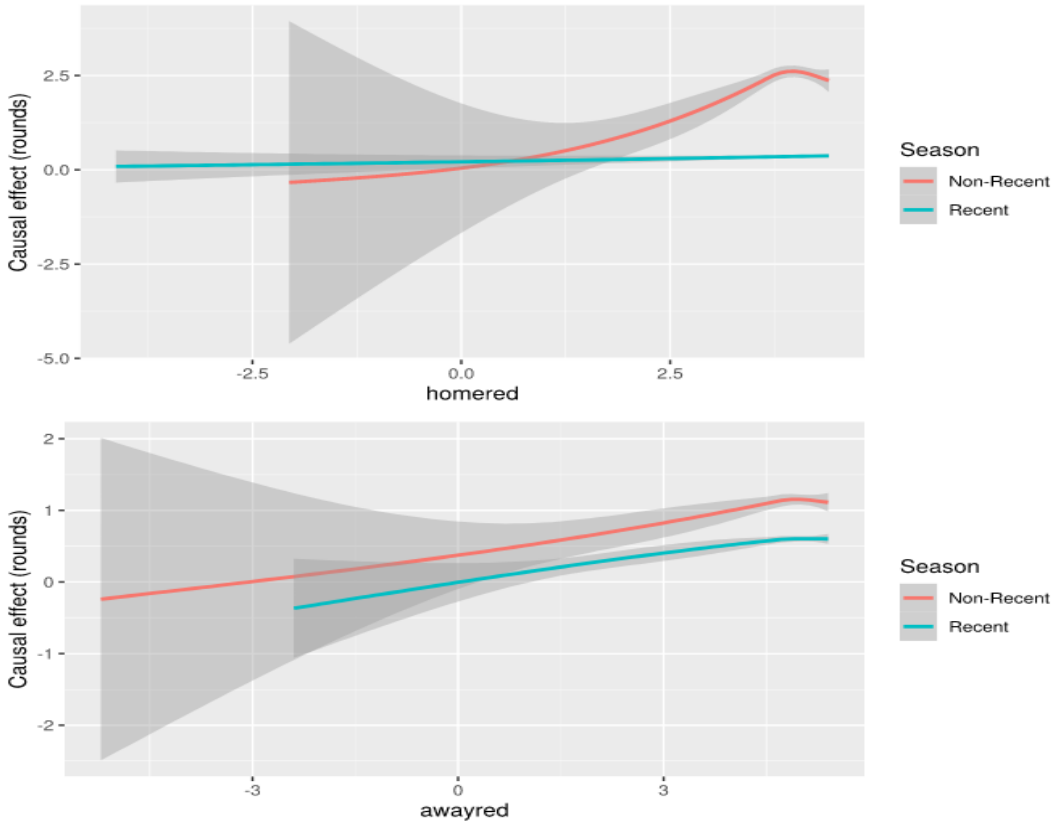

Figure 6. *Causal effect of variations in penalties*

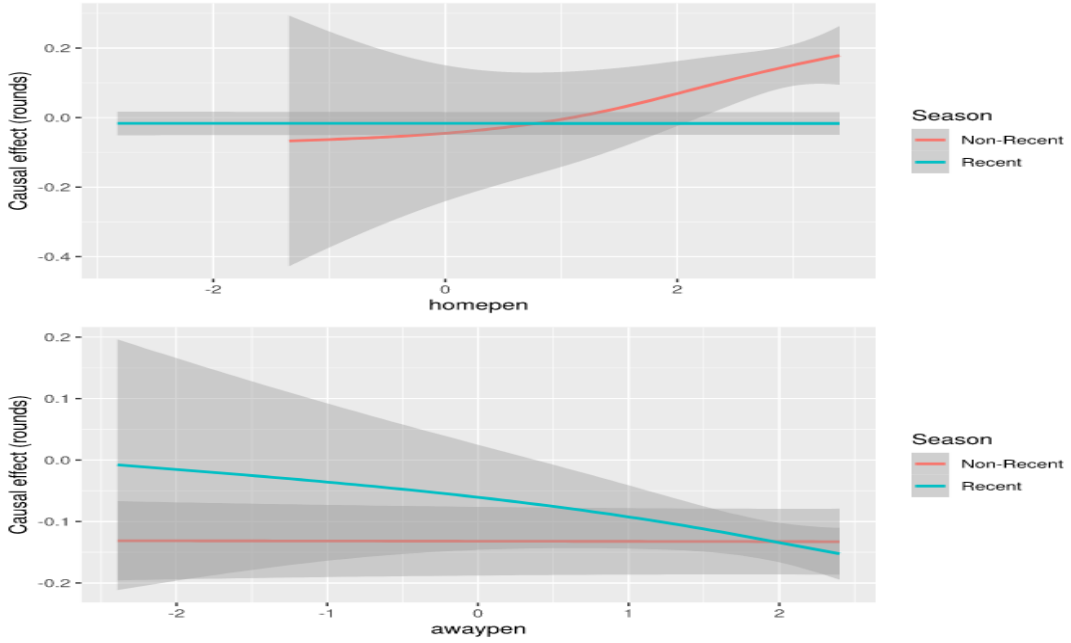

Figure 7. *Causal effects for home vs away referee decisions.*

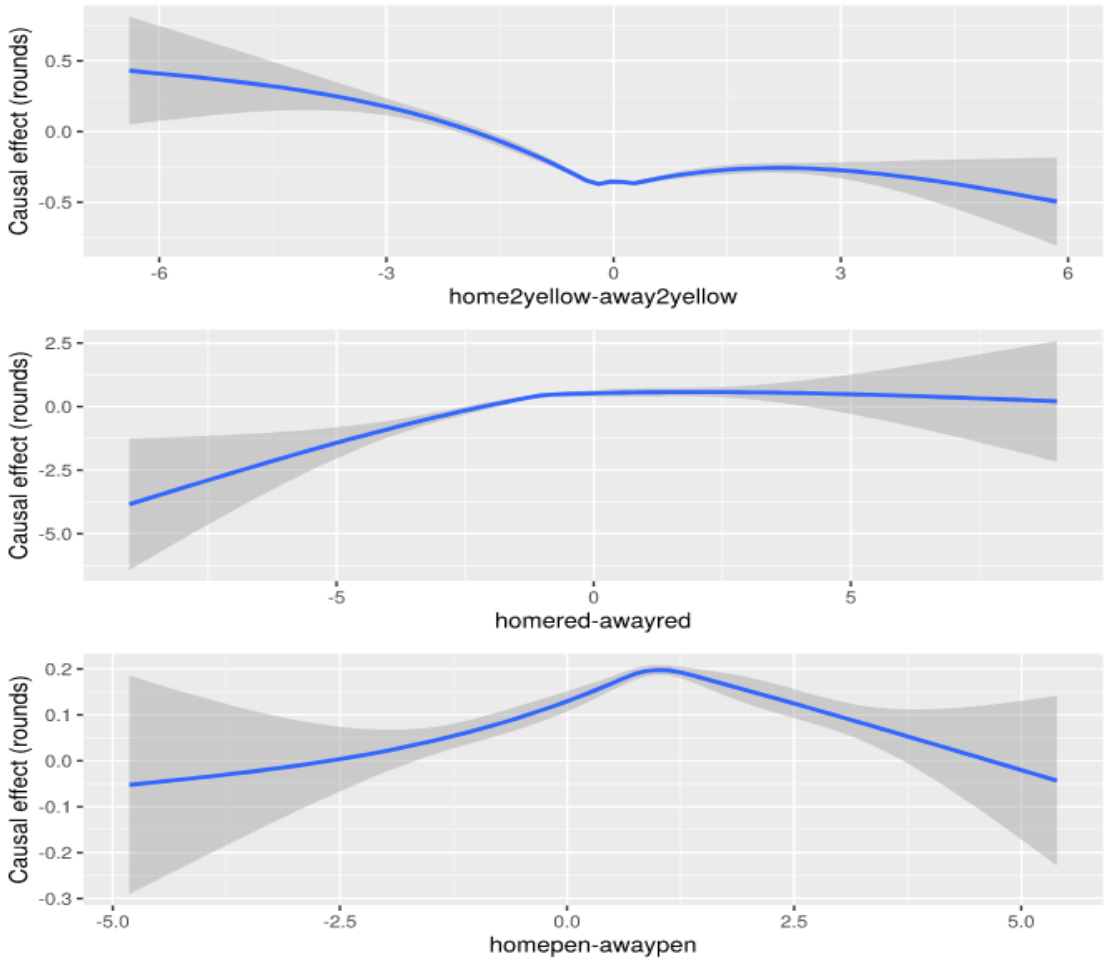

Figure 8 Causal effects for Brier Score

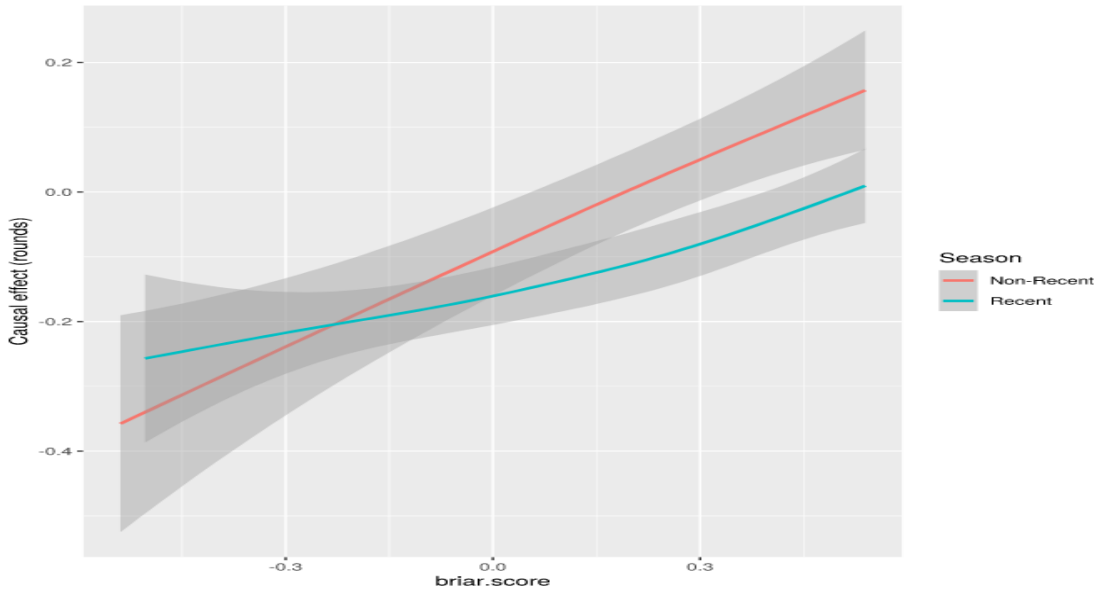

Figure 9 Estimates of determinants of the length of time between referee appoints using Causal Forest